COMPARAÇÕES MÚLTIPLAS PARA PARÂMETROS BINOMIAIS UTILIZANDO *BOOTSTRAP*

NÁDIA GIARETTA BIASE

2006

Nádia Giaretta Biase

COMPARAÇÕES MÚLTIPLAS PARA PARÂMETROS BINOMIAIS UTILIZANDO *BOOTSTRAP*

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Curso de Mestrado em Agronômia, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de "Mestre".

Orientador

Prof. Dr. Daniel Furtado Ferreira

LAVRAS

MINAS GERAIS-BRASIL

2006

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Biase, Nádia Giaretta.

Comparações múltiplas para parâmetros binomiais utilizando *bootstrap* / Nádia Giaretta Biase. - Lavras: UFLA, 2006.
68p. : il.

Orientador: Daniel Furtado Ferreira.
Dissertação (Mestrado) - UFLA.
Bibliografia.

1. Proporção binomial. 2. Método Monte Carlo. 3. *Bootstrap*.
I. Universidade Federal de Lavras. II.Título.

CDD-519.282
-519.52

Nádia Giaretta Biase

COMPARAÇÕES MÚLTIPLAS PARA PARÂMETROS BINOMIAIS UTILIZANDO *BOOTSTRAP*

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Curso de Mestrado em Agronômia, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de "Mestre".

APROVADA em 15 de fevereiro de 2006

Prof. Dr. Ednaldo Carvalho Guimarães	UFU
Prof. Dr. Augusto Ramalho de Morais	UFLA
Prof. Dr. Júlio Silvio de Sousa Bueno Filho	UFLA

Prof. Dr. Daniel Furtado Ferreira
UFLA
(Orientador)

LAVRAS
MINAS GERAIS-BRASIL

"A alegria está na luta, na tentativa,
no sofrimento envolvido.
Não na vitória propriamente dita."
(Mahatma Gandhi)

Dedico esta vitória:

Aos meus pais, José Luiz e Elisa, pelos inúmeros bons exemplos que propiciaram e que não pouparam esforços para minha formação e pelo infinito amor. Por tudo que sou, expresso minha gratidão e incondicional amor!

A minha irmã, Érica,
exemplo de determinação e coragem!

A minha irmã, Adriele,
exemplo de paciência e dedicação!

AGRADECIMENTOS

A Deus, força maior de todo ser humano, pela saúde, serenidade e força de vontade a mim concedidas para concluir mais uma etapa de estudo.

A Universidade Federal de Lavras, em especial ao Departamento de Ciências Exatas (DEX), pela oportunidade de realizar o mestrado.

À CAPES, pela concessão de bolsa de estudos.

Aos meus pais, José Luiz e Elisa, pelo grande apoio, incentivo, confiança e por se fazerem presentes nos momentos difíceis.

As minhas irmãs, Érica e Adriele, pelo amor e ternura, motivação e conselhos que me deram durante estes anos.

Ao professor Daniel Furtado Ferreira, pelos ensinamentos, dedicação e responsabilidade com que me orientou e, principalmente, pela amizade e disponibilidade em auxiliar-me a qualquer momento.

Ao professor Heyder Diniz Silva, da Universidade Federal de Uberlândia, pela amizade e orientação na iniciação científica, pelo incentivo inicial e por me fazer acreditar que era possível concretizar este sonho.

Aos professores do Departamento de Ciências Exatas, pelas condições que propiciaram para a realização do mestrado e aos funcionários do (DEX), pela eficiência e atenção prestadas.

A todos os meus colegas do curso, em especial às colegas Lívia e Verônica, pela amizade, companheirismo, troca de conhecimentos e atenção recebida durante estes anos.

À minha amiga, Maria Imaculada que, desde a graduação, tem demonstrado muito carinho, compreensão e preocupação comigo. Obrigada pela força e disposição em sempre me ajudar.

Ao meu amigo e futuro cunhado, Edivânio, que, por diversas vezes, foi tão prestativo, pelo carinho e amizade.

Às amigas de república, Gisele, Samantha, Andressa e Carla, pela amizade, paciência e alegria proporcionadas no convívio diário.

Aos amigos do Grupo Partilha e Perseverança (GPP), pelos inúmeros momentos de oração e alegrias compartilhadas e a todos os amigos de Lavras, em especial ao Itamar, Márcia, Muriel e Fabrícia, pela inesquecível

convivência.

Ao meu namorado, Leonardo, pela paciência, apoio e compreensão no decorrer destes anos.

A todos os meus familiares que acreditaram e colaboraram pelo meu sucesso profissional.

A todos aqueles que, de forma direta ou indireta, contribuíram para a realização desta etapa difícil, mas importante de minha vida, o meu sincero agradecimento.

Sumário

RESUMO

BIASE, Nádia Giaretta. **Comparações múltiplas para parâmetros binomiais utilizando *bootstrap*.** Lavras: UFLA, 2006. 68 p. Dissertação (Mestrado em Agronomia / Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, MG.*

A aplicação dos métodos de comparações múltiplas e a análise de variância não são alternativas viáveis para se comparar duas ou mais proporções binomiais, quando os experimentos são realizados considerando apenas repetições do evento de Bernoulli. Essa comparação pode ser feita por meio das técnicas de computação intensiva que utilizam *bootstrap* infinito. Este trabalho teve por objetivo avaliar a performance de dois testes de *bootstrap* envolvendo proporções binomiais, computando o erro tipo I por experimento e o poder. Esses dois testes de *bootstrap* infinito se diferenciam pelos estimadores de p_i utilizados. Em um dos testes foi considerado o estimador de máxima verossimilhança (MV) e, no outro, o estimador de Pan (Pan, 2002) que foram avaliados em diferentes configurações envolvendo número de populações e valores dos parâmetros n_i e p_i. O método de Monte Carlo foi utilizado para simular os experimentos, gerando-se 2.000 amostras para cada uma de duas etapas consideradas. Na primeira etapa, foram avaliadas as taxas de erro tipo I por experimento sob H_0 completa e parcial. As simulações sob a hipótese H_0 completa foram feitas para as combinações entre os valores dos parâmetros $p = 0,1;\ 0,5\ e\ 0,9$, números de populações $k = 2,\ 5\ e\ 10$ e tamanhos amostrais $n = 10,\ 30\ e\ 100$. Também foi avaliado o erro tipo I por experimento sob H_0 parcial, considerando uma diferença Δ entre os valores de p de dois grupos distintos. Numa segunda etapa, avaliou-se o poder dos testes sob H_0 parcial e sob H_1. Em ambas as etapas, as simulações foram realizadas adotando-se o valor nominal de significância de 1% e 5%. Os dois testes de *bootstrap* Pan e MV apresentaram excelentes performances, controlando o erro tipo I por experimento em níveis iguais ou inferiores aos valores nominais de significância e elevados valores de poder. Pelo fato de possuir uma performance melhor nas situações em que as proporções binomiais se afastam de 1/2 e os tamanhos amostrais são pequenos ($n \leq 10$), recomenda-se a utilização do teste *bootstrap* de Pan.

*__Comitê Orientador:__ Daniel Furtado Ferreira - UFLA. (Orientador)

ABSTRACT

BIASE, Nádia Giaretta. **Multiple comparison for binomial parameters using *bootstrap*.** Lavras: UFLA, 2006. 68 p. Dissertation (Master in Agronomy / Statistics and Agricultural Experimentation) - Federal University of Lavras, Lavras, MG.*

The multiple comparisons methods and the analysis of variance are not reliable alternatives for comparing two or more binomial proportions, when the experiments have only Bernoulli trails. Although, this comparison can be made using the intensive computational techniques named infinite bootstrap. This work aimed to evaluate the performance of two binomial proportions bootstrap tests computing the experimentwise type I error rates and the power. These two infinite bootstrap tests distinguished on the estimators of p_i. One of these tests considered the maximum likelihood estimator (ML) and the other the Pan's estimator (Pan, 2002) and they were evaluated in different configurations considering the number of populations and the parameters values, resultant of 2000 Monte Carlo simulations. In the first stage the experimentwise type I error rates were evaluated under complete null and partial H_0 hypotheses. The simulations under complete (H_0) were done in all combinations between parameters values $p = 0.1$; 0.5 and 0.9, number of populations $k = 2$, 5 and 10 and sample sizes $n = 10$, 30 and 100. The experimentwise type I error rate was also evaluated under partial H_0 considering a difference of Δ between values of p of distinct groups. In a second stage the powers of the tests were evaluated under partial H_0 and alternative hypotheses. Both simulations were done using 1% and 5% significance level. Pan´s and ML *bootstrap* tests showed excellent performance, because experimentwise error rate were always under their nominal levels. Powers of both procedures were high and they know best performance with extreme proportions ($p \neq 0.5$) and small sample sizes $n < 10$, when Pan´s bootstrap test is preferable.

*Guidance Committee: Daniel Furtado Ferreira - UFLA. (Adviser)

1 INTRODUÇÃO

Em várias situações reais, o pesquisador se depara com a necessidade de comparar duas ou mais proporções binomiais. A estratégia utilizada consiste em realizar estimação para a diferença das proporções das populações tomadas duas a duas ou em aplicar algum tipo de teste de hipótese. Quando o número de populações é maior do que dois, a segunda alternativa é, geralmente, aplicada por meio de uma análise de variação, principalmente se forem utilizadas repetições experimentais em delineamentos simples ou complexos. Inicialmente, é aplicado um teste F para a igualdade de todas as proporções e, posteriormente, se condicionada a rejeição dessa hipótese, é comum aplicar testes de comparações múltiplas, como, por exemplo, Tukey, Duncan, Sheffé e Student-Newman-Keuls (SNK).

A validade destes testes depende de algumas pressuposições, tais como normalidade dos resíduos, homogeneidade de variâncias e independência das observações. Em geral, a independência é garantida pela casualização e, mesmo que isso não tenha ocorrido, o teste ainda continua válido. Para o caso de testes sobre proporções binomiais não existe normalidade, a não ser de forma aproximada. Finalmente, pode-se constatar que a homogeneidade de variâncias também é um dos pressupostos não atendidos. Para as populações binomiais é bem conhecido o fato de a variância ser uma função da média. Assim, espera-se que as variâncias das diversas populações sejam, em geral, heterogêneas.

Uma das alternativas existentes para solucionar este problema são os modelos lineares generalizados. A inferência bayesiana é outra abordagem que vem sendo largamente empregada. Os modelos lineares generalizados constituem-se em uma generalização dos modelos lineares clássicos, em que a variável resposta possui distribuição de probabilidade pertencente à família exponencial. A inferência bayesiana depende de escolhas de modelos probabilísticos, baseados no conhecimento a priori dos pesquisadores sobre os parâmetros.

Por outro lado, testes de hipóteses e estimativas de parâmetros têm

sido realizados por meio de técnicas computacionais intensivas. Entre estas técnicas, o método de *bootstrap* tem se destacado, uma vez que possibilita obter a estimativa do parâmetro sem a necessidade de pressupor a distribuição do estimador. Conlon e Thomas (1990) introduziram uma técnica conhecida como *bootstrap* infinito, que possibilita realizar testes de hipóteses e estimar, por intervalo, parâmetros da binomial ou funções desses parâmetros que sejam de interesse.

Quando os experimentos sobre as populações binomiais são realizados sem considerar repetições experimentais, mas considerando apenas repetições do evento de Bernoulli, a análise de variância e os testes de comparações múltiplas ficam inviabilizados. Assim, as técnicas de computação intensiva que utilizam *bootstrap* infinito se tornam relevantes. Particularmente no caso de proporções nas quais os dados seguem distribuição binomial, dois estimadores do parâmetro de interesse podem ser utilizados, quais sejam, o estimador de máxima verossimilhança e o estimador de Pan (2002). Esse último tem como característica a utilização de quatro pseudo-observações, sendo duas delas consideradas como sucessos do evento de interesse.

O presente trabalho teve por objetivo realizar comparações múltiplas em populações binomiais utilizando *bootstrap* infinito e avaliar a sua performance computando-se o erro tipo I por experimento e o poder. Adicionalmente, o método de *bootstrap* infinito será avaliado considerando os estimadores de máxima verossimilhança e de Pan (2002) em diferentes configurações envolvendo número de populações e valores dos parâmetros n_i e p_i (tamanho da amostra e proporção da i-ésima população).

2 REFERENCIAL TEÓRICO

2.1 Procedimentos de comparações múltiplas

Os procedimentos de comparações múltiplas são largamente utilizados em diversas áreas da ciência. O assunto é tão vasto e importante que existem livros completos sobre estes métodos, inúmeras revistas científicas contendo artigos que abordam direta ou indiretamente o tema e, ainda, existe uma grande quantidade de trabalhos que citam estes procedimentos.

O objetivo principal em uma análise estatística de dados, em experimentos agronômicos, é estabelecer o maior número possível de informações sobre os tratamentos aplicados nas unidades experimentais (Petersen, 1977). A aplicação dos procedimentos de comparações múltiplas depende da natureza dos efeitos dos fatores em estudo. Quando os níveis destes fatores são quantitativos, a utilização de uma metolodogia de regressão é mais conveniente e, se os fatores são qualitativos com uma estruturação, é mais apropriado estabelecer comparações entre os níveis de um dos fatores por meio de contrastes, seguidas de um teste específico. No entanto, se os níveis dos fatores são qualitativos e não são estruturados, devem-se aplicar os procedimentos de comparações múltiplas (Machado et al., 2005).

Segundo Hochberg & Tanhane (1987), as comparações múltiplas entre efeitos de tratamentos são utilizadas na prática somente quando, na análise de variância, o teste F para a igualdade dos efeitos de tratamento é significativo. Os testes de comparações múltiplas servem como um complemento do teste F, para detectar possíveis diferenças entre os tratamentos (Banzatto & Kronka, 1989).

Quando os testes de comparações múltiplas são utilizados de maneira incoerente, pode haver perda de informações e redução da eficiência se procedimentos mais apropriados forem avaliados (Petersen, 1977). O freqüente uso inapropriado de comparações múltiplas deve-se ao ensinamento incorreto e, também, à resistência de não estatísticos em se

aventurarem no território desconhecido da especificação de contrastes (Pearce, 1993).

Uma situação em que se deve ter cuidado ao utilizar os procedimentos de comparações múltiplas é o caso de análises de ensaios fatoriais que freqüentemente têm aplicado estas comparações de maneira incorreta. Em experimentos deste tipo, independente dos fatores envolvidos serem quantitativos ou qualitativos, devem, em primeiro lugar, ser testada a significância dos efeitos dos fatores principais e das interações. Se os efeitos das interações são não significativos, então, toda a informação está contida nos efeitos dos fatores principais e, neste caso, as médias de cada nível de um fator em todos os níveis dos outros fatores podem ser comparados recorrendo-se a métodos de comparações múltiplas. Agora, se as interações forem significativas, existe dependência entre os efeitos dos fatores principais e, asssim sendo, não se podem estudar os efeitos principais isoladamente. Deve-se proceder o estudo dos efeitos de um dos fatores em cada nível do outro fator, por meio de contrastes ortogonais ou testes de comparações múltiplas (Petersen, 1977).

Apesar da facilidade de aplicação dos testes de comparações múltiplas, um aspecto a ser considerado quando se aplica esses testes é a ambigüidade dos seus resultados. Essa ambigüidade é um complicador adicional nas interpretações e nas decisões a serem tomadas pelo experimentador e decorre da possibilidade de que dois tratamentos, considerados como iguais a um terceiro, podem ser considerados diferentes entre si (Ramalho et al., 2000).

Qualquer que seja o procedimento de comparações múltiplas utilizado, a diferença observada entre quaisquer duas médias é comparada com um valor crítico apropriado em cada procedimento. Se essa diferença observada exceder o valor crítico, as duas médias são consideradas significativamente diferentes, caso contrário, não significativamente diferentes. Uma vez que as magnitudes dos valores críticos variam de procedimento para procedimento, resultados obtidos da aplicação de um procedimento a um grupo de dados irão diferir dos resultados obtidos se um outro procedimento é aplicado aos mesmos dados (Carmer & Swanson, 1973).

Os testes de comparações múltiplas mais utilizados na literatura são os de t de Student, Tukey, Student-Newman-Keuls, Duncan, Sheffé e outros que podem ser encontrados em vários livros voltados para a estatística experimental como os de Pimentel Gomes (1985), Banzatto e Kronka (1989) e Steel & Torrie (1980), entre outros.

O teste t de Student, conhecido também como critério da Diferença Mínima Significativa (LSD do inglês *Least Square Difference*) é realizado por meio da estatística:

$$LSD = t_{(\nu,\ \alpha/2)}\sqrt{\frac{2QME}{r}}$$

em que $t_{(\nu,\ \alpha/2)}$ é o quantil superior $100(\alpha/2)\%$ da distribuição t de Student com ν números de graus de liberdade; QME é o quadrado médio do resíduo e r é o número de repetições.

Este teste controla apenas o erro por comparação em um nível nominal máximo igual a α e, por este motivo, muitos pesquisadores recomendam o seu uso somente para realizar comparações planejadas inicialmente (Machado et al., 2005). Além disso, este teste apresenta a inconveniência de possuir a maior taxa de erro por experimento em relação aos demais testes de comparações múltiplas, como o teste de Tukey, Duncan e Student-Newman-Keuls (SNK)(Ramalho et al., 2000).

Um procedimento aplicado para preservar a taxa de erro por experimento consiste em utilizar o teste t protegido por Fisher. Este teste exige o cálculo do teste preliminar da hipótese nula global, baseado no valor observado da razão F obtida pelo quociente entre o quadrado médio do tratamento e o quadrado médio do erro, antes de se usar o teste LSD. Se o valor de F é significativo, então, o teste LSD pode ser aplicado normalmente e significa existir pelo menos uma diferença entre os tratamentos. Agora, se o valor de F é não significativo, nenhuma comparação de médias deve ser efetuada, eliminando a possibilidade de cometer o erro tipo I (Carmer & Swanson, 1973).

No entanto, o teste t protegido por Fisher não garante resultados satisfatórios, uma vez que, na maioria das situações reais, a hipótese H_0 é apenas parcialmente verdadeira, isto é, pelo menos um dos tratamen-

tos difere dos demais. Assim, o teste F terá significância com muita frequência nestas situações reais e a taxa de erro por experimento, incluindo comparações de médias homogêneas, não será controlada tendo valor superior ao nível de significância nominal adotado (Machado et al., 2005).

Para contornar este problema, existe outra alternativa para preservar a taxa de erro por experimento baseada na desigualdade de Bonferroni. Este procedimento consiste em alterar o nível nominal de significância α para a determinação do valor tabelado de t, dividindo-se o nível nominal de significância pelo número de comparações que serão realizadas. No caso de se realizarem todas as comparações múltiplas duas a duas, o nível de significância será:

$$\alpha_p = \frac{2\alpha}{k(k-1)}$$

em que p é o número de médias ordenadas abrangidas pelo contraste e k é igual ao número de níveis do fator que se deseja comparar.

Dessa forma, o valor crítico do teste de Bonferroni é:

$$LSDB = t_{(\nu,\ \alpha_p/2)}\sqrt{\frac{2QME}{r}}$$

em que $t_{(\nu,\ \alpha_p/2)}$ representa o quantil superior 100 $(\alpha_p/2)$ % da distribuição t de Student com ν números de graus de liberdade do resíduo.

Verifica-se que o teste LSD protege apenas contra o erro tipo I por comparação, já o teste LSD protegido por Fisher controla o erro tipo I por comparação em todos os casos e o erro tipo I por experimento sob H_0 completa, mas não controla o erro tipo I sob H_0 parcial. Finalmente, o teste LSDB controla o erro tipo I por comparação e por experimento com limite máximo das taxas iguais ao valor nomimal (Machado et al., 2005).

Ainda existe um outro método, conhecido como *false discorery rate* (FDR), proposto por Benjamim & Hochberg (1995), que estima taxas de erro tipo I em experimentos que envolvem comparações múltiplas. A FDR é definida pela esperança matemática da razão entre o número de erro tipo I cometido e o número total de hípoteses nulas rejeitadas. Este

procedimento estatístico é relativamente novo e limita o número de erros cometidos ao executar estas comparações. Estes autores recomendam a utilização deste método, pois ele garante valores de poder mais elevados do que o método de Bonferroni. O método de Bonferroni procura controlar a possibilidade de uma única rejeição da hipótese nula verdadeira entre todos os testes executados. Já o método que utiliza a FDR controla a proporção dos erros entre aqueles testes cuja a hipótese nula foi rejeitada. É um método rápido e fácil de computar e pode ser trivialmente adaptado a trabalhos com dados correlacionados.

O método de Tukey requer que todos os níveis de tratamento tenham o mesmo número de repetições e que as inferências de interesse sejam todo o conjunto de comparações duas a duas, ou seja, todos os pares de estimativas (Ramalho et al., 2000). A diferença mínima significativa (Δ) do teste é dada por:

$$\Delta = \ q_{(\alpha,\ k,\ \nu)}\sqrt{\frac{1}{2}\hat{V}(D)} = \ q_{(\alpha,\ k,\ \nu)}\sqrt{\frac{QMR}{r}}$$

em que:

$q_{(k,\ \nu,\ \alpha)}$ é o valor tabelado da amplitude estudentizada, no qual k representa o número de tratamentos, ν o número de graus de liberdade do erro e α o valor nominal de significância; $\hat{V}(D)$ é o estimador da variância de D, em que D é a diferença entre duas médias de tratamentos dada por: $D = \overline{Y}_{i.} - \overline{Y}_{j.}$; QME é o quadrado médio do resíduo e r é o número de repetições de cada tratamento.

Assim, todos os valores de D que superarem o valor de Δ serão considerados estatisticamente significativos e a hipótese $H_0 : \ \mu_i - \mu_j = 0$ deve ser rejeitada a α de significância, estabelecida em favor da hipótese alternativa H_1. O teste de Tukey é exato para testar a maior diferença entre as médias dos tratamentos e, nos demais casos, é considerado conservativo (Perecin e Malheiros, 1989). Para um mesmo valor de α e para $k > 2$, o valor de Δ é maior do que a estatística do teste LSD. Desse modo, o teste de Tukey é um procedimento que apresenta resultados mais conservativos do que o teste t de Student (Carmer & Swanson, 1973).

Este teste é útil em situações em que se deseja obter informações

preliminares no uso do experimento como um todo, na determinação de diferenças significativas ou para determinar intervalos de confiança para diferenças de médias populacionais. Em um programa de melhoramento vegetal, quando há muitas variedades, esse teste pode ser usado nas fases preliminares, quando o material é muito heterogêneo, para eliminar as piores ou selecionar as melhores (Perecin e Malheiros, 1989).

Segundo Ramalho et al. (2000), para o caso em que os tratamentos são desbalanceados, é recomendável, no lugar de r, usar a média harmônica, r_h, do número de repetições de cada tratamento dada por:

$$r_h = \frac{1}{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r_i}}$$

em que r_i é o número de repetições do tratamento i.

A finalidade do teste de Tukey é controlar a taxa de erro por experimento, sendo este teste bastante conservativo em relação à taxa de erro por comparação. O poder deste teste é baixo quando comparado com os demais testes de comparações múltiplas e apresenta uma redução drástica com o aumento do número de níveis dos tratamentos (Ramalho et al., 2000).

Enquanto os testes t de Student e de Tukey, exigem o cálculo de um único valor crítico, o teste de Student-Newman-Keuls (SNK) necessita do cálculo de $(k-1)$ valores críticos (Carmer & Swanson, 1973). O teste de Student-Newman-Keuls é um procedimento seqüencial baseado na amplitude estudentizada, válido para a totalidade dos contrastes de duas médias e, a princípio, exige a condição de balanceamento dos tratamentos (Perecin e Malheiros, 1989).

A diferença mínima significativa para o teste de SNK é definida por:

$$SNK_p = q_{(\alpha,\ p,\ \nu)}\sqrt{\frac{QME}{r}}$$

em que :

$q_{(\alpha,\ p,\ \nu)}$ são os valores tabelados da distribuição da amplitude estudentizada e dependem do nível α de significância, das $p = 2, 3, \ldots, k$ médias envolvidas pelo contraste e dos ν números de graus de liberdade do erro.

Um outro teste de comparação múltipla é o teste de Scheffé. Este teste pode ser aplicado a qualquer contraste de médias e, apesar de ser um teste mais conservativo que o teste t de Student, ele também é mais flexível, pois não exige a condição de ortogonalidade entre os contrastes e nem que estes sejam estabelecidos antes de se examinar os dados (Banzatto & Kronka, 1989). O seu valor crítico é calculado por:

$$S = \sqrt{(k-1)\sum_{i=1}^{k} \frac{c_i^2}{r} F_{(\alpha,\ k-1,\ \nu)} QME}$$

em que: k é o número de tratamentos do experimento; c_i são os coeficientes das médias do contraste em questão com $i = 1, 2, \ldots, k$; $F_{(\alpha,\ k-1,\ \nu)}$ é o quantil superior $100\alpha\%$ da distribuição F com $k-1$ e ν números de graus de liberdade do resíduo. Se, ao comparar o módulo da estimativa do contraste ($\hat{Y}$) com a estatística do teste, verificar-se que $|\hat{Y}| > S$, deve-se concluir que o contraste é significativo a α de significância, indicando que os grupos de médias confrontados no contraste diferem entre si.

O teste de Duncan tem as mesmas características que o teste de Tukey no que se refere às pressuposições exigidas para sua aplicação. A principal diferença entre estes testes é que, no teste de Duncan, o nível α de significância é alterado em função do número de médias abrangidas pelo contraste. Por essa razão, somente os contrastes entre médias ordenadas são considerados neste teste (Ramalho et al., 2000).

Para um contraste que abrange p médias ordenadas, o valor do nível de significância, proposto por Duncan (1955), considerado em cada passo da aplicação do teste, é dado por:

$$\alpha_p = 1 - (1-\alpha)^{p-1} \quad (2 \leq p \leq k)$$

Esse nível de significância proposto por Duncan (1955) fornece uma proteção separada para cada comparação par a par, em um nível nominal de significância α. Isso implica que o teste controla a taxa de erro por comparação, mas, não controla a taxa de erro por experimento (Ramalho et al., 2000). Este teste fornece resultados com maior poder de

discriminação que os do teste de Tukey, além de ser menos conservativo (Banzatto & Kronka,1989).

A diferença mínima (D_p) de Duncan é:

$$D_p = q_{(\alpha_p,\ p,\ v)}\sqrt{\frac{QME}{r}}$$

cujos termos já foram todos especificados anteriormente. Os valores de q podem ser encontrados em tabelas apropriadas em textos clássicos de estatística experimental (Steel & Torrie, 1980).

Observe-se que a expressão é semelhante ao teste de SNK. A diferença é que o nível de significância α do teste SNK é constante e, em ambos os testes, o número de médias abrangidas (p) varia em cada comparação.

Para muitas médias, como ocorre em programas de melhoramento, não há um procedimento ideal. Testes como Tukey, Bonferroni ou Scheffé tornam-se extremamente conservativos, ou seja, o nível de significância real para a maioria dos contrastes é muito mais baixo que o valor nominal. O inverso ocorre com os testes de Duncan e t de Student. O teste de Student-Newman-Keuls (SNK), embora muito trabalhoso, pode ser uma solução (Perecin & Malheiros, 1989).

O procedimento de Scott e Knott (SK) utiliza a razão de verossimilhança para testar a hipótese de que k tratamentos podem ser divididos em dois grupos que maximizam a soma de quadrados entre grupos. Quando o número (k) de tratamentos é grande, o número de grupos cresce exponencialmente, dificultando a aplicação do teste. É importante salientar que existem $2^{k-1} - 1$ partições possíveis das k médias em dois grupos distintos (Ramalho et al., 2000). Para contornar este problema, deve-se ordenar as médias dos tratamentos e, com isso, o número de partições possíveis passa a ser obtido por $k - 1$.

Para aplicar-se o método de Scott e Knott, deve-se proceder da seguinte maneira:

i- ordenar as k médias e dividir os tratamentos em dois grupos, para todas as $k - 1$ partições possíveis dos valores médios ordenados;

ii- determinar a soma de quadrados máxima entre dois grupos. Essa soma de quadrados será definida por B_0 e será estimada da seguinte

maneira:

$$B_0 = \frac{T_1^2}{K_1} + \frac{T_2^2}{K_2} - \frac{(T_1 + T_2)^2}{K_1 + K_2}$$

sendo T_1 e T_2 os totais dos dois grupos, com K_1 e K_2 tratamentos em cada um, isto é:

$$T_1 = \sum_{i=1}^{k_1} \overline{Y}_{(i)} \quad e \quad T_2 = \sum_{i=k_1+1}^{k} \overline{Y}_{(i)}$$

em que $Y_{(i)}$ é a média do tratamento da posição ordenada i;

iii- determinar o valor da estatística λ da seguinte forma:

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

em que:

π é uma constante que equivale a 3,141593;

B_0 é o valor da soma de quadrados máxima entre dois grupos tomados sobre todas as $(k-1)$ partições possíveis, com k número de tratamentos envolvidos no grupo de médias considerado;

σ_0^2 é o estimador de máxima verossimilhança de $\sigma_{\overline{Y}}^2$ dado por:

$$\sigma_0^2 = \frac{1}{k + \nu} \left[\sum_{i=1}^{k} (\overline{Y}_{(i)} - \overline{Y})^2 \right] + \nu s_{\overline{Y}}^2$$

sendo:

$\overline{Y}_{(i)}$ a média do tratamento i ordenada;

$\overline{Y}$ a média geral do experimento;

$s_{\overline{Y}}^2 = \frac{QME}{r}$ a variância da média;

ν o número de graus de liberdade associados a este estimador;

iv- se $\lambda \geq \chi^2_{(\alpha;\ k/(\pi-2))}$, rejeita-se a hipótese de que os dois grupos são idênticos em favor da hipótese alternativa de que os dois grupos diferem;

v- no caso de rejeitar essa hipótese, os dois subgrupos formados serão independentemente submetidos aos passos (i) a (iii), fazendo, respectivamente, $n = K_1$ e $n = K_2$. O processo em cada subgrupo se encerra ao aceitar H_0 no passo (iv) ou se cada subgrupo contiver apenas uma média.

2.2 Simulação pelo método de Monte Carlo

A simulação é um processo que tenta reproduzir, por meio de programas de computadores, o comportamento de um sistema real, com a finalidade de estudar seu funcionamento sob condições alternativas (Dachs, 1988).

Para estudar ou avaliar um teste estatístico, muitas vezes, torna-se bastante difícil obter, analiticamente, informações sobre o poder e taxas de erro tipo I. Uma maneira de se obter as informações desejadas é por meio de simulações (Cecchetti, 1999).

Em grande parte dos trabalhos que envolvem simulação, sempre está associado o termo "método de Monte Carlo", que se refere ao uso de técnicas computacionais que geram amostras de acordo com determinadas distribuições de probabilidades, visando estudar novos comportamentos de diferentes técnicas estatísticas que poderiam ser empregadas num problema específico (Dachs, 1988).

O nome Monte Carlo está relacionado com a cidade de mesmo nome, no Principado de Mônaco, onde ocorriam jogos de azar. Assim, o uso atual do nome Monte Carlo envolve todos os mecanismos de simulação que utilizam variáveis aleatórias em seu sistema ou modelo. Essas variáveis aleatórias eram geradas manualmente ou mecanicamente. Atualmente, usam-se computadores para gerá-las (Bussab e Moretttin, 2004). Com o crescente avanço dos computadores, torna-se mais fácil a prática de simulação de variáveis aleátorias ou de amostras baseadas em modelos estatísticos apropriados com parâmetros conhecidos, com o objetivo de verificar a adequação de determinada metodologia ou na realização de comparações entre métodos (Dachs, 1988).

Em síntese, o método de simulação Monte Carlo consiste em simular dados a partir de uma seqüência de números aleatórios, com a finalidade de se obter uma amostra da população. Admite-se também que todo processo simulado que contém um componente aleatório de qualquer distribuição faz parte deste método (Carari, 2004).

2.3 Tipos de erro e poder do teste

Quando se realiza um teste de hipóteses, o interesse maior está na tomada de decisões a partir da aceitação ou não da hipótese referente ao parâmetro populacional (Steel & Torrie, 1980).

É preciso ter sempre a consciência de que toda inferência realizada está sujeita a erros. Esses erros podem ser classificados em três categorias. A primeira delas é conhecida como erro tipo I e ocorre, por definição, quando rejeita-se a hipótese nula sendo ela verdadeira, com uma probabilidade α, isto é, $\alpha = P(erro\ tipo\ I) = P(rejeitar\ H_0\ |H_0\ é\ verdadeira)$ (Bussab & Morettin, 2004). De modo geral, a probabilidade α de se cometer um erro do tipo I é um valor arbitrário e recebe o nome de valor nominal de significância e esse é o único tipo de erro sob controle do pesquisador. Alternativamente, comete-se o erro tipo II quando não se rejeita a hipótese nula dado que ela é falsa e a probabilidade de cometê-lo é representado por β, ou seja, $\beta = P(erro\ tipo\ II) = P(não\ rejeitar\ H_0|H_0\ é\ falsa)$. Esse erro não pode ser controlado diretamente.

Esses dois tipos de erros estão de tal forma associados que se a probabilidade de ocorrência de um deles diminuir, automaticamente a probabilidade de ocorrência do outro aumenta (Banzatto & Kronka, 1989). Assim, manter controle conservativo da taxa de erro tipo I causa um aumento na probabilidade de ocorrência do erro tipo II. Portanto, é preciso tomar algumas precauções, de modo que não se deve aplicar um teste a um nível muito baixo de probabilidade para não aumentar exageradamente a probabilidade de ocorrência do erro tipo II. Usualmente não se conhecem valores fixos para o parâmetro sob a hipótese alternativa, o que dificulta a determinação do valor de β (Bussab & Morettin, 2004).

São muitas as dificuldades em se avaliar o erro tipo I nos testes de comparações múltiplas e, em grande parte destes testes, a atenção está centrada nesse tipo de erro. Entre outras, existem duas formas de medir esse erro (Steel e Torrie, 1980). A razão entre o número de erro tipo I (concluindo que $\mu_i \neq \mu_j$ quando $\mu_i = \mu_j$) e o número total de compara-

ções realizadas é definida como taxa de erro por comparação, chamada de "comparisonwise error rate" . Em outras palavras, é a probabilidade de se rejeitar uma hipótese verdadeira em todas as possíveis combinações de médias de tratamentos tomadas duas a duas, ou seja:

$$\frac{N\acute{u}mero\ de\ decisões\ erradas}{N\acute{u}mero\ total\ de\ decisões}.$$

e a razão entre o número de experimentos com um ou mais erros tipo I e o número total de experimentos é definida como a taxa de erro por experimento, chamada de "experimentwise error rate" , isto é:

$$\frac{N\acute{u}mero\ de\ experimentos\ com\ pelo\ menos\ uma\ decisão\ errada}{N\acute{u}mero\ total\ de\ experimentos}.$$

Ao aplicar procedimentos de comparações múltiplas, o pesquisador pode ter interesse na probabilidade de se cometer pelo menos um erro do tipo I em uma série de k comparações. Se os k testes forem todos independentes, a expressão que representa a probabilidade de não cometer o erro do tipo I em nenhum dos testes é $(1-\alpha)^k$. E $1-(1-\alpha)^k$ expressa a probabilidade de não haver erro do tipo I em, pelo menos, um dos testes. Portanto, esta é a probabilidade máxima, numa seqüência de testes, de cometer pelo menos um erro do tipo I (Machado et al., 2005)

O terceiro e último tipo de erro, conhecido como erro tipo III, refere-se à probabilidade de classificar um nível de tratamento como superior ao outro, quando, na verdade, o segundo nível supera o primeiro, isto é, rejeita-se corretamente a hipótese nula, dado que ela é falsa, a favor de uma hipótese alternativa errada (Ramalho et al., 2000).

O poder de um teste é definido pela probabilidade de se rejeitar a hipótese nula H_0 quando, na verdade, ela é falsa e é dada por $(1-\beta)$ (Mood et al., 1974). O controle da taxa de erro tipo I real, de forma a garantir que o nível de probabilidade desejado em um conjunto de várias comparações seja alcançado, leva a uma redução do poder dos testes. Os métodos que se baseiam nesse tipo de controle são considerados conservativos por garantirem proteção excessiva. Em um teste conservativo, a probabilidade de se encontrar um resultado significativo falso (erro tipo I) é inferior ao valor α estabelecido (Snedecor & Cochran, 1980). De

preferência, devem-se estabelecer amostras com tamanhos relativamente grandes para poder reduzir a probabilidade de se cometer o erro tipo II, fixando um baixo risco α para a probabilidade de se cometer o erro tipo I (Guerra & Donaire, 1982).

2.4 *Bootstrap*

O *bootstrap* é um método computacional desenvolvido recentemente e é usado para obter estimativas de parâmetros (Efron & Tibshirani, 1993).

A técnica de *bootstrap* consiste em reamostrar a amostra baseada na premissa de que, na ausência de qualquer outro conhecimento da população, os valores encontrados em uma amostra aleatória de tamanho n são os melhores guias da distribuição da população. A única diferença que existe entre *bootstrap* e teste de permutação ou aleatorização é que no *bootstrap* a amostragem é feita com reposição (Manly, 1998). Segundo Crowley (1992), esta técnica baseia-se na reamostragem de dados reais para mostrar algum padrão neles existente, possibilitando assim o cálculo da precisão das estimativas por meio de limites de confiança ou probabilidades.

A partir de uma amostra aleatória de tamanho n de uma população, é obtida, com reposição, uma nova amostra de tamanho n dessa amostra. Cada amostra obtida por reamostragem é uma amostra *bootstrap*. O processo de *bootstrap* é executado inúmeras vezes, de modo que as estimativas dos parâmetros sejam obtidas e essas estimativas geram uma distribuição denominada distribuição de *bootstrap*. Em alguns casos, a estimativa do parâmetro de interesse obtida na amostra original é utilizada como parâmetro da função de densidade da variável aleatória correspondente. Essa densidade é uma estimativa da verdadeira densidade populacional. Por meio dela são realizadas amostragens e em cada etapa é obtida uma estimativa do parâmetro de interesse. Repetido esse processo milhares de vezes, pode-se gerar a distribuição de *bootstrap* que, nesse caso particular, é denominada de *bootstrap* infinito (Conlon

& Thomas, 1990). Esses autores apresentaram este procedimento para a distribuição binomial, que é o foco de interesse deste trabalho.

O fato das medidas de precisão serem obtidas diretamente dos dados, não dependendo completamente do Teorema do Limite Central (TLC), favorece o método do *bootstrap* em suas aplicações (Efron & Tibshirani, 1993).

2.5 Distribuições

2.5.1 Distribuição Bernoulli

Quando é necessário descrever uma determinada população, os pesquisadores utilizam famílias de distribuições caracterizadas por parâmetros. Essas distribuições, dependendo da situação, podem ser discretas ou contínuas. Uma das mais simples distribuições de variáveis aleatórias discretas é a distribuição Bernoulli.

Nos experimentos em que o espaço amostral tem apenas dois resultados possíveis, sucesso ou fracasso, e a cada resposta está associada uma determinada probabilidade, usa-se a distribuição Bernoulli para representar o fenômeno (Ferreira, 2005). Como exemplo de variáveis com este tipo de distribuição, tem-se que o sexo do primeiro filho de um casal será feminino ou masculino, que uma determinada semente germinará ou não ou que, no lançamento de uma moeda, o resultado será cara ou coroa. Nestes casos, associando uma variável aleatória Y aos possíveis resultados, define-se que Y assumirá o valor 1 se ocorrer o sucesso e 0 se ocorrer o fracasso.

Assim, uma variável aleatória Y é definida como tendo uma distribuição Bernoulli se a função de probabilidade é dada por:

$$P(Y = y) = \begin{cases} p^y(1-p)^{1-y} & \text{para y=0 ou y=1} \\ 0 & \text{caso contrário} \end{cases}$$

em que o parâmetro p representa a probabilidade de ocorrer o sucesso e $q = (1 - p)$ a probabilidade de ocorrer o fracasso (Mood et al., 1974).

Se Y tem uma distribuição Bernoulli, então, a esperança matemática de Y é $E(Y) = p$ e a variância de Y é $Var(Y) = pq = p\ (1-p)$.

Os experimentos que resultam numa variável Bernoulli são chamados ensaios de Bernoulli. Se um ensaio for realizado n vezes, de modo que o resultado de um ensaio não tenha influência alguma sobre o outro, e ainda, se a probabilidade p de obter o sucesso for constante para cada ensaio, origina-se uma nova distribuição, conhecida como distribuição Binomial (Mood et al., 1974).

2.5.2 Distribuição binomial

Considere uma população em que a probabilidade de elementos portadores de uma certa característica (sucesso) é p e a probabilidade de não ocorrência (fracasso) é $q = 1 - p$. Desta população retire, com reposição, todas as amostras aleatórias simples (AAS) possíveis de tamanho n. Se Y representa o número de sucessos obtidos nas n tentativas independentes, então, a variável aleatória Y terá uma distribuição binomial, ou seja, $Y \sim B(n, p)$ (Steel e Torrie, 1980).

A distribuição de uma variável aleatoria Y é definida por Mood et al. (1974) como uma distribuição binomial se a função probabilidade de Y é dada por:

$$\begin{aligned} P(Y=y) &= \begin{cases} \binom{n}{y} p^y q^{n-y} & \text{para } y = 0, 1, 2 \ldots, n \\ 0 & \text{caso contrário} \end{cases} \\ &= \binom{n}{y} p^y q^{n-y} I_{\{0,1,\ldots,\ n\}}(y) \end{aligned}$$

em que o parâmetro p satisfaz $0 \leq p \leq 1$, $q = 1 - p$, $\binom{n}{y} = \dfrac{n!}{y!(n-y)!}$ e $I(\bullet)$ é a função indicadora.

A aplicação da distribuição binomial, que é considerada uma das mais importantes distribuições discretas, não é limitada à modelagem probabilística de certos fenômenos e abrange uma série de procedimentos de estimação e inferência (Ferreira, 2005).

Assim, quando deseja-se fazer inferência sobre uma amostra de tamanho n, a proporção de indivíduos portadores da característica na amostra é dada por: $\hat{p} = \frac{y}{n}$, em que y indica o total de sucessos na amostra de tamanho n . Com isso, a média e a variância da distribuição amostral das proporções, com parâmetros n e p, são dadas respectivamente por:

$$\mu_{\hat{p}} = p \quad e \quad \sigma^2_{\hat{p}} = \frac{p\ q}{n} = \frac{p\ (1-p)}{n}$$

Quando a probabilidade de sucesso em cada experimento amostral é constante e as tentativas são independentes, a distribuição amostral de $\hat{p} = \frac{y}{n}$ atende às condições da distribuição binomial. Além disso, de acordo com o Teorema do Limite Central (TLC), para n grande ($n > 30$), pode-se considerar a distribuição amostral de $\hat{p}$ como aproximadamente normal, principalmente se $p \to \dfrac{1}{2}$ (Bussab & Morettin, 2004).

2.6 Estimação

A estimação tem por finalidade utilizar dados amostrais para estimar parâmetros populacionais desconhecidos (Bussab & Morettin, 2004). Esta estimação pode ser feita de duas maneiras. A primeira, conhecida como estimação pontual, retorna uma única estimativa do parâmetro θ desconhecido, a partir de uma amostra aleatória. Mas, se o processo de estimação consiste em criar um intervalo de possíveis valores, admitindo-se que o valor real de um parâmetro θ desconhecido tem uma probabilidade específica de pertencer a este intervalo, então, define-se a estimação intevalar (Mood et al., 1974).

A partir de uma amostra $Y_1, Y_2, \ldots, Y_n$, de uma população que apresenta alguma característica de interesse, podem-se obter estimativas dos parâmetros populacionais, como, por exemplo, a média e a variância. Estimador de um parâmetro θ é qualquer função das observações da amostra e é representado por $\hat{\theta}$; a estimativa é o valor assumido pelo estimador na amostra. Neste caso, tem-se que, se μ é a média populacional, $\overline{Y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i$ é um estimador da média μ e $\overline{y}$ é uma estimativa de

μ. Várias amostras podem ser retiradas de uma mesma população, conseqüentemente podem-se obter inúmeras estimativas para o estimador.

Em função da existência de vários métodos e critérios para estimar parâmetros, pode ser difícil escolher, entre vários estimadores, aquele que melhor representa um mesmo parâmetro. Existem certas propriedades que um estimador pode ou não possuir e que permitem decidir se é melhor ou não do que outro (Mood et al., 1974). Dentre estas propriedades, Ferreira (2005) define as que são mais desejáveis no processo de inferência:

i- um estimador $\hat{\theta}$ é considerado um estimador não viesado do parâmetro θ se $E(\hat{\theta}) = \theta$;

ii- estimador eficiente $(\hat{\theta})$ é aquele, dentre os estimadores não viesados de θ, que possui a menor variância;

iii- um estimador $\hat{\theta}$ de um parâmetro θ de uma população é consistente quando sua distribuição se torna mais concentrada à medida que n tende a infinito. Um estimador $\hat{\theta}$ com consistência simples é aquele que converge em probabilidade para o parâmetro à medida que n tende a infinito, isto é, além de não ser viesado, sua variância tende a zero com o aumento da amostra n.

2.6.1 Estimador de máxima verossimilhança

Dentre os métodos de estimação pontual existentes, o mais importante e de grande aplicação na teoria estatística é o da máxima verossimilhança (MV). Ele foi introduzido por Fisher, em 1922 e consiste em se obter valores do parâmetro desconhecido que maximizam a função de densidade de uma amostra particular observada (Bussab & Morettin, 2004).

A função de verossimilhança de uma amostra aleatória com n variáveis $Y_1, Y_2, \ldots, Y_n$, denotada por $L(\theta)$, é definida como sendo a densidade

conjunta de n variáveis, ou seja, $f(y_1, y_2, \ldots, y_n; \theta)$, e deve ser considerada como uma função de θ (Mood et al., 1974). Em razão dos valores amostrais $Y_1, Y_2, \ldots, Y_n$ serem independentes, é possível definir a densidade conjunta ou função de verossimilhança, $L(\theta)$, pelo produtório das densidades de cada Y_i $(i = 1, 2, \ldots, n)$. Assim, a função de verossimilhança $L(\theta)$, é definida por:

$$L(\theta) = f(y_1; \theta) f(y_2; \theta) \ldots f(y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

O estimador de máxima verossimilhança de θ é o valor de $\hat{\theta}$ que maximiza $L(\theta)$. Para se obter o estimador de máxima verossimilhança $\hat{\theta}$, deve-se tomar a primeira derivada de $L(\theta)$ em relação ao parâmetro θ, igualar a zero e resolver para θ. A solução é o estimador de máxima verossimilhança (MV) (Ferreira, 2005).

Os estimadores de máxima verossimilhança possuem, dentre as propriedades citadas anteriormente, uma outra propriedade que permite obter estimadores para funções de parâmetros. Essa propriedade é conhecida como propriedade da invariância e é definida por: seja $\hat{\theta}$ um estimador de máxima verossimilhança de θ, com função densidade $f(y; \theta)$, em que θ é um vetor unidimensional. Se $\tau(\bullet)$ é uma função de θ que possui um único valor estimado, então, o estimador de $\tau(\theta)$ é $\tau(\hat{\theta})$ (Mood et al., 1974).

Por meio desta propriedade é possível obter estimadores de máxima verossimilhança para funções importantes. Particularmente, quando têm-se proporções binomiais, o estimador de máxima verossimilhança do parâmetro p_i, obtido a partir de uma amostra aleatória de tamanho n_i, é dado por:

$$\hat{p}_i = \frac{y_i}{n_i}$$

em que y_i representa o número de sucessos do evento na amostra n_i, $i = 1, 2, \ldots, k$.

2.6.2 Estimador de Pan para o parâmetro binomial

Em inferência, é muito comum construir intervalo de confiança para o parâmetro binomial p, utilizando o intervalo de Wald. O intervalo de Wald para a proporção p de uma amostra de tamanho n é um método de estimação intervalar que baseia-se na aproximação assintótica normal e apresenta, como parâmetro estimado, o estimador de máxima verossimilhança obtido anteriormente, é e dado por:

$$IC_{(1-\alpha)}(p_i) : \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{V(\hat{p}, n)}$$

em que $z_{\frac{\alpha}{2}}$ é o quantil superior $100(\alpha/2)\%$ da distribuição normal padrão e $V(\hat{p}, n) = \frac{\hat{p}(1-\hat{p})}{n}$.

Em estudo envolvendo estimação intervalar com proporções binomiais, Agresti & Coull (1998) verificaram que o intervalo de Wald não apresentava resultados razoáveis para determinado tamanho de amostras e como alternativa para melhorar sua performance, propuseram uma modificação no estimador de p, obtendo, assim, um novo estimador conhecido como add-4. O estimador modificado de p proposto por estes autores consiste em adicionar quatro pseudo-observações na amostra da população, das quais, duas são consideradas como sucesso e duas como fracasso do evento de interesse, e é dado por:

$$\tilde{p} = \frac{y+2}{n+4}$$

Substituindo o estimador $\tilde{p}$ no lugar do estimador $\hat{p}$ no intervalo de Wald, obtém-se o intervalo add-4:

$$IC_{(1-\alpha)}(p) : \tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{V(\tilde{p}, n+4)}$$

em que: $V(\tilde{p}, n+4) = \dfrac{\tilde{p}(1-\tilde{p})}{n+4}$.

Apesar deste método ser simples e de performance extremamente satisfatória, Pan (2002) notou que ainda poderia existir progresso na performance do método de estimação intervalar add-4. Verificou que um teste t é melhor do que um teste z quando tem-se a média de uma

distribuição normal com variância desconhecida e que o teste t é mais conservativo do que o teste z, isto é, que o teste t seria mais indicado para manter o erro tipo I dentro do nível nominal específicado. Além disso, observou que, no intervalo de Wald, a variância de $\hat{p}$ é substituída por sua estimativa $V(\hat{p}, n)$ no lugar de $V(p, n)$. Com isso, Pan (2002) sugeriu aproximar a distribuição de $\tilde{p}$ por uma distribuição t de Student com ν número de graus de liberdade corrigidos pelo método de Satterthwaite (1941), dados por:

$$\nu = \frac{2V(\tilde{p},\ n+4)^2}{\Omega(\tilde{p},\ n+4)}$$

em que $\Omega(\tilde{p},\ n{+}4)$ equivale a $var[V(\tilde{p},\ n{+}4)]$, que é obtida calculando-se os quatro primeiros momentos de X. Esta expressão pode ser encontrada em Carari (2004).

Assim, o intervalo de confiança sugerido por Pan (2002) é:

$$IC_{(1-\alpha)}(p) : \tilde{p} \pm t_{(\nu,\ \frac{\alpha}{2})}\sqrt{V(\tilde{p},\ n+4)}$$

em que $t_{(\nu,\ \frac{\alpha}{2})}$ refere-se ao quantil superior $100(\alpha/2)\%$ da distribuição t de Student com ν números de graus de liberdade.

Após avaliar este método por simulação, Pan (2002) comprovou suas idéias. O intervalo de confiança, obtido por meio do estimador $\tilde{p}$ e do teste t, apresentou performance um pouco melhor do que o intervalo add-4. Essa melhora ocorreu principalmente nos casos em que os valores de $\tilde{p}$ eram próximos de 0 ou de 1 e os valores de n eram pequenos.

3 METODOLOGIA

Para a realização deste trabalho foram feitas simulações Monte Carlo com o intuito de avaliar as taxas de erro tipo I e poder do teste para k populações binomiais, com parâmetros p_i e n_i, referentes à i-ésima população. As simulações foram realizadas gerando 2.000 amostras para diferentes situações de duas etapas consideradas.

Em uma primeira etapa, foram avaliadas as taxas de erro tipo I por experimento considerando hipóteses H_0 completa: H_0: $p_1 = p_2 = \ldots = p_k$ e hipóteses H_0 parcial: $p_1 = p_2 = \ldots = p_i \neq p_{i+1} = p_{i+2} = \ldots = p_k$. As simulações foram feitas considerando-se os valores dos parâmetros $p = 0,1;\ 0,5\ e\ 0,9$ para H_0 completa. Para a hipótese H_0 parcial, considerou-se uma diferença entre os parâmetros $p_1 = p_2 = \ldots = p_i$ e $p_{i+1} = p_{i+2} = \ldots = p_k$, denominada de Δ, estipulada por $\Delta = 0,01;\ 0,05;\ 0,1;\ 0,2;\ 0,3;\ 0,4;\ 0,5;\ 0,6;\ 0,7;\ 0,8\ e\ 0,9$. Para a realização destas simulações considerou-se que o valor do parâmetro p dentro de um dos grupos foi de 0,01 e foram feitas as combinações entre os valores de Δ, número de populações binomiais $k = 2,\ 5\ e\ 10$, tamanho das amostras $n = 10,\ 30\ e\ 100$, e valor nominal de significância α igual a 5% e a 1%. Foram simuladas também, algumas situações em que os valores de p se aproximavam de 0,5. Para estes casos, consideraram se os valores de $\Delta = 0,01$; 0,1 e 0,4, e admitiu-se que os valores de p no primeiro grupo foram de 0,30; 0,45 e 0,5. Foram realizadas 2.000 simulações para cada uma das 102 situações, num total de 204.000 experimentos simulados nesta primeira etapa.

Na segunda etapa, o mesmo procedimento de simulação foi realizado para medir o poder sob a hipótese H_0 parcial, e sob a hipótese H_1 ($p_1 \neq p_2 \neq p_3 \neq \ldots \neq p_k$). Em cada simulação realizada para avaliar o poder sob a hipótese H_1, foi feita a combinação entre os mesmos números de populações binomiais (k), tamanhos amostrais (n), diferença entre p_k e p_1 dado por (Δ) e valor nominal de significância (α). Admitiu-se também que a diferença entre quaisquer duas proporções, $p_i's$ consecutivas é dada por: $\delta = \Delta/(k-1)$.

Para avaliar o poder ao considerar a hipótese H_0 parcial, estabeleceu-se a formação de dois grupos dististos, G_1 e G_2. Nas situações em que o número de populações binomiais foi igual a 5 ($k = 5$), o grupo G_1 foi constituído pelas proporções binomias $p_1,\ p_2,\ e\ p_3$ e o grupo G_2 pelas proporções binomiais restantes $p_4\ e\ p_5$. No caso de $k = 10$, definiu-se que as cinco primeiras proporções binomiais pertenceriam ao primeiro grupo e as demais ao segundo grupo. O número de comparações, nestas situações específicas, foi dado pela multiplicação do total de proporções pertencentes ao grupo G_1, com o número total de proporções do grupo G_2. Foram considerados os mesmos números de populações binomiais, tamanhos amostrais e valor nominal de significância da primeira etapa e os mesmos valores de Δ estabelecidos na segunda etapa sob H_1. Com isso, chegou-se ao total de 348.000 simulações (174 situações x 2.000 experimentos).

Em uma amostra aleatória $Y_1, Y_2, \ldots, Y_k$, em que y_i representa o número de sucessos observados na i-ésima população de tamanho n_i, o estimador para p_i de Pan (2002) que foi utilizado é dado por:

$$\tilde{p}_i = \frac{y_i + 2}{n_i + 4} \tag{1}$$

e o estimador de máxima verossimilhança é:

$$\hat{p}_i = \frac{y_i}{n_i}. \tag{2}$$

Foram consideradas todas as m comparações múltiplas da família de testes de hipóteses definidos de forma geral para a l-ésima comparação por:

$$H_0^{(l)} : p_i = p_h \qquad 1 \leq h \neq i \leq k \tag{3}$$

sendo $l = 1, 2, \ldots, m$ e $m = \frac{k(k-1)}{2}$.

Para o par de proporções $(p_i^{(j)}, p_h^{(j)})$ a seguinte estatística foi definida:

$$q_{ih}^{(j)} = \frac{max(p_i^{(j)}, p_h^{(j)}) \ - \ min(p_i^{(j)}, p_h^{(j)})}{\sqrt{\hat{V}(p_i^{(j)}, n_i^{(j)}) + \hat{V}(p_h^{(j)}, n_h^{(j)})}} \tag{4}$$

em que $p_i^{(j)}$ é dado pelo estimador da equação (1), quando $j = 1$ e $p_i^{(j)}$ é dado pelo estimador da equação (2), quando $j = 2$; $n_i^{(j)} = n_i + 2$ *se* $j = 1$ ou $n_i^{(j)} = n_i$ *se* $j = 2$; e

$$\hat{V}(p_i^{(j)},\ n_i^{(j)}) = \frac{p_i^{(j)}(1 - p_i^{(j)})}{n_i^{(j)}} \tag{5}$$

O valor da equação (4) foi obtido em cada simulação realizada e para cada par de proporções (i, h). Para impor a hipótese nula H_0 de igualdade das k proporções, foi obtido um único estimador combinando os k estimadores $p_i^{(j)}$ e aplicado o método de *bootstrap* infinito. Para isso, foi considerada a função de probabilidade conjunta estimada das k populações binomiais independentes por:

$$\hat{P}^{(j)}(Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_k = y_k) = \prod_{i=1}^{k} \binom{n_i}{y_i} p^{(j)y_i}(1 - p^{(j)})^{n_i - y_i} \tag{6}$$

Ao impor a hipótese nula H_0, determinou-se um estimador comum dos $p_i's$ que, sob H_0, é dado pelo parâmetro p, sendo obtido pela média ponderada:

$$p^{(j)} = \frac{\sum_{i=1}^{k} p_i^{(j)}(n_i - 1)}{n - k} \tag{7}$$

em que $p_i^{(j)}$ refere-se ao estimador da equação (1) quando $j = 1$ e $p_i^{(j)}$ ao estimador da equação (2) quando $j = 2$ e $n = \sum_{i=1}^{k} n_i$.

Dessa densidade foram realizadas B amostras aleatórias de *bootstrap*. A b-ésima amostra é dada por: $y_{1b},\ y_{2b},\ \ldots,\ y_{kb}$. Nesta amostra, ao par i e h é aplicada a expressão (4) e o valor resultante é representado por $q_{ihb}^{(j)}$.

Para todos os m pares na b-ésima amostra de *bootstrap*, foi considerada a estatística:

$$\Omega_b^{(j)} = \ max\{q_{12b}^{(j)},\ q_{13b}^{(j)},\ \ldots,\ q_{(k-1)kb}^{(j)}\}$$

e formado o conjunto:

$$\Omega^{(j)} = \{\Omega_1^{(j)},\ \Omega_2^{(j)},\ \ldots,\ \Omega_B^{(j)}\} = \bigcup_{b=1}^{B} \Omega_b^{(j)} \tag{8}$$

Os p-valores denominados ajustados (Ferreira et al., 2005) são dados por:

$$P_{ih,g}^{(j)} = \frac{1}{B} \sum_{b=1}^{B} I(\Omega_b^{(j)} \geq q_{ih}^{(j)}) \tag{9}$$

em que $I(\bullet)$ é uma função indicadora.

Os p-valores de cada par de populações foram comparados com os p-valores nominais (α) de 1% e 5%. Para os casos em que o p-valor foi menor ou igual a α, então, a hipótese nula ($H_0^{(l)}$) correspondente foi rejeitada. Assim, o erro tipo I ou o poder foram computados em um número M de simulações Monte Carlo realizadas em cada configuração.

A proporção de experimentos com pelo menos uma rejeição de alguma das m hipóteses nulas verdadeiras nas M simulações realizadas é a taxa de erro tipo I por experimento e a proporção de rejeições corretas das hipóteses nulas falsas é o poder. Para determinar se as taxas de erro tipo I estavam próximas aos valores nominais estabelecidos, foram calculados um limite superior e um limite inferior, para que estes valores pudessem assumir. Os intervalos de 99% confiança para o níveis nominais de significância 1% e 5%, em porcentagem, foram, respectivamente, de [0,5188; 1,727] e [3,8282; 6,3914], obtidos por meio do intervalo de confiança exato para proporções (Leemis & Trivedi, 1996). Dessa forma, aqueles valores que não pertenceram a esse intervalo foram considerados diferentes dos níveis nominais de significância.

Os resultados obtidos foram utilizados para comparar o desempenho dos testes de *bootstrap* utilizando os estimadores de máxima verossimilhanca (MV) e de Pan (Pan, 2002).

4 RESULTADOS E DISCUSSÃO

4.1 Erro tipo I

Os dois testes de *bootstrap* utilizados neste trabalho foram incialmente avaliados para o erro tipo I por experimento. Essencialmente, estes testes se diferenciam pelos estimadores de p_i utilizados. Em um deles, foi considerado o estimador de máxima verossimilhança e, no outro, o estimador de Pan (Pan, 2002). Também foram avaliadas duas formas de computar o erro tipo I: uma sob H_0 completa e a outra sob H_0 parcial.

4.1.1 Erro tipo I sob H_0 completa

Na Tabela 1 são apresentadas as taxas, em porcentagem de erro tipo I por experimento sob H_0 completa dos dois testes de *bootstrap*, em função do número de populações (k), do tamanho da amostra (n) e dos valores dos parâmetros (p), considerando o valor nominal de significância de 5%. Os dois testes foram identificados por Pan e MV, fazendo referência aos estimadores de p que os diferencia. Todos os resultados são médias de 2.000 simulações Monte Carlo com 2.000 reamostragens de *bootstrap* em cada uma delas.

Pode-se observar, de maneira geral, que houve controle do erro tipo I por experimento, pois nenhum valor superou significativamente (P<0,01) o nível nominal de 5%. Em muitos casos, o que ocorreu foram taxas significativamente (P<0,01) menores do que 5%, indicando que os testes eram conservativos nestas situações. Houve tendências de os testes apresentarem melhores resultados, ou seja, taxas de erro tipo I iguais ao valor nominal, para valores de p próximos a 0,5, mesmo para tamanhos de amostras pequenos. Houve também uma melhor performance do teste *bootstrap* de MV, pois ocorreram menos casos em que o teste foi conservativo quando comparado com o teste *bootstrap* de Pan.

Tabela 1: Erro tipo I por experimento (%) sob H_0 completa, para diferentes números de populações (k), diferentes tamanhos de amostras (n) e diferentes valores do parâmetro (p) para os estimadores de Pan (Pan, 2002) e máxima verossimilhança (MV) ao valor nominal de $0,05$.

		$p = 0,1$		$p = 0,5$		$p = 0,9$	
k	n	Pan	MV	Pan	MV	Pan	MV
2	10	$1,00^*$	$1,00^*$	4,10	4,05	$0,65^*$	$0,65^*$
2	30	$2,35^*$	3,95	5,10	4,90	$2,30^*$	3,95
2	100	5,55	5,85	5,05	4,90	5,10	5,20
5	10	$0,00^*$	$0,15^*$	$3,35^*$	$3,30^*$	$1,50^*$	$1,00^*$
5	30	$2,40^*$	3,90	5,25	5,15	$2,50^*$	4,20
5	100	4,00	4,55	5,15	5,25	4,20	4,30
10	10	$0,00^*$	$1,35^*$	$2,55^*$	$2,55^*$	$0,00^*$	$0,80^*$
10	30	$1,05^*$	$3,15^*$	4,40	4,20	$0,70^*$	$3,00^*$
10	100	4,45	5,20	5,20	5,20	$3,50^*$	4,35

* significativamente (P<0,01) inferior a 5%

O teste *bootstrap* de Pan foi conservativo com $k = 2, 5$ e 10 para $n = 10$ e 30 com $p = 0,1$ ou $p = 0,9$. Para $p = 0,5$, este teste foi conservativo para $k = 5$ e 10 com $n = 10$. Com $n = 100$, somente para $k = 10$ e $p = 0,9$ o teste em questão foi conservativo. Em todas as demais situações, o tamanho do teste foi não significativamente diferente do valor nominal de 5%.

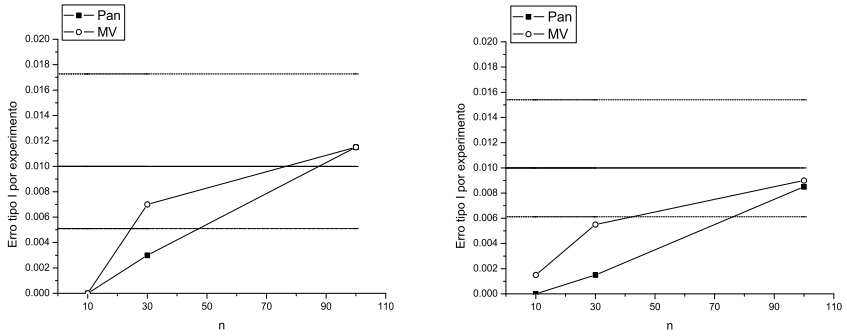
Estes resultados são surpreendentes, uma vez que não houve casos em que as taxas de erro tipo I tenham superado significativamente o valor nominal de significância de 5%. Este fato mostra um controle do erro tipo I, embora, em amostras pequenas (10 ou 30) e para $p_i's$ afastados de 0,5, apresente excesso de conservadorismo, isto é, taxas significativamente inferiores ao valor nominal. Isso, provavelmente, pode afetar de maneira indesejável o poder, ou seja, causando a sua redução.

Para grandes amostras ($n = 100$), praticamente os dois testes tiveram tamanho não diferente significativamente do valor nominal, exceto com maior número de populações ($k = 10$) com $p = 0,9$ para o teste de Pan.

Assim, se p afasta-se 0,5 e k é grande, o tamanho da amostra deve ser bem maior para que o teste de Pan tenha tamanho igual ao nominal.

Para o valor nominal de significância de 1%, os resultados da taxa de erro tipo I por experimento foram bastante similares aos observados para 5%. Assim, todos os resultados ou não diferiram significativamente (P>0,01) do valor nominal de 1% ou foram significativamente (P<0,01) inferiores.

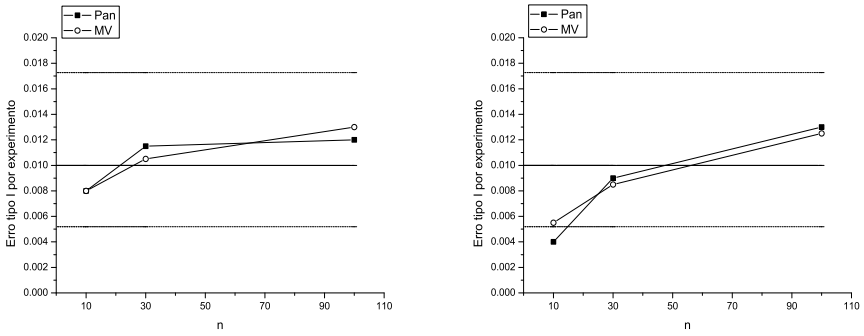
Nas Figuras 1 (a) e (b) e 2 (a) e (b) são apresentados alguns destes resultados. Para $p = 0,1$ e $k = 2$ e 10, pode-se observar, na Figura 1, (a) e (b), que o teste *bootstrap* de Pan foi significativamente conservativo quando $n \leq 30$ e apresentou tamanho não significativamente (P>0,01) diferente do nominal quando $n = 100$. O teste *bootstrap* de MV para $n = 10$ e $k = 2$ e também para $k = 10$ e $n = 10$ e 30 foi conservativo e, nos demais casos, apresentou tamanho não significativamente (P>0,01) diferente do valor nominal de significância de 1%.



(a) (b)

Figura 1: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 2$ e (b) $k = 10$, com $p = 0,1$ e para $\alpha = 1\%$, considerando a hipótese H_0 completa.

Para $p = 0,5$ (Figura 2), somente o teste *bootstrap* de Pan na situação de $k = 10$ e $n = 10$ foi conservativo. Todas as demais situações de ambos os testes apresentaram taxas não diferentes significativamente (P>0,01) do valor nominal de 1%.



(a) (b)

Figura 2: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 2$ e (b) $k = 10$, com $p = 0,5$ e para $\alpha = 1\%$, considerando a hipótese H_0 completa.

4.1.2 Erro tipo I sob H_0 parcial

Nas Tabelas 2 e 3, os erros tipo I por experimento sob H_0 parcial foram apresentados em função de k, n e Δ para $\alpha = 5\%$. Na Tabela 2 estão os valores de Δ inferiores a 0,5 e na Tabela 3 os demais valores de Δ. O que se observa, de maneira geral, é que houve controle do erro tipo I em todos os casos, tendo, na grande maioria, ambos os testes de *bootstrap* sido conservativos. Os casos em que os testes tiveram erros tipo I não significativamente (p>0,01) diferentes do valor nominal foram com $k = 10$, $n = 30$, $\Delta = 0,05$ e 0,1 para o teste de *bootstrap* MV, $k = 10$, $n = 10$, $\Delta = 0,2$; 0,3 e 0,4 para o teste de MV.

É conveniente salientar que, para garantir que o espaço paramétrico de $p_i's$ não fosse violado, utilizou-se a estratégia de fixar os valores de p em 0,01 no primeiro grupo e de $0,01+\Delta$ no segundo. Assim, quando Δ é pequeno (0,01 e 0,05), os valores de p em ambos os grupos estão afastados de 0,5 e espera-se, como aconteceu sob H_0 completa, que os testes sejam mais conservativos. Isso realmente foi constatado nas Tabelas 2 e 3.

Os testes *bootstrap* de Pan e MV apresentaram comportamentos similares em relação ao controle do erro tipo I, sendo classificados como conservativos em quase todas as situações simultaneamente. Se forem comparados com as taxas de erro observadas, verifica-se uma pequena

vantagem para o teste de *bootstrap* MV, cujos valores estavam mais próximos de $\alpha = 5\%$ e, em alguns poucos casos, não significativamente (p>0,01) diferentes de $\alpha = 5\%$.

Tabela 2: Erro tipo I, por experimento (%), sob H_0 parcial para diferentes números de populações (k), tamanhos de amostras (n) e diferenças entre os parâmetros p de cada grupo (Δ), para os estimadores de Pan (Pan, 2002) e máxima verossimilhança (MV) ao valor nominal de $0,05$.

		$\Delta = 0,01$		$\Delta = 0,05$		$\Delta = 0,1$	
k	n	Pan	MV	Pan	MV	Pan	MV
5	10	$0,00^*$	$0,00^*$	$0,00^*$	$0,10^*$	$0,05^*$	$1,15^*$
5	30	$0,00^*$	$0,25^*$	$0,10^*$	$1,30^*$	$0,60^*$	$2,20^*$
5	100	$0,00^*$	$0,85^*$	$0,90^*$	$1,25^*$	$0,95^*$	$0,90^*$
10	10	$0,00^*$	$0,15^*$	$0,00^*$	$1,10^*$	$0,05^*$	$3,40^*$
10	30	$0,00^*$	$1,15^*$	$0,10^*$	4,95	$0,75^*$	5,15
10	100	$0,05^*$	$2,40^*$	$1,15^*$	$2,05^*$	$1,90^*$	$1,75^*$
		$\Delta = 0,2$		$\Delta = 0,3$		$\Delta = 0,4$	
k	n	Pan	MV	Pan	MV	Pan	MV
5	10	$0,20^*$	$0,20^*$	$0,50^*$	$3,40^*$	$0,95^*$	$2,70^*$
5	30	$1,45^*$	$2,10^*$	$0,70^*$	$0,45^*$	$1,00^*$	$0,75^*$
5	100	$1,50^*$	$1,30^*$	$0,80^*$	$0,65^*$	$0,65^*$	$0,65^*$
10	10	$0,30^*$	6,15	$0,85^*$	5,95	$2,25^*$	5,35
10	30	$2,70^*$	$3,25^*$	$2,45^*$	$1,60^*$	$2,05^*$	$1,75^*$
10	100	$2,20^*$	$1,75^*$	$1,65^*$	$1,50^*$	$1,95^*$	$1,85^*$

* significativamente (P<0,01) inferior a 5%

Para o valor nominal de significância de 1%, observou-se o mesmo comportamento geral dos testes obtidos para $\alpha = 5\%$. Na Figura 3 (a) e (b) estão apresentadas as taxas de erros observadas para os testes *bootstrap* de Pan e MV, considerando $\Delta = 0,05$, em função de n para $k = 5$ e 10, respectivamente. Conforme já comentado para $\alpha = 5\%$, o teste de *bootstrap* MV apresentou performance um pouco superior, pois os níveis de significância foram superiores aos do teste de Pan, mas inferiores (con-

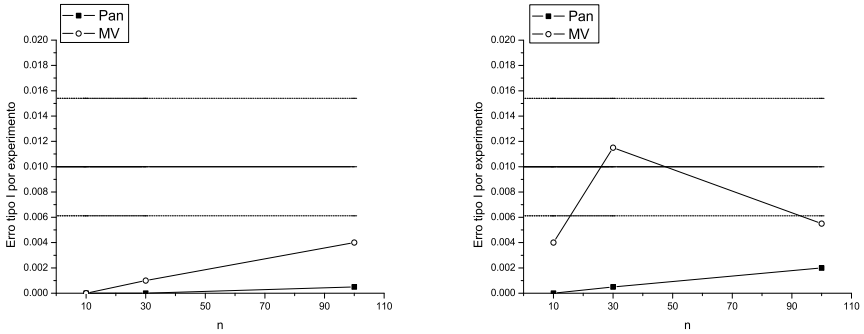
servativo) ou iguais (ideal) ao nível nomimal. Com um número menor de populações, os testes de *bootstrap* foram mais conservativos se fixado um mesmo tamanho de amostra, neste caso com $p_i's$ pequenos (próximos 0,01) e Δ pequeno (0,05).

Tabela 3: Erro tipo I, por experimento (%), sob H_0 parcial para diferentes números de populações (k), tamanhos de amostras (n) e diferenças entre os parâmetros p de cada grupo (Δ), para os estimadores de Pan (Pan, 2002) e Máxima Verossimilhança (MV) ao valor nominal de $0,05$.

		$\Delta = 0,5$		$\Delta = 0,6$		$\Delta = 0,7$	
k	n	Pan	MV	Pan	MV	Pan	MV
5	10	$1,20^*$	$1,80^*$	$0,90^*$	$1,50^*$	$0,55^*$	$0,70^*$
5	30	$0,75^*$	$0,55^*$	$1,05^*$	$1,00^*$	$1,00^*$	$1,10^*$
5	100	$0,95^*$	$0,75^*$	$0,55^*$	$0,55^*$	$0,45^*$	$0,45^*$
10	10	$2,00^*$	$2,95^*$	$1,70^*$	$2,20^*$	$0,70^*$	$0,80^*$
10	30	$2,05^*$	$1,60^*$	$1,80^*$	$1,70^*$	$1,20^*$	$1,65^*$
10	100	$1,60^*$	$1,35^*$	$1,40^*$	$1,30^*$	$1,15^*$	$1,35^*$
		$\Delta = 0,8$		$\Delta = 0,9$			
k	n	Pan	MV	Pan	MV		
5	10	$0,10^*$	$0,10^*$	$0,00^*$	$0,00^*$		
5	30	$0,35^*$	$0,50^*$	$0,05^*$	$0,15^*$		
5	100	$0,55^*$	$0,65^*$	$0,20^*$	$0,55^*$		
10	10	$0,15^*$	$0,15^*$	$0,00^*$	$0,00^*$		
10	30	$0,65^*$	$0,95^*$	$0,05^*$	$0,15^*$		
10	100	$0,85^*$	$0,95^*$	$0,40^*$	$2,00^*$		

* significativamente (P<0,01) inferior a 5%

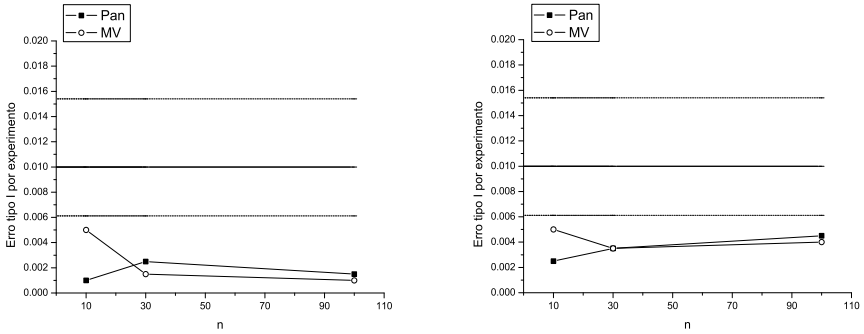
Na Figura 4 (a) e (b) são apresentados os erros tipo I, por experimento, dos dois testes, para $\Delta = 0,5$ e $k = 5$ e 10. Nesta situação, ambos os testes foram conservativos, independente do tamanho amostral e do número de populações. Somente para pequenas amostras é que houve uma performance um pouco melhor do teste MV em relação ao de Pan. Para grandes amostras, tanto para $k = 5$ quanto para $k = 10$, os testes



(a) (b)

Figura 3: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 5$ e (b) $k = 10$, com $\Delta = 0,05$ e para $\alpha = 1\%$, considerando a hipótese H_0 parcial.

tenderam a se igualar com relação às taxas observadas de erro tipo I, por experimento.

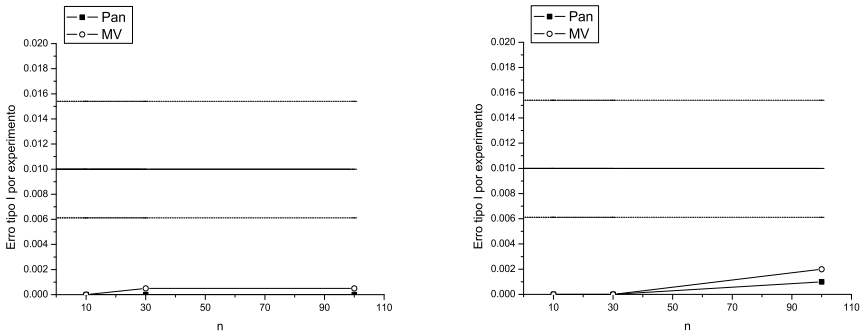


(a) (b)

Figura 4: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 5$ e (b) $k = 10$, com $\Delta = 0,5$ e para $\alpha = 1\%$, considerando a hipótese H_0 parcial.

Na Figura 5 (a) e (b) estão apresentadas as taxas de erro tipo I, por experimento, sob H_0 parcial dos dois testes para $\Delta = 0,9$ e com $k = 5$ e 10. Novamente, em todas as situações de n, os testes de bootstrap apresentaram-se conservativos. Este caso particular foi o mais conservativo de todos, provavelmente, por causa de um dos grupos possuir

$p^{(1)} = 0,01$ e o outro $p^{(2)} = 0,91$. Valores afastados de 0,5 geram situações em que os testes binomiais são mais pernósticos. O teste bootstrap MV mostrou-se um pouco superior ao teste de *bootstrap* de Pan.



(a) (b)

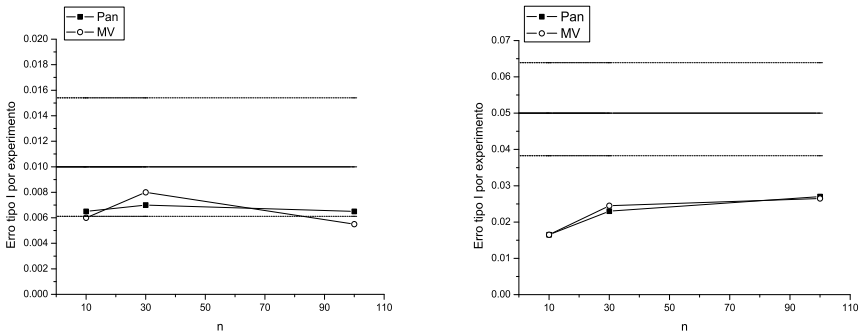
Figura 5: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 5$ e (b) $k = 10$, com $\Delta = 0,9$ e para $\alpha = 1\%$, considerando a hipótese H_0 parcial.

Procurando avaliar situações em que os valores de p se aproximavam de $0,5$, ainda sob H_0 parcial, foram feitas simulações adicionais, nas quais se avaliou o erro tipo I por experimento, considerando os dois níveis nominais de significância de 1% e 5%.

Na Figura 6 (a) e (b) são apresentados os erros tipo I, por experimento, sob H_0 parcial, em que um dos grupos possuía valores de p iguais a $0,5$ e o outro, valores iguais a $0,51$ para $\alpha = 1\%$ e 5%, respectivamente. Para $\alpha = 1\%$, todos os erros tipo I dos testes foram não significativamente (P>0,01) diferentes do valor nominal, exceto para o teste de *bootstrap* MV, com $n = 100$, que, neste caso, foi um pouco conservativo. Este resultado é diferente daqueles observados em casos semelhantes, em que os $p_i's$ de um dos grupos se afastavam grandemente de $0,5$, sendo, em geral, menos conservativo ou, até mesmo, não conservativo.

Para $\alpha = 5\%$ (Figura 6 (b)), os resultados foram todos conservativos, embora menos conservativos do que os casos semelhantes sob H_0 parcial, com um dos grupos apresentando valores de p afastados de $0,5$. Verificou-se que o tamanho da amostra quase não influenciou as taxas de erro

tipo I, por experimento. Com o aumento do tamanho das amostras de 10 para 30, houve um pequeno aumento das taxas de erro tipo I, por experimento, em ambos os testes e, para grandes amostras ($n \geq 30$), essas taxas permaneceram constantes. Os testes MV e Pan apresentaram resultados bastante parecidos.



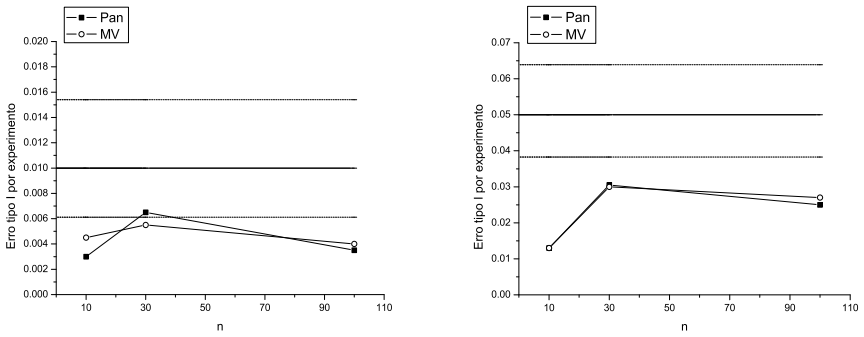
(a) (b)

Figura 6: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), com $k = 10$, $\Delta = 0,01$; $p^{(1)} = 0,50$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$, considerando a hipótese H_0 parcial.

Para a diferença entre os dois grupos um pouco maior ($\Delta = 0,1$) e valores de p de um dos grupos iguais a 0,45 e do outro 0,55, as taxas de erro tipo I por experimento foram apresentadas na Figura 7 (a) e (b), para $\alpha = 1\%$ e 5%, respectivamente. Em ambos os casos ($\alpha = 1\%$ ou 5%), os testes apresentaram resultados parecidos e conservativos, exceto o teste *bootstrap* de Pan com $\alpha = 1\%$ e $n = 30$, que não diferiu significativamente (P>0,01) do valor nominal de 1%. Em ambos os casos, os testes foram menos conservativos do que quando aplicados em situações semelhantes com pelo menos um dos grupos com valores de p afastados de 0,5. Houve melhorias nas taxas, ou seja, estas aproximaram-se mais dos respectivos valores de α, com o aumento de n de 10 para 30. De 30 para 100, em alguns casos, houve até certa redução das taxas.

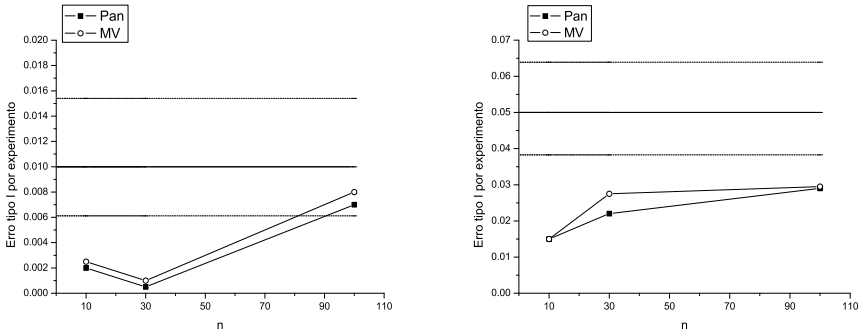
Finalmente, para $\Delta = 0,4$, com valores de p de um dos grupos iguais a 0,3, as taxas de erros para $\alpha = 1\%$ e 5% foram apresentadas na Figura

8 (a) e (b).



(a) (b)

Figura 7: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), com $k = 10$, $\Delta = 0,1$; $p^{(1)} = 0,45$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$, considerando a hipótese H_0 parcial.



(a) (b)

Figura 8: Taxas de erro tipo I, por experimento, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n) com $k = 10$, $\Delta = 0,4$; $p^{(1)} = 0,3$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$, considerando a hipótese H_0 parcial.

Novamente, os testes apresentaram valores bem parecidos de taxas de erro tipo I, por experimento e foram conservativos, exceto para $\alpha = 1\%$ e $n = 100$, em ambos os testes (Figura 8 (a)). Nesse caso, o tamanho das amostras teve uma maior influência nas taxas de erro tipo I, por experimento, dos dois testes de *bootstrap*, principalmente para $n = 100$,

considerando $\alpha = 1\%$, em que as taxas de ambos os testes não diferiram significativamente do valor nominal de 1%.

De maneira geral, os testes foram menos conservativos para $\alpha = 1\%$. Quando a diferença foi maior ($\Delta = 0,4$) entre grupos, o teste MV foi ligeiramente menos conservativo, principalmente para amostras intermediárias ($n = 30$) e $\alpha = 5\%$ e uniformemente menos conservativo em relação a n quando a significância nominal é de 1%.

4.2 Poder

Várias avaliações dos testes de *bootstrap* de MV e Pan foram realizadas para mensurar o poder. Nestas avaliações, foram considerados vários tamanhos de amostras (n), número de populações (k) e várias diferenças entre a maior e a menor proporção binomial das k populações (Δ). Foi simulada também uma situação em que dois grupos possuiam os mesmos valores de p internamente e que diferiam entre si por uma quantidade (Δ) específica. Esta última situação foi chamada de H_0 parcial. As comparações entre populações de grupos diferentes foram utilizadas para avaliar o poder. Estas duas situações são discutidas separadamente nas subseções 4.2.1 e 4.2.2.

4.2.1 Poder sob H_1

Na Tabela 4 são apresentados os valores de poder dos testes de *bootstrap* Pan e MV, em função de k, n e Δ para $\alpha = 5\%$, sendo Δ a diferença a ser detectada entre os valores de p de duas populações distintas. Se o valor de Δ for muito pequeno ($\Delta = 0,01$), os valores de poder dos testes são menores ou iguais ao nível nominal de significância adotado, mesmo para amostras grandes ($n = 100$).

Para valores pequenos e moderados de Δ ($\Delta \leq 0,3$), quando n é pequeno ($n = 10$), o poder, em alguns casos, pode ser igual ou inferior ao valor nominal $\alpha = 0,05$. Isso aconteceu para $\Delta = 0,1$ e $n = 10$, com

todos os valores de k dos testes estudados e para $\Delta = 0,3$, $n = 10$, $k = 5$, com o teste de Pan e $\Delta = 0,3$, $n = 10$, $k = 10$, com ambos os testes.

Tabela 4: Poder sob H_1, para diferentes números de populações (k), tamanhos de amostras (n) e diferenças entre a maior e a menor proporção binomial (Δ), para os estimadores de Pan (Pan, 2002) e máxima verossimilhança (MV), a $0,05$ de nível nominal.

		$\Delta = 0,01$		$\Delta = 0,1$		$\Delta = 0,3$	
k	n	Pan	MV	Pan	MV	Pan	MV
2	10	0,00	0,00	1,25	1,25	33,50	33,60
2	30	0,10	0,15	19,80	35,30	93,90	95,05
2	100	2,05	5,25	88,15	89,70	100,00	100,00
5	10	0,00	0,00	0,00	0,70	5,15	9,90
5	30	0,00	0,00	2,85	8,30	69,90	77,60
5	100	0,05	0,70	61,70	66,85	100,00	100,00
10	10	0,00	0,00	0,00	0,55	1,35	5,50
10	30	0,00	0,20	0,90	4,05	53,65	62,35
10	100	0,00	0,30	42,80	50,80	100,00	100,00
		$\Delta = 0,5$		$\Delta = 0,7$		$\Delta = 0,9$	
k	n	Pan	MV	Pan	MV	Pan	MV
2	10	81,45	81,45	97,90	95,20	100,00	64,35
2	30	100,00	100,00	100,00	100,00	100,00	95,45
2	100	100,00	100,00	100,00	100,00	100,00	100,00
5	10	37,50	39,15	82,60	79,85	99,50	65,65
5	30	99,25	99,85	100,00	100,00	100,00	95,65
5	100	100,00	100,00	100,00	100,00	100,00	100,00
10	10	17,80	19,35	63,15	60,45	97,00	59,80
10	30	97,55	98,40	100,00	100,00	100,00	95,85
10	100	100,00	100,00	100,00	100,00	100,00	100,00

Pode-se observar também que, para $\Delta = 0,1$, os valores de poder de ambos os testes foram baixos, mesmo quando o tamanho das amostras era grande ($n \geq 30$). Isso ocorreu em ambos os testes de *bootstrap*, com $n = 30$ e 100, para todos os valores de k. Para $\Delta = 0,3$, verificou-se

que os valores de poder dos testes de *bootstrap* Pan e MV foram baixos apenas para tamanhos amostrais iguais a 30, considerando $k = 5$ e 10. Para grandes amostras ($n = 100$), a performance dos testes se igualou e se aproximou de 100%.

Para pequenas ou moderadas diferenças Δ, o poder aumenta consideravelmente com o aumento de n de 10 para 30 ou para 100. Também pode-se observar um grande efeito do número de populações no sentido de reduzir o poder. Assim, fixado um tamanho de amostra, um valor de Δ e o teste, o aumento de k provoca grandes reduções no valor do poder, principalmente se Δ é moderado ou pequeno. Esta redução é consideravelmente menor, em geral, se o valor de n é maior.

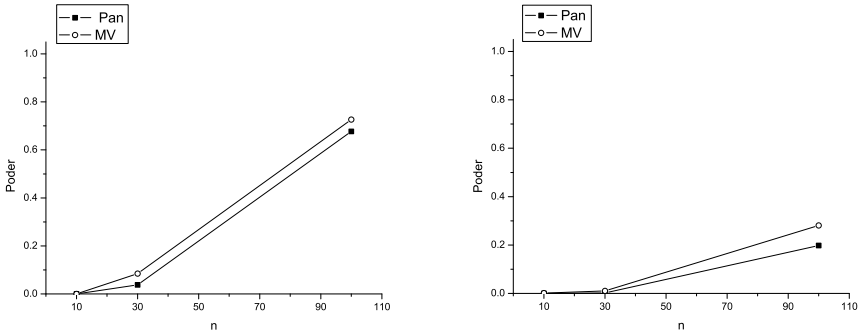
Assim, é importante aumentar os tamanhos das amostras se o pesquisador tem a intenção de comparar um maior número de populações. O que ocorre, na prática, é o contrário, ou seja, o aumento do número de níveis de tratamento em geral é acompanhado de reduções no número de repetições.

O teste *bootstrap* de MV foi quase sempre superior ao teste *bootstrap* de Pan em relação ao poder. Eles tendem a igualar as performances quando n aumenta. Quando os Δ eram grandes ($\Delta \geq 0,5$), em geral, os testes também tenderam a igualar a performance. No entanto, quando as diferenças foram muito grandes ($\Delta = 0,9$), uma das populações aproximava-se de 0 e a outra de 1, houve uma inversão de performances, o teste Pan tornou-se superior ao teste de MV. Isso provavelmente ocorre devido ao fato apontado por outros pesquisadores (Agresti & Coull, 1998; Pan, 2002) que o estimador das proporções add-4 possui melhores propriedades do que o estimador de máxima verossimilhança quando p afasta-se de 0,5 e o valor de n não é muito grande.

Se os valores de (Δ) são grandes ou muito grandes ($\Delta \geq 0,5$), os valores de poder aproximam-se de 100%, principalmente se $n \geq 30$. Para o caso de $n = 10$ e $k \geq 5$, conforme já foi salientado, há uma considerável redução de poder. Se o pesquisador almeja detectar pequenas diferenças ($\Delta \leq 0,1$), é recomendável utilizar tamanhos amostrais maiores de que 100, principalmente se o valor de k for superior a 2. Essa recomendação

é fundamentada no pressuposto de que haverá um maior poder para detectar tais diferenças entre os parâmetros binomiais.

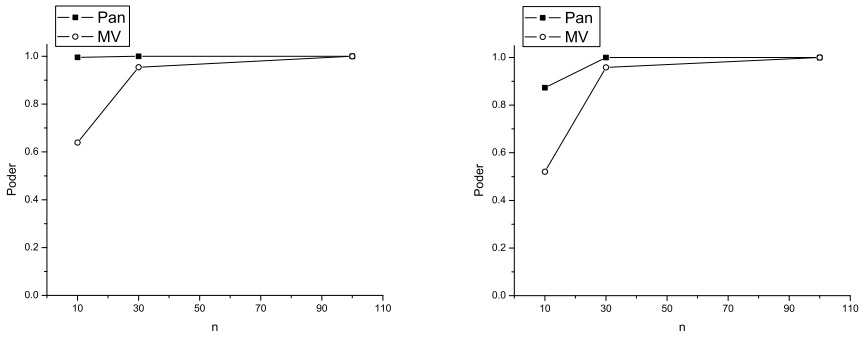
Na Figura 9 (a) e (b), estão apresentados os valores de poder em função de n para $\alpha = 1\%$ e $\Delta = 0,1$ com $k = 2$ e 10, respectivamente. A comparação do poder dos testes Pan e MV mostra a superioridade do teste MV. O aumento de n provoca aumentos consideráveis de poder, principalmente se k é pequeno ($k = 2$). A comparação dos valores de poder para $k = 2$ (Figura 9 (a)) com $k = 10$ (Figura 9 (b)) mostra que há uma redução expressiva do poder dos testes fixado um mesmo valor de n. Estes resultados são basicamente similares aos observados para $\alpha = 5\%$.



(a) (b)

Figura 9: Poder, sob H_1, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), diferença $\Delta = 0,1$ e número de populações iguais (a) $k = 2$ e (b) $k = 10$, para $\alpha = 1\%$.

Como houve uma semelhança muito grande entre os resultados de $\alpha = 1\%$ com os de $\alpha = 5\%$, apenas mais uma situação foi apresentada na Figura 10 (a) e (b). Este caso ilustra com clareza o efeito no poder de uma situação em que o valor de p de uma das populações se aproxima de 0 e o valor de p da outra, de 1. Assim como havia ocorrido para $\alpha = 5\%$, o teste Pan foi superior ao teste de MV para $n \leq 30$, em ambos os casos ($k = 2$) e ($k = 10$). Na Figura 10 (b) pode-se observar uma redução expressiva de poder com o aumento de $k = 2$ (Figura 10 (a)) para $k = 10$, em ambos os testes com $n \leq 30$.



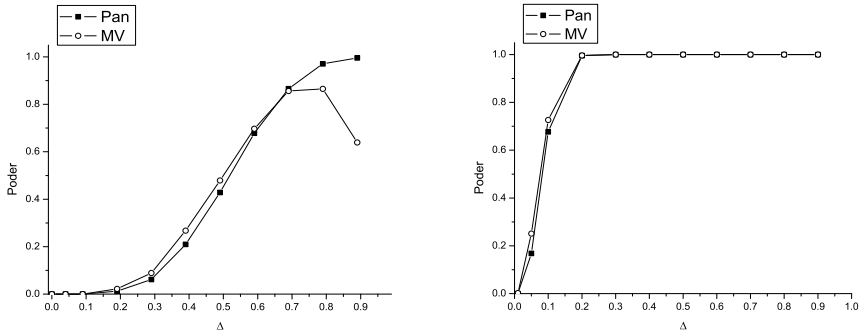
(a) (b)

Figura 10: Poder, sob H_1, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), diferença $\Delta = 0,9$ e número de populações iguais (a) $k = 2$ e (b) $k = 10$, para $\alpha = 1\%$.

Uma outra situação retratada para $\alpha = 1\%$ refere-se ao poder dos testes expresso em função da diferença Δ. Na Figura 11 (a) e (b) são apresentadas situações para $k = 2$ e $n = 10$ e 100, respectivamente. Em ambos os casos, há um incremento do poder com aumento Δ, o que é esperado pela teoria (Mood et al., 1974).

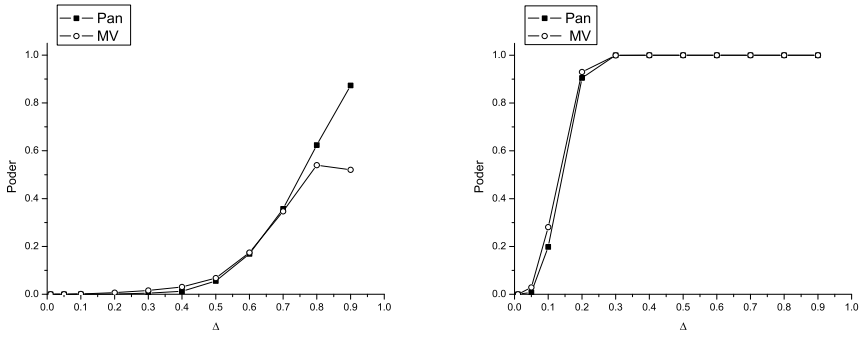
O teste MV para $n = 10$ (Figura 11 (a)) apresentou redução no poder a partir de $\Delta = 0,8$, contrariando o que é esperado pela teoria, em função da baixa qualidade do estimador, quando os valores de p se aproximam de 0 ou 1 e as amostras são pequenas. O teste Pan nesta mesma situação apresentou curva de poder estimada condizente com o esperado, ou seja, monótona crescente.

Para amostras grandes (Figura 11 (b)), os valores de poder foram similares nos dois testes e atingiram 100% rapidamente para $\Delta \geq 0,2$. Se forem comparadas as curvas de poder da Figura 11 (a) e (b) percebe-se que há uma taxa de crescimento maior quando n é maior. Por exemplo, com $n = 10$ e $\Delta = 0,2$, os valores de poder são próximos de zero nos dois testes e com $n = 100$ e mesmo valor de Δ, os valores de poder são iguais a 100%. O teste MV teve seu problema de queda de poder eliminado com $n = 100$, mostrando que o tamanho de amostra tem um importante papel na qualidade do estimador e, portanto, na performance do teste.



(a) (b)

Figura 11: Poder, sob H_1, dos testes de *bootstrap* Pan e MV, em função da diferença Δ, com $k = 2$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$, para $\alpha = 1\%$.



(a) (b)

Figura 12: Poder, sob H_1, dos testes de *bootstrap* Pan e MV, em função da diferença Δ, com $k = 10$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$, para $\alpha = 1\%$.

Na Figura 12 (a) e (b) estão as curvas de poder estabelecidas, em função de Δ para os dois testes, considerando $\alpha = 1\%$, $k = 10$ e $n = 10$ e 100, respectivamente. O que se percebe é um comportamento semelhante ao relatado para o caso de $k = 2$ (Figura 11 (a) e (b)). No entanto, se forem comparadas as curvas de poder para um dado teste sendo fixados Δ e n, o que se verifica é que há uma redução de poder com o aumento de k. Quando n e Δ são grandes, os valores de poder atingem 100% e esse efeito não existe mais. Novamente, com $n = 10$ e $\Delta > 0,7$, o teste

MV apresentou as mesmas deficiências relatadas anteriormente, sendo superado pelo teste de Pan.

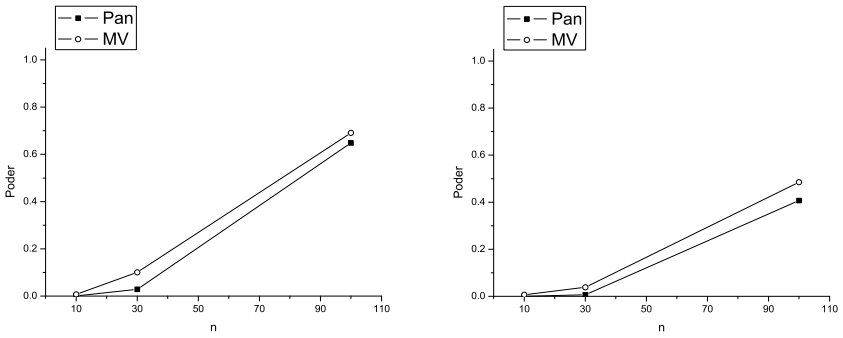
4.2.2 Poder sob H_0 parcial

Devido à grande similaridade entre os comportamentos dos testes simulados para o valor nominal de significância de 1 e 5%, foram apresentados apenas os resultados para $\alpha = 1\%$. Na Figura 13 (a) e (b) são apresentados os valores de poder, sob H_0 parcial, para os testes *bootstrap* de Pan e MV, considerando $\Delta = 0,1$, em função de n e com $k = 5$ e 10, respectivamente, para $\alpha = 1\%$. Para avaliar o poder sob a hipótese H_0 parcial, considerou-se a formação de dois grupos, tendo, no primeiro grupo, os valores de p sido fixados em 0,01 e, no segundo, os valores das proporções binomiais p foram dados por $0,01 + \Delta$ que, neste caso, são iguais a 0,11.

Pode-se observar, na Figura 13 (a) e (b), que os valores de poder do teste *bootstrap* de MV foram sempre superiores aos valores de poder do teste *bootstrap* de Pan, independente do tamanho amostral e do número de populações. Com o aumento dos valores de n houve um crescimento expressivo do poder de ambos os testes, tendo este crescimento sido maior para $k = 5$.

Comparando-se os valores de poder dos testes Pan e MV na Figura 13 (a) e (b) para um valor fixo de n, o que se observa é uma grande redução do poder com o aumento de $k = 5$ (Figura 13 (a)) para $k = 10$ (Figura 13 (b)). Este comportamento foi semelhante ao observado para os valores de poder sob a hipótese H_1. Se os valores de poder dos testes Pan e MV, sob a hipótese H_0 parcial, forem comparados aos valores de poder dos mesmos testes sob a hipótese H_1, fixado um valor de n, um α e um teste, verifica-se que o poder dos testes sob a hipótese H_0 parcial é superior ao poder dos testes sob a hipótese H_1.

Na Figura 14 (a) e (b) são apresentados os valores de poder para a diferença $\Delta = 0,5$ entre os dois grupos, em função de n e para $\alpha = 1\%$, com número de populações $k = 5$ e 10, respectivamente. Nesta situação,



(a) (b)

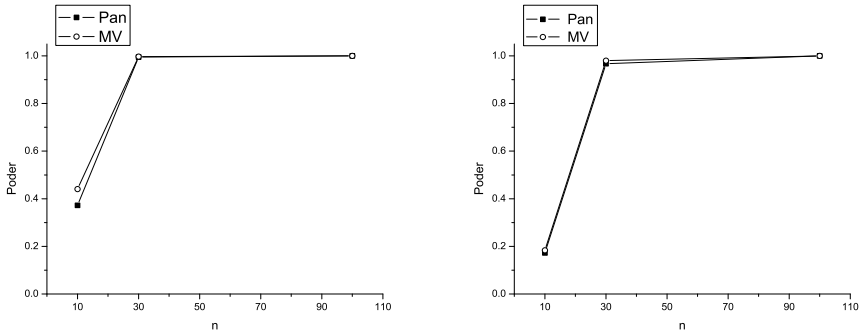
Figura 13: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), diferença $\Delta = 0,1$ e número de populações iguais (a) $k = 5$ e (b) $k = 10$, para $\alpha = 1\%$.

os valores de p de um dos grupos eram próximos a 0,5 e os outros valores de p próximos a 0. Em ambos os casos ($k = 5$ e $k = 10$), os testes apresentaram resultados parecidos e somente para $k = 5$ e $n \leq 30$, o teste *bootstrap* de MV foi superior ao teste de Pan, embora a diferença seja muito pequena.

Na Figura 14 (b) pode-se observar que a performance dos dois testes tendeu a se igualar para todos os valores de n e houve uma redução considerável para $n \leq 30$ dos valores de poder dos testes, com o aumento do número de populações de $k = 5$ para $k = 10$.

Para o valor de p do primeiro grupo próximo a 0, $p^{(1)} = 0,01$ e o valor de p do segundo grupo próximo a 1, $p^{(2)} = 0,91$, os valores de poder são apresentados na Figura 15 (a) e (b), para $k = 5$ e 10, respectivamente, com $\alpha = 1\%$. Neste caso, o teste Pan foi superior ao teste MV para $n \leq 30$, tanto para $k = 5$ como para $k = 10$. Pode-se verificar, para $n \leq 30$, uma pequena redução do poder dos testes ao aumentar o número de populações de $k = 5$ (Figura 15 (a)) para $k = 10$ (Figura 15 (b)). Sob H_0 parcial, foi observado o mesmo padrão de resposta ocorrido sob H_1, em que os valores de poder, se a diferença entre os valores de p é grande ($\Delta = 0,9$), têm tendência de se aproximarem de 100%, se $n \geq 30$.

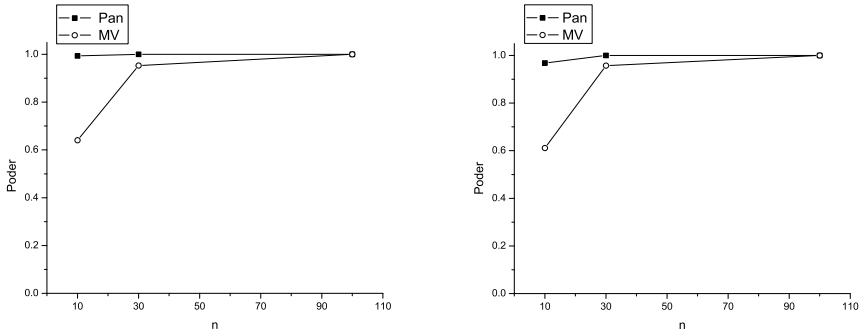
Pode-se observar, de maneira geral, que o poder dos testes, sob a



(a) (b)

Figura 14: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), diferença $\Delta = 0,5$ e número de populações iguais (a) $k = 5$ e (b) $k = 10$, para $\alpha = 1\%$.

hipótese H_0 parcial, foi superior ao poder sob a hipótese H_1, principalmente para números de populações intermediários ($k = 5$) e $\alpha = 1\%$.



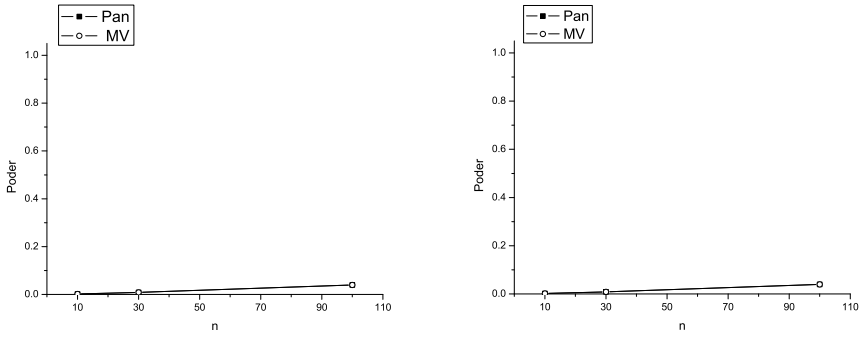
(a) (b)

Figura 15: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), diferença $\Delta = 0,9$ e número de populações iguais (a) $k = 5$ e (b) $k = 10$, para $\alpha = 1\%$.

Como os testes de hipóteses sobre proporções binomiais que utilizam o estimador de máxima verossimilhança são pouco eficientes quando $p \to 0$ ou $p \to 1$ e n é pequeno, poderia ser questionada a validade dos procedimentos utilizados. Também poderia se pensar que o estimador de Pan pudesse ser beneficiado e os resultados de poder do teste base-

ado neste estimador pudessem ser resultantes deste fato, pois, foi fixado, para o primeiro grupo, o valor de p em 0,01. Assim, buscou-se simular situações de H_0 parcial em que algumas das diferenças Δ utilizadas anteriormente fossem adotadas mas que os valores $p^{(1)}$ e $p^{(2)}$ estivessem o mais próximo de 0,5 quanto fosse possível.

Na Figura 16 (a) e (b) são apresentados os valores de poder em que um dos grupos possuía $p^{(1)} = 0,45$ e o outro, $p^{(2)} = 0,55$, para $\alpha = 1\%$ e 5%, respectivamente. Nesta situação, o que se observa é que ambos os testes tiveram comportamentos semelhantes. Em ambos os casos ($\alpha = 1\%$ e $\alpha = 5\%$), os dois testes apresentaram valores de poder iguais e próximos de 0 para $n \leq 30$. Com o aumento do tamanho das amostras, os valores de poder dos testes teve um pequeno acréscimo. Se forem comparados estes resultados com aqueles da Figura 13 (b), pode-se observar que houve uma drástica redução de poder se for fixado o mesmo valor de n.

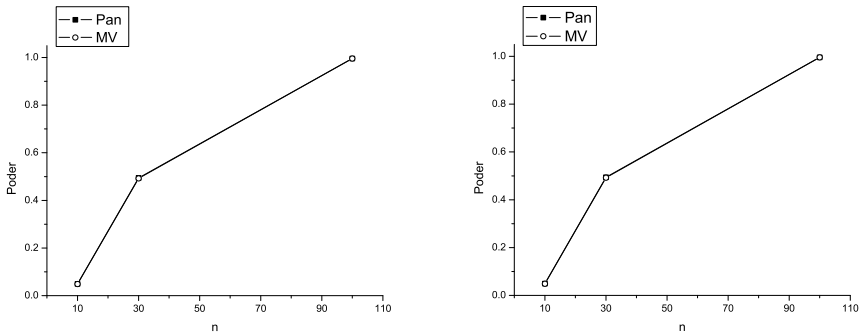


(a) (b)

Figura 16: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), $k = 10$, $\Delta = 0,1$; $p^{(1)} = 0,45$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$.

Na Figura 17 (a) e (b), foram apresentados os valores de poder para $\alpha = 1\%$ e 5%, respectivamente, em função de n para $\Delta = 0,4$. Novamente, para todos os valores de n, os dois testes apresentaram performance similar para valores de p próximos a 0,5 nas duas situações ($\alpha = 1\%$ e $\alpha = 5\%$). Houve um crescimento considerável dos valores

de poder de ambos os testes com o aumento do tamanho das amostras, principalmente se $n \geq 30$. No entanto, se forem comparados os valores de poder dos testes Pan e MV, sendo fixados os valores de n e α, o que se observa é um crescimento considerável do poder ao aumentar o valor de $\Delta = 0,1$ para $\Delta = 0,4$.



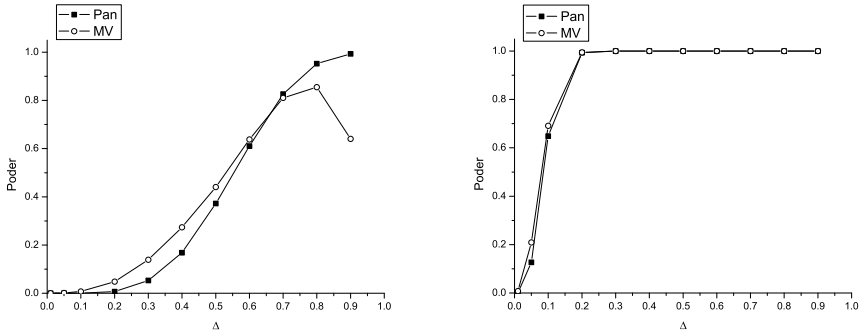
(a) (b)

Figura 17: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função dos tamanhos amostrais (n), $k = 10$, $\Delta = 0,4$; $p^{(1)} = 0,30$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$.

Na Figura 18 (a) e (b) são apresentadas as curvas de poder, em função de Δ, para $\alpha = 1\%$, $k = 5$ e $n = 10$ e 100, respectivamente. Os dois testes avaliados apresentaram um crescimento do poder com o aumento de Δ. Para diferenças maiores entre os valores de p dos dois grupos avaliados ($\Delta \geq 0,8$), com $n = 10$, observou-se um decréscimo do poder do teste MV. Na Figura 18 (b), é possível vizualizar o efeito do tamanho das amostras. Mesmo com valores de Δ pequeno ($\Delta \leq 0,2$), o poder dos testes foi superior aos valores de poder observados para $n = 10$ (Figura 18 (a)), se for fixado um teste e um valor de Δ.

Além disso, para grandes amostras, o teste de MV não apresentou redução do poder e as performances de ambos os testes aproximaram-se de 100% para $\Delta > 0,2$. A possível causa da redução do poder do teste MV com $n = 10$ e $\Delta \geq 0,8$ foi atribuída à proximidade de 0 ou de 1 dos parâmetros $p^{(1)}$ e $p^{(2)}$, respectivamente. A deficiência desse estimador, quando $p \to 0$ ou $p \to 1$ e n é pequeno, refletiu no poder do

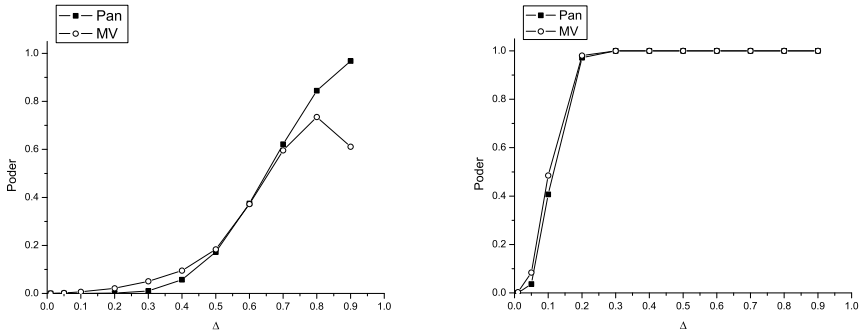
teste associado, como já havia sido preconizado.



(a) (b)

Figura 18: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função da diferença Δ, com $k = 5$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$ para $\alpha = 1\%$.

Finalmente, para $k = 10$, na Figura 19 (a) e (b) são apresentados os valores de poder para $\alpha = 1\%$, em função de Δ e com $n = 10$ e 100, respectivamente. Novamente, os testes apresentaram curvas de poder estimadas bem parecidas, se comparadas a Figura 18 (a) e (b) com a Figura 19 (a) e (b). Estes resultados também foram condizentes com os apresentados na Figura 12 (a) e (b), sob a hipótese H_1.



(a) (b)

Figura 19: Poder, sob H_0 parcial, dos testes de *bootstrap* Pan e MV, em função da diferença Δ, com $k = 10$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$ para $\alpha = 1\%$.

De maneira geral, o teste *bootstrap* de MV apresentou uma redução

no poder para diferenças entre grupos de proporções binomiais, maiores do que 0,8, com $n = 10$ e em todos os valores de k ($k = 2$, 5 e 10). Para $n = 10$, o poder dos testes Pan apresentou curva monótona não decrescente, sob as hipóteses H_0 parcial e H_1, de acordo com o esperado pela teoria. Quando o tamanho das amostras é grande ($n = 100$) e $\Delta \geq 0,2$, os valores de poder dos testes Pan e MV atingiram 100% e verificou-se uma redução do poder de ambos os testes com o aumento de k, se fixado o teste, o valor de n e de Δ.

4.3 Considerações finais

A performance de dois testes de comparações múltiplas de proporções binomiais foi comparada por simulação Monte Carlo. Os dois testes se diferenciaram basicamente no estimador das proporções, sendo um deles o estimador de máxima verossimilhança e o outro o estimador add-4 de Pan (Pan, 2002). Avaliaram-se as taxas de erro tipo I por experimento e o poder dos testes em diferentes situações que contemplavam os tamanhos amostrais, o número de populações, os valores de p e os níveis de significância nominais.

O erro tipo I por experimento foi controlado em nível inferior ou, no máximo, igual ao valor nominal α adotado em ambos os testes. Em nenhuma situação, houve comportamento liberal (erro tipo I superior a α) de ambos os testes. O teste *bootstrap* Pan foi, em geral, mais conservativo e esperava-se que o seu poder fosse inferior ao poder do teste *bootstrap* MV. Isso não ocorreu ou, quando ocorreu, as diferenças observadas no poder a favor do teste *bootstrap* MV foram quase sempre inexpressivas. Este comportamento foi o mesmo, tanto para $\alpha = 5\%$ quanto para $\alpha = 1\%$.

O erro tipo I por experimento também foi avaliado sob H_0 parcial nas comparações intragrupos. Os testes apresentaram taxas de erro tipo I quase sempre inferiores ao valor nominal α adotado e em nenhuma configuração houve resultados que classificassem os testes como liberais. Nestes casos, o que ocorreu foi uma acentuada redução das taxas de erro tipo I dos testes quando estes eram comparados com a situação de

H_0 completa. O poder, nesta mesma situação de H_0 parcial medido nas comparações intergrupos, foi até maior que o observado sob H_1, contrariando novamente a expectativa.

O poder dos dois testes mostrou, via de regra, uma pequena superioridade do teste *bootstrap* MV, embora nas situações em que o teste *bootstrap* Pan foi superior, esta superioridade tenha sido muito expressiva. Isso ocorreu quando os valores de Δ eram grandes ($\Delta \geq 0,8$) e n menores ($n \leq 30$), mas, a causa foi atribuída ao afastamento de p do valor $\frac{1}{2}$, em que a normal não se ajusta adequadamente à binomial. Estes fatos ocorreram tanto sob H_1 quanto sob H_0 parcial. Se o valor de n é grande ($n = 100$), os valores de poder de ambos os testes se igualam, mesmo para grandes valores de Δ ($\Delta \geq 0,8$).

É importante enfatizar que os valores de poder de ambos os testes são relativamente baixos se as amostras são pequenas ($n \leq 30$). Nos experimentos de avaliação de sementes quanto ao vigor e poder germinativo são utilizadas quatro repetições com 25 sementes para cada tratamento. Assim, os tamanhos amostrais são iguais a 100 para cada população (tratamento), não havendo nenhum problema com o poder dos testes. Já no caso de ensaios com insetos, os tamanhos amostrais são, em geral, limitados em valores iguais ou inferiores a 30, o que causa uma redução no poder dos testes em detectar diferenças significativas entre os tratamentos.

Um outro interessante fato constatado refere-se ao efeito do número k de populações no poder dos testes. Verificou-se que o aumento de k provocava uma expressiva redução no poder, fixados os demais fatores estudados. Esta redução é cada vez menor à medida que o valor n cresce. Em geral, os pesquisadores, ao aumentarem o número de populações (tratamentos), tendem a reduzir o número de repetições para manter o mesmo número de parcelas no experimento. Portanto, o efeito pode ser indesejado quando os testes *bootstrap* para proporções forem utilizados. Assim, deve-se atentar para este fato antes de se realizar algum tipo de planejamento de experimento em que serão avaliadas proporções binomiais.

5 CONCLUSÕES

Os testes de comparações múltiplas em populações binomiais apresentaram excelentes performances, controlando o erro tipo I, por experimento, em níveis iguais ou inferiores aos valores de significância e curvas de poder com padrão correspondente ao esperado pela teoria.

Nas diferentes configurações avaliadas recomenda-se a utilização do teste *bootstrap* de Pan em conseqüência da melhor performance em relação ao poder nas situações em que as proporções binomiais se afastam de 1/2 e os tamanhos amostrais são pequenos ($n \leq 10$).

6 REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A.; COULL, B. A. Approximate is better than "exact" for interval estimation of binomial proportions. **American Statistician**, Alexandria, v. 52, n. 2, p. 119-126, May 1998.

BANZATO, D. A.; KRONKA, S. N. **Experimentação Agrícola**. Jaboticabal: FUNEP, 1989. 247 p.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society Serie B - Methodological**, oxford, v. 57, n. 1, p. 289-300, 1995.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 5. ed. São Paulo: Saraiva, 2004. 526 p.

CARMER, S. G.; SWANSON, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. **Journal American Statistical Association**, Washington, v. 68, n. 341, p. 66-74, Mar.1973.

CARARI, M. L. **Intervalo de confiança para a diferença de duas proporções binomiais utilizando *Bootstrap* infinito**. 2004. 66 p. Dissertação (Mestrado em Agronomia) - Universidade Federal de Lavras, Lavras, MG.

CECCHETTI, D. **Poder e taxas de erro tipo I de quatro critérios multivariados para o teste de igualdade de efeitos de tratamentos avaliados por meio do método de Monte Carlo**. 1999. 56 p. Dissertação (Mestrado em Agronomia) - Universidade Federal de Lavras, Lavras, MG.

CROWLEY, P. H. Resampling methods for computation-intensive data analysis in ecology and evolution. **Annual Review of Ecology and Systematics**, Palo Alto, v.23, p.405 - 447, 1992.

CONLON, M.; THOMAS, R.G. A new confidence interval for the difference of two binomial proportions. **Computational Statistics & Data Analysis**, Amsterdam, v. 9, n. 2, p. 237-241, Mar. 1990.

DACHS, J. N. **Estatística computacional:** uma introdução em turbo Pascal. Rio de Janeiro: Livros Técnicos e Científicos Editora Ltda., 1988. 236p.

DUNCAN, D. B. Multiple range and multiple F test. **Biometrics**, Washington, v. 11, n. 1, p. 1-41, Mar. 1955.

EFRON, B.; TIBSHIRANI, R. **An introduction to the *bootstrap***. New York: Chapmam & Hall, 1993. 436 p.

FERREIRA, D. F.; DEMÉTRIO, C. G. B.; MANLY, B. F. J.; MACHADO, A. A. Aplicações dos métodos de *bootstrap* nos procedimentos de comparações múltiplas. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 50.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 11., Londrina, 2005. **Anais...** Londrina: ISBN, 2005. p. 136.

FERREIRA, D. F. **Estatística Básica** 1. ed. Lavras: UFLA, 2005. 664 p.

GUERRA, M. J.; DONAIRE, D. **Estatística indutiva:** teoria e exercícios. 2. ed. São Paulo: Liv. Ciência e Tecnologia, 1982. 311 p.

HOCHBERG, Y.; TANHANE, A. C. **Multiple comparison precedures**. New York: J. Wiley & Sons, 1987. 450 p.

LEEMIS, L. M.; TRIVED, K. S. A comparison of approximate interval estimators for the Bernoulli parameter. **The American Statistician**, Alexandria, v. 50, n. 1, p. 63-68, Feb. 1996.

MACHADO, A. A.; DEMÉTRIO, C. G. B.; FERREIRA, D. F.; SILVA, J. G. C. Estatística experimental: uma abordagem fundamentada no planejamento e no uso de recursos computacionais. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 50.; SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 11., Londrina, 2005. **Curso...** Londrina: ISBN, 2005. 290 p.

MANLY, B. F. J. **Randomization, *bootstrap* and Monte Carlo**

methods in biology. 2. ed. London: Chapman-Hall, 1998. 399 p.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. New York: Jonh Wiley & Sons, 1974. 564 p.

PAN, W. Approximate confidence intervals for one proportion and two proportions. **Computational Statistics & Data Analysis**, Amsterdam, v. 40, n. 1, p. 143-157, July 2002.

PEARCE, S. C. Data analysis in agricultural experimentation III. Multiple comparisons. **Experimental Agriculture**, Cambridge, v. 29, n. 1, p. 1-8, Jan. 1993.

PERECIN, D.; MALHEIROS, E. B. Curso: Procedimentos para comparações múltiplas. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔNICA, 3; REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 34, Lavras, 1989. **Curso...** Lavras: ESAL, 1989. 67 p.

PETERSEN, R. G. Use and misure of multiple comparison procedures. **Agronomy Journal**, Madison, v. 69, p. 205-208, Mar./Apr. 1977.

PIMENTEL GOMES, F. **Curso de estatística experimental**. 11. ed. Piracicaba: Livraria Nobel, 1985. 468 p.

RAMALHO, M. A. P.; FERREIRA D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2000. 303 p.

SAS INSTITUTE SAS/STAT. **SAS/IML Software**: usage and reference, version 6. Cary, 1990. 501 p.

SATTERTWAITE, F. F. Synthesis of variance. **Psychometrika**, Dover, v. 6, p. 309-316, 1941.

SNEDCOR, G. W.; COCHRAN, W. G. **Statistical methods**. Iwoa: University of Ames, 1980. 507 p.

STEEL, R. G. D.; TORRIE, J. H. **Principles and procedures of statistics**. 2. ed. New York: McGraw-Hill Book, 1980. 633 p.

ANEXOS

ANEXO 1: Programa ulilizado para computar o erro tipo I por experimento dos testes de *bootstrap* Pan e MV, sob H_0 completa.

```
proc iml;
   /*Definir parâmetros para a simulação*/
   k=2;M=2000;B=2000;
   ni=j(k,1,10);
   n = sum(ni);
   print k M B;
   print ni;
   yi=j(k,1,0);
   yib1=j(k,1,0);
   yib2=j(k,1,0);
   pi=j(k,1,0.5);
   *pi[k]=0.9;
   print pi;
   pi1 = j(k,1,0);
   pi2 = j(k,1,0);
   pi1b = j(k,1,0);
   pi2b = j(k,1,0);
   qih = j(k,k,0);/* acima da diagonal qih1 e abaixo qih2*/
   nsig = j(4,1,0); /* nsig[1]= est 1, alpha = 0.05, nsig[2]= est 1,
                       alpha = 0.01, nsig[3]= est 2, alpha = 0.05,
                       nsig[4]= est 2, alpha =0.01*/
   alpha5 = 0.05;alpha1=0.01;
   /*Iniciar simulação Monte Carlo*/
   do i=1 to M;
      /*Gerar as amostras populacionais e estimação dos pi´s*/
         do pop = 1 to k;
            yi[pop] = RANBIN(0,ni[pop],pi[pop]);
            pi1[pop] = (yi[pop]+2)/(ni[pop]+4);
            pi2[pop] = yi[pop]/ni[pop];
         end;
         print yi pi1 pi2;
```

```
/*obter qih na amostra original*/
do ii = 1 to k-1;
    do iii = ii + 1 to k;
        v11 = pi1[ii]*(1-pi1[ii])/(ni[ii]+2);
        v12 = pi1[iii]*(1-pi1[iii])/(ni[iii]+2);
        qih1 = (max(pi1[ii],pi1[iii])-
        min(pi1[ii],pi1[iii]))/(v11+v12)**0.5;
        *print qih1 v11 v12;
        v21 = pi2[ii]*(1-pi2[ii])/ni[ii];
        v22 = pi2[iii]*(1-pi2[iii])/ni[iii];
        if (v21=0 & v22=0) then qih2 = (max(pi2[ii],pi2[iii])-
        min(pi2[ii],pi2[iii]));
        else qih2 = (max(pi2[ii],pi2[iii])-
        min(pi2[ii],pi2[iii]))/(v21+v22)**0.5;
        *print qih2 v21 v22;
        qih[ii,iii] = qih1;
        qih[iii,ii] = qih2;
    end;
end;
print qih;
/* Reamostragem infinita das k amostras impondo H0 com
estimativas iguais dos pi´s */
pi1m = 0;
pi2m = 0;
do pop = 1 to k;
    pi1m = pi1m + pi1[pop]*(ni[pop]-1)/(n - k);
    pi2m = pi2m + pi2[pop]*(ni[pop]-1)/(n - k);
end;
do pop = 1 to k;
    yib1[pop] = RANBIN(0,ni[pop],pi1m);
    if pi2m = 0 then yib2[pop] = 0;
    else if pi2m = 1 then yib2[pop] = ni[pop];
    else yib2[pop] = RANBIN(0,ni[pop],pi2m);
```

```
    pi1b[pop] = (yib1[pop]+2)/(ni[pop]+4);
    pi2b[pop] = yib2[pop]/ni[pop];
end;
print pi1m pi2m yib1 yib2 pi1b pi2b;
/* obter qihb na amostra de bootstrap */
omeg1=-1;
omeg2=-1;
do ii = 1 to k-1;
    do iii = ii + 1 to k;
        v11b = pi1b[ii]*(1-pi1b[ii])/(ni[ii]+2);
        v12b = pi1b[iii]*(1-pi1b[iii])/(ni[iii]+2);
        qih1b = (max(pi1b[ii],pi1b[iii])-
        min(pi1b[ii],pi1b[iii]))/(v11b+v12b)**0.5;
        *print qih1b v11b v12b;
        v21b = pi2b[ii]*(1-pi2b[ii])/ni[ii];
        v22b = pi2b[iii]*(1-pi2b[iii])/ni[iii];
        if (v21b=0 & v22b=0) then qih2b =
        (max(pi2b[ii],pi2b[iii]) -
        min(pi2b[ii],pi2b[iii]));
        else qih2b = (max(pi2b[ii],pi2b[iii])-
        min(pi2b[ii],pi2b[iii]))/(v21b+v22b)**0.5;
        *print qih2b v21b v22b;
        if qih1b>omeg1 then omeg1=qih1b;
        if qih2b>omeg2 then omeg2=qih2b;
    end;
end;
do ii = 1 to k-1;
    do iii = ii + 1 to k;
        if omeg1>=qih[ii,iii] then pvalb[ii,iii]=
        pvalb[ii,iii]+1/B;
        if omeg2>=qih[iii,ii] then pvalb[iii,ii]=
        pvalb[iii,ii]+1/B;
    end;
```

```
        end;
        print omeg1 omeg2 pvalb;
        end;/*fim do bootstrap*/
        achou5m1=1;
        achou1m1=1;
        achou5m2=1;
        achou1m2=1;
        do ii = 1 to k-1;
            do iii = ii + 1 to k;
                if pvalb[ii,iii]<=alpha5 then achou5m1 = 0;
                if pvalb[ii,iii]<=alpha1 then achou1m1 = 0;
                if pvalb[iii,ii]<=alpha5 then achou5m2 = 0;
                if pvalb[iii,ii]<=alpha1 then achou1m2 = 0;
            end;
        end;
        if achou5m1 = 0 then nsig[1]=nsig[1]+1/M;
        if achou1m1 = 0 then nsig[2]=nsig[2]+1/M;
        if achou5m2 = 0 then nsig[3]=nsig[3]+1/M;
        if achou1m2 = 0 then nsig[4]=nsig[4]+1/M;
    end;/*fim da Monte Carlo*/
    *print achou5m1 achou1m1 achou5m2 achou1m2;
    print nsig;
quit;
```

ANEXO 2: Programa ulilizado para computar o poder dos testes de *bootstrap* Pan e MV, sob H_1.

```
proc iml;
    /*Definir parâmetros para a simulação*/
    k=2;M=2000;B=2000;delta=0.01;
    ni=j(k,1,10);
    n = sum(ni);
    print k M B;
```

```
    print ni;
    yi=j(k,1,0);
    yib1=j(k,1,0);
    yib2=j(k,1,0);
    pi=j(k,1,0.01);
    diff=delta/(k-1);
    do i=2 to k;
       pi[i]=pi[i-1]+diff;
    end;
    print delta diff;
    print pi;
    pi1 = j(k,1,0);
    pi2 = j(k,1,0);
    pi1b = j(k,1,0);
    pi2b = j(k,1,0);
    qih = j(k,k,0);/* acima da diagonal qih1 e abaixo qih2*/
    nsig = j(4,1,0);/* nsig[1]= est 1, alpha = 0.05, nsig[2]= est 1,
                        alpha = 0.01, nsig[3]= est 2, alpha = 0.05,
                       nsig[4]= est 2, alpha = 0.01*/
    alpha5 = 0.05;alpha1=0.01;
/*Iniciar simulação Monte Carlo*/
    do i=1 to M;
          /*Gerar as amostras populacionais e estimação dos pi´s*/
          do pop = 1 to k;
             yi[pop] = RANBIN(0,ni[pop],pi[pop]);
             pi1[pop] = (yi[pop]+2)/(ni[pop]+4);
             pi2[pop] = yi[pop]/ni[pop];
          end;
          *print yi pi1 pi2;
          /*obter qih na amostra original*/
          do ii = 1 to k-1;
             do iii = ii + 1 to k;
                v11 = pi1[ii]*(1-pi1[ii])/(ni[ii]+2);
```

```
        v12 = pi1[iii]*(1-pi1[iii])/(ni[iii]+2);
        qih1 = (max(pi1[ii],pi1[iii]) -
        min(pi1[ii],pi1[iii]))/(v11+v12)**0.5;
        *print qih1 v11 v12;
        v21 = pi2[ii]*(1-pi2[ii])/ni[ii];
        v22 = pi2[iii]*(1-pi2[iii])/ni[iii];
        if (v21=0 & v22=0) then qih2 = (max(pi2[ii],pi2[iii]) -
        min(pi2[ii],pi2[iii]));
        else qih2 = (max(pi2[ii],pi2[iii]) -
        min(pi2[ii],pi2[iii]))/(v21+v22)**0.5;
        *print qih2 v21 v22;
        qih[ii,iii] = qih1;
        qih[iii,ii] = qih2;
    end;
 end;
print qih;
 /* Reamostragem infinita das k amostras impondo H0
 com estimativas iguais dos pi´s */
pi1m = 0;
pi2m = 0;
do pop = 1 to k;
    pi1m = pi1m + pi1[pop]*(ni[pop]-1)/(n - k);
    pi2m = pi2m + pi2[pop]*(ni[pop]-1)/(n - k);
end;
do pop = 1 to k;
    yib1[pop] = RANBIN(0,ni[pop],pi1m);
    if pi2m = 0 then yib2[pop] = 0;
    else if pi2m = 1 then yib2[pop] = ni[pop];
    else yib2[pop] = RANBIN(0,ni[pop],pi2m);
    pi1b[pop] = (yib1[pop]+2)/(ni[pop]+4);
    pi2b[pop] = yib2[pop]/ni[pop];
end;
print pi1m pi2m yib1 yib2 pi1b pi2b;
```

```
    /* obter qihb na amostra de bootstrap */
    omeg1=-1;
    omeg2=-1;
    do ii = 1 to k-1;
        do iii = ii + 1 to k;
            v11b = pi1b[ii]*(1-pi1b[ii])/(ni[ii]+2);
            v12b = pi1b[iii]*(1-pi1b[iii])/(ni[iii]+2);
            qih1b = (max(pi1b[ii],pi1b[iii]) -
            min(pi1b[ii],pi1b[iii]))/(v11b+v12b)**0.5;
            print qih1b v11b v12b;
            v21b = pi2b[ii]*(1-pi2b[ii])/ni[ii];
            v22b = pi2b[iii]*(1-pi2b[iii])/ni[iii];
            if (v21b=0 & v22b=0) then qih2b =
            (max(pi2b[ii],pi2b[iii])- min(pi2b[ii],pi2b[iii]));
            else qih2b = (max(pi2b[ii],pi2b[iii]) -
            min(pi2b[ii],pi2b[iii]))/(v21b+v22b)**0.5;
            print qih2b v21b v22b;
            if qih1b>omeg1 then omeg1=qih1b;
            if qih2b>omeg2 then omeg2=qih2b;
        end;
    end;
    do ii = 1 to k-1;
        do iii = ii + 1 to k;
            if omeg1>=qih[ii,iii] then pvalb[ii,iii] =
            pvalb[ii,iii]+1/B;
            if omeg2>=qih[iii,ii] then pvalb[iii,ii] =
            pvalb[iii,ii]+1/B;
        end;
    end;
end;/*fim do bootstrap*/
print omeg1 omeg2 ;
achou5m1=1;
achou1m1=1;
```

```
        achou5m2=1;
        achou1m2=1;
        print pvalb;
        do ii = 1 to 1;
            do iii = k to k;
                if pvalb[ii,iii]<=alpha5 then achou5m1 = 0;
                if pvalb[ii,iii]<=alpha1 then achou1m1 = 0;
                if pvalb[iii,ii]<=alpha5 then achou5m2 = 0;
                if pvalb[iii,ii]<=alpha1 then achou1m2 = 0;
            end;
        end;
        if achou5m1 = 0 then nsig[1]=nsig[1]+1/M;
        if achou1m1 = 0 then nsig[2]=nsig[2]+1/M;
        if achou5m2 = 0 then nsig[3]=nsig[3]+1/M;
        if achou1m2 = 0 then nsig[4]=nsig[4]+1/M;
    end;/*fim da Monte Carlo*/
    print nsig;
quit;
```

ANEXO 3: Programa ulilizado para computar o erro tipo I, por experimento, e o poder dos testes de *bootstrap* Pan e MV, sob H_0 parcial.

```
proc iml;
    /*Definir parâmetros para a simulação*/
    k=5;M=2000;B=2000;delta=0.01;g=2;
    *ccont=0;
    ni=j(k,1,10);
    n = sum(ni);
    print k M B;
    print ni;
    yi=j(k,1,0);
    yib1=j(k,1,0);
    yib2=j(k,1,0);
```

```
pi=j(k,1,0.01);
gr=j(k,1,1);
if k=5 then
do;
   g1=3;
   g2=2;
   ncomp=g1*g2;
end;
else if k=10 then
do;
   g1=5;g2=5;
   ncomp=g1*g2;
end;/*específicos*/
diff=delta/(k-1);
do i=2 to k;
   if i<=g1 then pi[i]=pi[i-1];
   else if i=g1+1 then pi[i]=pi[i-1]+delta;
   else if i>g1+1 then pi[i]=pi[i-1];
   if i>g1 then gr[i]=2;
   print i;
end;
print gr;
print delta diff ;
print pi;
pi1 = j(k,1,0);
pi2 = j(k,1,0);
pi1b = j(k,1,0);
pi2b = j(k,1,0);
qih = j(k,k,0);/* acima da diagonal qih1 e abaixo qih2*/
nsig = j(4,1,0); /* nsig[1]= est 1, alpha = 0.05, nsig[2]= est 1,
                    alpha = 0.01, nsig[3]= est 2, alpha = 0.05,
                    nsig[4]= est 2, alpha = 0.01*/
nsigpod = j(4,1,0); /* nsig[1]= est 1, alpha = 0.05, nsig[2]= est 1,
```

```
                    alpha = 0.01, nsig[3]= est 2, alpha = 0.05,
                    nsig[4]= est 2, alpha = 0.01*/
alpha5 = 0.05;alpha1=0.01;
 /*Iniciar simulação Monte Carlo*/
do i=1 to M;
   /*Gerar as amostras populacionais e estimação dos pi´s*/
   do pop = 1 to k;
      yi[pop] = RANBIN(0,ni[pop],pi[pop]);
      pi1[pop] = (yi[pop]+2)/(ni[pop]+4);
      pi2[pop] = yi[pop]/ni[pop];
   end;
   *print yi pi1 pi2;
   /*obter qih na amostra original*/
   do ii = 1 to k-1;
      do iii = ii + 1 to k;
         v11 = pi1[ii]*(1-pi1[ii])/(ni[ii]+2);
         v12 = pi1[iii]*(1-pi1[iii])/(ni[iii]+2);
         qih1 = (max(pi1[ii],pi1[iii]) -
         min(pi1[ii],pi1[iii]))/(v11+v12)**0.5;
         *print qih1 v11 v12;
         v21 = pi2[ii]*(1-pi2[ii])/ni[ii];
         v22 = pi2[iii]*(1-pi2[iii])/ni[iii];
         if (v21=0 & v22=0) then qih2 = (max(pi2[ii],pi2[iii]) -
         min(pi2[ii],pi2[iii]));
         else qih2 = (max(pi2[ii],pi2[iii]) -
         min(pi2[ii],pi2[iii]))/(v21+v22)**0.5;
         *print qih2 v21 v22;
         qih[ii,iii] = qih1;
         qih[iii,ii] = qih2;
      end;
    end;
   *print qih;
   /* Reamostragem infinita das k amostras impondo H0
```

```
com estimativas iguais dos pi´s*/
pi1m = 0;
pi2m = 0;
do pop = 1 to k;
   pi1m = pi1m + pi1[pop]*(ni[pop]-1)/(n - k);
   pi2m = pi2m + pi2[pop]*(ni[pop]-1)/(n - k);
end;
do pop = 1 to k;
   yib1[pop] = RANBIN(0,ni[pop],pi1m);
   if pi2m = 0 then yib2[pop] = 0;
   else if pi2m = 1 then yib2[pop] = ni[pop];
   else yib2[pop] = RANBIN(0,ni[pop],pi2m);
   pi1b[pop] = (yib1[pop]+2)/(ni[pop]+4);
   pi2b[pop] = yib2[pop]/ni[pop];
end;
/* obter qihb na amostra de bootstrap */
omeg1=-1;
omeg2=-1;
do ii = 1 to k-1;
   do iii = ii + 1 to k;
      v11b = pi1b[ii]*(1-pi1b[ii])/(ni[ii]+2);
      v12b = pi1b[iii]*(1-pi1b[iii])/(ni[iii]+2);
      qih1b = (max(pi1b[ii],pi1b[iii]) -
      min(pi1b[ii],pi1b[iii]))/(v11b+v12b)**0.5;
      *print qih1b v11b v12b;
      v21b = pi2b[ii]*(1-pi2b[ii])/ni[ii];
      v22b = pi2b[iii]*(1-pi2b[iii])/ni[iii];
      if (v21b=0 & v22b=0) then qih2b =
      (max(pi2b[ii],pi2b[iii]) - min(pi2b[ii],pi2b[iii]));
      else qih2b = (max(pi2b[ii],pi2b[iii]) -
      min(pi2b[ii],pi2b[iii]))/(v21b+v22b)**0.5;
      *print qih2b v21b v22b;
      if qih1b>omeg1 then omeg1=qih1b;
```

```
        if qih2b>omeg2 then omeg2=qih2b;
      end;
  end;
  do ii = 1 to k-1;
      do iii = ii + 1 to k;
         if omeg1>=qih[ii,iii] then pvalb[ii,iii] = pvalb[ii,iii]+1/B;
         if omeg2>=qih[iii,ii] then pvalb[iii,ii] = pvalb[iii,ii]+1/B;
      end;
  end;
end;/*fim do bootstrap*/
achou5m1=1;
achou1m1=1;
achou5m2=1;
achou1m2=1;
*print pvalb;
do ii = 1 to k-1;
   do iii = ii+1 to k;
   if (pvalb[ii,iii]<=alpha5) & (gr[ii]=gr[iii]) then achou5m1 = 0;
   if (pvalb[ii,iii]<=alpha1) & (gr[ii]=gr[iii]) then achou1m1 = 0;
   if (pvalb[iii,ii]<=alpha5) & (gr[ii]=gr[iii]) then achou5m2 = 0;
   if (pvalb[iii,ii]<=alpha1) & (gr[ii]=gr[iii]) then achou1m2 = 0;
   /*poder*/
   if abs(gr[iii] - gr[ii])>1e-5 then
      do;
      *ccont=ccont+1;
      *tt=gr[ii];
      *ttt=gr[iii];
      *print tt ttt;
      if (pvalb[ii,iii]<=alpha5) then
      nsigpod[1]=nsigpod[1]+1/(M*ncomp);
      if (pvalb[ii,iii]<=alpha5) then
      nsigpod[2]=nsigpod[2]+1/(M*ncomp);
      if (pvalb[iii,ii]<=alpha5) then
```

```
            nsigpod[3]=nsigpod[3]+1/(M*ncomp);
            if (pvalb[iii,ii]<=alpha5) then
            nsigpod[4]=nsigpod[4]+1/(M*ncomp);
            end;
         end;
      end;
      if achou5m1 = 0 then nsig[1]=nsig[1]+1/M;
      if achou1m1 = 0 then nsig[2]=nsig[2]+1/M;
      if achou5m2 = 0 then nsig[3]=nsig[3]+1/M;
      if achou1m2 = 0 then nsig[4]=nsig[4]+1/M;
   end;/*fim da Monte Carlo*/
   *print nsig nsigpod ccont ncomp M;
   print nsig nsigpod;
quit;
```