

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**  
**CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA.**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA**  
**DEPARTAMENTO DE ESTATÍSTICA**

**Uma Modificação da Extensão do Algoritmo *AID***  
**para Modelos Lineares Generalizados Usando**  
**Reamostragem *Bootstrap*.**

Cátia Valéria Presotti  
Bacharel em Estatística

Orientadora: **Profa. Dra. Maria Cecília Mendes Barreto**

**SÃO CARLOS**  
**2006**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**  
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA.  
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA  
DEPARTAMENTO DE ESTATÍSTICA

**Uma Modificação da Extensão do Algoritmo *AID*  
para Modelos Lineares Generalizados Usando  
Reamostragem *Bootstrap*.**

Cátia Valéria Presotti

Dissertação apresentada ao Departamento  
de Estatística da Universidade Federal de  
São Carlos como parte dos requisitos  
necessários para obtenção do título de  
Mestre em Estatística.

**SÃO CARLOS**  
**2006**

**Ficha catalográfica elaborada pelo DePT da  
Biblioteca Comunitária da UFSCar**

P934me

Presotti, Cátia Valéria.

Uma modificação da extensão do algoritmo AID para modelos lineares generalizados usando reamostragem Bootstrap / Cátia Valéria Presotti. -- São Carlos : UFSCar, 2006.

93 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Análise de regressão. 2. Método de agrupamento. 3. Modelos lineares generalizados. 4. Distribuição binomial. 5. Método bootstrap. 6. Função – poder. I. Título.

CDD: 519.536 (20<sup>a</sup>)

*Dedico com amor este trabalho aos meus queridos pais, Leonildo e Teresinha, pelo  
incondicional amor e apoio na realização dos meus sonhos...  
Ao meu irmão Leandro, pelo incentivo, companheirismo, dedicação, amizade..  
Ao meu namorado Frank, pela compreensão, paciência, amizade.....*

## A vitória faz parte do sonho

*O sabor da vitória está na realização daquilo que você sonhou em conquistar. Pode existir alegria maior do que a realização de um sonho do que a conquista daquilo que tão intensamente se desejou e, finalmente, conseguiu-se obter? A sensação de vitória é mais completa não só quando nos empenhamos em realizar alguma coisa mas também quando sonhamos com ela e, finalmente, a conseguimos.*

*Autor desconhecido*

*“O sábio que se mostra orgulhoso e pedante revela que não sabe honrar a ciência”*

Dr. Augusto Guimarães Chaves

Livro O Homem que Calculava – Malba Tahan

## **Agradecimentos**

*A Deus e a Nossa Senhora, que me abençoam e dão forças diariamente, para que eu possa seguir em frente sempre.*

*A meus pais, Leonildo e Teresinha, que não pouparam esforços e amor para que eu cumprisse mais esta etapa de minha formação;*

*Ao meu irmão Leandro, pelo amor, por estar sempre do meu lado em todos os momentos da minha vida, me incentivando, orientando, apoiando...*

*Ao meu namorado Frank, pelo amor, companheirismo e paciência em todos os momentos e pela sua ajuda no programa das simulações .*

*À minha orientadora **Profa. Dra. Maria Cecília Mendes Barreto**, que com entusiasmo de mestre, a sabedoria e a tranqüilidade de um sábio, incentivou todos os nossos sonhos, hoje concretizados. Pelo seu carinho, confiança, amizade, paciência, minha eterna gratidão.*

*Aos meus avós paternos e maternos: Armelinda e Santo Presotti; Euvira e José Mecca e a todos os meus familiares pelo apoio e orações para realização deste trabalho.*

*Aos meus amigos Fabiola, Carla, Denilton e Erlandison pela amizade e pelo apoio na hora das simulações.*

*Aos meus queridos amigos e colegas de mestrado, pelo carinho depositado nesses dois anos de mestrado.*

*Às minhas duas grandes amigas de graduação, Ana Paula e Edilaine, pela força e pelo constante incentivo ao longo de minha caminhada.*

*Aos meus professores do Mestrado: Dr. José Galvão Leite, Dra. Maria Cecília Mendes Barreto, Dr. Luis Milan, Dr. Carlos Alberto Diniz e Dr. Josemar Rodrigues.*

*Aos funcionários do Departamento de Estatística, em particular ao Robson, pela amizade ao longo da graduação e mestrado e a Dona Luiza, pelo seu constante sorriso.*

*A todos aqueles que, de forma direta ou indireta, colaboraram para que eu pudesse realizar da melhor forma possível este trabalho.*

**Meu muito obrigado!!!**

## Resumo

Uma das situações mais freqüentes encontradas por pesquisadores de diversas áreas do conhecimento é formar grupos de indivíduos que sejam, de alguma maneira, similares entre si. A análise de agrupamento é um conjunto de técnicas estatísticas que identificam subgrupos ou classes distintas de indivíduos mutuamente excludentes com base nas similaridades existentes entre os indivíduos, ou seja, os mais semelhantes pertencem ao mesmo grupo. No caso específico de uma única variável resposta e diversas explicativas, vários procedimentos podem ser utilizados, entre eles: contrastes de médias, comparações múltiplas ou técnicas aglomerativas, como teste de SCOTT-KNOTT e a técnica AID - (*Automatic Interaction Detector*). As técnicas de agrupamento de médias podem não ser adequadas na aplicação em dados com distribuição diferente da normal. Nesse caso, utilizou-se a extensão do algoritmo AID, no qual se baseia o método de SCOTT-KNOTT. Essa extensão é usada em modelos lineares generalizados e adota, como medida de homogeneidade de grupos, uma estatística baseada na função desvio que tem assintoticamente distribuição qui-quadrado. Neste trabalho, apresentam-se o método de reamostragem *bootstrap* adaptado para a extensão do algoritmo AID, sua curva poder simulada e a curva poder simulada assintótica, considerando um delineamento inteiramente ao acaso, com tamanho da amostra binomial  $n_b$ , com  $r$  tratamentos e  $n_r$  repetições por tratamento. Os resultados do estudo por simulação indicam que, conforme aumenta-se o tamanho da amostra binomial, o poder simulado dos dois testes aumenta rapidamente. Por outro lado, esse comportamento é mais acentuado conforme aumenta-se o número de tratamentos. Vale ressaltar que, para o caso extremo em que o tamanho da binomial é igual a 1 e o número de tratamentos igual a 8 e 12, a curva poder simulada do teste *bootstrap* é destacadamente superior à curva poder do teste assintótico para o número de repetições por tratamento estudado.

# Abstract

One of the most frequently situation found by researchers is to find groups of similar individuals. The cluster analysis is a set of statistical techniques that identify mutually exclusive subgroups or classes over individuals, based on their similarity. When then main is to group means of treatments we can use contrasts, multiple comparitions or clustering techniques as the SCOTT-KNOTT test and AID (Automatic Interaction Detector) technique. In this work we focus on the comparition of the simulated power function of the asymptotic test and also of the bootstrap test for the extension of the AID algorithm for generalized linear models. The bootstrap power function over main the asymptotic power when the number of binomial sample is equal to one, and the number of treatments is equal to 8 and 12, in a completed randomized experiment with a single factor for binomial variables.

# Sumário

<b>1. Introdução</b>	<b>1</b>
<b>2. Análise de agrupamento</b>	<b>4</b>
2.1 Teste de SCOTT-KNOTT .....	5
2.2 Principais diferenças entre os testes de comparações múltiplas com relação ao SCOTT –KNOTT.....	9
2.3 Técnica - AID ( <i>Automatic Interaction Detection</i> ).....	10
<b>3. Uma extensão do método SCOTT e KNOTT em modelos lineares generalizados</b>	<b>13</b>
3.1 Modelos lineares generalizados.....	13
3.2 Uma extensão de algoritmo AID em modelos lineares generalizados.....	24
3.3 Exemplo da aplicação da extensão do algoritmo AID em um experimento com couve chinesa ( <i>Brassic pekinensis</i> ).....	29
<b>4. Testes usando métodos de computação intensiva</b>	<b>35</b>
4.1 Reamostragem <i>bootstrap</i> .....	36
4.1.1 Erro padrão <i>Bootstrap</i> .....	36
4.1.2 Teste de hipótese para duas amostras.....	39
4.2 Função poder <i>bootstrap</i> .....	43
<b>5 Uma proposta do teste <i>bootstrap</i> na extensão do algoritmo AID</b>	<b>45</b>
5.1 Desenvolvimento do algoritmo para o teste <i>bootstrap</i> .....	45

5.2	Determinação do tamanho da simulação para o teste <i>bootstrap</i> e o teste assintótico.....	47
5.3	Exemplo da aplicação do teste <i>bootstrap</i> no experimento com couve chinesa ( <i>Brassica pekinensis</i> ).....	52
<b>6.</b>	<b>Um estudo via simulação para as curvas poder do teste bootstrap e do teste assintótico</b>	<b>56</b>
6.1	Tamanho da amostra binomial igual a 1.....	56
6.2	Tamanho da amostra binomial igual a 10.....	62
6.3	Tamanho da amostra binomial igual a 100.....	67
6.4	Comparação geral.....	72
<b>7.</b>	<b>Conclusões e propostas futuras</b>	<b>75</b>
	<b>Apêndice</b>	<b>77</b>
	Apêndice A: Programa em <i>R</i> para a primeira partição usando o teste assintótico no experimento com couve chinesa ( <i>Brassica pekinensis</i> ) para primeira partição.....	77
	Apêndice B: Programa em <i>R</i> para determinar o tamanho da simulação de $S_A$ do teste assintótico.....	80
	Apêndice C: Programa em <i>R</i> para determinar o tamanho da simulação de $B$ do teste bootstrap.....	83
	Apêndice D: Programa em <i>R</i> para determinar o tamanho da simulação $S$ .....	87
	<b>Referências</b>	<b>91</b>

# Lista de Figuras

<b>Figura 1 -</b>	Árvore de divisão dicotômica, considerando como partição ótima a maior redução e como regra de parada a obtenção de 4 grupos.....	32
<b>Figura 2 -</b>	Árvore de divisão dicotômica, considerando como partição ótima a maior redução e como regra de parada, a partição ótima com um nível descritivo superior a um valor pré-fixado $\alpha=0,01$ .....	34
<b>Figura 3 -</b>	Gráfico do p-valor médio em função das 1000 simulações para a obtenção do ponto de equilíbrio $S_A$ .....	49
<b>Figura 4 -</b>	Gráfico referente à simulação com 500 reamostragem bootstrap para a obtenção do ponto de equilíbrio $B$ .....	50
<b>Figura 5 -</b>	Gráfico do p-valor médio em função das 1000 simulações para a obtenção do $S_B$ .....	51
<b>Figura 6 -</b>	Árvore de divisão dicotômica, considerando como partição ótima a maior redução e como regra de parada, a partição ótima com um nível descritivo superior a um valor pré-fixado $\alpha=0,01$ .....	54
<b>Figura 7 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=1$ , $R=3,5,8,12$ tratamentos e $n=5$ por tratamento.....	59
<b>Figura 8 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=1$ , $r=3,5,8,12$ tratamentos e $n_r=10$ por tratamento.....	61
<b>Figura 9 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=10$ , $R=3,5,8,12$ tratamentos e $n=5$ por tratamento.....	64
<b>Figura 10 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=10$ , $R=3,5,8,12$ tratamentos e $n=10$ por tratamento.....	66
<b>Figura 11 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=10$ , $R=3,5,8,12$ tratamentos e $n=5$ por tratamento.....	69

<b>Figura 12 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=10$ , $R=3,5,8,12$ tratamentos e $n=10$ por tratamento.....	71
<b>Figura 13 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=1,10,100$ , $R=3,5,8,12$ tratamentos e $r=5$ por tratamento e <b>-*</b> <i>PoderBoot</i> e <b>-●</b> <i>PoderAssint</i> .....	73
<b>Figura 14 -</b>	Curva Poder do Teste Assintótico e do Teste <i>Bootstrap</i> para o tamanho da binomial $n_b=1,10,100$ , $R=3,5,8,12$ tratamentos e $n=10$ por tratamento e <b>-*</b> <i>PoderBoot</i> e <b>-●</b> <i>PoderAssint</i> .....	74

# Capítulo 1

## Introdução

Uma das situações mais freqüentes encontradas por pesquisadores de diversas áreas do conhecimento é formar grupos de indivíduos que sejam, de alguma maneira, similares entre si. A análise de agrupamento é um conjunto de técnicas estatísticas que identificam subgrupos ou classes distintas de indivíduos mutuamente excludentes com base nas similaridades existentes entre os indivíduos, ou seja, os mais semelhantes pertencem ao mesmo grupo. Em particular, no teste de SCOTT-KNOTT esses conceitos são aplicados após a análise de variância (ANOVA), quando da rejeição da hipótese nula, isto é, quando existe diferença entre médias. Feita a análise, buscam-se métodos que explicam a variável resposta desses tratamentos de maneira mais detalhada. Vários procedimentos podem ser utilizados, dentre eles: contrastes de médias, comparações múltiplas ou técnicas aglomerativas como teste de SCOTT-KNOTT e a técnica AID - (*Automatic Interaction Detector*).

O teste proposto por Scott e Knott (1974) tem a finalidade de separar as médias em grupos distintos, por meio da minimização da variação dentre os grupos e maximização da variação entre grupos. Os testes de comparações múltiplas apresentam, algumas vezes, resultados de difícil interpretação, por exemplo, os tratamentos A e B são iguais entre si, bem como os tratamentos B e C são iguais entre si. Entretanto, os tratamentos A e C são diferentes. Segundo Morgan e Sonquist (1963), a técnica AID é uma técnica de agrupamento de médias que se baseia na análise de variância para segmentar as observações em grupos distintos, para os quais podem ser desenvolvidos modelos de previsão causais.

As técnicas de agrupamento de médias podem não ser adequadas na aplicação em dados com distribuição diferente da normal. A extensão do algoritmo AID, desenvolvido por Barreto (1993), para modelos lineares generalizados adota como

medida de homogeneidade de grupos uma estatística baseada na função desvio que tem distribuição assintótica qui-quadrado.

Entretanto, usando métodos de computação intensiva, tem sido possível apresentar soluções mais satisfatórias, principalmente quando as estatísticas envolvidas não têm forma fechada para sua distribuição. Em particular, o método *bootstrap* tem como idéia principal fazer reamostragens a partir da amostra original. Essa técnica é especialmente útil para lidar com problemas estatísticos que envolvem amostras de tamanho pequeno e/ou estimadores cuja distribuição (exata ou assintótica) ainda não foi obtida. A reamostragem é feita com reposição do conjunto de dados original, como se estivesse amostrando da população. Repete-se esse procedimento um número  $B$  grande de vezes e, para cada reamostra, calculam-se as quantidades de interesse. Os  $B$  valores calculados são utilizados para estimar a distribuição desconhecida dessa quantidade.

Por outro lado, a função poder indica a probabilidade de se rejeitar a hipótese nula,  $H_0$ , quando ela é falsa. Pode-se interpretar como a chance de detectar uma real diferença entre as médias, ou seja, detectar a diferença, caso ela realmente exista.

Propomos, neste trabalho, usar a reamostragem *bootstrap* a partir de um algoritmo desenvolvido especialmente para a situação e obter a curva poder simulada do teste *bootstrap*. O estudo da curva poder simulada do teste *bootstrap* e da curva poder simulada do teste assintótico permite comparar ambos os testes e verificar aquele que é mais poderoso. Considera-se, assim, um delineamento inteiramente ao acaso, com tamanho da amostra binomial  $n_b$ , com  $R$  tratamentos e  $n_r$ ,  $r=1...R$ , repetições por tratamento.

Ao iniciar o capítulo 2, apresentam-se alguns conceitos sobre análise de agrupamento, entre eles o teste de SCOTT-KNOTT e a técnica AID, além de vantagens da utilização do teste SCOTT-KNOTT em relação aos testes de comparações múltiplas.

A extensão do método de SCOTT-KNOTT em modelos lineares generalizados é enfocada no capítulo 3. Estudos e resultados relativos à extensão do algoritmo AID complementam o capítulo.

O capítulo 4, por sua vez, trata exclusivamente do procedimento de reamostragem *bootstrap*, descrevendo de maneira detalhada a obtenção do erro padrão *bootstrap*, o problema da comparação de duas amostras e o teste de hipótese *bootstrap* para duas amostras.

No capítulo 5 apresentam-se as propostas de reamostragem *bootstrap* para obtenção do algoritmo do teste *bootstrap* e traz, ainda, o planejamento de estudo via simulação para determinação do tamanho da simulação para o teste *bootstrap* e o teste assintótico e uma comparação entre o teste *bootstrap* e o assintótico com um conjunto de dados reais.

Os resultados alcançados por meio deste estudo e considerando um delineamento inteiramente ao acaso, com tamanho da amostra binomial  $n_b=1, 10$  e  $100$ , com  $R=3, 5, 8$  e  $12$  tratamentos e  $n_r=5$  e  $10$  repetições por tratamento, a obtenção das curvas poder para o teste *bootstrap* e o teste assintótico, além de alguns comentários, são encontrados no capítulo 6.

Finalmente, o capítulo 7 apresenta as principais conclusões extraídas deste estudo, acompanhadas de algumas propostas para estudos futuros.

## Capítulo 2

# Análise de Agrupamento

A análise de agrupamento é um conjunto de técnicas estatísticas que identificam classes ou subgrupos distintos de indivíduos mutuamente excludentes com base nas similaridades existentes entre os indivíduos, ou seja, de maneira que os mais semelhantes pertençam ao mesmo grupo. Quando o conjunto de dados é composto por uma variável resposta e diversas variáveis explicativas categorizadas ou tratamentos, essas técnicas podem ser aplicadas após a análise de variância (ANOVA), quando da rejeição da hipótese nula, ou seja, quando existem diferenças entre médias. Feita a análise, buscam-se métodos que expliquem a variável resposta desses tratamentos de maneira mais detalhada. Vários procedimentos podem ser utilizados, entre eles: contrastes de médias, comparações múltiplas ou técnicas aglomerativas, como o teste de SCOTT-KNOTT e a técnica AID.

O teste proposto por Scott e Knott (1974) é utilizado após a rejeição da igualdade de médias em uma análise de variância e tem a finalidade de separar as médias em grupos distintos, por meio da minimização da variação dentro os grupos e maximização da variação entre grupos. Os testes de comparações múltiplas apresentam, algumas vezes, resultados de difícil interpretação, por exemplo, os tratamentos A e B são iguais entre si, bem como os tratamentos B e C são iguais entre si. Entretanto, os tratamentos A e C são diferentes. Como a técnica AID utilizada no teste SCOTT-KNOTT procura grupos homogêneos, essa situação não é encontrada.

Segundo Morgan e Sonquist (1963), a técnica AID – (*Automatic Interaction Detector*) é uma técnica de agrupamento de médias, que se baseia na análise de variância para segmentar as observações em grupos distintos para os quais podem ser desenvolvidos modelos de previsão causais. A partir da técnica AID outros métodos foram desenvolvidos, entre eles ID3 – *Interactive Dichotomizer 3*, C4.5, CHAID – *Chi-square Automatic Interaction Detection*, CART – *Classification and Regression Trees* e

QUEST – *Quick Unbiased Efficient Statistical Tree*, segundo Rodrigues (2005). Neste trabalho, é abordada somente a técnica AID e o teste de SCOTT-KNOTT, nos quais a extensão do algoritmo AID é baseada.

Na seção 2.1, é abordado o teste de SCOTT-KNOTT, na seção 2.2, são apresentadas as principais diferenças entre os testes de comparações múltiplas em relação ao SCOTT-KNOTT e, na seção 2.3, é introduzida a técnica - AID (Automatic Interaction Detection).

## 2.1 - Teste de SCOTT-KNOTT

O teste de SCOTT-KNOTT corresponde a uma análise de agrupamentos univariada que usa soma de quadrados como medida de distância entre grupos. Assim, após a rejeição da hipótese de igualdade de médias no teste  $F$ , o teste de SCOTT-KNOTT propõe a aplicação sucessiva do procedimento descrito abaixo, até que uma regra de parada seja acionada.

Tendo por base o modelo

$$Y_{rj} = \mu + \tau_r + \varepsilon_{rj}, \begin{cases} r = 1, \Lambda, R \\ j = 1, \Lambda, n_r \end{cases} \quad (2.1)$$

onde,

$Y_{rj}$  é a  $j$ -ésima resposta no  $r$ -ésimo tratamento;

$\mu$  é o efeito comum;

$\tau_r$  é efeito do  $r$ -ésimo tratamento;

$\varepsilon_{rj}$  é o erro aleatório, com distribuição  $N(0, \sigma^2)$ ;

$R$  é o número total de tratamentos;

$n_r$  é o número de repetições por tratamento.

o teste de SCOTT-KNOTT usa uma estatística da razão de verossimilhança obtida sob a hipótese de que os dois subgrupos de uma partição são iguais.

De acordo com Scott e Knott (1974), Silva (1998), Santos (2000) e Borges (2002), quando o número de repetições por tratamento é constante,  $n_r=n$ , o procedimento de SCOTT-KNOTT pode ser executado utilizando-se os seguintes passos:

- (i) Ordenam-se as médias dos tratamentos para se obter todas as partições possíveis dos tratamentos em dois subgrupos. A expressão usual do número de partições é  $R-1$ , onde  $R$  é o número de tratamentos considerados.
- (ii) Determina-se todas as somas de quadrado ( $B_{0i}$ ) entre dois subgrupos das  $R-1$  partições possíveis pela expressão

$$B_{0i} = \frac{T_{1i}^2}{r_{1i}} + \frac{T_{2i}^2}{r_{2i}} - \frac{(T_{1i} + T_{2i})^2}{r_{1i} + r_{2i}} \quad \text{com } i=1,2,\dots,R-1 \quad (2.2)$$

onde,

$B_{0i}$  é a soma de quadrado da  $i$ -ésima partição;

$T_{1i}$  é a soma total dos valores das médias dos tratamentos pertencentes ao subgrupo 1 da  $i$ -ésima partição;

$T_{2i}$  é a soma total dos valores das médias dos tratamentos pertencentes ao subgrupo 2 da  $i$ -ésima partição;

$r_{1i}$  é o número de tratamentos do subgrupo 1 da  $i$ -ésima partição;

$r_{2i}$  é o número de tratamentos do subgrupo 2 da  $i$ -ésima partição.

Seja  $B_0$  a maior soma de quadrados entre dois subgrupos de médias,

$$B_0 = \max_{i \in \{1, K, R-1\}} B_{0i};$$

- (iii) Obtém-se o valor da estatística  $\lambda$  do teste pela seguinte expressão

$$\lambda = \frac{\pi}{2(\pi - 2)} * \frac{B_0}{\hat{\sigma}_0^2} \quad (2.3)$$

onde  $\pi$  é constante de valor 3.141593;

$\hat{\sigma}_0^2$  é dado por:

$$\hat{\sigma}_0^2 = \frac{\left[ \sum_{r=1}^R \left( \bar{y}_r - \bar{y} \right)^2 + vQME \right]}{(R + v)}$$

sendo:

$\bar{y}_r$  é a média do tratamento  $r$ ;

$\bar{y}$  é a média geral de todos os tratamentos;

$QME$  é o quadrado médio do erro;

$R$  é o número de tratamentos considerados;

$v = R(n-1)$  é o número de graus de liberdade do erro;

$n$  é o número de repetições por tratamento.

A estatística da razão de verossimilhança  $\lambda$  tem, sob a hipótese de que todas as médias dos subgrupos da partição em questão são iguais, uma distribuição aproximadamente  $\chi^2$  com  $v_0 = \frac{R}{\pi - 2}$  graus de liberdade, segundo Scott e Knott (1974).

- (iv) Se  $\lambda > \chi_\alpha^2(v_0)$ , onde  $\chi_\alpha^2(v_0)$  corresponde ao quantil superior  $\alpha$  de uma distribuição qui-quadrado com  $v_0$  graus de liberdade, rejeita-se a hipótese de que os dois subgrupos são iguais em favor da hipótese alternativa de que os mesmos são diferentes entre si;
- (v) Repetem-se os passos de (i) a (iv) até não se rejeitar a hipótese de formação de novas partições dentro dos subgrupos considerados.

O teste de SCOTT-KNOTT corresponde a um método de agrupamento divisivo hierárquico como a técnica AID.

A escolha do  $\alpha$  apropriado deve levar em conta o número de divisões que serão feitas. Segundo Scott e Knott (1974), de uma maneira geral, para  $j$  grupos homogêneos

em algum estágio, a probabilidade de se ter pelo menos uma divisão significativa é de, no máximo,  $\alpha^* = 1 - (1 - \alpha)^J$ .

Para melhor esclarecimento do processo, seja um experimento com médias ordenadas **A1, A2, A3, A4, A5, A6** que serão divididas em dois grupos distintos. Pela simplificação, temos  $R-1=5$  partições das médias ordenadas, em 2 grupos distintos. As possíveis partições são:

1ª partição: <b>A1</b>	<i>versus</i>	<b>A2, A3, A4, A5, A6</b>
2ª partição: <b>A1, A2</b>	<i>versus</i>	<b>A3, A4, A5, A6</b>
3ª partição: <b>A1, A2, A3</b>	<i>versus</i>	<b>A4, A5, A6</b>
4ª partição: <b>A1, A2, A3, A4</b>	<i>versus</i>	<b>A5, A6</b>
5ª partição: <b>A1, A2, A3, A4, A5</b>	<i>versus</i>	<b>A6</b>

Suponha que a 4ª partição produziu o maior valor  $B_0$ . A seguir, obtém-se o valor da estatística  $\lambda$ , verificando sua significância, ou seja, se  $\lambda > \chi^2_\alpha(v_0)$  a um nível  $\alpha$  pré-fixado. Se foi significativa, repete-se o processo nos subgrupos formados. Considera-se que a 4ª partição é a melhor, ou seja, **A1, A2, A3, A4 vs A5, A6**. Toma-se, então, o primeiro subgrupo **A1, A2, A3, A4** e formam-se novos subgrupos da seguinte forma:

1ª partição: <b>A1</b>	<i>versus</i>	<b>A2, A3, A4</b>
2ª partição: <b>A1, A2</b>	<i>versus</i>	<b>A3, A4</b>
3ª partição: <b>A1, A2, A3</b>	<i>versus</i>	<b>A4</b>

Busca-se, novamente, a partição que maximiza  $B_0$  e obtém-se o novo valor da estatística  $\lambda$ , verificando sua significância, ou seja, se  $\lambda > \chi^2_\alpha(v_0)$  a um nível  $\alpha$  pré-fixado. Se foi significativa, repete-se o processo nos subgrupos formados até que não tenha mais subgrupos a serem formados. Agora, procede-se à segunda partição do subgrupo **A5, A6**, se a partição **A5 vs A6** for significativa, caso contrário, **A5** e **A6** pertencerão ao mesmo grupo. Podemos considerar, como um exemplo, que temos ao final 3 grupos: (**A1, A2, A3**), (**A4**) e (**A5, A6**).

O teste de SCOTT-KNOTT é de fácil aplicação, sendo um método hierárquico, no qual os grupos em qualquer estágio são relacionados com aqueles dos estágios prévios.

No mesmo artigo, Scott e Knott (1974) apresenta uma modificação, denominada “teste de SCOTT-KNOTT modificado”, que permite uma pequena variação quanto ao critério para a determinação da partição entre grupos. Nesse caso, os itens (i) e (ii) do procedimento SCOTT-KNOTT são feitos de modo análogo, sendo calculada, no item (iii), a estatística  $\lambda$  para todas as  $R-1$  partições, e não apenas para a partição em que ocorre a maior soma de quadrados entre grupos ( $B_0$ ). Escolhe-se a partição em que a estatística  $\lambda$  teve o menor valor significativo, ou seja, o menor valor de  $\alpha$  tal que,  $\lambda > \chi^2_{\alpha}(v_0)$ . Prossegue-se de forma semelhante nos subgrupos formados por esta partição, até não haver significância da estatística  $\lambda$  de divisão dos grupos, ou seja,  $\lambda < \chi^2_{\alpha}(v_0)$ .

## **2.2 - Principais Diferenças entre os Testes de Comparações Múltiplas com Relação ao SCOTT-KNOTT**

O método de comparação múltipla é uma das estratégias mais utilizadas pelos pesquisadores por estar disponível na maioria dos *softwares* estatísticos. Entretanto, muitas vezes, os pesquisadores têm muitas dificuldades ao aplicar os testes, para interpretar os resultados, principalmente quando há um número grande de tratamentos e seu interesse está em uma separação real de grupos de médias, sem ambigüidade nos resultados (FERREIRA *et. al*, 1999). Com o objetivo de eliminar essas ambigüidades, utiliza-se o método aglomerativo de SCOTT-KNOTT, que possui poder elevado e taxas de erro tipo I quase sempre de acordo com os níveis nominais, segundo Silva (1998).

Segundo Borges (2002), deve-se ficar atento, porque existem situações em que se aplica o procedimento de comparações múltiplas de forma inadequada, principalmente em análises de experimentos agrônômicos conduzidos em esquema fatorial, no qual cada nível de um fator é combinado com os demais níveis dos outros fatores. Em experimentos desse tipo, independentemente de os fatores envolvidos serem qualitativos ou quantitativos, devem ser primeiramente testadas as significâncias dos

efeitos principais e das interações. Se a interação não for significativa, então toda informação está contida nos efeitos principais, sendo que as médias de cada nível dos fatores podem ser comparadas recorrendo a métodos mais adequados. Comparar as médias dos efeitos principais, ignorando a existência de possíveis interações significativas, pode conduzir a conclusões incorretas.

Alguns cuidados devem ser tomados ao selecionar um procedimento estatístico, para que o mesmo seja apropriado para os dados que estão sendo analisados. O uso incorreto dos testes de comparações múltiplas pode resultar em perda de informação e em redução da eficiência, quando procedimentos mais apropriados são avaliados (BORGES, 2002).

A principal diferença entre os testes de comparações múltiplas e o teste de SCOTT-KNOTT é que este último não faz mais comparações de médias duas a duas e, sim, comparações entre grupos de médias.

O estudo realizado por Ferreira *et. al*, (1999) caracteriza essa vantagem pois, a partir de um conjunto de dados abordados em Von Pinho *et. al*, (1997), verificou-se claramente que o teste SCOTT-KNOTT apresentou uma real separação dos grupos de médias, que em nenhum caso foi maior do que sete separações. Por outro lado, os testes de TUKEY e DUNCAN apresentaram até 15 letras do alfabeto na descrição das médias de tratamentos gerando, com isso, ambigüidades nos resultados.

### **2.3 - Técnica - AID (*Automatic Interaction Detection*)**

A técnica AID (*Automatic Interaction Detection*), proposta por Morgan e Sonquist (1963) como uma técnica de análise de agrupamentos, consiste em dividir o conjunto de dados sucessivamente em grupos, combinando ou não as classes das variáveis preditoras ou níveis dos fatores, de modo a produzir grupos que se diferenciem ao máximo com respeito à média da variável resposta. O processo de subdivisão é aplicado enquanto as divisões obtidas contribuem para explicar a variação dos dados e, ao final, tem-se um conjunto de grupos que podem ser identificados por alguns níveis dos fatores ou variáveis preditoras e suas interações mais importantes.

Segundo Bussab (1979), ao se procurar grupos em um conjunto de dados, uma solução seria examinar todas as partições possíveis do conjunto de dados e escolher aquela que é ótima segundo algum critério. Na prática, esse procedimento de busca pode se reduzir a um número limitado de grupos diminuindo o grande número de alternativas que deveríamos examinar.

Por esse motivo, podemos adotar o método hierárquico divisivo, que consiste nos seguintes passos:

- i. No início, todas as classes da variável preditora estão em um único grupo que será dividido em dois, de acordo com um critério de partição ótima;
- ii. Em seguida, um dos dois grupos resultantes é escolhido de acordo com um critério de seleção para ser dividido em dois grupos;
- iii. Os passos anteriores são repetidos sucessivamente até que um número final de grupos seja alcançado ou algum outro critério de parada seja satisfeito.

Em geral, para cada problema temos um conjunto de regras que determinam as partições. Pode-se dizer que a regra de parada pode ser classificada como livre se nenhuma condição é imposta às partições, ou então como restrita, caso certas condições sejam levadas em consideração. Dentre as restrições, se os grupos formados tiverem elementos consecutivos com relação a alguma ordem fixada no conjunto  $1, 2, \dots, R$  de índices das observações, a regra é monotônica. Se, por outro lado, os grupos são formados por outro tipo de restrição, então a regra é complexa.

Seja  $T(I, \gamma)$  o conjunto das partições possíveis do conjunto de índices  $\{I=1, 2, \dots, R\}$ , usando a regra  $\gamma$  de partição dicotômica. O algoritmo desenvolve-se por meio de divisões dicotômicas nos tratamentos marginais. Seja  $I(X_j) = \{1, 2, \dots, r_j\}$ ,  $j=1, \dots, p$  o conjunto de índices marginais criados pelo preditor  $X_j$  sobre todo o conjunto  $I$ . Então,  $\psi(I(X_j), \gamma_j)$  é o conjunto de todas as possíveis partições dicotômicas do conjunto  $I$  com a regra marginal  $\gamma_j$ . Logo, a classe de todas as partições dicotômicas possíveis do conjunto  $I$  no algoritmo AID é dada por

$$\Psi(I, \gamma) = \prod_{j=1}^p \psi(I(X_j), \gamma_j)$$

Uma vez estabelecido o objetivo de obter-se grupos homogêneos internamente, é de interesse imediato saber que partições levam ao agrupamento mais homogêneo. Podemos dizer que o problema é semelhante à classificação cruzada múltipla em análise de variância, uma vez que a técnica aglomerativa tenta agrupar tratamentos que têm médias muito próximas. Como cada tratamento é determinado de maneira única pelos níveis dos fatores, os grupos são estabelecidos por meio das relações entre a variável resposta e os fatores. No caso específico de análise de variância com um fator, quando as  $n$  observações estão em um único grupo, a soma de quadrados total,  $SQT$ , é considerada como medida de homogeneidade, ou mesmo como medida de dispersão, uma vez que representa a variabilidade de todas as observações em relação à média geral  $\bar{y}$ . A  $SQT$  pode ser decomposta em duas somas de quadrados,  $SQE$  e  $SQTrat$ , que são denominadas, respectivamente, variação entre tratamentos e variação dentro de tratamentos. Esta última representa uma medida de ajuste do modelo de médias através da medida de distância ou discrepância entre os valores observados e os valores ajustados. Ela é usada como medida de homogeneidade na técnica AID e é uma das opções para se construir o critério de partição ótima.

No próximo capítulo, estudaremos a extensão do algoritmo AID, na qual se baseia o método de SCOTT-KNOTT, desenvolvido por Barreto (1993), Barreto e Demétrio (1995;1998) para modelos lineares generalizados (McCULLAGH e NELDER, 1989), adotando como medida de homogeneidade de grupos uma estatística baseada na função desvio ou *deviance*.

## Capítulo 3

# Uma extensão do método SCOTT e KNOTT em modelos lineares generalizados

As técnicas de agrupamento de médias podem não ser adequadas à aplicação em dados com distribuição diferente da normal. Neste capítulo, apresenta-se uma extensão do algoritmo *Automatic Interaction Detection* (AID), no qual se baseia o método de SCOTT-KNOTT, desenvolvido por Barreto (1993), Barreto e Demétrio (1995;1998), para modelos lineares generalizados (McCULLAGH e NELDER,1989), adotando como medida de homogeneidade de grupos uma estatística baseada na função desvio. Essa extensão pode ser aplicada a uma grande variedade de modelos que possuem diversas variáveis preditoras, tais como de regressão e de análise de variância, de regressão logística, log-lineares, dentre outros.

Na seção 3.1, são descritos os modelos lineares generalizados e suas propriedades; na seção 3.2, uma extensão do algoritmo AID em modelos lineares generalizados e, finalmente, na seção 3.3, apresenta-se um exemplo da aplicação da extensão do algoritmo AID em um conjunto de dados reais.

### 3.1 Modelos lineares generalizados

Em delineamentos de experimentos, muitas vezes, percebemos que não é razoável supor que a variável resposta tem distribuição normal. Em 1972, McCullagh e Nelder (1989) propuseram um modelo linear mais geral, que permite que a variável resposta pertença a alguma distribuição da família exponencial.

Essa generalização incorporou diversos modelos já existentes, dentre eles, modelo de médias, modelo de regressão múltipla, regressão logística, modelos loglineares e alguns modelos de sobrevivência.

De acordo com McCullagh e Nelder (1989), Paula (2000) ou Demétrio (2002), o modelo linear generalizado é composto por 3 partes distintas:

1) **Componente sistemático** é uma função linear das variáveis regressoras ou explicativas, denominado preditor linear, análoga à definida em modelos de regressão, análise de variância, análise de covariância, regressão logística e modelos loglineares;

2) **Componente aleatório** é estabelecido a partir da variável resposta de interesse,  $Y_{ij}$ , que pode ser contínua ou discreta, pertencendo a distribuições da família exponencial;

3) **Função de ligação** faz a conexão entre o componente aleatório e o componente sistemático, ou seja, relaciona a média da variável resposta  $Y_{ij}$ ,  $\mu_i$ , ao preditor linear, isto é,

$$\eta_i = g(\mu_i),$$

sendo  $g(\cdot)$  uma função monótona, derivável.

A densidade de probabilidade da variável resposta  $Y_{ij}$  para variáveis contínuas ou a probabilidade funcional para respostas discretas pode ser expressa como

$$f(y) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (3.1)$$

para algumas funções  $a$ ,  $b$ , e  $c$  que determinam a distribuição específica. Para  $\phi$  ou  $a(\phi)$  fixado, a distribuição é da família exponencial de um parâmetro. As funções  $a$  e  $c$  são tais que  $a(\phi) = \frac{\phi}{w}$  e  $c = c\left(y, \frac{\phi}{w}\right)$ , onde  $w$  é um peso conhecido para cada observação.

Se  $\mu$  representa a média de  $Y$ , então a variância é expressa como uma função da média e é dada da seguinte forma;

$$\mu = E(Y) = b'(\theta)$$

$$\sigma^2 = Var(Y) = a(\phi)b''(\theta)$$

onde  $b'(\theta)$  é referente à derivada primeira e  $b''(\theta) = d\mu/d\theta$  é referente à derivada segunda em função de  $\mu$ , e é representada por  $V(\mu)$ . Logo, a variância de  $Y$  pode ser escrita como:

$$Var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$$

Tem-se que  $\lambda(\theta, \phi; y) = \ln f_Y(y; \theta, \phi)$  é o logaritmo da função de verossimilhança considerado como uma função de  $\theta$  e de  $\phi$ , dado  $y$ . A média e a variância da variável aleatória  $Y$  podem também ser obtidas facilmente a partir das relações conhecidas (Demétrio, 2002):

$$E(U) = E\left[\frac{d\lambda}{d\theta}\right] = 0$$

e

$$Var(U) = E(U^2) = E(-U') = E\left[-\frac{d^2\lambda}{d\theta^2}\right]$$

sendo  $U$  chamada função score.

Portanto, usando (3.1)

$$\lambda = \frac{1}{a(\phi)}[y\theta - b(\theta)] + c(y; \phi),$$

tem-se

$$U = \frac{d\lambda}{d\theta} = \frac{1}{a(\phi)} [y - b'(\theta)]$$

e

$$U' = \frac{d^2\lambda}{d\theta^2} = \frac{-1}{a(\phi)} b''(\theta).$$

Logo,

$$E(U) = \frac{1}{a(\phi)} [E(Y) - b'(\theta)] = 0 \rightarrow E(Y) = b'(\theta)$$

e

$$\left. \begin{aligned} Var(U) &= -E(U') = \frac{1}{a(\phi)} b''(\theta) \\ Var(U) &= E(U^2) = \frac{1}{[a(\phi)]^2} Var(Y) \end{aligned} \right\} \rightarrow Var(Y) = a(\phi) b''(\theta)$$

Na Tabela 1 estão identificadas as funções  $a$ ,  $b$ ,  $c$ ,  $\mu(\theta)$  e  $V(\mu)$ , os parâmetros  $\theta$  e  $\phi$  para as distribuições normal, binomial e Poisson.

**Tabela 1:** Identificadores da família exponencial para algumas distribuições.

Distribuição	$a(\phi)$	$\theta$	$b(\theta)$	$c(y; \phi)$	$\mu(\theta)$	$V(\mu)$
<b>Normal</b> $N(\mu, \sigma^2)$	$\sigma^2$	$\mu$	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[ \frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right]$	$\theta$	$1$
<b>Poisson</b> $P(\mu)$	$1$	$\ln \mu$	$e^\theta$	$-\ln y!$	$e^\theta$	$\mu$
<b>Binomial</b> $B(m, \pi)^*$	$1$	$\ln \left( \frac{\pi}{1-\pi} \right)$	$m \ln(1 + e^\theta)$	$\ln \binom{m}{y}$	$m \frac{e^\theta}{1+e^\theta}$	$\frac{1}{m} \mu(m-\mu)$

\*  $m\pi = \mu = E(Y)$

Para uma amostra aleatória simples, o logaritmo da função de verossimilhança é parametrizado em termos das médias  $\mu_i$ , do parâmetro de dispersão  $\phi$ , da variável resposta  $y_i$ , e do peso à priori  $w_i$  e é expresso por:

$$L(y, \mu, \phi) = \sum_i \log(f(y_i, \mu, \phi))$$

onde o somatório refere se a todas as observações da amostra.

Segundo Demétrio (2002), o ajuste de um modelo linear generalizado é determinado pelo vetor  $\hat{\beta}$  de estimativas dos parâmetros, o usando o método da máxima verossimilhança.

O logaritmo da função de verossimilhança para um conjunto de observações independentes  $y_1, \dots, y_n$  é dado pela soma das contribuições individuais, isto é,

$$\lambda = \lambda(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \lambda(\theta_i; y_i) = \sum_{i=1}^n \left\{ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right\}$$

sendo que  $E(Y_i) = \mu_i$ ,  $\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$  e  $\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i)$ .

Uma propriedade da família exponencial de distribuições diz respeito ao fato de seus elementos satisfazerem as condições de regularidade suficientes para assegurar que o máximo global do logaritmo da função de verossimilhança  $\lambda(\boldsymbol{\theta}; \mathbf{y})$  seja dado unicamente pela solução do sistema de equações  $\mathbf{U}_{\boldsymbol{\theta}} = \frac{d\lambda}{d\boldsymbol{\theta}} = \mathbf{0}$ , temos que a função escore é dada por

$$U_j = \sum_{i=1}^n \frac{\partial \lambda(\theta_i; y_i, \phi)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \lambda_i}{\partial \beta_j}$$

mas,

$$\begin{aligned}
\lambda &= f(\theta_1, \theta_2, \Lambda, \theta_i, \Lambda, \theta_n) \\
&\downarrow \\
\theta_i &= \int V_i^{-1} d\mu_i = q(\mu_i) \\
&\downarrow \\
\mu_i &= g^{-1}(\eta_i) = h(\eta_i) \\
&\downarrow \\
\eta_i &= \sum_{j=1}^p x_{ij} \beta_j
\end{aligned}$$

e, pela regra da cadeia,

$$\begin{aligned}
U_j &= \sum_{i=1}^n \frac{d\lambda_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\
&= \sum_{i=1}^n \frac{1}{a_i(\phi)} [y_i - b'(\theta_i)] \frac{1}{\frac{d\mu_i}{d\theta_i}} \frac{d\mu_i}{d\eta_i} x_{ij} \quad (3.2)
\end{aligned}$$

Temos, que

$$E(Y_i) = \mu_i = b'(\theta_i)$$

e

$$Var(Y_i) = a_i(\phi) \frac{d\mu_i}{d\theta_i} = a_i(\phi) V(\mu_i) \Rightarrow \frac{d\mu_i}{d\theta_i} = V(\mu_i)$$

logo

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{1}{a_i(\phi)} [y_i - \mu_i] \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij} \quad (3.3)$$

Em geral, as equações  $U_j=0$ ,  $j=1,2,\dots,p$ , não são lineares e têm que ser resolvidas numericamente por processos iterativos. O método iterativo de Newton-Raphson, para a solução de uma equação  $f(x)=0$ , é baseado na aproximação de Taylor para a função  $f(x)$  nas vizinhanças do ponto  $x_0$ , ou seja,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) = 0$$

obtendo

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

ou, de uma forma mais geral

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}$$

sendo  $x^{(m+1)}$  o valor de  $x$  no passo  $(m+1)$ ,  $x^{(m)}$  o valor de  $x$  no passo  $m$ ,  $f(x^{(m)})$  a função  $f(x)$  avaliada em  $x^{(m)}$  e  $f'(x^{(m)})$  a derivada da função  $f(x)$  avaliada em  $x^{(m)}$ .

Considerando que desejamos obter a solução do sistema de equações

$$\mathbf{U}_\beta = \frac{d\lambda}{d\beta} = \mathbf{0}, \quad (3.3),$$

e usando a versão multivariada do método de Newton-Raphson

$$\beta^{(m+1)} = \beta^{(m)} + (\mathbf{I}_0^{-1})^{(m)} \mathbf{U}^{(m)}$$

sendo  $\beta^{(m+1)}$  e  $\beta^{(m)}$  os vetores de parâmetros estimados nos passos  $m$  e  $(m+1)$ ,

$\mathbf{U}^{(m)}$  o vetor escore, isto é, o vetor de derivadas parciais de 1ª ordem de  $f(x)$ , com elementos  $\frac{\partial \lambda}{\partial \beta_j}$ , avaliado no passo  $m$  e  $(\mathbf{I}_0^{-1})^{(m)}$  a inversa da negativa da matriz de

derivadas parciais de 2ª ordem de  $f(x)$ , com elementos  $\frac{-\partial^2 \lambda}{\partial \beta_j \partial \beta_k}$ , avaliada no passo  $m$ .

Quando as derivadas de 2ª ordem são obtidas facilmente, o método de Newton-Raphson é bastante útil. Acontece, porém, que isso nem sempre ocorre e, no caso dos

modelos lineares generalizados, usa-se o método escore de Fisher que, em geral, é mais simples. Esse método envolve a substituição da matriz de derivadas parciais de 2ª ordem pela matriz de valores esperados das derivadas parciais, isto é, a substituição da matriz de informação observada,  $\mathbf{I}_0$ , pela matriz de informação esperada de Fisher,  $\mathfrak{I}$ . Logo,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{I}^{-1})^{(m)} \mathbf{U}^{(m)} \quad (3.4)$$

sendo que  $\mathfrak{I}$  tem elementos dados por  $\mathfrak{I}_{jk} = E\left[\frac{-\partial^2 \lambda}{\partial \beta_j \partial \beta_k}\right] = E\left[\frac{\partial \lambda}{\partial \beta_j} \frac{\partial \lambda}{\partial \beta_k}\right]$ , que é a matriz de covariâncias dos  $U_j$ 's.

Multiplicando ambos os lados de (3.4) por  $\mathfrak{I}^{(m)}$  tem-se,

$$\mathfrak{I}^{(m)} \boldsymbol{\beta}^{(m+1)} = \mathfrak{I}^{(m)} \boldsymbol{\beta}^{(m)} + \mathbf{U}^{(m)}$$

Assim, usando (3.3), obtemos

$$\begin{aligned} \mathfrak{I}_{jk} = E(U_j U_k) &= \sum_{i=1}^n \frac{1}{[a_i(\phi)]^2} E(Y_i - \mu_i)^2 \frac{1}{[V(\mu_i)]^2} \left[\frac{d\mu_i}{d\eta_i}\right]^2 x_{ij} x_{ik} \\ &= \sum_{i=1}^n \frac{1}{[a_i(\phi)]^2} a_i(\phi) V(\mu_i) \frac{1}{[V(\mu_i)]^2} \left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij} x_{ik} \\ &= \sum_{i=1}^n \frac{1}{a_i(\phi)} \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{ij} x_{ik} \end{aligned}$$

e fazendo  $a_i(\phi) = \frac{\phi}{w_i}$ , com  $\phi > 0$ , constante,  $w_i$  peso *a priori* e  $w_i = \frac{w_i}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2$ ,

tem-se

$$\mathfrak{I} = \frac{1}{\phi} \mathbf{X}^T \mathbf{W} \mathbf{X}$$

com elementos  $\mathfrak{S}_{jk} = \sum_{i=1}^n \frac{1}{\phi} x_{ij} W_i x_{ik}$ ,  $\mathbf{X}$  a matriz do modelo e  $\mathbf{W} = \text{diag}(W_1, \dots, W_n)$ . No

caso das funções de ligação canônicas  $W_i = w_i V(\mu_i)$ , pois  $\frac{d\theta_i}{d\mu_i} = \frac{d\eta_i}{d\mu_i} = V^{-1}(\mu_i)$ .

Além disso, re-arranjando os termos de  $U_j$  (3.2) tem-se

$$U_j = \sum_{i=1}^n \frac{w_i}{\phi} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} = \sum_{i=1}^n \frac{1}{\phi} x_{ij} W_i \frac{d\eta_i}{d\mu_i} (y_i - \mu_i)$$

Após vários cálculos, que podem ser encontrados em Demétrio (2002), tem-se

$$\boldsymbol{\beta}^{(m+1)} = \left( \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)} \quad (3.5)$$

que tem a forma da solução das equações normais, para o modelo linear obtida pelo método dos quadrados mínimos ponderados, exceto que nesse caso a solução  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$  é obtida por processo numérico iterativo. É importante observar que a expressão (3.5) independe de  $\phi$ .

O algoritmo proposto por Nelder e Wedderburn (1972) para ajuste de modelos lineares generalizados corresponde à obtenção de estimativas de máxima verossimilhança dos parâmetros  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$  no preditor  $\boldsymbol{\eta}$  por mínimos quadrados ponderados iterativos. O algoritmo de estimação pode ser resumido da seguinte forma:

1. Obter as estimativas  $\eta_i^{(m)} = \sum_{j=1}^p x_{ij} \beta_j^{(m)}$  e  $\mu_i^{(m)} = g^{-1}(\eta_i^{(m)})$ ;

2. Obter a variável dependente ajustada  $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)})$  e os

$$\text{pesos } W_i^{(m)} = \frac{w_i}{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2};$$

3. Calcular (3.5), voltar ao passo (1) com  $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m+1)}$  e repetir o processo até convergência, obtendo  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$ .

Se  $l(\tilde{y}, \tilde{y})$  é o logaritmo da função de verossimilhança expressa como uma função dos valores preditos da média  $\tilde{\mu}$ , e o vetor  $\tilde{y}$  de valores da variável resposta, então, a função desvio (*scaled deviance*) é definida por

$$D^*(\tilde{y}, \hat{\tilde{\mu}}) = 2 \left[ l(\tilde{y}, \tilde{y}) - l(\hat{\tilde{\mu}}, \tilde{y}) \right] \quad (3.6)$$

Considerando a expressão (3.6), a função desvio pode ser reescrita por:

$$\begin{aligned} D^*(\tilde{y}; \hat{\tilde{\mu}}) &= 2 \left[ l(\tilde{y}; \tilde{y}) - l(\hat{\tilde{\mu}}; \tilde{y}) \right] \\ &= \sum_{i=1}^n 2w_i \left\{ y_i \left( \tilde{\theta}_i - \hat{\theta}_i \right) - \left[ b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right] \right\} / \phi \\ &= D(\tilde{y}; \hat{\tilde{\mu}}) / \phi \end{aligned}$$

onde  $\phi$  é o parâmetro de dispersão;

$D(\tilde{y}; \hat{\tilde{\mu}})$  é chamado desvio (*deviance*) para o modelo sob estudo;

$l(\tilde{y}; \tilde{y})$  é o logaritmo da função de verossimilhança, calculado no ponto em que os valores ajustados são iguais às observações;

$l\left(\hat{\mu}; \tilde{y}\right)$  é o logaritmo da função de verossimilhança, calculado no ponto de máxima verossimilhança;

$\tilde{\theta} = \theta\left(\tilde{y}\right)$  é a estimativa de máxima verossimilhança do parâmetro canônico sob o modelo completo, isto é, que considera um parâmetro para cada observação;

$\hat{\theta} = \theta\left(\hat{\mu}\right)$  é a estimativa de máxima verossimilhança do parâmetro canônico sob o modelo de interesse.

O número de graus de liberdade associados ao desvio é definidos por,  $U = n - p$ , onde  $n$  é o número de observações e  $p$  é o número de parâmetros linearmente independentes no modelo.

Segundo Demétrio (2002), a qualidade do ajuste de um modelo linear generalizado é avaliada por meio da função desvio e da estatística de  $X^2$  de Pearson generalizada, cuja expressão é dada por:

$$X^2 = \sum_{i=1}^n w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

sendo  $V(\hat{\mu}_i)$  a função de variância estimada sob o modelo que está sendo ajustado aos dados.

Para dados provenientes das distribuições binomial e Poisson em que  $\phi=1$  e  $X^2$  é a estatística original de Pearson, comumente usada na análise dos modelos logístico e loglinear para tabelas multidimensionais, que pode ser descrita na forma

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

sendo  $O_i$  a frequência observada e  $E_i$  a frequência esperada.

Para as distribuições não-normais, temos apenas resultados assintóticos, isto é, a distribuição  $\chi_{n-p}^2$  pode ser usada somente como uma aproximação que, em muitos casos, pode ser pobre. Além disso,  $\chi^2$  tem como desvantagem o fato de tratar os  $y_i$ 's simetricamente.

### 3.2 - Uma extensão do algoritmo AID em modelo lineares generalizados

A extensão da técnica AID em modelos lineares generalizados possibilita uma nova versão do teste de SCOTT-KNOTT para agrupamentos de níveis de fatores, quando da rejeição de sua igualdade. Ao se procurar grupos homogêneos em um conjunto de dados, uma alternativa seria examinar todas as partições possíveis do conjunto de dados e escolher aquela que é ótima segundo algum critério. O algoritmo AID reduz o número de partições a serem examinadas, calculando a estatística de teste em todas as partições quando os tratamentos estão ordenados por seu valor esperado. O valor máximo dessa estatística é usado no teste assintótico, para verificar se os dois grupos da partição são estatisticamente diferentes.

Considerando o modelo de um ensaio completamente ao acaso de  $R$  tratamentos com  $n_r$  repetições por tratamentos  $r=, \dots, R$ , o desvio do modelo associado é expresso por:

$$D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_{1+R}\right) = \sum_{r=1}^R \sum_{j=1}^{n_r} 2w_{rj} \left[ y_{rj} (\tilde{\theta}_{rj} - \hat{\theta}_r) - b(\tilde{\theta}_{rj}) + b(\hat{\theta}_r) \right]$$

ou pode ser reescrito da seguinte forma:

$$D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_{1+R}\right) = 2 \left\{ \sum_{r=1}^R \sum_{j=1}^{n_r} y_{rj} \tilde{\theta}_{rj} - r \sum_{r=1}^R n_r \bar{y}_r \hat{\theta}_r \right\}$$

onde  $\tilde{\theta}_{rj} = \theta(y_{rj})$  e  $\tilde{\theta}_{rj} = \theta(\hat{\mu}_{rj})$  são as estimativas dos parâmetros canônicos. O

número de graus de liberdade associado é  $\sum_{r=1}^R n_r - R$ .

O desvio do mesmo ensaio, considerando todas as médias iguais entre si,  $D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_1\right)$ , tem  $\sum_{r=1}^R n_r - 1$  graus de liberdade, e é expresso por:

$$D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_1\right) = \sum_{r=1}^R \sum_{j=1}^{n_r} 2w_{rj} \left[ y_{rj} (\tilde{\theta}_{rj} - \hat{\theta}) - b(\tilde{\theta}_{rj}) + b(\hat{\theta}) \right]$$

No caso específico de partição dicotômica do conjunto de índices  $I = \{1, 2, \dots, R\}$ , segundo a regra  $\gamma$ , cada partição  $d$  forma dois conjuntos de índices de tratamentos constituídos por  $J(d)$  e  $L(d)$ . Seja  $D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_d\right)$  o desvio do modelo de dois grupos definidos pela partição  $d$ , expresso por:

$$D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_d\right) = \sum_{r \in J(d)} \sum_{j=1}^{n_r} 2w_{rj} \left[ y_{rj} (\tilde{\theta}_{rj} - \hat{\theta}_{J(d)}) - b(\tilde{\theta}_{rj}) + b(\hat{\theta}_{J(d)}) \right] + \sum_{r \in L(d)} \sum_{j=1}^{n_r} 2w_{rj} \left[ y_{rj} (\tilde{\theta}_{rj} - \hat{\theta}_{L(d)}) - b(\tilde{\theta}_{rj}) + b(\hat{\theta}_{L(d)}) \right] \quad (3.7)$$

onde  $\hat{\theta}_{J(d)}$  é a estimativa do parâmetro canônico, quando a  $r$ -ésimo tratamento pertence ao conjunto  $J(d)$ , e  $\hat{\theta}_{L(d)}$ , quando pertence ao conjunto  $L(d)$ .

Com relação ao desvio (3.7), que é uma medida de discrepância obtida pela partição  $d$ , temos que

$$D\left(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}_d\right) = 2 \left\{ \sum_{r=1}^R \sum_{j=1}^{n_r} y_{rj} \tilde{\theta}_{rj} - n_{J(d)} \bar{y}_{J(d)} \hat{\theta}_{J(d)} - n_{L(d)} \bar{y}_{L(d)} \hat{\theta}_{L(d)} \right\} \quad (3.8)$$

A expressão (3.8) é, então, a discrepância medida em termos do desvio de um modelo de um ensaio completamente ao acaso em que as médias dos tratamentos formam dois grupos definidos pela partição  $d$ .

Barreto (1993), Barreto e Demétrio (1995;1998) propõem usar a diferença

$$\Delta\left(\begin{matrix} \hat{\mu}_d \\ \sim \end{matrix}\right) = D\left(\begin{matrix} y, \hat{\mu}_1 \\ \sim \end{matrix}\right) - D\left(\begin{matrix} y, \hat{\mu}_d \\ \sim \end{matrix}\right) \quad (3.9)$$

como uma medida de homogeneidade de dois grupos, uma vez que  $D\left(\begin{matrix} y, \hat{\mu}_d \\ \sim \end{matrix}\right)$  é uma medida de discrepância do modelo de dois grupos. Quanto menor o desvio  $D\left(\begin{matrix} y, \hat{\mu}_d \\ \sim \end{matrix}\right)$ , mais heterogêneos serão os dois grupos. Desse modo, quanto maior  $\Delta\left(\begin{matrix} \hat{\mu}_d \\ \sim \end{matrix}\right)$ , mais heterogêneos serão os grupos formados pela partição  $d$ .

Para a escolha da divisão mais promissora, Barreto (1993), Barreto e Demétrio (1995; 1998) propõem usar a estatística  $\Delta\left(\begin{matrix} \hat{\mu}_d \\ \sim \end{matrix}\right)$  (3.9). Assim, no conjunto de todas as partições possíveis, escolhe-se aquela que produz a maior redução  $\Delta\left(\begin{matrix} \hat{\mu}_d \\ \sim \end{matrix}\right)$ , ou seja,

$$\Delta\left(\begin{matrix} \hat{\mu}_{d_0} \\ \sim \end{matrix}\right) = \underset{d}{\text{máx}} \left\{ \Delta\left(\begin{matrix} \hat{\mu}_d \\ \sim \end{matrix}\right) : d \text{ entre as partições possíveis} \right\} \quad (3.10)$$

A partição mais fina aceitável das observações é aquela definida para cada tratamento formando um grupo, enquanto a mais elementar é aquela que corresponde a um único grupo. Um caso intermediário é aquele em que os  $R$  tratamentos estão divididos em  $G$  grupos.

O desvio do modelo de  $G$  grupos,  $D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_{1+G}\right)$ , varia entre o desvio do modelo que corresponde à participação mais fina com  $R$  grupos,  $D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_{1+R}\right)$ , e o desvio do modelo correspondente à partição mais rudimentar com um único grupo,  $D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_1\right)$ .

Analogamente, a estatística  $R^2$  generalizada dada por

$$R^2(G) = \frac{D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_1\right) - D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_{1+G}\right)}{D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_1\right)}$$

corresponde, então, à proporção do desvio explicada pelo modelo de  $G$  grupos com relação ao modelo mais rudimentar, como  $0 \leq R^2(1) \leq R^2(G) \leq R^2(R) \leq 1$ . Uma forma reduzida no intervalo  $[0,1]$ , denominada coeficiente de determinação é dada por

$$R_*^2(G) = \frac{D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_1\right) - D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_{1+G}\right)}{D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_1\right) - D\left(\underset{\sim}{y}; \underset{\sim}{\hat{\mu}}_{1+R}\right)} = \frac{R^2(G)}{R^2(R)}$$

que podemos interpretar como a proporção padronizada do desvio, explicada na formação de tais modelos.

Logo, a extensão do algoritmo AID em um modelo linear generalizado (BARRETO, 1993) é descrita por meio da definição dos seguintes elementos:

**a) Critérios de seleção:**

- a<sub>1</sub>) Escolher como grupo a ser dividido aquele que produz a maior redução de  $\Delta\left(\hat{\mu}_d\right)$ ;
- a<sub>2</sub>) Escolher como grupo a ser dividido aquele que produz, no passo seguinte, a maior redução de  $\Delta\left(\hat{\mu}_d\right)$ .

**b) Critérios de partição ótima:**

- b<sub>1</sub>) Aquela que produz a maior redução de  $\Delta\left(\hat{\mu}_d\right)$ ;
- b<sub>2</sub>) Aquela que produz a redução  $\Delta\left(\hat{\mu}_d\right)$  com o menor nível descritivo.

**c) Regras de parada:**

- c<sub>1</sub>) O número de elementos nos grupos resultantes é menor que algum número pré-fixado;
- c<sub>2</sub>) O número total de divisões não alcança um certo máximo;
- c<sub>3</sub>) O nível descritivo do máximo da redução  $\Delta\left(\hat{\mu}_d\right)$  não é maior que um valor pré-fixado;
- c<sub>4</sub>) O valor de  $R_*^2$  alterou-se menos do que um valor pré-fixado em relação à divisão imediatamente anterior.

Tanto na adoção do critério de partição ótima ( $b_2$ ) como na adoção da regra de parada ( $c_3$ ), é importante o conhecimento da distribuição da estatística  $\Delta\left(\hat{\mu}_{d_0}\right)$ , que é o máximo entre as reduções  $\Delta\left(\hat{\mu}_d\right)$  possíveis.

No caso da distribuição normal e regra de divisão livre, Scott e Knott (1974) verificaram que a distribuição assintótica de (3.10) é proporcional a uma distribuição qui-quadrado com  $R/(\pi - 2)$  graus de liberdade e parâmetro de proporcionalidade  $2(\pi - 2)/\pi$ . Quando o número de tratamentos é menor do que 10, Bussab (1976) verificou que essa aproximação é boa, principalmente quando o número de observações por tratamento é o mesmo.

No caso da distribuição binomial, Barreto (1993), Barreto e Demétrio (1995) verificaram, por simulação, que o máximo da redução de  $\Delta\left(\hat{\mu}_d\right)$  também apresenta distribuição aproximadamente proporcional a uma distribuição qui-quadrado com  $R/(\pi - 2)$  graus de liberdade e parâmetro de proporcionalidade igual a  $\pi / 2(\pi - 2)$ . Esses resultados foram verificados para  $R$  variando de 3 a 10, sendo que a qualidade da aproximação é maior conforme  $R$  aumenta. Resultados análogos são apresentados por Barreto e Demétrio (1998) no caso da distribuição Poisson.

### **3.3 - Exemplo da aplicação da extensão do algoritmo AID em um experimento com couve chinesa (*Brassic pekinensis*)**

O experimento foi instalado e conduzido em casa de vegetação, em vasos de 8  $dm^3$ , com uma planta por vaso, no Setor de Olericultura e Plantas Aromático-Medicinais, que pertence ao Departamento de Produção Vegetal de FCAV – UNESP, Jaboticabal – SP, no período de novembro de 2001 a fevereiro de 2002, conforme relatado por Presotti (2002). Com o objetivo de avaliar quatro doses de potássio (100, 150, 200 e 250 mg de *potássio* por  $dm^3$  de solo) e três híbridos (*Komachi (K)*, *Taibyó, (T)* e *Nangoku, (N)*) de couve-chinesa, num esquema fatorial, que aqui é considerado como 12 tratamentos com 6 repetições cada, os sintomas visuais de deficiência de potássio nas folhas foram observados e anotadas, sua presença (1) ou ausência (0). Os dados e as médias por tratamentos estão descritos na Tabela 2.

**Tabela 2:** Resultados da presença (1) ou ausência (0) do sintoma visual de deficiência de potássio nas folhas de couve chinesa (*Brassica pekinensis*) e as médias dos tratamentos.

Tratamentos (Doses; híbridos)	Repetições						Médias
	I	II	III	IV	V	VI	
1 (100,K)	1	1	1	1	1	1	6/6
2 (100,T)	0	1	1	1	1	1	5/6
3 (100,N)	0	1	0	0	0	0	1/6
4 (150,K)	1	1	1	1	1	1	6/6
5 (150,T)	1	0	1	0	0	0	2/6
6 (150,N)	0	0	0	0	0	0	0/6
7 (200,K)	1	1	1	1	1	1	6/6
8 (200,N)	0	0	0	1	1	0	2/6
9 (200,T)	0	0	0	0	0	0	0/6
10 (250,K)	0	1	1	1	0	1	4/6
11 (250,T)	0	0	0	1	0	0	1/6
12 (250,N)	0	0	0	0	0	0	0/6

A análise do desvio correspondente ao modelo que supõe distribuição binomial para a variável resposta, presença/ausência de deficiência de potássio, e a função de ligação logística encontram-se na Tabela 3.

**Tabela 3:** Análise do desvio da deficiência de potássio em um delineamento com 12 tratamentos e 6 repetições.

Fonte de variação	DF	Deviance	p-valor
Tratamento	11	60.18	<.0,001
Resíduos	60	39.13	

Como existe diferença entre tratamentos devido ao p-valor se inferior a uma  $\alpha=0,05$ , utilizou-se a extensão do algoritmo AID em modelos lineares generalizados para se fazer o agrupamento de médias, considerando dois critérios diferentes.

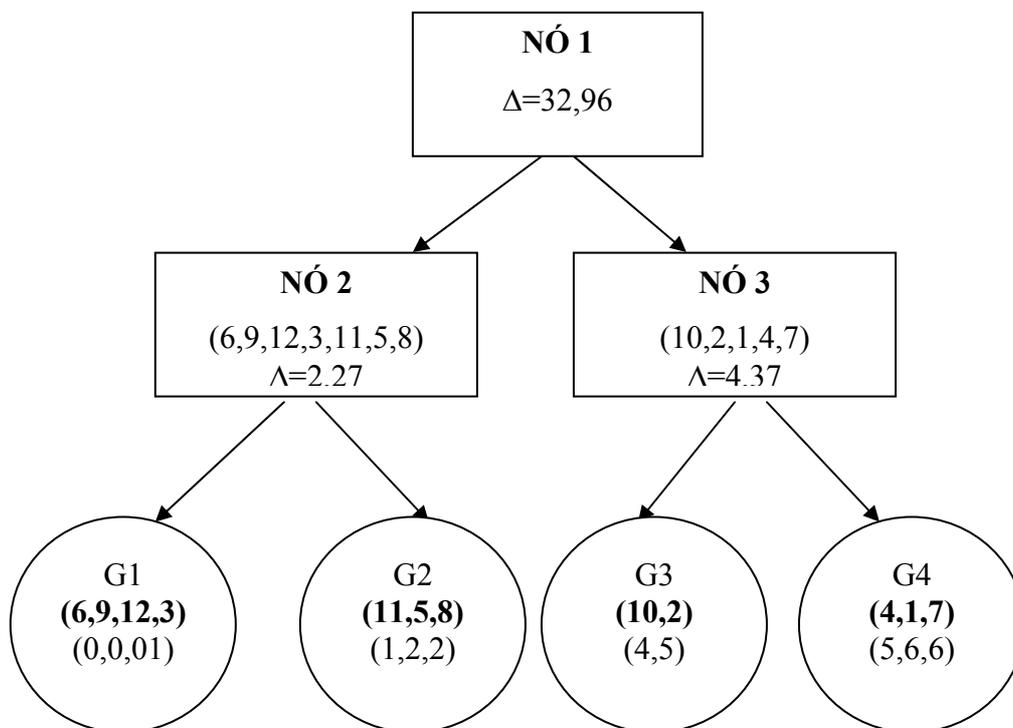
Primeiramente, o algoritmo de divisão dicotômica foi aplicado considerando como critério de partição ótima a que produz a maior redução de  $\Delta\left(\hat{\mu}_d\right) (a_l)$ ,

aquela que produz a maior redução de  $\Delta\left(\hat{\mu}_d\right) (b_i)$  e como regra de parada a obtenção de 4 grupos finais ( $c_2$ ). Os principais elementos para a construção da árvore encontram-se na Tabela 4 e a árvore na Figura 1.

**Tabela 4:** Resumo dos resultados do algoritmo de divisão dicotômica considerando como critério de seleção a maior redução, como critério de partição ótima a que produz a maior redução, e como regra de parada a obtenção de 4 grupos.

NÓ	$\frac{2(\pi - 2)\Delta_0}{\pi}$	$\frac{R}{(\pi - 2)}$	Grupos
<b>1</b>	32,96	10,51	{6,9,12,3,11,5,8} e {10,2,1,4,7}
<b>2</b>	2,27	6,13	{6,9,12,3} e {11,5,8}
<b>3</b>	4,37	4,38	{10,2} e {1,4,7}

O algoritmo começa pelo nó (1) com a apresentação da estatística corrigida da partição que apresentou a maior redução no passo seguinte, o número de graus de liberdade associados e a partição dada pelos tratamentos {6,9,12,3,11,5,8} e {10,2,1,4,7}. A seguir, examinando os nós (2) e (3), esse último é escolhido por apresentar a maior redução: Até esse passo temos 3 grupos: {6,9,12,3,11,5,8}, {10,2} e {1,4,7}. O algoritmo prossegue examinando o nó (2), com isso, neste passo, tem-se os 4 grupos formados: {6,9,12,3}, {11,5,8}, {10,2} e {1,4,7}.



**Figura 3** - Árvore de divisão dicotômica, considerando como partição ótima a maior redução e como regra de parada a obtenção de 4 grupos.

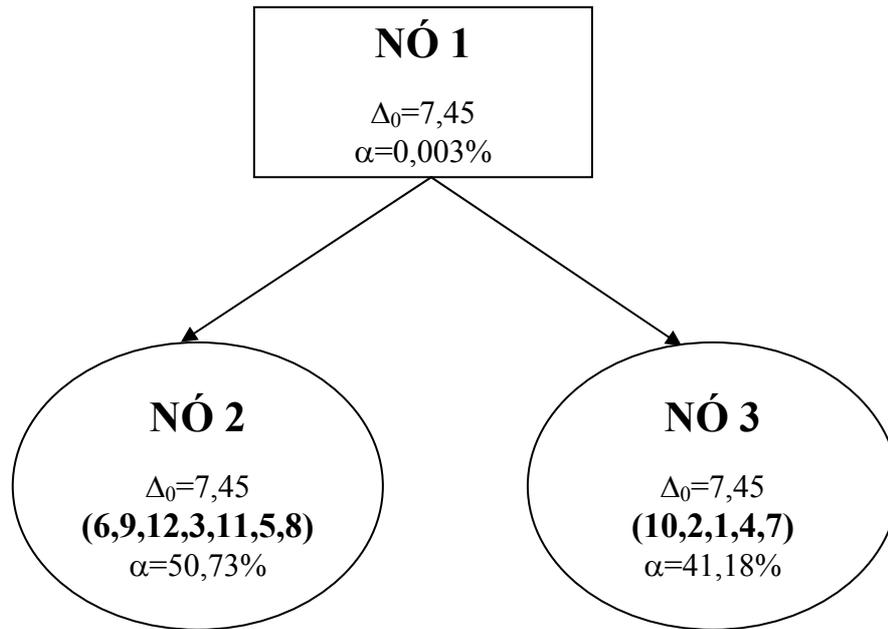
Com relação à estatística  $R_*^2(G)$ , a aplicação  $(43,36 + 3,13 + 6,01) / 60,18 = 87,24\%$  de explicação do modelo.

Uma outra forma de aplicar o algoritmo de divisão dicotômica nos resultados do experimento com couve chinesa é, por exemplo, utilizar como critério de seleção a maior redução de  $\Delta\left(\hat{\mu}_d\right)(a_1)$ , como critério de partição ótima a que produz a partição com o menor nível descritivo ( $b_2$ ) e, como regra de parada, o nível descritivo do máximo da redução  $\Delta\left(\hat{\mu}_d\right)$  não é maior que 1% ( $c_3$ ). Na Tabela 5, estão os principais elementos para a construção da árvore cuja forma final está na Figura 4.

**Tabela 5:** Resumo dos resultados do algoritmo de divisão dicotômica, considerando como critério de seleção ( $a_1$ ), como critério de partição ótima ( $b_2$ ) e, como regra de parada, o nível descritivo do máximo da redução não é maior que 1% ( $c_3$ ).

NÓ	$\Delta_0$	$\frac{R}{(\pi - 2)}$	p-valor	Grupos
<b>1</b>	4,53	10,51	0,003%	(6,9,12,3,11,5,8) e (10,2,1,4,7)
<b>2</b>	7,45	6,13	50,73%	(6,9,12) e (3,11,5,8)
<b>3</b>	6,01	4,38	41,18%	(10,2) e (1,4,7)

A formação da árvore começou pelo nó (1) da Tabela 5, com o cálculo da estatística  $\Delta_0$  cujo valor foi de 4,53 e o número de graus de liberdade associado é  $\frac{R}{(\pi - 2)}$  igual a 10,51, e o nível descritivo é  $\alpha = 0,003\%$ . Para a escolha da próxima partição dicotômica, calculou-se o valor da estatística  $\Delta_0$  no passo seguinte: para o nó (2), seu valor foi de 7,45 com  $\frac{R}{(\pi - 2)}$  igual a 6,13 graus de liberdade e nível descritivo de 50,73%, e para o nó (3),  $\Delta_0$  igual a 6,01 com  $\frac{R}{(\pi - 2)}$  igual a 4,38 graus de liberdade e nível descritivo de 41,18%. Com esses resultados, chegou-se à conclusão de que esse nó não deve ser dividido. Ao final do algoritmo, obtiveram-se os grupos: {6,9,12,3,11,5,8} e {10,2,1,4,7}. Adotando-se o nível de significância individual de 1% de probabilidade, o nível geral é de  $1 - (1 - 0,01)^2 = 1,99\%$ . Na Figura 2, encontra-se no dendograma correspondente a Tabela 5.



**Figura 2** - Árvore de divisão dicotômica, considerando como partição ótima a maior redução e, como regra de parada, a partição ótima com um nível descritivo superior a um valor pré-fixado  $\alpha=0,01$ .

No Apêndice A, encontra-se o programa referente ao software *R 2.2.0 project – CRAN*, feito especialmente para obtenção desses resultados.

A utilização de teste assintótico pode ficar restrita se os tamanhos de amostras forem menores que os indicados. Por outro lado, métodos de computação intensiva apresentam-se como uma técnica muito útil, por não exigir que os estimadores envolvidos tenham distribuição (exata ou assintótica) conhecida.

No próximo capítulo, apresentam-se uma revisão do método de reamostragem *bootstrap*, detalhando o problema de duas amostras, e o teste de hipótese *bootstrap* para duas amostras, além de uma revisão sobre a função poder.

## Capítulo 4

# Testes usando métodos de computação intensiva

Proposto por Efron em 1979, o método de reamostragem *bootstrap* foi inicialmente utilizado para encontrar o erro padrão de um estimador de um parâmetro. Sendo um método de computação intensiva, apresenta-se como uma técnica muito útil para lidar com problemas estatísticos envolvendo amostras de tamanho pequeno ou moderado, além daqueles que envolvem estimadores cuja distribuição (exata ou assintótica) ainda não foi obtida. O *bootstrap* tem como idéia principal fazer reamostragem com reposição a partir da amostra aleatória simples original.

O teste de hipótese via método reamostragem *bootstrap* pode ser utilizado para o caso de uma ou duas amostras. Neste estudo, utiliza-se a situação de duas amostras,  $\mathbf{z}$  e  $\mathbf{y}$ , que pertencem às distribuições  $F$  e  $G$ , respectivamente, e independentes entre si, e deseja-se testar a hipótese de nulidade  $H_0 : F=G$ . O teste de hipótese *bootstrap* inicia-se com a estatística teste  $\hat{\theta} = \hat{\mu}_F - \hat{\mu}_G$  com a diferença das médias. Por conveniência, assume-se que se a hipótese nula  $H_0$  não é verdadeira, temos valores maiores que  $|\hat{\theta}| = \hat{\mu}_F - \hat{\mu}_G$ , do que quando  $H_0$  é falso.

Por outro lado, a função poder indica a probabilidade de rejeitar a hipótese nula  $H_0$ , quando ela é falsa. Pode-se interpretar como a chance de detectar uma real diferença entre as médias, ou seja, detectar a diferença caso ela realmente exista.

Na seção 4.1, é apresentado o método de reamostragem *bootstrap*, detalhando o cálculo do erro padrão *bootstrap*, a situação de duas amostras e o teste de hipótese *bootstrap*. Já na seção 4.2, aborda-se a função poder *bootstrap*. Os conceitos sobre

bootstrap apresentados neste capítulo são baseados em Davison e Hinkley (1997) e Efron e Tibshirani (1993).

## 4.1 Método de Reamostragem *Bootstrap*

O *bootstrap* é um método de computação intensiva que pode ser utilizado para avaliar a acurácia de estimativas estatísticas. A idéia primeira de Efron (1979) foi fazer reamostragens a partir da amostra original. A reamostragem é feita com reposição do conjunto de dados original, como se estivéssemos amostrando da população. Repete-se esse procedimento um número  $B$  grande de vezes, isto é, onde  $B$  é o número de reamostragens *bootstrap* e, para cada reamostra, calculam-se as quantidades de interesse. Os  $B$  valores calculados são utilizados para estimar a distribuição desconhecida dessa quantidade.

### 4.1.1 Erro Padrão *Bootstrap*

Suponha que se deseja estimar o parâmetro  $\theta = t(F)$  a partir de uma amostra aleatória simples  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  de uma distribuição de probabilidade desconhecida  $F$ . A estimação de  $F$  é obtida por meio da distribuição empírica  $\hat{F}$ , que corresponde a uma distribuição discreta que atribui probabilidade  $1/n$  em cada um dos valores observados de  $x_i, i=1, 2, \dots, n$  indicado por  $\mathbf{x}$ .

Uma forma de estimar  $\theta = t(F)$  é usar a mesma função  $t$  em  $\hat{F}$ . Essa abordagem é conhecida como princípio *plug-in*. Esse princípio é um método simples de estimação de parâmetros a partir da amostra, ou seja, relaciona o estimador com a função empírica  $\hat{F}$ . A estimativa *plug-in* para o parâmetro  $\theta = t(F)$  é definido por

$$\hat{\theta} = t(\hat{F})$$

Em outras palavras, calcula-se a estimativa de  $\theta = t(F)$  como uma função da distribuição de probabilidade  $F$ , pela mesma função de distribuição empírica  $\hat{F}$ ,  $\hat{\theta} = t(\hat{F})$ .

Uma amostra *bootstrap*  $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$  é definida como sendo uma amostra aleatória com reposição de tamanho  $n$  retirada de  $\hat{F}$ . Por exemplo: se  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , uma possível amostra *bootstrap* é dada por  $\mathbf{x}^* = (x_4^*, x_3^*, x_2^*, x_2^*)$ . A notação é

$$\hat{F} \rightarrow (x_1^*, x_2^*, \dots, x_n^*),$$

onde  $\mathbf{x}^*$  não é o conjunto de dados originais de  $\mathbf{x}$ , mas uma versão aleatorizada, ou reamostrada de  $\mathbf{x}$ . Nessa amostra *bootstrap*, calcula-se a replicação *bootstrap* de  $\hat{\theta}$ ,  $\hat{\theta}^* = s(\mathbf{x}^*)$ . Se  $s(\mathbf{x})$  é a média amostral então  $s(\mathbf{x}^*)$  é a média para o conjunto de

$$\text{dados bootstrap } \bar{x}^* = \frac{\sum_{i=1}^n x_i^*}{n}.$$

A estimativa do erro padrão *bootstrap* de  $\hat{\theta}$ ,  $se_b(\hat{\theta})$ , possui uma estimativa *plug-in* usando a distribuição empírica no lugar da distribuição desconhecida  $F$ .

Para se obter uma estimativa *bootstrap* do erro padrão de  $\hat{\theta}$  selecionam-se  $B$  amostras *bootstrap* independentes, calcula-se as correspondentes replicações *bootstrap*  $\hat{\theta}^*$ ,  $\hat{\theta}^*(b)$  e o erro padrão de  $\hat{\theta}$  a partir do desvio padrão empírico das replicações. O erro padrão *bootstrap*,  $\hat{se}_B(\hat{\theta})$ , é calculado como a raiz quadrada da variância calculada entre as estimativas *bootstrap*  $\hat{\theta}^*(b)$ , isto é,

$$\hat{se}_B = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(\bullet)]^2 / (B-1) \right\}^{1/2} \quad (4.1)$$

onde  $\hat{\theta}^*(\bullet) = \sum_{b=1}^B \frac{\hat{\theta}^*(b)}{B}$

O algoritmo 1 é uma descrição mais explícita do procedimento *bootstrap* para calcular o erro padrão .

→ **Algoritmo 1**

1. Seleccionam-se  $B$  amostras *bootstrap* independentes  $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \dots, \mathbf{x}^{*B}$ , cada uma contendo  $n$  elementos, retirados com reposição, a partir do conjunto de dados original  $\mathbf{x}$ ;
2. Calcula-se, em cada amostra *bootstrap*, a estimativa correspondente

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \quad b = 1, \dots, B$$

3. Estima-se o erro padrão  $se_b(\hat{\theta})$  usando as  $B$  replicações, por meio da expressão (4.1).

O número de replicações bootstrap  $B$ , para calcular o erro padrão normalmente está entre 25 e 200. Como  $B \rightarrow \infty$ ,  $\hat{se}_B$  se aproxima da estimativa *plug-in* de  $se_F(\hat{\theta})$ , mostra-se que o limite de  $\hat{se}_B$ , quando  $B$  tende ao infinito, é igual à estimativa *bootstrap* ideal de  $se_F(\hat{\theta})$ , ou seja,  $\lim_{B \rightarrow \infty} \hat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*)$ .

#### 4.1.2 Teste da hipótese *bootstrap* para duas amostras

Sejam  $\mathbf{z} = (z_1, z_2, \dots, z_m)$  uma amostra aleatória simples da distribuição  $F$  de tamanho  $m$  e  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  uma amostra aleatória simples de tamanho  $n$  da distribuição de  $G$ , independentes entre si. Então, os dados observados correspondem ao vetor  $\mathbf{x} = (\mathbf{z}, \mathbf{y})$  de comprimento  $m+n$ .

Sejam  $\hat{F}$  e  $\hat{G}$  as distribuições empíricas baseadas em  $\mathbf{z}$  e  $\mathbf{y}$ , respectivamente. Considerando o modelo de probabilidade  $P$  como um par de distribuições, ou seja,  $P=(F, G)$ , então, a estimativa natural de  $P=(F, G)$  é  $\hat{P} = (\hat{F}, \hat{G})$

Obtido  $\hat{P}$ , a definição de amostra *bootstrap*  $\mathbf{x}^*$  é simplesmente:  $\hat{P} \rightarrow \mathbf{x}^*$ , ou seja,  $P \rightarrow \mathbf{x}$ . Na situação de duas amostras, tem-se  $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*)$  onde:

$$\hat{F} \rightarrow \mathbf{z}^* \quad \text{independente de} \quad \hat{G} \rightarrow \mathbf{y}^*$$

O tamanho das amostras de  $\mathbf{z}^*$  e  $\mathbf{y}^*$  são os mesmos que  $\mathbf{z}$  e  $\mathbf{y}$ , respectivamente. Sendo  $\theta$  um parâmetro dado por  $\theta = \mu_z - \mu_y = E_F(z) - E_G(y)$ , pode-se usar como estimador de  $\theta$  a expressão

$$\begin{aligned} \hat{\theta} &= \hat{\mu}_z - \hat{\mu}_y \\ \hat{\theta} &= \bar{z} - \bar{y} \end{aligned} \tag{4.2}$$

Se  $\hat{\theta}$  é muito maior em módulo que  $\theta$ , então, a diferença entre  $\mu_z$  e  $\mu_y$  em módulo é grande.

A replicação *bootstrap* indicada por  $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*)$  corresponde a uma amostra aleatória simples, com reposição de tamanho  $m+n$  dos elementos do vetor original  $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ .

O erro padrão de  $\hat{\theta}$  pode ser escrito como  $se_p(\hat{\theta})$ , que indica sua dependência do mecanismo de probabilidade desconhecido  $P=(F,G)$ . A estimativa *bootstrap* de  $se_p(\hat{\theta})$  é a estimativa *plug-in*

$$se_{\hat{P}}(\hat{\theta}^*) = \left\{ \text{var}_{\hat{P}}(\bar{z}^* - \bar{y}^*) \right\}^{1/2}$$

O fato de que  $\hat{\theta}^*$  é computado de duas amostras,  $\mathbf{z}^*$  e  $\mathbf{y}^*$ , não afeta a definição (4.2).

Deseja-se testar a hipótese nula  $H_0=F=G$  que não existe nenhuma diferença entre  $F$  e  $G$ .

Se a diferença entre  $F$  e  $G$  for descrita em termos de  $\mu_F$  e  $\mu_G$ , define-se  $H_0: \mu_F = \mu_G$ . Se  $H_0$  é verdadeiro, então, pode-se dizer que não existe diferença entre o comportamento aleatório de  $\mathbf{z}$  ou  $\mathbf{y}$ .

O teste de hipótese *bootstrap*  $H_0$  inicia-se com a estatística de teste  $\hat{\theta}$ , com a diferença das médias. Por conveniência, assume-se que se a hipótese nula  $H_0$  não é verdadeira, esperamos observar valores maiores de  $|\hat{\theta}|$  quando  $H_0$  é falso. Assim, se  $\mu_F > \mu_G$  ou  $\mu_F \leq \mu_G$ , espera-se que  $|\hat{\theta}| = \bar{z} - \bar{y}$  seja grande. Por outro lado, valores grandes de  $\hat{\theta}$ , são uma forte evidência de que  $H_0$  não é verdadeiro.

Tendo observado  $\hat{\theta}$ , o nível de significância alcançado ou nível descritivo do teste (*Achieved Significance Level-ASL*) é definido por

$$ASL = \Pr ob_{H_0} \left\{ \left| \hat{\theta}^* \right| \geq \left| \hat{\theta} \right| \right\}$$

ou seja, é a probabilidade, sob  $H_0$ , de que os valores de  $\left| \hat{\theta} \right|$ , nas reamostragens *bootstrap*, sejam maiores do que o valor em módulo na amostra original e  $\hat{\theta}^*$  corresponde à variável aleatória estudada, em termos da amostra *bootstrap*.

Formalmente, pede-se escolher uma probabilidade  $\alpha$  pequena, como 0.05 ou 0,01, e se rejeita  $H_0$  se o ASL for menor que  $\alpha$ . Se o ASL for maior que  $\alpha$ , então aceita-se  $H_0$ , não existe diferença entre  $F$  e  $G$ .

O algoritmo *bootstrap* para duas amostras é descrito da seguinte forma:

→ **Algoritmo 2**

1. Esboçam-se  $B$  amostras de tamanho  $n+m$  com reposição de  $\mathbf{x}$ . Sejam as  $n$  primeiras observações  $\mathbf{z}^{*b}$  e as  $m$  observações restantes  $\mathbf{y}^{*b}$ , para  $b=1,2,\dots,B$ .
2. Calcula-se  $t(\cdot)$  em cada amostra,  $t(\mathbf{x}^{*b}) = \bar{\mathbf{z}}^{*b} - \bar{\mathbf{y}}^{*b}$ ,  $b = 1, 2, \dots, B$ .
3. Calcula-se  $\hat{ASL}_{boot}$  aproximado por  $\hat{ASL}_{boot} = \# \left\{ t(\mathbf{x}^{*b}) \geq |t_{obs}| \right\} / B$ , onde  $t_{obs} = t(\mathbf{x})$  o valor observado da estatística na amostra inicial.

Testes mais exatos podem ser obtidos usando outras estatísticas, em vez de  $t(\mathbf{x}) = \bar{z} - \bar{y}$ . Por exemplo:

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\bar{\sigma} \sqrt{1/n + 1/m}} \quad (4.3)$$

$$\text{onde } \bar{\sigma} = \left\{ \frac{\left[ \sum_{i=1}^n (z_i - \bar{z})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 \right]}{[n + m - 2]} \right\}^{1/2}.$$

Para as amostras *bootstrap* calculam-se as análogas, neste caso, supõe-se que o desvio padrão é igual nas duas populações. O algoritmo 3 testa a hipótese nula de que as duas populações são idênticas, isto é,  $F = G$ .

Se quiséssemos testar se apenas as médias são iguais, poderíamos usar uma forma análoga à estatística (4.3) para duas amostras, onde supomos que as variâncias são iguais, isto é,

$$t(\mathbf{x}^{*b}) = \frac{\bar{z}^{*b} - \bar{y}^{*b}}{\bar{\sigma}^{*b} \sqrt{1/n + 1/m}}$$

$$\text{onde } \bar{\sigma}^{*b} = \left\{ \frac{\left[ \sum_{i=1}^n (z_i^{*b} - \bar{z}^{*b})^2 + \sum_{i=1}^n (y_i^{*b} - \bar{y}^{*b})^2 \right]}{[n + m - 2]} \right\}^{1/2}.$$

A estatística

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\sqrt{\bar{\sigma}_1^2 / n + \bar{\sigma}_2^2 / m}} \quad (4.4)$$

$$\text{onde } \bar{\sigma}_1^2 = \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{(n-1)}, \quad \bar{\sigma}_2^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(m-1)}.$$

Uma outra versão de teste *bootstrap* para comparação de duas amostras considera os desvios de cada elemento de cada amostra em relação à média comum,  $\bar{\mathbf{x}}$ .

O algoritmo 3 calcula a estatística de teste bootstrap para comparação de duas amostras usando esse conceito.

→ **Algoritmo 3**

1. Sejam  $\hat{F}$  a probabilidade dos pontos  $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, 2, \dots, n$ , e  $\hat{G}$  a probabilidade dos pontos  $\tilde{y}_i = y_i - \bar{y} + \bar{x}, i = 1, 2, \dots, m$ , onde  $\bar{z}$  e  $\bar{y}$  são as médias dos grupos e  $\bar{x}$  é a média da amostra combinada.
2. Formam-se  $B$  conjuntos *bootstrap*  $\mathbf{x}^{*b} = (\mathbf{z}^{*b}, \mathbf{y}^{*b})$  onde  $\mathbf{z}^{*b}$  é uma amostra com reposição de  $\tilde{z}_1, \tilde{z}_2, \dots, \tilde{z}_n$  e  $\mathbf{y}^{*b}$  é uma amostra com reposição de  $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_m$ .
3. Calculam-se  $t(\cdot)$  definido em (4.4) em cada conjunto de dados,
4. O valor aproximado de  $\hat{ASL}_{boot}$  é dado por  $\hat{ASL}_{boot} = \# \left\{ t(\mathbf{x}^{*b}) \geq |t_{obs}| \right\} / B$ , onde  $t_{obs} = t(\mathbf{x})$  é o valor observado da estatística (4.4).

## 4.2 - Função poder *bootstrap*

A função poder indica a probabilidade de se rejeitar a hipótese nula  $H_0$ , dado possíveis valores do parâmetro de interesse, ou seja  $P[\text{rejeitar } H_0 | \theta]$ . Pode-se interpretar como a chance de detectar uma real diferença entre as médias, ou seja, detectar a diferença, caso ela realmente exista.

Neste trabalho, obtêm-se por meio de simulação, as funções poder do teste referente à extensão do algoritmo AID, usando a distribuição assintótica (3.10) e o teste *bootstrap*.

As curvas poder simuladas foram obtidas para as seguintes configurações: o número de tratamentos iguais a  $R$ , com  $n_r$  repetições por tratamentos, sendo os tamanhos de amostras binomiais  $n_b$  iguais a 1,10,100, o parâmetro de interesse indicado por  $\theta = p_1 - p_2$ , onde  $p_1$  corresponde à probabilidade da binomial dos tratamentos que estão no primeiro subgrupo da partição, e  $p_2$ , no segundo subgrupo e o erro do tipo I foi fixado em 5%. Para o teste assintótico e o teste *bootstrap* foram simuladas para os valores  $\theta=0.00,0.05,0.10,0.15,0.20,0.25,0.30,0.35,0.40$  e  $0.45$ , correspondendo a, por exemplo,  $p_1=0.50$  e  $p_2=0.50,0.45,0.40,0.35,0.30,0.25,0.10$  e  $0.05$ .

No próximo capítulo, será apresentada uma proposta de teste *bootstrap* na extensão do algoritmo AID e estudos sobre o tamanho da simulação para o teste *bootstrap* e o teste assintótico.

## Capítulo 5

# Uma proposta do teste *bootstrap* na extensão do algoritmo AID

Como foi visto no capítulo 4, uma das principais vantagens do uso de reamostragem *bootstrap* é o fato de ela não requerer o conhecimento da distribuição da variável de interesse ou do estimador do parâmetro. Trata-se, portanto, de um processo alternativo e eficiente em várias situações, principalmente quando as amostras são de tamanho pequeno.

Neste capítulo, é apresentada uma proposta do teste *bootstrap* na extensão do algoritmo AID, sendo detalhado, na seção 5.1, o desenvolvimento do algoritmo para o teste *bootstrap*, na seção 5.2, o estudo sobre tamanhos das simulações para o teste *bootstrap* e o teste assintótico, e na 5.3, uma comparação entre o teste *bootstrap* e o teste assintótico com conjunto de dados reais.

### 5.1 Desenvolvimento do algoritmo para o teste *bootstrap*.

Utilizando o método de reamostragem *bootstrap* apresentado no capítulo 4, foi elaborado um algoritmo para se obter o teste *bootstrap*. Considerando um experimento com um fator de classificação com  $R$  tratamentos,  $n_r$  repetições por tratamento e tamanho da amostra binomial igual  $n_b$ , em um modelo inteiramente ao acaso, a análise deve ser iniciada com o teste para verificar se os tratamentos são diferentes ou não entre si. Dado que são diferentes, podemos iniciar o processo para a obtenção da primeira partição mais promissora.

O algoritmo para o teste *bootstrap* tem início com a ordenação dos dados pelo seu valor esperado e, em seguida, obtém-se a maior *deviance* entre as partições do

modelo original. Cada reamostra bootstrap é comparada com a deviance do modelo original. No total, são obtidas  $B$  amostras *bootstrap*, o p-valor *bootstrap* corresponde à diferença entre o número de vezes em que a maior *deviance bootstrap* é maior que a *deviance* do modelo original, dividido por  $B$ . A seguir, detalha-se o algoritmo do teste *bootstrap*.

→ **Algoritmo do teste *bootstrap***

- (i) Ordenam-se os tratamentos por seu valor esperado;
- (ii) Após a ordenação, calcula-se a *deviance* dos  $R-1$  modelos correspondentes às partições dicotômicas;
- (iii) Seja  $\Delta_0$  a maior deviance do modelo original;
- (iv) Faz-se uma reamostragem *bootstrap* nos dados originais;
- (v) Ordenam-se os dados reamostrados, por seu valor esperado;
- (vi) Calcula-se a *deviance* dos  $R-1$  modelos correspondentes às partições dicotômicas;
- (vii) Obtém-se a maior *deviance*  $\Delta_0^{(1)}$  da reamostra *bootstrap*;
- (viii) Repetem-se os passos (iv) a (vii) um número de  $B-1$  vezes, obtendo  $B$  valores para  $\Delta_0(\bullet)$  correspondendo à maior *deviance*;
- (ix) Compara-se a maior deviance do modelo original com a maior deviance de cada um dos modelos reamostrados, usando um contador para obter o número de vezes em que  $\Delta_0(\bullet)$  é maior ou igual a  $\Delta_0^{(b)}$ ,  $b=1, \dots, B$ ;
- (x) Obtém-se o p-valor dividindo o número obtido em (ix) por  $B$ .

A partir deste algoritmo, foi realizado o estudo sobre o tamanho da simulação para a obtenção das curvas poder, que é apresentado na seção 5.2.

## 5.2 Determinação do tamanho da simulação para o teste *bootstrap* e o teste assintótico

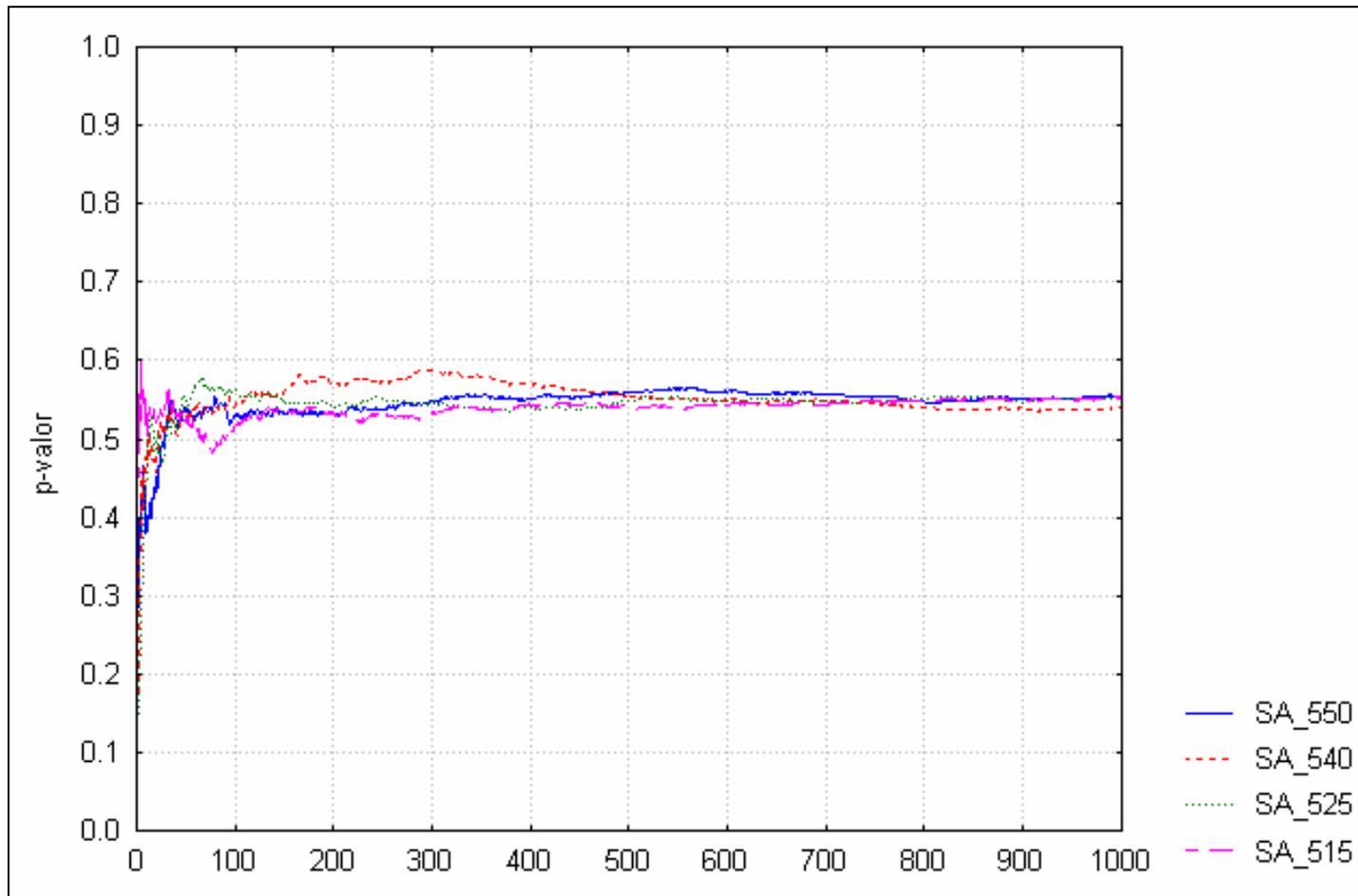
Considerando  $n_b$  o tamanho da amostra binomial,  $R$  o número de tratamentos e  $n_r$  o número de repetições por tratamentos, seja  $\theta$  a diferença entre  $p_1$  e  $p_2$ , onde  $p_1$  corresponde à probabilidade da binomial dos tratamentos que estão no primeiro subgrupo da partição e,  $p_2$ , no segundo subgrupo.

O estudo inicial sobre o tamanho da simulação para a construção da curva poder do teste assintótico consistiu em calcular a probabilidade de rejeitar  $H_0$ , dado que  $\theta = 0,10$ , considerando o tamanho da simulação  $S_A = 2,3,4,\dots,1500$ . Esse procedimento foi repetido 4 vezes com sementes distintas: 515, 525, 540 e 550. O programa para a determinação do tamanho da simulação  $S_A$  programado no *software R 2.2.0 project – CRAN*, encontra-se no Apêndice B. Na Figura 2, encontram-se as 4 curvas simuladas. Pode-se observar que o valor da simulação estabiliza-se a partir de  $S_A$  igual a 900. Em todas as curvas, adotou-se  $S_A = 1000$ .

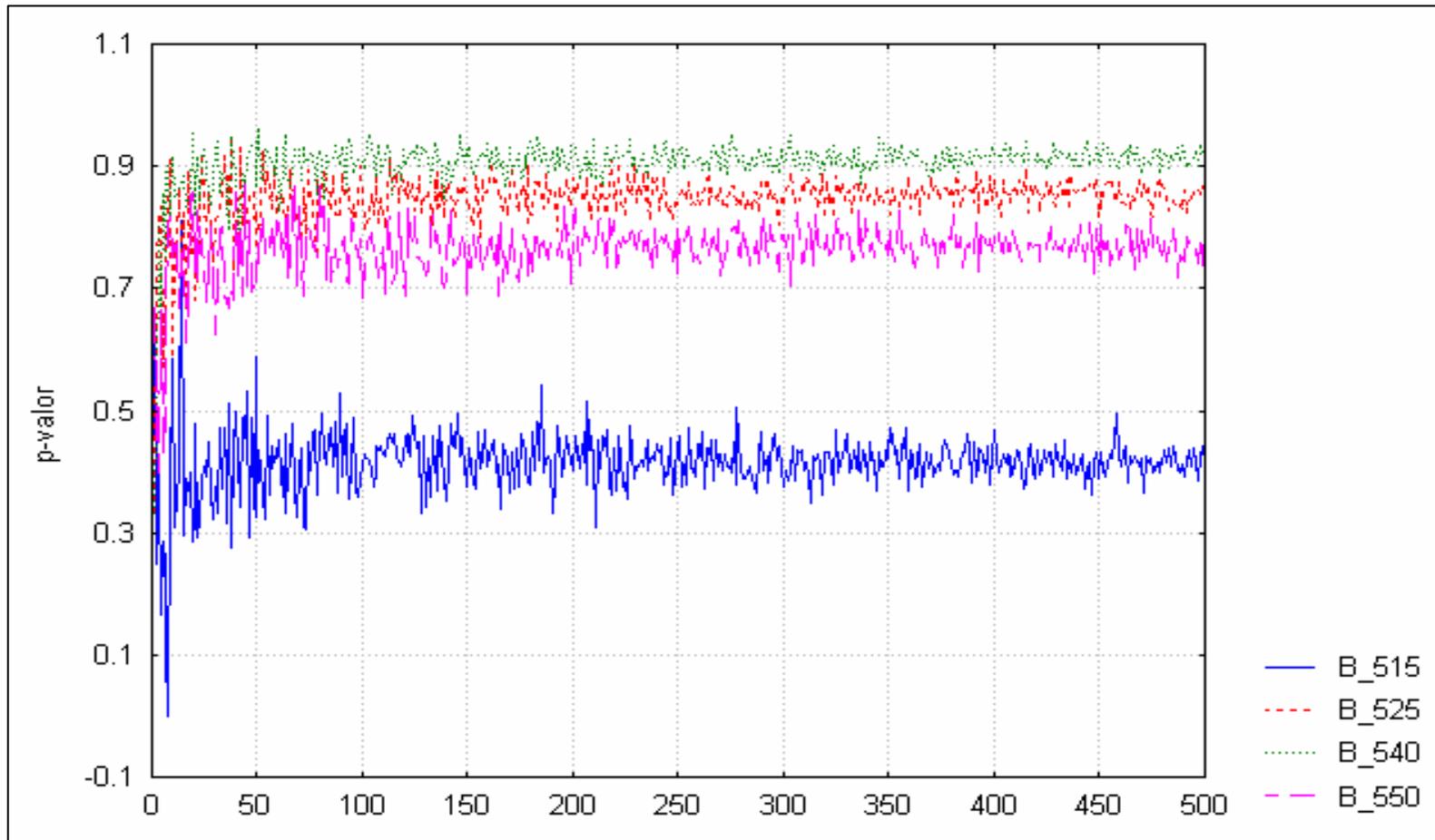
Para determinar o número necessário de reamostragens *bootstrap*, calculou-se a probabilidade de rejeitar  $H_0$ , dado que  $\theta = 0,10$ , considerando o tamanho da simulação  $B=2,3,4,\dots,500$ . De maneira análoga ao teste assintótico, esse procedimento repetiu-se 4 vezes com sementes distintas: 515, 525, 540 e 550. O programa para a determinação do tamanho da simulação  $B$ , programado no *software R 2.2.0 project – CRAN*, encontra-se no Apêndice C. A partir da Figura 3, pode-se observar que as 4 curvas se estabilizam a partir de  $B$  igual a 200. Então, em todas as curvas, adotou-se  $B$  igual a 250.

A partir dos resultados obtidos na reamostragem *bootstrap*, calculou-se o ponto de equilíbrio da simulação do teste *bootstrap*. De maneira análoga ao teste assintótico, a curva poder simulada foi repetida 4 vezes com sementes distintas: 515, 525, 540 e 550, considerando o tamanho da simulação  $S_B=2,3,4,\dots,1500$ .

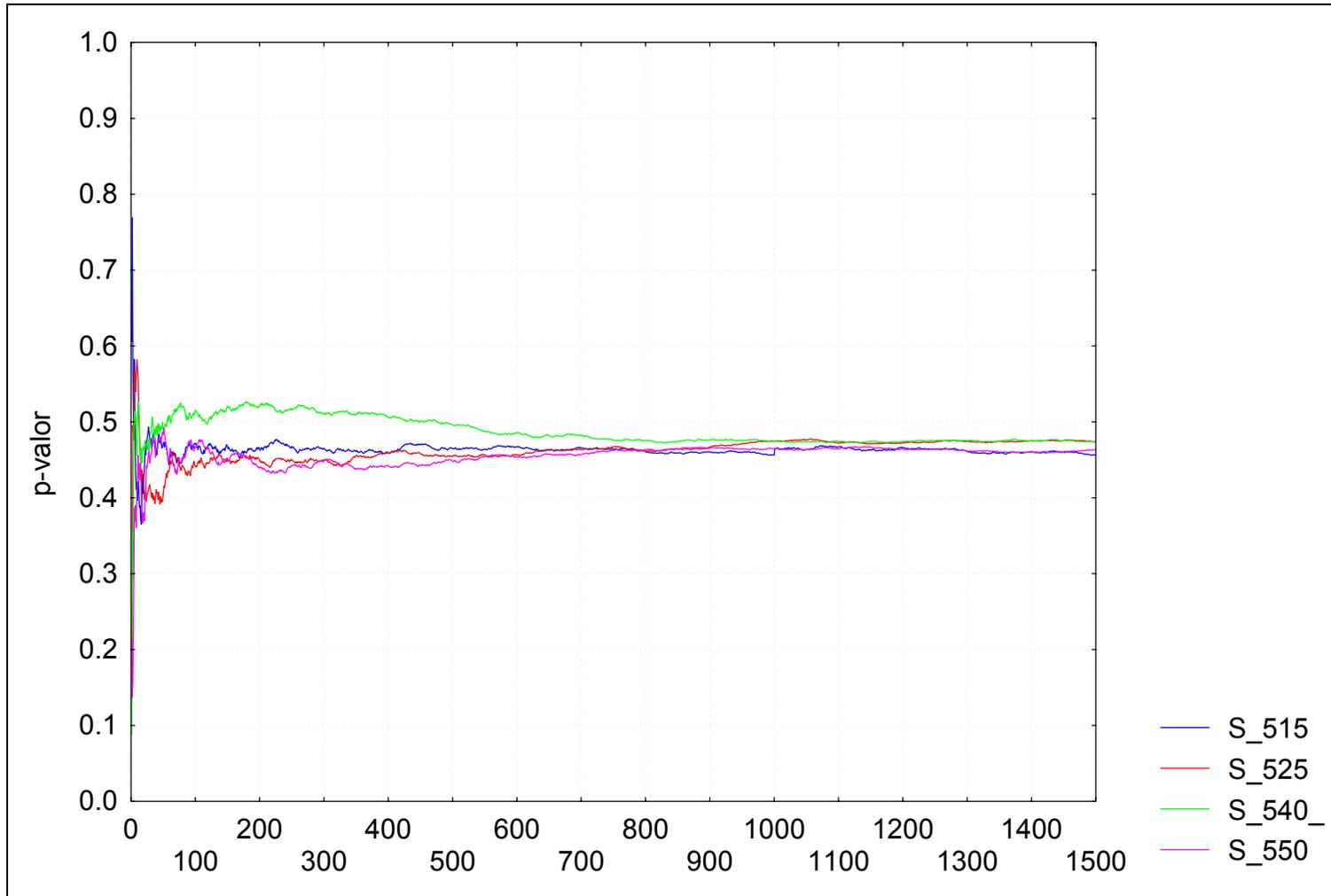
O programa desenvolvido especialmente para a determinação do tamanho do algoritmo do teste *bootstrap*, foi programado no software *R 2.2.0 project -CRAN* e encontra-se no Apêndice D. A partir da Figura 4, pode-se observar que as 4 curvas estabilizam-se a partir de  $S_B$  igual a 950. Então, em toda as curvas, adotou-se  $S_B$  igual a 1000. Tem-se que a notação a partir de agora será  $S = S_A = S_B = \text{máximo}$ , ou seja,  $S$  é o valor máximo obtido para a simulação  $S=1000$ .



**Figura 3** - Gráfico do p-valor médio em função das 1000 simulações para a obtenção do ponto de equilíbrio  $S_A$ .



**Figura 4** - Gráfico referente à simulação com 500 reamostragem *bootstrap* para a obtenção do ponto de equilíbrio  $B$ .



**Figura 5-** Gráfico do p-valor médio em função das 1000 simulações para a obtenção do  $S_B$ .

### 5.3 Exemplo de aplicação do teste *bootstrap* no experimento com couve chinesa.

Na seção 3.3, os resultados do experimento com couve chinesa (Tabela 2) foram analisados usando um modelo linear generalizado para distribuição binomial e função de ligação logística, cuja análise do desvio para a presença da deficiência de potássio encontra-se na Tabela 3.

O segundo exemplo da construção da árvore de divisão dicotômica apresentada na seção 3.3, resumida na Tabela 5 e ilustrada na Figura 5, considerou como critério de seleção ( $a_1$ ), como critério de partição ótima ( $b_2$ ) como regra de parada, o nível descritivo do máximo da redução  $\Delta\left(\hat{\mu}_d\right)$  não é maior que 1% ( $c_3$ ).

Vamos refazer, então, a árvore de divisão dicotômica considerando como regra de parada o nível descritivo do máximo da redução  $\Delta\left(\hat{\mu}_d\right)$  não é maior que 1% usando o teste *bootstrap*.

Na Tabela 8, estão os resultados da construção da árvore de divisão dicotômica usando o teste *bootstrap* no cálculo do nível descritivo do máximo da redução  $\Delta\left(\hat{\mu}_d\right)$  como regra de parada, enquanto que, na Figura 6, está ilustrada a árvore obtida.

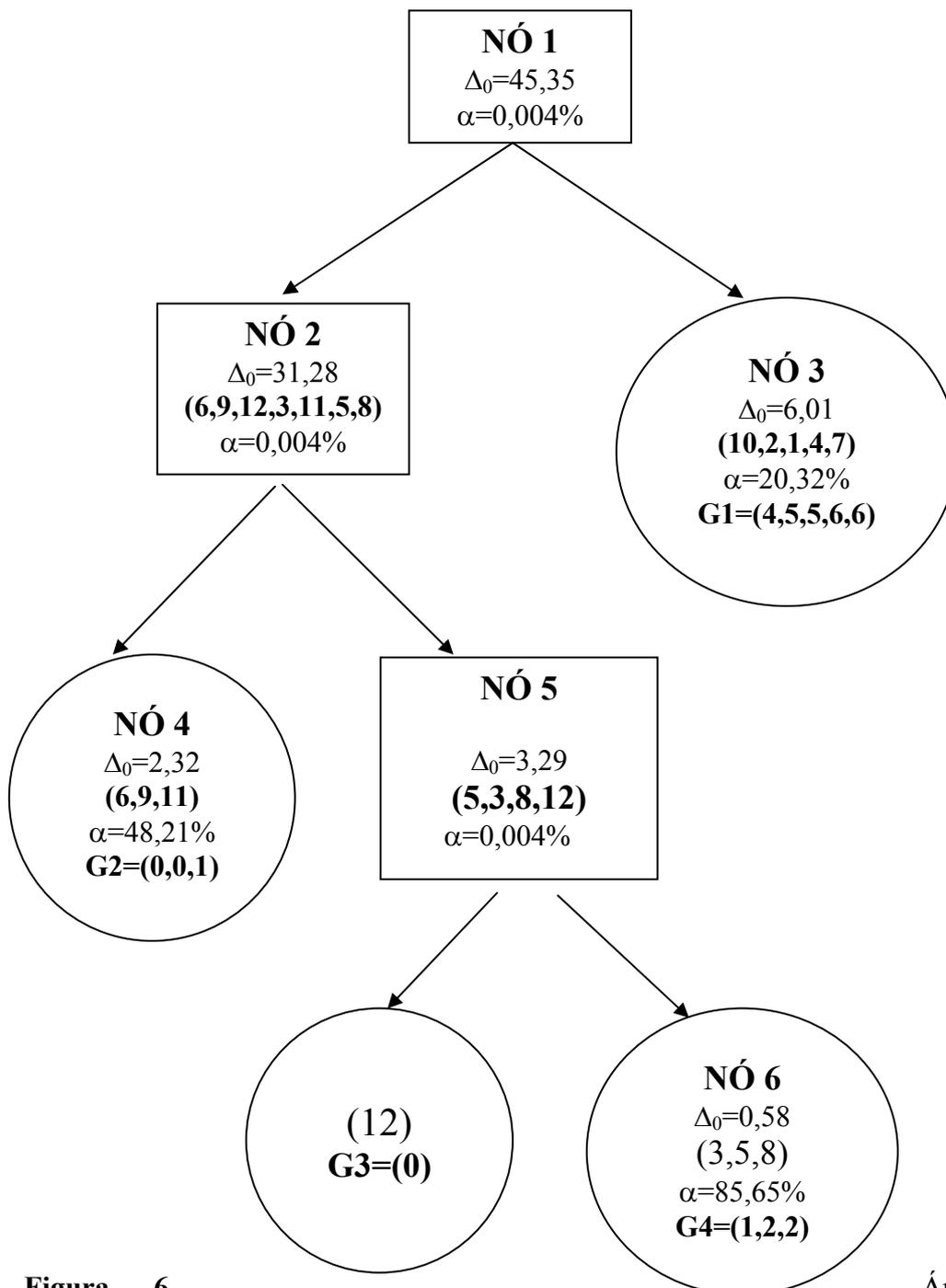
**Tabela 8:** Resumo dos resultados do algoritmo de divisão dicotômica considerando como critério de seleção ( $a_1$ ), como critério de partição ótima ( $b_2$ ), e como regra de parada, o nível descritivo do máximo da redução não é maior que 1% ( $c_3$ ).

NÓ	$\Delta_0$	$\frac{R}{(\pi - 2)}$	p-valor	Grupos
1	45,35	10,51	0,004%	(6,9,12,3,11,5,8) e (10,2,1,4,7)
2	31,28	6,13	0,004%	(11,6,9) e (5,3,8,12)
3	6,01	4,37	20,32%	(10,2) e (1,4,7)
4	2,32	2,62	48,21%	(11) e (6,9)
5	3,29	3,50	0,004%	(12) e (3,5,8)
6	0,58	2,62	85,65%	(3) e (5,8)

A formação da árvore começou pelo nó (1) da Tabela 8, com o cálculo da estatística  $\Delta_0$ , cujo valor foi de 45,35 e o número de graus de liberdade associado é

$\frac{R}{(\pi-2)}$  igual a 10,51, e o nível descritivo é  $\alpha = 0,004\%$ . Para a escolha da próxima partição dicotômica, calculou-se o valor da estatística  $\Delta_0$  no passo seguinte: para o nó (2), seu valor foi de 31,28 com  $\frac{R}{(\pi-2)}$  igual a 6,13 graus de liberdade e nível descritivo de  $\alpha = 0,004\%$  e, para o nó (3),  $\Delta_0$  igual a 2,32 com  $\frac{R}{(\pi-2)}$  igual a 2,62 graus de liberdade e nível descritivo de 48,21%. Desse modo, a segunda divisão escolhida deve ser feita no nó (2). Como o nível descritivo é inferior a 1% de probabilidade, examinam-se as partições no passo seguinte: para o nó (4),  $\Delta_0$  é igual a 3,39 com  $\frac{R}{(\pi-2)}$  igual a 3,50 graus de liberdade e nível descritivo de 0,004% e, para o nó (5),  $\Delta_0$  é igual a 0,58 com  $\frac{R}{(\pi-2)}$  igual a 2,62 graus de liberdade e nível descritivo de 85,65%. Com esses resultados, chegou-se à conclusão de que esse nó não deve ser dividido. Ao examinar o nó (4), deve-se aceitar a divisão em dois grupos e passa-se a examinar o nó (6). O valor de  $\Delta_0$  é igual a 4,37 com  $\frac{R}{(\pi-2)}$  igual a 4,37 graus de liberdade e nível descritivo de 20,32%. Portanto, o nó (6) não deve ser mais dividido.

Ao final do algoritmo obtiveram-se os grupos: {12}, {3,5,8} e {10,2,1,4,7}. Adotando-se o nível de significância individual de 1% de probabilidade, o nível geral é de  $1 - (1 - 0,01)^3 = 2,97\%$ . Na Figura 6, encontra-se no dendograma correspondente à Tabela 8.



**Figura 6** - Árvore de divisão dicotômica, considerando como partição ótima a maior redução e como regra de parada, a partição ótima com um nível descritivo superior a um valor pré-fixado  $\alpha=0,01$ .

Ao comparar os dois testes, verifica-se que os grupos formados pelo teste *bootstrap* e o teste assintótico são distintos, ou seja, o teste *bootstrap* possui 4 grupos:

{10,2,1,4,7}, {6,9,11}, {12} e {3,5,8}, onde os grupos representam os seguintes híbridos: {(250,K), (100,T), (100,K), (150,K), (200,K)}, {(150,N), (200,N), (250,T)}, {(250,N)} e {(100,N), (150,T), (200,T)}; e o teste assintótico, dois grupos: {6,9,12,3,11,5,8} e {10,2,1,4,7}, onde os grupos representam os seguintes híbridos: {(150,N), (200,N), (250,T), (250,N), (100,N), (150,T), (200,T)} e {(250,K), (100,T), (100,K), (150,K), (200,K)}.

No próximo capítulo, apresentam-se os gráficos simulados referentes às curvas poder para as seguintes configurações:  $R=3,5,8$  e  $12$  tratamentos, com  $n_r=5$  e  $10$  repetições por tratamento e os tamanhos de amostras binomiais  $n_b=1,10,100$  para o teste bootstrap e teste assintótico. O erro do tipo I foi fixado em  $\alpha=5\%$ .

## Capítulo 6

# Estudo Via Simulação para as Curvas Poder do Teste *Bootstrap* e Teste Assintótico

Conforme apresentado no capítulo 5, os pontos de equilíbrio para o tamanho da reamostragem *bootstrap* e o tamanho da simulação para obtenção da curva poder foram, respectivamente,  $S_B=250$  e  $S=1000$ . Usando esses parâmetros, neste capítulo, apresenta-se o estudo via simulação para as curvas poder do teste *bootstrap* e do teste assintótico, na situação específica de delineamentos inteiramente ao acaso, com  $R=3, 5, 8$  e  $12$  tratamentos,  $n=n_r=5$  e  $10$  repetições por tratamento e o valor da amostra binomial  $n_b=1, 10$  e  $100$ .

### 6.1 – Tamanho da Amostra Binomial igual a 1

Nesta seção, apresentam-se os resultados obtidos, via simulação, para o teste assintótico e o teste *bootstrap*. As Tabelas 9 e 10 e as Figuras 7 e 8 contêm algumas características obtidas por meio da simulação, referentes ao desempenho das curvas poder considerando o tamanho da amostra binomial igual  $n_b=1$ ,  $R=3, 5, 8$  e  $12$  tratamentos, com  $n=5$  e  $10$  repetições por tratamento. As curvas poder *bootstrap* e assintótica foram colocadas nas mesmas tabela e figura com o objetivo de verificar qual dos testes é mais poderoso.

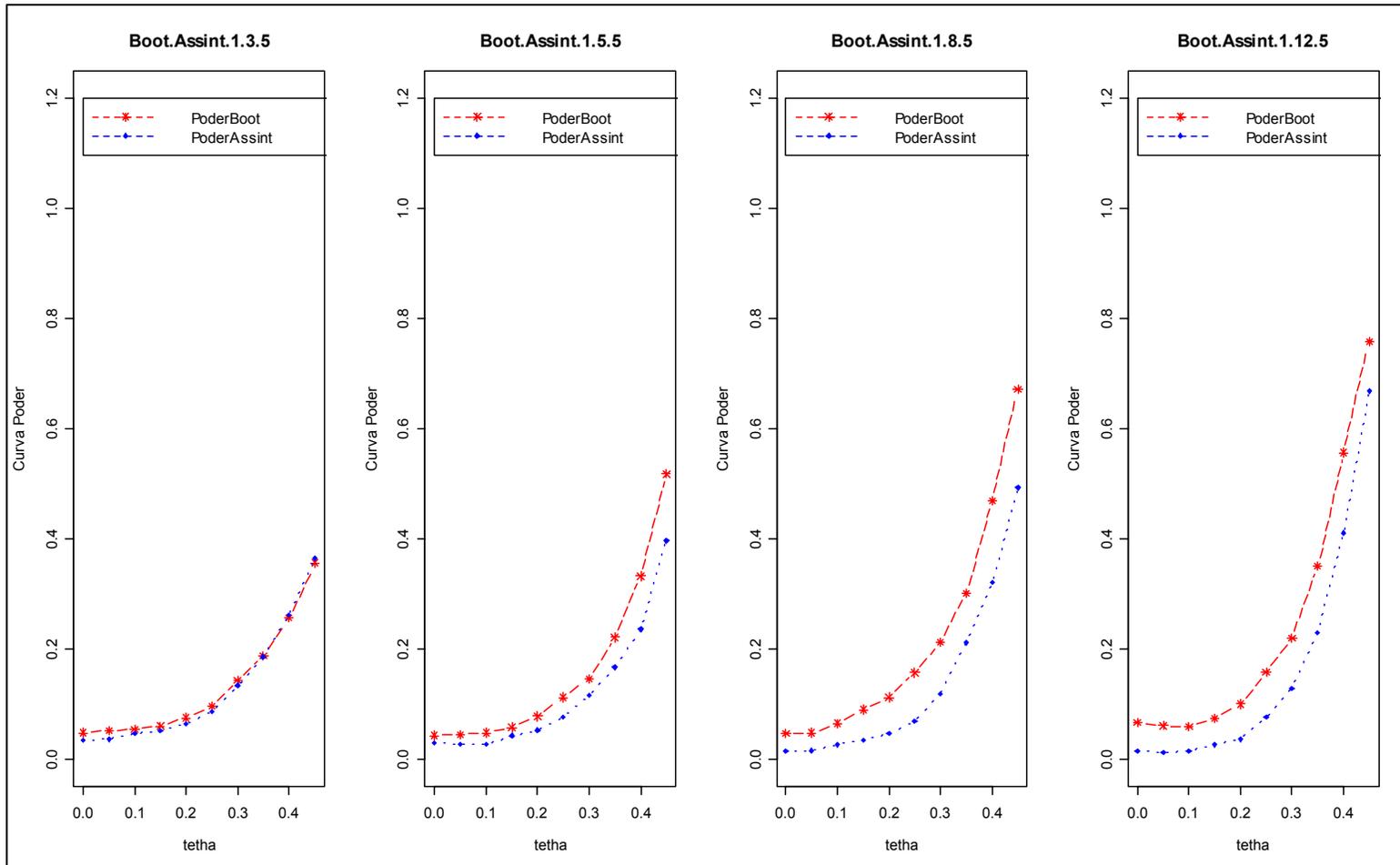
Pode-se notar, por meio da Tabela 9, que para a primeira configuração correspondendo a  $n_b=1$ ,  $R=3$ ,  $n=5$ , as curvas poder do teste assintótico e do teste *bootstrap* são praticamente iguais. O mesmo não acontece com as demais configurações em que a curva poder simulada *bootstrap* domina a curva poder simulada assintótica.

Por outro lado, ao se comparar às curvas poder simuladas *bootstrap* pode-se verificar, por meio da Figura 7, que o poder aumenta à medida que se aumenta o número de tratamentos. Esse comportamento, entretanto, é mais moderado entre as curvas poder simuladas assintótica.

A Tabela 10 corresponde às mesmas configurações da Tabela 7, exceto no número de repetições por tratamento,  $n=10$ . A análise de seus resultados é análoga à da Tabela 9, ressaltando-se que as curvas poder para  $n=10$  dominam as curvas poder para  $n=5$ . Além disso, as curvas poder assintóticas são praticamente iguais às curvas poder *bootstrap* quando  $\theta$  está próximo de  $0,45$ .

**Tabela 9** : Curvas poder para o teste *bootstrap* e o teste assintótico, calculadas a partir do tamanho da binomial igual a  $n_b = 1$ , com  $R=3,5,8,12$  tratamentos e  $n=5$  repetições por tratamento, para os valores de  $\theta$ .

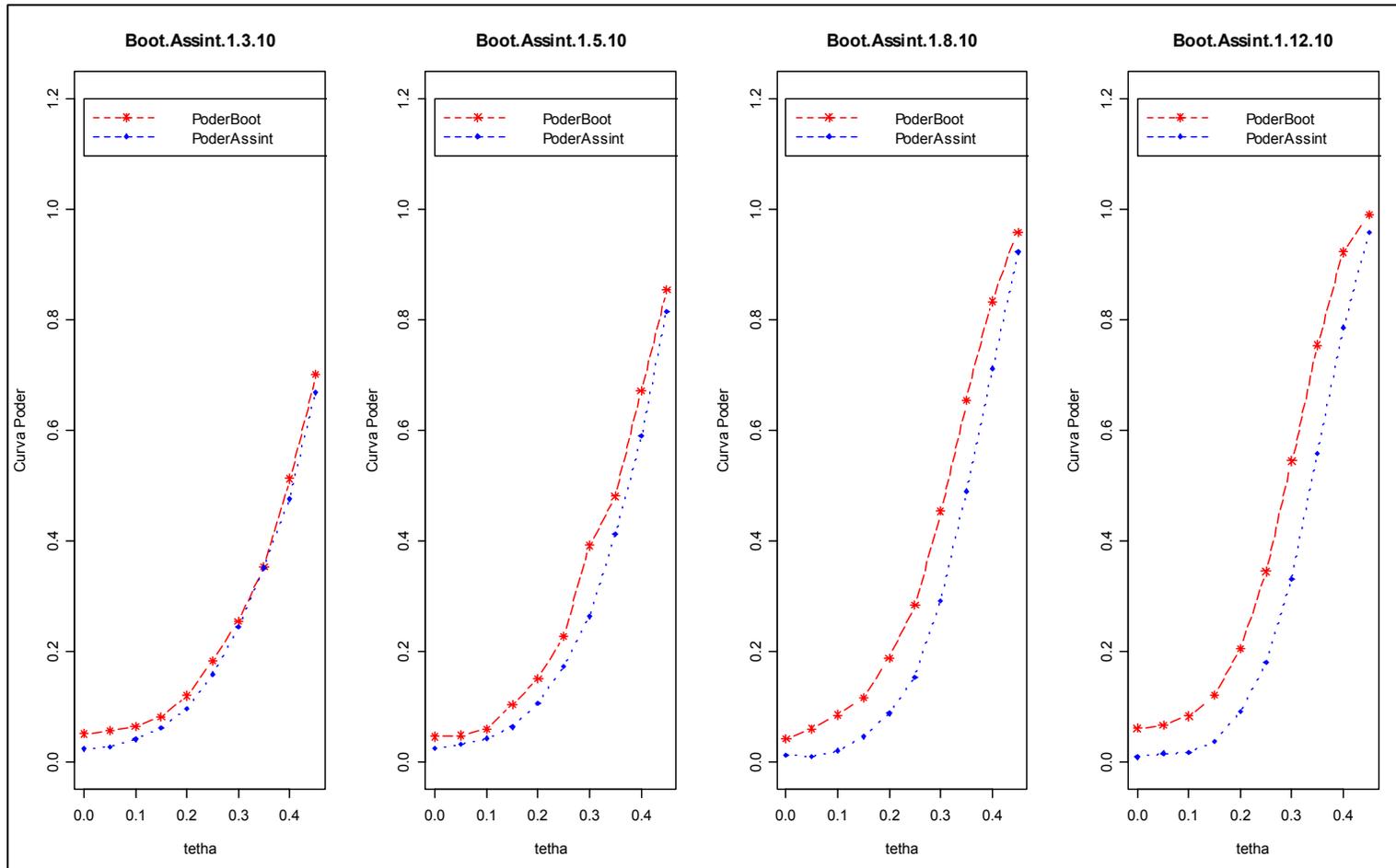
$P_1$	$P_2$	$\theta$	$n_b = 1, R=3 \text{ e } n=5$		$n_b = 1, R=5 \text{ e } n=5$		$n_b = 1, R=8 \text{ e } n=5$		$n_b = 1, R=12 \text{ e } n=5$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.50	0.00	0.049	0.0360	0.044	0.031	0.048	0.0160	0.067	0.016
0.5	0.45	0.05	0.053	0.0370	0.046	0.028	0.049	0.0170	0.062	0.013
0.5	0.40	0.10	0.055	0.0480	0.049	0.028	0.066	0.0270	0.060	0.016
0.5	0.35	0.15	0.061	0.0530	0.059	0.044	0.091	0.0360	0.075	0.027
0.5	0.30	0.20	0.076	0.0650	0.079	0.054	0.113	0.0480	0.101	0.037
0.5	0.25	0.25	0.097	0.0870	0.113	0.077	0.158	0.0700	0.159	0.077
0.5	0.20	0.30	0.144	0.1350	0.147	0.117	0.213	0.1200	0.220	0.129
0.5	0.15	0.35	0.188	0.1870	0.222	0.168	0.302	0.2120	0.351	0.231
0.5	0.10	0.40	0.257	0.2620	0.333	0.237	0.469	0.3210	0.557	0.411
0.5	0.05	<b>0.45</b>	<b>0.357</b>	<b>0.3650</b>	<b>0.518</b>	<b>0.397</b>	<b>0.671</b>	<b>0.4930</b>	<b>0.758</b>	<b>0.668</b>



**Figura 7** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=1$ ,  $R=3,5,8,12$  tratamentos e  $n=5$  por tratamento.

**Tabela 10:** Curvas poder para o teste *bootstrap* e o teste assintótico, calculadas a partir do tamanho da binomial igual a  $n_b = 1$ , com  $R=3,5,8,12$  tratamentos e  $n=10$  repetições por tratamento, para os valores de  $\theta$ .

$P_1$	$P_2$	$\theta$	$n_b = 1, R=3 \text{ e } n=10$		$n_b = 1, R=5 \text{ e } n=10$		$n_b = 1, R=8 \text{ e } n=10$		$n_b = 1, R=12 \text{ e } n=10$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.50	0.00	0.052	0.025	0.047	0.026	0.043	0.013	0.062	0.010
0.5	0.45	0.05	0.058	0.028	0.049	0.033	0.061	0.011	0.068	0.017
0.5	0.40	0.10	0.065	0.042	0.060	0.044	0.086	0.022	0.084	0.018
0.5	0.35	0.15	0.083	0.063	0.104	0.064	0.117	0.047	0.122	0.038
0.5	0.30	0.20	0.121	0.097	0.152	0.107	0.189	0.089	0.206	0.092
0.5	0.25	0.25	0.183	0.160	0.228	0.173	0.284	0.154	0.345	0.181
0.5	0.20	0.30	0.255	0.245	0.392	0.264	0.454	0.292	0.545	0.332
0.5	0.15	0.35	0.354	0.352	0.481	0.413	0.654	0.490	0.754	0.558
0.5	0.10	0.40	0.513	0.476	0.671	0.590	0.833	0.712	0.922	0.786
0.5	0.05	<b>0.45</b>	<b>0.701</b>	<b>0.668</b>	<b>0.854</b>	<b>0.814</b>	<b>0.958</b>	<b>0.922</b>	<b>0.989</b>	<b>0.957</b>



**Figura 8** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=1$ ,  $R=3,5,8,12$  tratamentos e  $n=10$  por tratamento.

## 6.2 - Tamanho da Amostra Binomial igual a 10

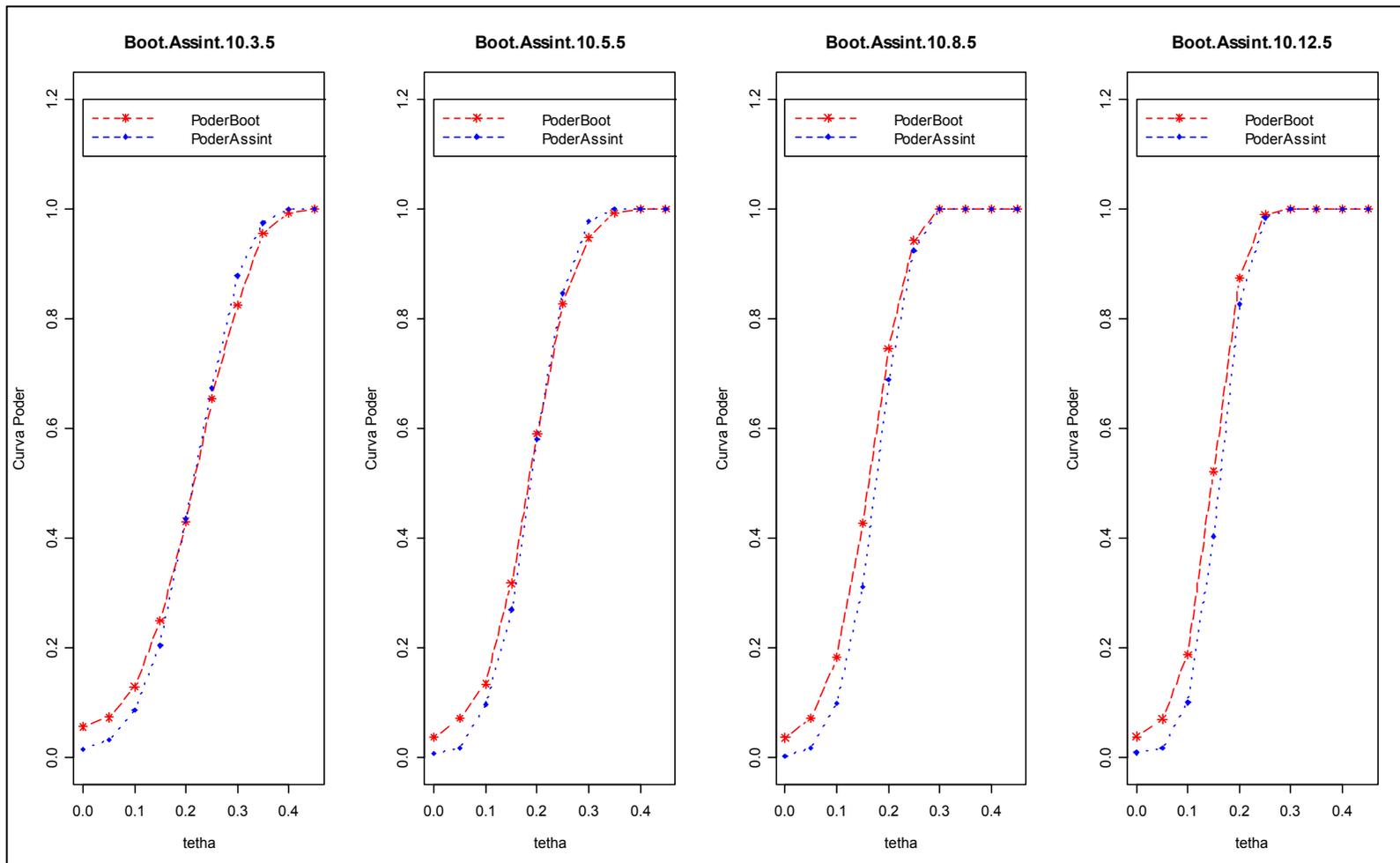
As Tabelas 11 e 12 e as Figuras 9 e 10 apresentam alguns resultados referentes ao desempenho das curvas poder, em que são considerados os tamanhos das amostras binomiais  $n_b=10$ , para delineamentos inteiramente ao acaso com  $R=3, 5, 8$  e  $12$  tratamentos,  $n=5$  e  $10$  repetições por tratamento.

Pode-se reparar, por meio da Tabela 11 e da Figura 9, que quando o número de repetições por tratamento é igual  $n=5$  e os valores de  $\theta$  estão entre  $0,0 \leq \theta \leq 0,15$ , a curva poder simulada do teste *bootstrap* sempre domina a curva poder simulada do teste assintótico, sendo, entretanto, o poder pequeno em ambos os casos. Já para valores de  $\theta$  acima  $0,20$  o poder dos dois testes aumenta rapidamente, apresentando uma discreta diferença do aumento do poder, conforme o número de tratamentos aumenta. Observa-se, também, por meio da Figura 9, uma melhora no poder do teste *bootstrap* e do teste assintótico quando o número de tratamentos aumenta, ou seja, quando  $R=5, 8$  e  $12$ , mas neste caso, os dois são praticamente iguais.

A Tabela 12 e a Figura 10 apresentam as curvas poder simuladas dos testes *bootstrap* e assintótico quando o número de repetições por tratamento é igual  $n=10$ . De uma maneira geral, o comportamento das curvas poder simuladas é análogo ao obtido na Tabela 12 e na Figura 10, sendo que existe um domínio acentuado da curva poder *bootstrap* para valores de  $\theta$  entre  $0,0$  e  $0,10$ . Por outro lado, o poder já está próximo de 1 para valores de  $\theta$  acima de  $0,25$ , considerando  $R=3$  e  $5$ . Para valores de  $R=8$  e  $12$ , o poder aproxima-se de 1 para valores de  $\theta$  superiores ou iguais a  $0,20$ .

**Tabela 11:** Curvas poder para o teste *bootstrap* e o teste assintótico, calculadas a partir do tamanho da binomial igual a  $n_b = 10$ , com  $R=3,5,8,12$  tratamentos e  $n=5$  repetições por tratamento, para os valores de  $\theta$ .

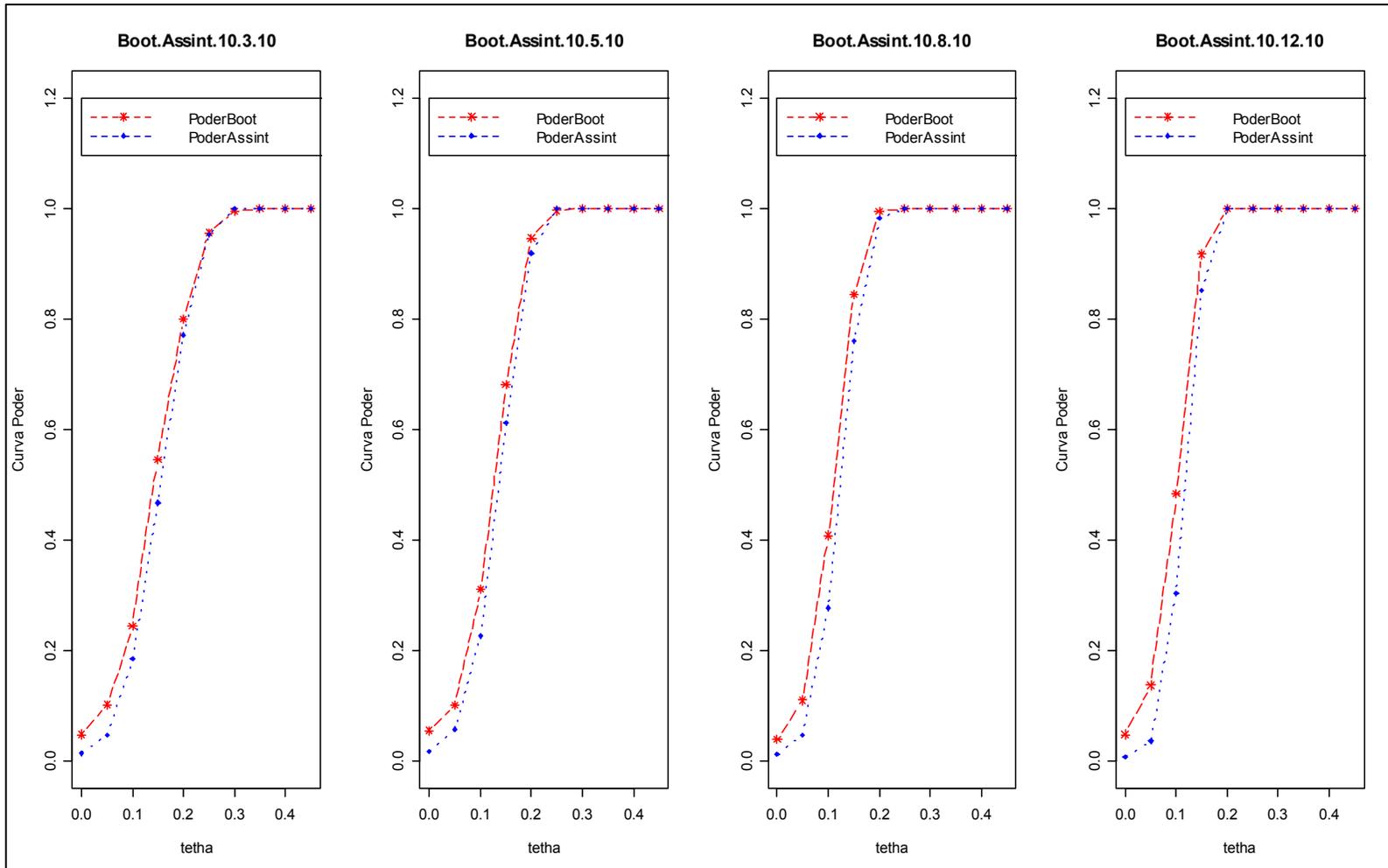
$P_1$	$P_2$	$\theta$	$n_b = 10, R=3 \text{ e } n=5$		$n_b = 10, R=5 \text{ e } n=5$		$n_b = 10, R=8 \text{ e } n=5$		$n_b = 10, R=12 \text{ e } n=5$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.50	<b>0.00</b>	<b>0.057</b>	<b>0.016</b>	<b>0.038</b>	<b>0.008</b>	<b>0.037</b>	<b>0.003</b>	<b>0.039</b>	<b>0.010</b>
0.5	0.45	<b>0.05</b>	<b>0.074</b>	<b>0.033</b>	<b>0.072</b>	<b>0.018</b>	<b>0.073</b>	<b>0.019</b>	<b>0.071</b>	<b>0.019</b>
0.5	0.40	<b>0.10</b>	<b>0.130</b>	<b>0.088</b>	<b>0.135</b>	<b>0.098</b>	<b>0.183</b>	<b>0.100</b>	<b>0.189</b>	<b>0.101</b>
0.5	0.35	<b>0.15</b>	<b>0.250</b>	<b>0.205</b>	<b>0.318</b>	<b>0.271</b>	<b>0.428</b>	<b>0.312</b>	<b>0.522</b>	<b>0.404</b>
0.5	0.30	<b>0.20</b>	<b>0.431</b>	<b>0.436</b>	<b>0.590</b>	<b>0.580</b>	<b>0.745</b>	<b>0.689</b>	<b>0.874</b>	<b>0.826</b>
0.5	0.25	0.25	0.655	0.673	0.827	0.846	0.942	0.924	0.990	0.984
0.5	0.20	0.30	0.825	0.878	0.947	0.978	1.000	1.000	1.000	1.000
0.5	0.15	0.35	0.955	0.974	0.992	0.999	1.000	1.000	1.000	1.000
0.5	0.10	0.40	0.993	0.999	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.05	0.45	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



**Figura 9** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=10$ ,  $R=3,5,8,12$  tratamentos e  $n=5$  por tratamento.

**Tabela 12:** Curvas poder para o teste *bootstrap* e o teste assintótico, calculada a partir do tamanho da binomial igual a  $n_b = 10$ , com  $R=3,5,8,12$  tratamentos e  $n=10$  repetições por tratamento, para os valores de  $\theta$ .

$P_1$	$P_2$	$\theta$	$n_b=10, R=3 \text{ e } n=10$		$n_b=10, R=5 \text{ e } n=10$		$n_b=10, R=8 \text{ e } n=10$		$n_b=10, R=12 \text{ e } n=10$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.50	<b>0.00</b>	<b>0.056</b>	<b>0.018</b>	<b>0.056</b>	<b>0.0180</b>	<b>0.041</b>	<b>0.014</b>	<b>0.049</b>	<b>0.009</b>
0.5	0.45	<b>0.05</b>	<b>0.102</b>	<b>0.059</b>	<b>0.102</b>	<b>0.0590</b>	<b>0.111</b>	<b>0.048</b>	<b>0.138</b>	<b>0.037</b>
0.5	0.40	<b>0.10</b>	<b>0.312</b>	<b>0.227</b>	<b>0.312</b>	<b>0.2270</b>	<b>0.409</b>	<b>0.278</b>	<b>0.484</b>	<b>0.304</b>
0.5	0.35	0.15	0.682	0.613	0.682	0.6130	0.844	0.760	0.917	0.852
0.5	0.30	<b>0.20</b>	0.946	0.919	0.946	0.9190	<b>0.994</b>	<b>0.982</b>	<b>1.000</b>	<b>1.000</b>
0.5	0.25	<b>0.25</b>	<b>0.996</b>	<b>0.999</b>	<b>0.996</b>	<b>0.9990</b>	1.000	1.000	1.000	1.000
0.5	0.20	0.30	1.000	1.000	1.000	1.0000	1.000	1.000	1.000	1.000
0.5	0.15	0.35	1.000	1.000	1.000	1.0000	1.000	1.000	1.000	1.000
0.5	0.10	0.40	1.000	1.000	1.000	1.0000	1.000	1.000	1.000	1.000
0.5	0.05	0.45	1.000	1.000	1.000	1.0000	1.000	1.000	1.000	1.000



**Figura 10** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=10$ ,  $R=3,5,8,12$  tratamentos e  $n=10$  por tratamento.

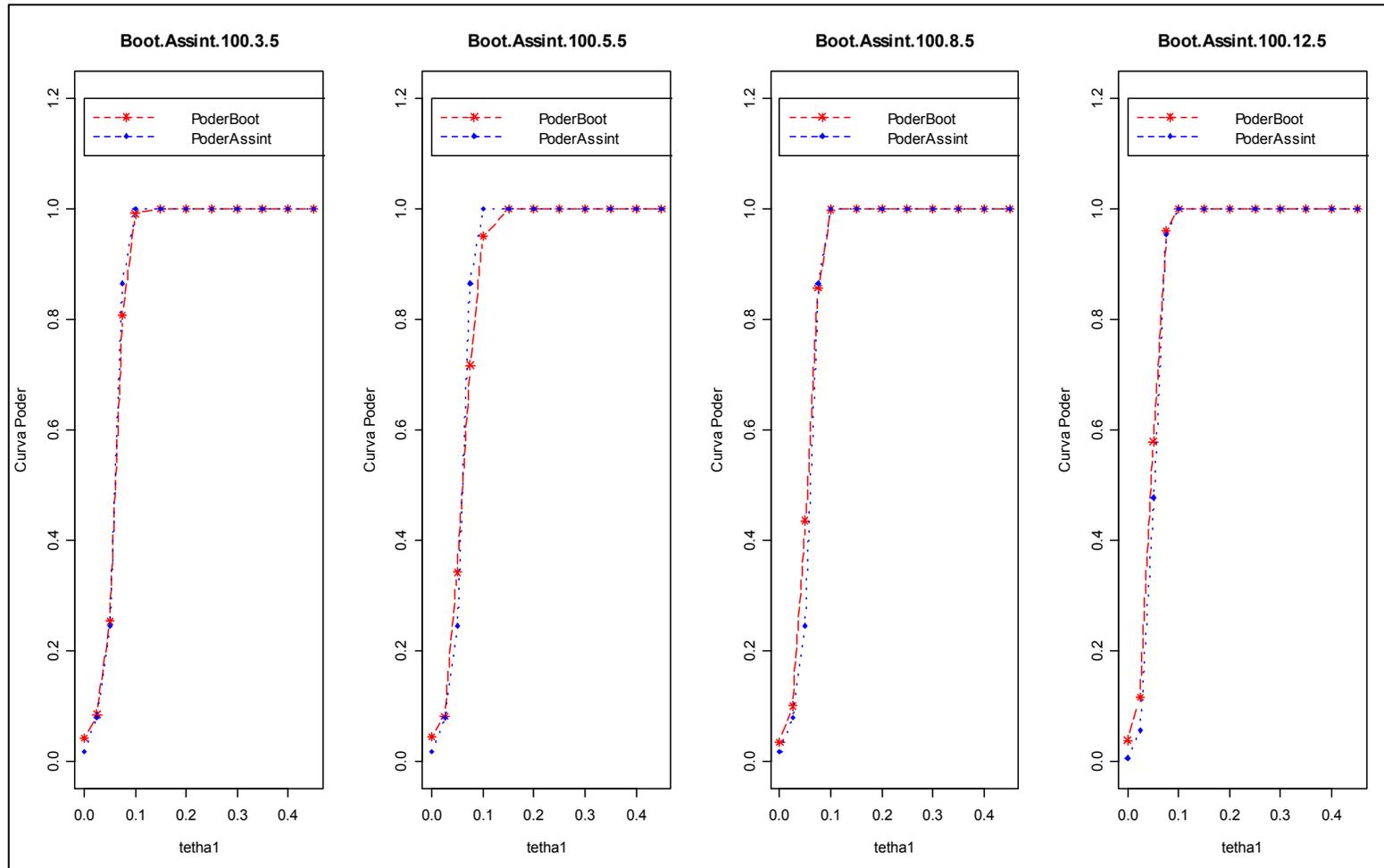
### **6.3 - Tamanho da Amostra Binomial igual a 100**

As curvas poder simuladas apresentadas nas Tabelas 13 e 14 e nas Figuras 11 e 12 referem-se a delineamentos em que o tamanho da binomial é igual a 100.

A análise detalhada desses elementos revela que o poder dos dois testes é alto e que são praticamente semelhantes, não havendo distinção entre eles em relação ao número de tratamentos e número de repetições por tratamento.

**Tabela 13:** Curvas poder para o teste *bootstrap* e o teste assintótico, calculadas a partir do tamanho da binomial igual à  $n_b = 100$ , com  $R=3,5,8,12$  tratamentos e  $n=5$  repetições por tratamento, para os valores de  $\theta$ .

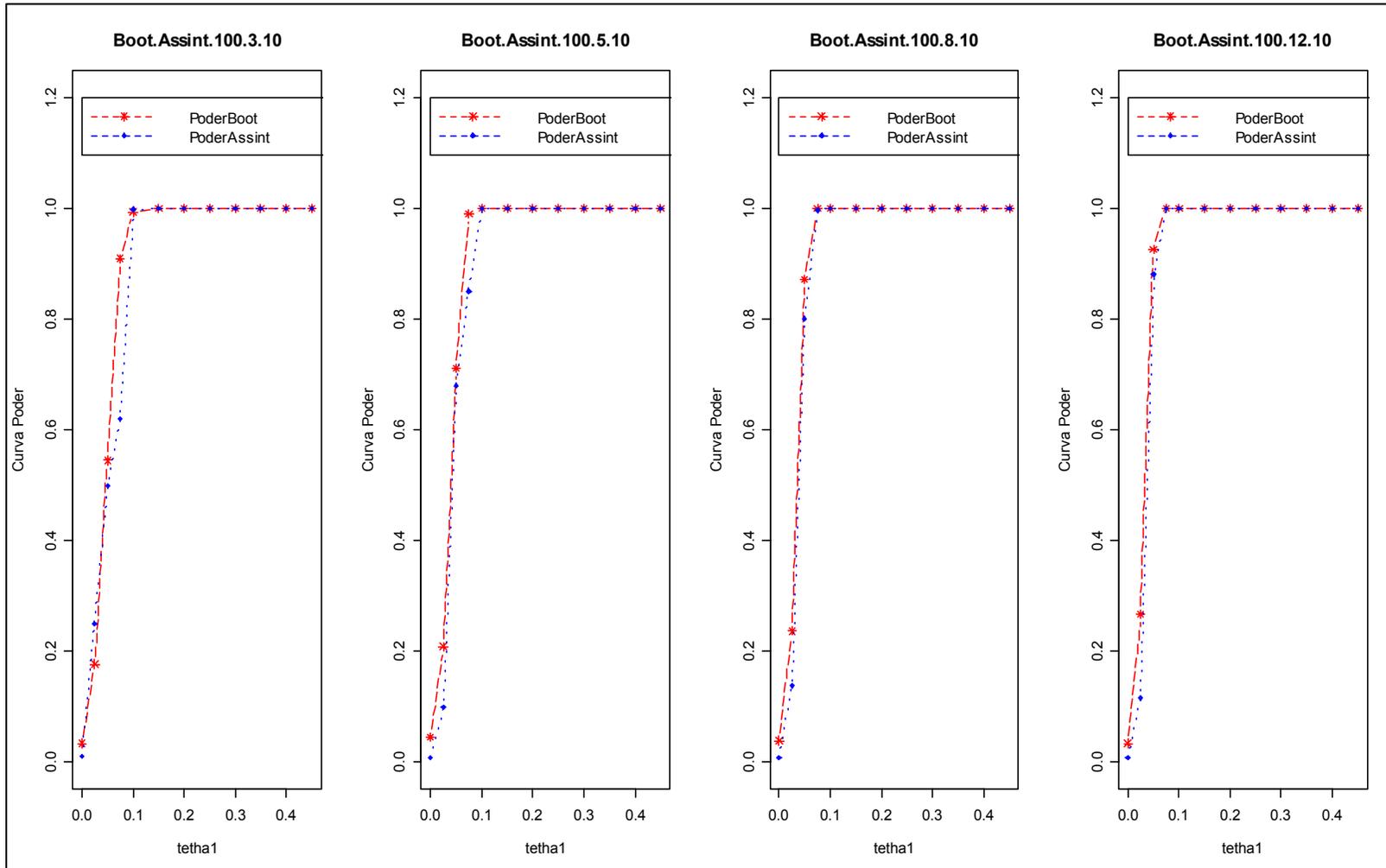
$P_1$	$P_2$	$\theta$	$n_b=100, R=3 \text{ e } n=5$		$n_b=100, R=5 \text{ e } n=5$		$n_b=100, R=8 \text{ e } n=5$		$n_b=100, R=12 \text{ e } n=5$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.500	0.000	0.043	0.018	0.045	0.013	0.036	0.010	0.039	0.007
0.5	0.475	0.025	0.086	0.080	0.082	0.057	0.101	0.039	0.117	0.057
0.5	0.450	0.050	0.254	0.246	0.343	0.316	0.436	0.374	0.579	0.478
0.5	0.425	0.075	0.807	0.865	0.717	0.743	0.856	0.874	0.960	0.953
0.5	0.400	0.100	0.991	1.000	0.950	0.970	0.998	0.999	1.000	1.000
0.5	0.350	0.150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.300	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.250	0.250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.200	0.300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.150	0.350	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.100	0.400	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.050	0.450	0.043	0.018	1.000	1.000	1.000	1.000	1.000	1.000



**Figura 11** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=10$ ,  $R=3,5,8,12$  tratamentos e  $n=5$  por tratamento.

**Tabela 14:** Curvas poder para o teste *bootstrap* e o teste assintótico, calculadas a partir do tamanho da binomial igual à  $n_b = 100$ , com  $R=3,5,8,12$  tratamentos e  $n=10$  repetições por tratamento, para os valores de  $\theta$ .

$P_1$	$P_2$	$\theta$	$n_b=100, R=3 e n=10$		$n_b=100, R=5 e n=10$		$n_b=100, R=8 e n=10$		$n_b=100, R=12 e n=10$	
			PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint	PoderBoot	PoderAssint
0.5	0.500	0.000	0.034	0.011	0.046	0.009	0.039	0.009	0.034	0.009
0.5	0.475	0.025	0.177	0.094	0.209	0.100	0.237	0.138	0.268	0.116
0.5	0.450	0.050	0.545	0.499	0.711	0.680	0.871	0.801	0.926	0.88
0.5	0.425	0.075	0.909	0.908	0.989	0.850	1.000	0.996	1.000	1.000
0.5	0.400	0.100	0.993	0.998	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.350	0.150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.300	0.200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.250	0.250	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.200	0.300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.150	0.350	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.100	0.400	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.5	0.050	0.450	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



**Figura 12** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=10$ ,  $R=3,5,8,12$  tratamentos e  $n=10$  por tratamento.

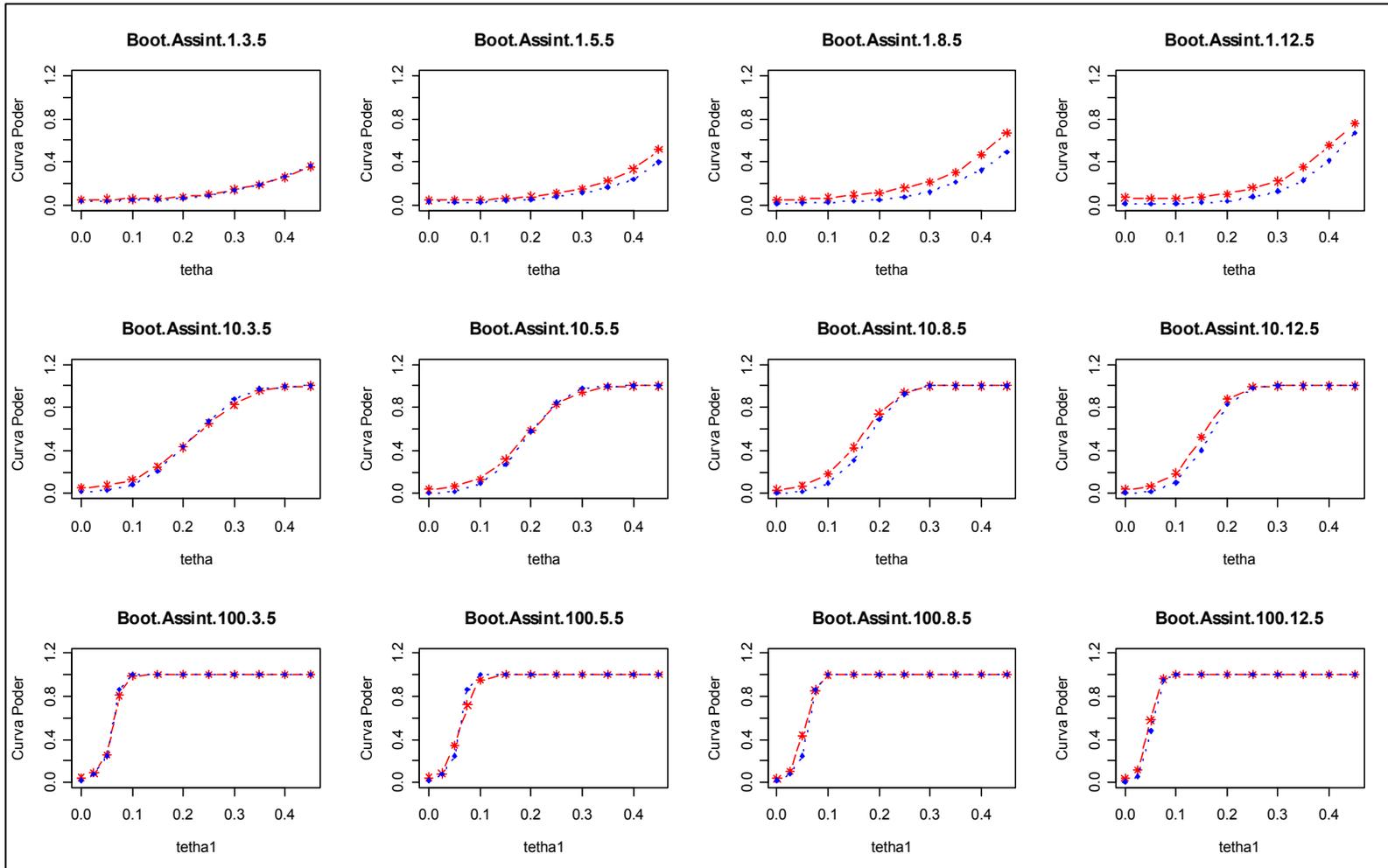
## 6.4 Comparação geral

As Figuras 13 e 14 procuram reunir as curvas poder simuladas dos testes *bootstrap* e assintótico, considerando, respectivamente, o número de repetições por tratamento igual a 5 e 10.

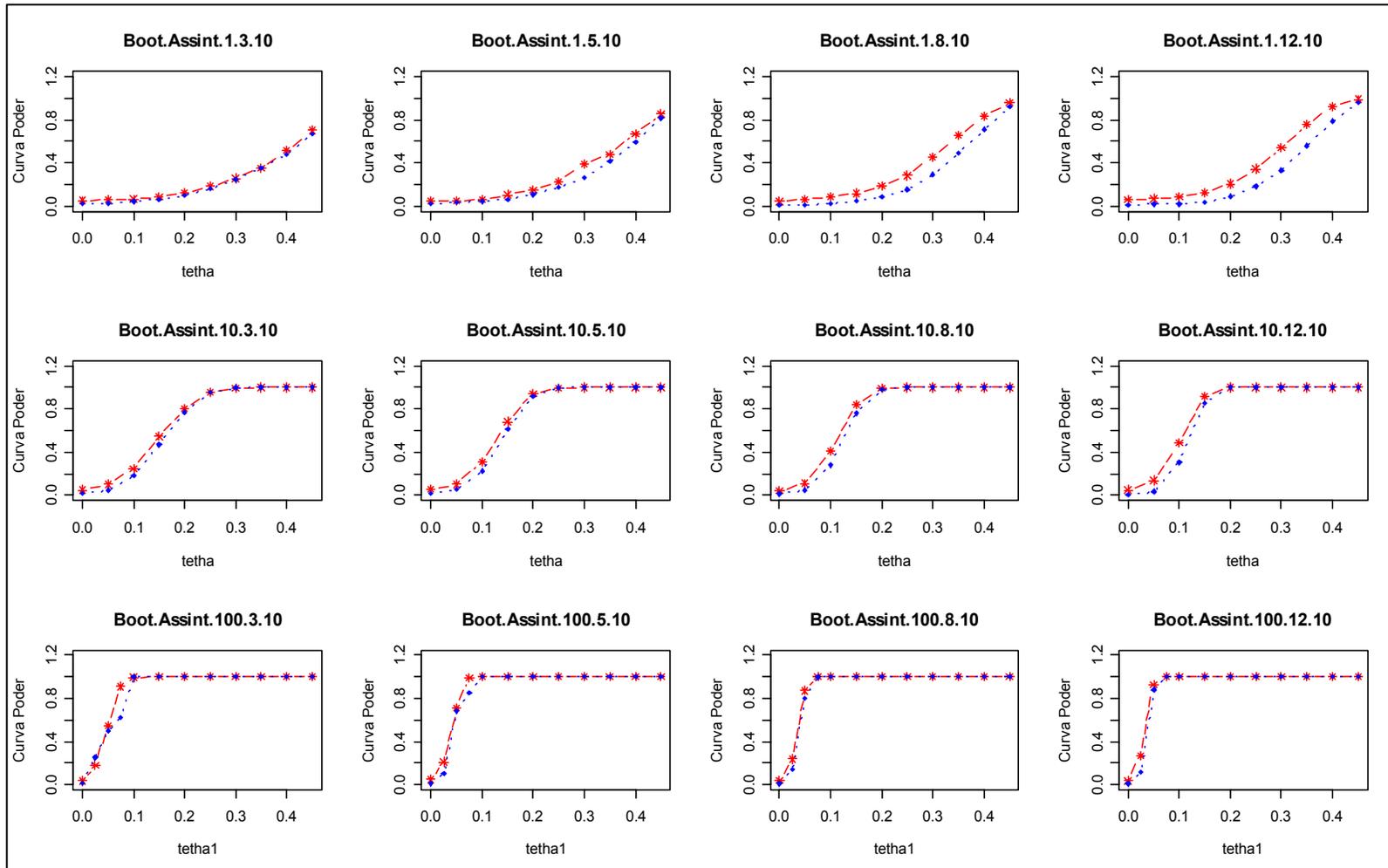
A análise conjunta evidencia que, conforme aumenta o tamanho da amostra binomial, o poder simulado dos dois testes aumenta rapidamente. Por outro lado, esse comportamento é mais acentuado conforme aumenta-se o número de tratamentos.

Vale ressaltar que, para o caso extremo em que o tamanho da binomial é igual a 1 e o número de tratamentos igual a 8 e 12, a curva poder simulada *bootstrap* é destacadamente superior à curva poder simulada assintótica para o número de repetições por tratamentos estudados.

A dificuldade encontrada neste trabalho deu-se em relação ao tempo de simulação do teste *bootstrap* em comparação ao assintótico. Quando se aumentou o número de tratamentos, o tempo de simulação praticamente dobrou. Por exemplo: ao considerar  $R=5$  e  $n=10$ , não importando o tamanho da amostra binomial, e  $R=12$  e  $n=10$ , tem-se uma diferença média de 15 horas em cada ponto da curva simulada. Portanto, considerada a simulação com a seguinte configuração  $n_b=1$ ,  $r=12$  e  $n=10$ , o tempo de simulação foi aproximadamente de 10 dias.



**Figura 13** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=1,10,100$ ,  $R=3,5,8,12$  tratamentos e  $n=5$  por tratamento e -\*- PoderBoot e -●- PoderAssint.



**Figura 14** - Curva Poder do Teste Assintótico e do Teste *Bootstrap* para o tamanho da binomial  $n_b=1, 10, 100$ ,  $R=3, 5, 8, 12$  tratamentos e  $n=10$  por tratamento e -\*- PoderBoot e -●- PoderAssint

## Capítulo 7

# Conclusões e propostas futuras

Realizou-se, neste trabalho, um amplo estudo via simulação, buscando avaliar o comportamento das curvas poder do teste assintótico e do teste *bootstrap*, quando da utilização da extensão do algoritmo AID em modelos lineares generalizados para delineamentos inteiramente ao acaso. Foram levados em consideração 4 diferentes números de tratamentos ( $R$ ), 3 diferentes tamanhos das amostras binomiais ( $n_b$ ) e 2 números de repetições por tratamento ( $n_r$ ), ou seja,  $R=3,5,8, e 12$ ,  $n_b=1,10 e 10$  e  $n=5 e 10$ .

Verificou-se que, quando o tamanho da amostra binomial é igual a 1 e o número de repetições por tratamento  $n=5$ , o poder aumenta na medida em que se aumenta o número de tratamentos. Esse comportamento, entretanto, é mais moderado entre as curvas poder simuladas assintóticas. Ressalta-se, também, que as curvas poder para  $n=10$  dominam as curvas poder para  $n=5$ .

Ao considerar o tamanho da amostra binomial igual a 10, o número de repetições por tratamento igual  $n=5$  e valores de  $\theta$  entre  $0,0$  e  $0,15$ , a curva poder simulada do teste *bootstrap* sempre domina a curva poder simulada do teste assintótico, sendo o poder pequeno em ambos os casos. Já para valores de  $\theta$  acima de  $0,20$ , o poder dos dois testes aumenta rapidamente, apresentando uma discreta diferença do aumento do poder conforme o número de tratamentos aumenta. Com relação ao tamanho da binomial igual  $n_b=10$ , as curvas poder dos testes *bootstrap* e assintótico possuem, de uma maneira geral, um comportamento análogo ao obtido para  $n=5$ , sendo que existe um domínio acentuado da curva poder *bootstrap* para valores de  $\theta$  entre  $0,0$  e  $0,10$ . Por outro lado, o poder já está próximo de 1 para valores de  $\theta$  acima de  $0,25$ , considerando

$R=3$  e  $5$  e, para valores de  $R=8$  e  $12$ , o poder aproxima-se de 1 para valores de  $\theta$  superiores ou iguais a  $0,20$ .

Finalmente, quando o tamanho da amostra binomial é igual a 100, observa-se que as curvas poder dos testes assintóticos e *bootstrap* possuem um poder alto e são praticamente semelhantes, não havendo distinção entre eles em relação ao número de tratamentos e número de repetições por tratamento.

Desse modo, de uma maneira geral, os dois testes têm o mesmo poder simulado. O teste *bootstrap* é mais poderoso que o teste assintótico quando o tamanho da binomial é pequeno ( $n_b=1$ ) em todas as situações, inclusive considerando o número de tratamentos ( $R$ ) estudados. Além disso à medida que o número de repetições por tratamento aumenta, mais poderoso é o teste *bootstrap*.

O experimento com couve chinesa foi analisado e as árvores de divisão dicotômica foram apresentadas nas seções 3.3 (assintótica) e 5.3 (*bootstrap*). Os grupos formados foram distintos apesar de se usar o mesmo critério de seleção, o mesmo critério de partição ótima e a mesma regra de parada. Como o teste *bootstrap* para essa situação é mais poderoso, podemos assegurar que os grupos formados pelo teste *bootstrap* realmente apresentam diferença significativa.

Para continuidade deste trabalho, sugere-se o estudo via simulação para as curvas poder do teste assintótico e do teste *bootstrap* para as seguintes distribuições: Poisson e Gama.

# Apêndice

## Apêndice A: Programa em R para a primeira partição usando o teste assintótico no experimento com couve chinesa (*Brassica pekinensis*) para primeira partição.

```
semente<-515
set.seed(semente)
k<-12          # número de tratamento
ni<-6          # número de repetições
n<-k*ni       # vetor completo = número de tratamento * número de
  repetições
t<-gl(12,6)    # vetor tratamento
x<-c(1,1,1,1,1,1,0,1,1,1,1,1,0,1,0,0,0,0,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,
  0,0,0,1,1,1,1,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,0,1,0,0,0,
  ,0,0,0,0,0)
# vetor dados originais da couve
dado<-data.frame(t=t,x=x) # dados
m1<-c(x[1:6])      # divisão de 6 em 6
m2<-c(x[7:12])    # divisão de 6 em 6
m3<-c(x[13:18])   # divisão de 6 em 6
m4<-c(x[19:24])   # divisão de 6 em 6
m5<-c(x[25:30])   # divisão de 6 em 6
m6<-c(x[31:36])   # divisão de 6 em 6
m7<-c(x[37:42])   # divisão de 6 em 6
m8<-c(x[43:48])   # divisão de 6 em 6
m9<-c(x[49:54])   # divisão de 6 em 6
m10<-c(x[55:60]) # divisão de 6 em 6
m11<-c(x[61:66]) # divisão de 6 em 6
m12<-c(x[67:72]) # divisão de 6 em 6
somal<-numeric() # soma
somal[1]<-sum(m1)  # soma
somal[2]<-sum(m2)  # soma
somal[3]<-sum(m3)  # soma
somal[4]<-sum(m4)  # soma
somal[5]<-sum(m5)  # soma
somal[6]<-sum(m6)  # soma
somal[7]<-sum(m7)  # soma
somal[8]<-sum(m8)  # soma
somal[9]<-sum(m9)  # soma
somal[10]<-sum(m10) # soma
somal[11]<-sum(m11) # soma
somal[12]<-sum(m12) # soma
somal # vetor soma
smaior<-somal[1]  # inicio da ordem da posição
pos<-1
for (j in 2:12){
  if (smaior < somal[j]){
    smaior<-somal[j]
    pos<-j
  }
}
```

```

smaior
pos
soma2<-soma1
soma3<-order(soma2)
soma3          # vetor da posição
soma4<-numeric()
#----- posição 1 do vetor soma3
if (soma3[1]==1){
  for (i in 1:6){
    soma4[i]<-m1[i]
  }
}
.
.
.
if (soma3[1]==12){
  for (i in 1:6){
    soma4[i]<-m12[i]
  }
}
#----- posição 2 do vetor soma3
.
.
.
#----- posição 12 do vetor soma3
if (soma3[12]==1){
  j<-1
  for (i in 67:72){
    soma4[i]<-m1[j]
    j<-j+1
  }
}
.
.
.
if (soma3[12]==12){
  j<-1
  for (i in 67:72){
    soma4[i]<-m12[j]
    j<-j+1
  }
}
soma4          # vetor dos dados
###calculo inicial da deviance para os dados ordenados ###
M1<-rep(1, times=6)
M11<-rep(2, times=66)          #vetor de 1 e 2
c1<-c(M1,M11)
.
.
.
M11<-rep(1, times=66)
M111<-rep(2, times=6)          #vetor de 1 e 2
c11<-c(M11,M111)
d<data.frame(t=t, dadd=soma4, v1=c1, v2=c2, v3=c3, v4=c4, v5=c5, v6=c6, v7=c7,
  v8=c8, v9=c9, v10=c10, v11=c11) # dados com os vetores de 1 e 2
funglm<-function(d){ # funcao de geracao da deviance
  glm1<-glm(dadd~v1, binomial(link = "logit"), d)
}
.
.
.

```

```

glm11<-glm(dadd~v11,binomial(link = "logit"),d)
tabela1<-anova(glm1)
.
.
.
  tabela11<-anova(glm11)
dev<c(tabela1[2,2],tabela2[2,2],tabela3[2,2],tabela4[2,2],tabela5[2,
2],tabela6[2,2],tabela7[2,2],tabela8[2,2],tabela9[2,2],tabela10[2,2
],tabela11[2,2])
return(dev)
}
y<-funglm(d)          # vetor da deviance
Y
fmaior<-function(y){ # ordenacao do vetor da devianca para obtencao
do maior valor
maior<-numeric()
maior<-y[1]
for (j in 2:11){
  if (maior < y[j]){
    maior<-y[j]
  }
}
return(maior) }
maior<-fmaior(y)
maior          # maior valor da deviance
  semente<-515
set.seed(semente)
Sa<-2
devmaior<-numeric()
niveldesc<-numeric()
pvalorassint<-numeric()

```

## Apêndice B: Programa em R para determinar o tamanho da simulação $S_A$ para o teste assintótico.

```

for (l in 1:1000) {
k<-5 # numero de tratamento
ni<-10 # numero de repeticoes
n<-k*ni # vetor completo = numero de tratamento * numero de
repeticoes
t<-gl(5,10) # vetor tratamento
v<-rbinom(1,0.50, n=20) # geracao da binomial
z<-rbinom(1,0.50, n=30) # geracao da binomial
x<-c(v, z) # vetor binomial com os tratamentos
dado<-data.frame(rep=t, dad=x) # dados
m1<-c(x[1:10]) # divisao de 10 em 10
m2<-c(x[11:20]) # divisao de 10 em 10
m3<-c(x[21:30]) # divisao de 10 em 10
m4<-c(x[31:40]) # divisao de 10 em 10
m5<-c(x[41:50]) # divisao de 10 em 10
soma1<-numeric() # soma
soma1[1]<-sum(m1) # soma
soma1[2]<-sum(m2) # soma
soma1[3]<-sum(m3) # soma
soma1[4]<-sum(m4) # soma
soma1[5]<-sum(m5) # soma
soma1 # vetor soma
smaior<-soma1[1] # inicio da ordem da posicao
pos<-1
for (j in 2:5){
if (smaior < soma1[j]){
smaior<-soma1[j]
pos<-j
}
}
smaior
pos
soma2<-soma1
soma3<-order(soma2)
soma3 # vetor da posicao
soma4<-numeric()
#----- posicao 1 do vetor soma3
if (soma3[1]==1){
for (i in 1:10){
soma4[i]<-m1[i]
}
}
.
.
.
if (soma3[1]==5){
for (i in 1:10){
soma4[i]<-m5[i]
}
}
}
#----- posicao 2 do vetor soma3
.
.
.
#----- posicao 5 do vetor soma3

```

```

if (soma3[5]==1){
  j<-1
  for (i in 41:50){
    soma4[i]<-m1[j]
    j<-j+1
  }
}
.
.
.
if (soma3[5]==5){
  j<-1
  for (i in 41:50){
    soma4[i]<-m5[j]
    j<-j+1
  }
}
soma4 # vetor dos dados
####calculo inicial da deviance para os dados ordenados #
M1<-rep(1, times=10)
M11<-rep(2, times=40) #vetor de 1 e 2
c1<-c(M1,M11)
M2<-rep(1, times=20)
M22<-rep(2, times=30) #vetor de 1 e 2
c2<-c(M2,M22)
M3<-rep(1, times=30)
M33<-rep(2, times=20) #vetor de 1 e 2
c3<-c(M3,M33)
M4<-rep(1, times=40)
M44<-rep(2, times=10) #vetor de 1 e 2
c4<-c(M4,M44)
d<-data.frame(rep=t, dadd=soma4, v1=c1, v2=c2, v3=c3, v4=c4) # dados com
os vetores de 1 e 2
funglm<-function(d) { # funcao de geracao da deviance
  glm1<-glm(dadd~v1, binomial(link = "logit"), d)
  glm2<-glm(dadd~v2, binomial(link = "logit"), d)
  glm3<-glm(dadd~v3, binomial(link = "logit"), d)
  glm4<-glm(dadd~v4, binomial(link = "logit"), d)
  tabela1<-anova(glm1)
  tabela2<-anova(glm2)
  tabela3<-anova(glm3)
  tabela4<-anova(glm4)
  dev<-c(tabela1[2,2], tabela2[2,2], tabela3[2,2], tabela4[2,2])
  return(dev)
}
y<-funglm(d) # vetor da deviance

fmaior<-function(y) { # ordenacao do vetor da devianca
  para obtencao do maior valor
  maior<-numeric()
  maior<-y[1]
  for (j in 2:4){
    if (maior < y[j]){
      maior<-y[j]
    }
  }
  return(maior)
}
maior<-fmaior(y)

```

```
maior
devmaior[l]<-maior          # maior valor da deviance
gl<-k/(pi-2)               # grau de liberdade qui-quadrado
niveldesc[l]<-1-pchisq(devmaior[l],gl)
pvalorassint[l]<-sum(niveldesc)/Sa
Sa<-Sa+1}
devmaior
niveldesc
pvalorassint
```

## Apêndice C: Programa em R para determinar tamanho da simulação de B para o teste bootstrap.

```

semente<-515
set.seed(semente)
k<-5 # número de tratamento
ni<-10 # número de repetições
n<-k*ni # vetor completo=número de tratamento* número de repetições
t<-gl(5,10) # vetor tratamento
v<-rbinom(1,0.45, n=20) # geração da binomial
z<-rbinom(1,0.55, n=30) # geração da binomial
x<-c(v,z) # vetor binomial com os tratamentos
dado<-data.frame(rep=t,dad=x) # dados
m1<-c(x[1:10]) # divisão de 10 em 10
m2<-c(x[11:20]) # divisão de 10 em 10
m3<-c(x[21:30]) # divisão de 10 em 10
m4<-c(x[31:40]) # divisão de 10 em 10
m5<-c(x[41:50]) # divisão de 10 em 10
soma1<-numeric() # soma
soma1[1]<-sum(m1) # soma
soma1[2]<-sum(m2) # soma
soma1[3]<-sum(m3) # soma
soma1[4]<-sum(m4) # soma
soma1[5]<-sum(m5) # soma
soma1 # vetor soma
smaior<-soma1[1] # início da ordem da posição
pos<-1
for (j in 2:5){
  if (smaior < soma1[j]){
    smaior<-soma1[j]
    pos<-j
  }
}
smaior
pos
soma2<-soma1
soma3<-order(soma2)
soma3 # vetor da posição
soma4<-numeric()
#----- posição 1 do vetor soma3
if (soma3[1]==1){
  for (i in 1:10){
    soma4[i]<-m1[i]
  }
}
.
.
.
if (soma3[1]==5){
  for (i in 1:10){
    soma4[i]<-m5[i]
  }
}
#-----posição 2 do vetor soma3
.
.
.
#----- posição 5 do vetor soma3
if (soma3[5]==1){

```

```

    j<-1
    for (i in 41:50){
      soma4[i]<-m1[j]
      j<-j+1
    }
  }
  .
  .
  .
if (soma3[5]==5){
  j<-1
  for (i in 41:50){
    soma4[i]<-m5[j]
    j<-j+1
  }
}
soma4                                # vetor dos dados
####calculo inicial da deviance para os dados ordenados ##
M1<-rep(1, times=10)
M11<-rep(2, times=40)                 #vetor de 1 e 2
c1<-c(M1,M11)
M2<-rep(1, times=20)
M22<-rep(2, times=30)                 #vetor de 1 e 2
c2<-c(M2,M22)
M3<-rep(1, times=30)
M33<-rep(2, times=20)                 #vetor de 1 e 2
c3<-c(M3,M33)
M4<-rep(1, times=40)
M44<-rep(2, times=10)                 #vetor de 1 e 2
c4<-c(M4,M44)
d<-data.frame(rep=t, dadd=soma4, v1=c1, v2=c2, v3=c3, v4=c4)           #
  dados com os vetores de 1 e 2
funglm<-function(d){                 # função de geração da deviance
  glm1<-glm(dadd~v1, binomial(link = "logit"), d)
  glm2<-glm(dadd~v2, binomial(link = "logit"), d)
  glm3<-glm(dadd~v3, binomial(link = "logit"), d)
  glm4<-glm(dadd~v4, binomial(link = "logit"), d)
  tabela1<-anova(glm1)
  tabela2<-anova(glm2)
  tabela3<-anova(glm3)
  tabela4<-anova(glm4)
  dev<-c(tabela1[2,2], tabela2[2,2], tabela3[2,2], tabela4[2,2])
  return(dev)
}
y<-funglm(d)                          # vetor da deviance
fmaior<-function(y){ # ordenação do vetor da deviance para obtenção
  do maior valor
  maior<-numeric()
  maior<-y[1]
  for (j in 2:4){
    if (maior < y[j]){
      maior<-y[j]
    }
  }
  return(maior)
}
maior<-fmaior(y)
maior                                # maior valor da deviance
#####Bootstrap#####
pboot<-numeric()

```

```

pb<-numeric()
B<-numeric()
B<-2
for (h in 1:500){
  xb<-numeric(n)
  cont<-numeric()
  bsoma1<-numeric()
  vmaior<-numeric(B)
  for(j in 1:B) {
    xb<-sample(soma4, replace=TRUE) # reamostragem
    b1<-c(xb[1:10])
    b2<-c(xb[11:20])
    b3<-c(xb[21:30]) # divisão de 10 em 10
    b4<-c(xb[31:40])
    b5<-c(xb[41:50])
    bsoma1[1]<-sum(b1)
    bsoma1[2]<-sum(b2)
    bsoma1[3]<-sum(b3) # soma
    bsoma1[4]<-sum(b4)
    bsoma1[5]<-sum(b5)
    bsoma1
    bsmaior<-bsoma1[1]
    bpos<-1
    for (i in 2:5){
      if (bsmaior < bsoma1[i]){
        bsmaior<-bsoma1[i]
        bpos<-i
      }
    }
    bsmaior # vetor da maior soma dentro do boot
    bpos
    bsoma2<-bsoma1
    bsoma3<-order(bsoma2)
    bsoma3
    bsoma4<-numeric()
    #####inicio da ordenação do vetor bsoma4 #####
    #----- posição 1 do vetor soma3
    if (bsoma3[1]==1){
      for (i in 1:10){
        bsoma4[i]<-b1[i]
      }
    }
    .
    .
    .
    if (bsoma3[1]==5){
      for (i in 1:10){
        bsoma4[i]<-b5[i]
      }
    }
    }
    #----- posição 2 do vetor soma3
    .
    .
    .
    #----- posição 5 do vetor soma3
    if (bsoma3[5]==1){
      ij<-1
      for (i in 41:50){
        bsoma4[i]<-b1[ij]

```

```

        ij<-ij+1
    }
}
.
.
.
if (bsoma3[5]==5){
  ij<-1
  for (i in 41:50){
    bsoma4[i]<-b5[ij]
    ij<-ij+1
  }
}
bsoma4 # vetor ordenado dos dados
bd<-data.frame(rep=t,yy=bsoma4,v1=c1,v2=c2,v3=c3,v4=c4) # dados
com o vetor ordenado e o de 1 e 2
  bfunglm<-function(bd) { # função para o inicio do calculo
da deviance para os dados ordenados dentro do boot
  bglm1<-glm(yy~v1,binomial(link = "logit"),bd)
  bglm2<-glm(yy~v2,binomial(link = "logit"),bd)
  bglm3<-glm(yy~v3,binomial(link = "logit"),bd)
  bglm4<-glm(yy~v4,binomial(link = "logit"),bd)
  btabela1<-anova(bglm1)
  btabela2<-anova(bglm2)
  btabela3<-anova(bglm3)
  btabela4<-anova(bglm4)
  bdev<-c(btabela1[2,2],btabela2[2,2],btabela3[2,2],btabela4[2,2])
  return(bdev)
}
by<-bfunglm(bd)
by # vetor da deviance dentro do boot
bfmaior<-function(by){
  bmaior<-numeric()
  bmaior<-by[1]

  for (i in 2:4){
    if (bmaior < by[i]){
      bmaior<-by[i]
    }
  }
  return(bmaior)
} # fim do da função bfmaior
bmaior<-bfmaior(by)
bmaior # vetor do maior valor da deviance
vmaior[j]<-bmaior
dev<-maior # valor da maior deviance do modelo
original
for (j in 1:B) { # para obter os 0 e 1
  if (vmaior[j]>=dev)
    (cont[j]=1)
  else (cont[j]=0)
} # for do cont
} # for do reamostragem
cont # valor de 0 e 1 quando o valor for maior ou menor que o
valor da deviance do modelo original
pb[h]<-(1+sum(cont))/(B+1)
pboot<-(1-pb) # p-valor para B
B<-B+1
} # for do h
pboot # vetor dos p-valores

```

## Apêndice D: Programa em R para determinar o tamanho da simulação S.

```
semente<-515
set.seed(semente)
S<-2
spvalor<-numeric()
pvalor<-numeric(S)
for (is in 1:1000) {
  k<-5
  ni<-10
  n<-k*ni
  t<-gl(5,10)
  v<-rbinom(1,0.45, n=20)
  z<-rbinom(1,0.55, n=30)
  x<-c(v, z)
  dado<-data.frame(rep=t, dad=x)
  m1<-c(x[1:10])
  m2<-c(x[11:20])
  m3<-c(x[21:30])
  m4<-c(x[31:40])
  m5<-c(x[41:50])
  soma1<-numeric()
  soma1[1]<-sum(m1)
  soma1[2]<-sum(m2)
  soma1[3]<-sum(m3)
  soma1[4]<-sum(m4)
  soma1[5]<-sum(m5)
  soma1
  smaior<-soma1[1]
  pos<-1
  for (j in 2:5){
    if (smaior < soma1[j]){
      smaior<-soma1[j]
      pos<-j
    }
  }
  smaior
  pos
  soma2<-soma1
  soma3<-order(soma2)
  soma3
  soma4<-numeric()
#----- posição 1 do vetor soma3
  if (soma3[1]==1){
    for (i in 1:10){
      soma4[i]<-m1[i]
    }
  }
  .
  .
  .
  if (soma3[1]==5){
    for (i in 1:10){
      soma4[i]<-m5[i]
    }
  }
#----- posição 2 do vetor soma3
```

```

.
.
.
#----- posição 5 do vetor soma3
  if (soma3[5]==1){
    j<-1
    for (i in 41:50){
      soma4[i]<-m1[j]
      j<-j+1
    }
  }
.
.
.
  if (soma3[5]==5){
    j<-1
    for (i in 41:50){
      soma4[i]<-m5[j]
      j<-j+1
    }
  }
  soma4
###calculo inicial da deviance para os dados ordenados #
M1<-rep(1, times=10)
M11<-rep(2, times=40)
c1<-c(M1,M11)
M2<-rep(1, times=20)
M22<-rep(2, times=30)
c2<-c(M2,M22)
M3<-rep(1, times=30)
M33<-rep(2, times=20)
c3<-c(M3,M33)
M4<-rep(1, times=40)
M44<-rep(2, times=10)
c4<-c(M4,M44)
d<-data.frame(rep=t, dadd=soma4, v1=c1, v2=c2, v3=c3, v4=c4)
funglm<-function(d) {
  glm1<-glm(dadd~v1, binomial(link = "logit"), d)
  glm2<-glm(dadd~v2, binomial(link = "logit"), d)
  glm3<-glm(dadd~v3, binomial(link = "logit"), d)
  glm4<-glm(dadd~v4, binomial(link = "logit"), d)
  tabela1<-anova(glm1)
  tabela2<-anova(glm2)
  tabela3<-anova(glm3)
  tabela4<-anova(glm4)
  dev<-c(tabela1[2,2], tabela2[2,2], tabela3[2,2], tabela4[2,2])
  return(dev)
}
y<-funglm(d)
fmaior<-function(y) {
  maior<-numeric()
  maior<-y[1]
  for (j in 2:4){
    if (maior < y[j]){
      maior<-y[j]
    }
  }
  return(maior)
}
maior<-fmaior(y)

```

```

maior
#####Bootstrap#####
B<-250
xb<-numeric(n)
cont<-numeric()
bsoma1<-numeric()
vmaior<-numeric(B)
for(j in 1:B) { xb<-sample(soma4, replace=TRUE)
  b1<-c(xb[1:10])
  b2<-c(xb[11:20])
  b3<-c(xb[21:30])
  b4<-c(xb[31:40])
  b5<-c(xb[41:50])
  bsoma1[1]<-sum(b1)
  bsoma1[2]<-sum(b2)
  bsoma1[3]<-sum(b3)
  bsoma1[4]<-sum(b4)
  bsoma1[5]<-sum(b5)
  bsoma1
  bsmaior<-bsoma1[1]
  bpos<-1
  for (i in 2:5){
    if (bsmaior < bsoma1[i]){
      bsmaior<-bsoma1[i]
      bpos<-i
    }
  }
  bsmaior
  bpos
  bsoma2<-bsoma1
  bsoma3<-order(bsoma2)
  bsoma3
  bsoma4<-numeric()
#----- posição 1 do vetor soma3
  if (bsoma3[1]==1){
    for (i in 1:10){
      bsoma4[i]<-b1[i]
    }
  }
.
.
.
  if (bsoma3[1]==5){
    for (i in 1:10){
      bsoma4[i]<-b5[i]
    }
  }
#----- posição 2 do vetor soma3
.
.
#----- posição 5 do vetor soma3
if (bsoma3[5]==1){
  ij<-1
  for (i in 41:50){
    bsoma4[i]<-b1[ij]
    ij<-ij+1
  }
}
.

```

```

.
.
  if (bsoma3[5]==5) {
    ij<-1
    for (i in 41:50) {
      bsoma4[i]<-b5[ij]
      ij<-ij+1
    }
  }
  bsoma4
  bd<-data.frame(rep=t,yy=bsoma4,v1=c1,v2=c2,v3=c3,v4=c4)
  bfunglm<-function(bd) {
    bglm1<-glm(yy~v1,binomial(link = "logit"),bd)
    bglm2<-glm(yy~v2,binomial(link = "logit"),bd)
    bglm3<-glm(yy~v3,binomial(link = "logit"),bd)
    bglm4<-glm(yy~v4,binomial(link = "logit"),bd)
    btabela1<-anova(bglm1)
    btabela2<-anova(bglm2)
    btabela3<-anova(bglm3)
    btabela4<-anova(bglm4)
    bdev<-
c(btabela1[2,2],btabela2[2,2],btabela3[2,2],btabela4[2,2])
    return(bdev)
  }
  by<-bfunglm(bd)
  bfmaior<-function(by) {
    bmaior<-numeric()
    bmaior<-by[1]
    for (i in 2:4) {
      if (bmaior < by[i]) {
        bmaior<-by[i]
      }
    }
    return(bmaior)
  }
  bmaior<-bfmaior(by)
  bmaior
  vmaior[j]<-bmaior
  j<-j+1
}
vmaior
#-----
dev<-maior
for (j in 1:B){if (vmaior[j]>=dev) (cont[j]=1) else (cont[j]=0) }#
cont
cont
pboot<-(1+sum(cont))/(B+1)
pvalor[is]<-pboot
if(any(pvalor[is]>=0.05)) conts[is]<-1 else conts[is]<-0
poder<-sum(conts)/1000
poderboot<-1-poder

pvalor[is]<-pboot
spvalor[is]<-sum(pvalor)/is
is<-is+1
}
spvalor

```

## Referências

BARRETO,M.C.M; DEMÉTRIO,C.G.B. Um estudo da distribuição assintótica da medida de homogeneidade entre grupos da extensão da técnica de agrupamento AID considerando a distribuição de Poisson. *Revista Matemática Estatística*, São Paulo,v.16, p.191-207, 1998.

BARRETO,M.C.M; DEMÉTRIO,C,G,B.Uma extensão da técnica AID em modelos lineares generalizados. *Revista Matemática Estatística*, São Paulo, v.30, p.141-154, 1995.

BARRETO,M.C.M. *Uma extensão da técnica AID em modelos lineares generalizados*. Piracicaba, 1993, 200p. Tese de Doutorado - Escola Superior de Agronomia Luiz de Queiroz, Universidade de São Paulo.

BORGES.L.C.*Poder e taxas de erro I dos testes Scott-Knott sob distribuições normal e não normais dos resíduos*. 2002. 96p. Dissertação de Mestrado em Agronomia – Centro de Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Lavras.

DAVINSON.A.C; HINKLEY.D.V. *Bootstrap Methods and their Application*. 2 ed. Cambridge. 1997. 582p.

DEMÉTRIO.C.G.B. *Modelos Lineares Generalizados em Experimentação Agrônômica*. Apostila.113p. 2002. ESALQ – Universidade São Paulo campus Piracicaba.

EFRON, B. *Bootstrap Methods: Another look at the Jackknife Annals of Statistics*. V.7, p.126,1979.

EFRON,B; TIBSHIRANI. R. J. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993. 436p.

FERREIRA. D F; MUNIZ, J. A, AQUINO, L. H.. Comparações Múltiplas em Experimentos com grande número de Tratamentos – Utilização do teste de SCOTT-KNOTT. *Ciência e Agrotecnologia.*, Lavras, v.23, n.3, p. 745-742, 1999.

GARCIA. C.A. *Precisão relativa e Intervalo de Confiança Bootstrap para a Variância Populacional de uma Distribuição Normal Baseada na Amostragem por Conjuntos Ordenados*. 2004. 127p. Tese de Mestrado – Universidade Federal de São Carlos.

McCULLAGH, P. ;NELDER,S.A. *Generalized linear models*. 2 ed. Londres: Chapman and Hall, 1989. 511p.

MORGAN, J. N, SONQUIST, J.A. Problems in the analysis of survey data. *J. Am. Stat. Assoc.*, v.58, p.415-434, 1963.

PAULA.G.A. *Modelos de Regressão com Apoio Computacional*. 292p. 2000. São Paulo: IME- Universidade São Paulo campus São Paulo.

PRESOTTI, L.E.. *Adubação Potássica na Cultura da Couve-Chinesa (Brassica pekinensis)*. Trabalho de Graduação – Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, 34p, 2002.

RODRIGUES, M. A. S. *Árvores de Classificação*. Trabalho de Graduação – Universidade dos Açores, 32p, 2005

SANTOS. C. *Novas alternativas de testes de agrupamento avaliadas por meio de simulação Monte Carlo*. 2000 85p. Dissertação de Mestrado em Agronomia – Centro de Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Lavras.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, v. 30, n.2 p.507-512, 1974.

SILVA. E. C. *Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo*. 1998. 54p. Dissertação de Mestrado em Agronomia –

Centro de Estatística e Experimentação Agropecuária, Universidade Federal de Lavras,  
Lavras.

VON. PINHO, R. G.; VON PINHO, E. V. R.; FRAGA, A. C.; MACHADO, J. Avaliação de fontes de resistência á ramulose do algodoeiro causada por *Colletotrichum gossypi* var. *cephalosporioides*. *Ciência e Agrotecnologia*, Lavras, v.21, n.3, p.253-259, 1997.