COMPARAÇÕES MÚLTIPLAS PARA PROPORÇÕES BINOMIAIS UTILIZANDO *BOOTSTRAP*

Nádia Giaretta BIASE[1]
Daniel Furtado FERREIRA[1]

- RESUMO: 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. Por isso, este trabalho teve por objetivo avaliar a *performance* de dois testes de *bootstrap* envolvendo proporções binomiais, computando-se o erro tipo I por experimento e o poder dos testes. 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 e foram avaliados em diferentes configurações envolvendo número de populações e valores dos parâmetros, resultante de 2.000 simulações Monte Carlo. Observou-se excelente *performance* nos testes de *bootstrap* Pan e MV, controlando o erro tipo I por experimento em níveis iguais ou inferiores aos valores nominais de significância e elevados valores de poder. Recomenda-se a utilização do teste *bootstrap* de Pan por ter uma *performance* melhor nas situações em que as proporções binomiais se afastam de 1/2 e quando os tamanhos amostrais são pequenos ($n \leq 10$).

- PALAVRAS-CHAVE: Evento de Bernoulli; método Monte Carlo; estimador de máxima verossimilhança e de Pan.

1 Introdução

Em várias situações reais, o pesquisador se depara com a necessidade de comparar duas ou mais proporções binomiais, como por exemplo, nos ensaios de vigor e poder germinativo de sementes. Quando o número de populações é maior que dois, testes de hipóteses são geralmente aplicados por meio de uma análise

[1]Departamento de Ciências Exatas, Universidade Federal de Lavras - UFLA, Caixa Postal 37, CEP: 37200-000, Lavras, MG, Brasil. E-mail: *nadiabiase@yahoo.com.br / danielff@ufla.br*

de variância, quando forem utilizadas repetições experimentais em delineamentos simples ou complexos e os dados forem supostamente normais. Segundo Hochberg e Tanhane (1987) inicialmente deve-se aplicar um teste F para a igualdade média de todas as proporções e, posteriormente, se condicionado a rejeição dessa hipótese, é comum aplicar testes de comparações múltiplas como, por exemplo, Tukey, Scheffé e Student-Newman-Keuls (SNK).

A validade destes tesses depende de algumas pressuposições, tais como, normalidade dos resíduos, homogeneidade de variâncias e independência das observações. Em geral, a independência é garantida pela aleatorização, ainda que esta independência não tenha sido alcançada, garante a validade dos testes. A normalidade dos resíduos não existe, a não ser de forma aproximada e a homogeneidade de variâncias também é outra pressuposição não atendida para o caso de testes sobre proporções binomiais. 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, a não ser sob a hipótese nula de igualdade das proporções das populações.

Para contornar este problema, modelos lineares generalizados e inferência bayesiana são abordagens que vem sendo largamente empregadas. Os modelos lineares generalizados constituem-se em uma generalização dos modelos lineares clássicos em que a variável resposta pode envolver uma variedade de distribuições de probabilidades pertencentes à família exponencial, sem necessariamente ter uma distribuição normal (Lindsey, 1997). 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. De acordo com Manly (1998) esse método consiste em reamostrar os dados da amostra baseada na premissa de que na ausência de qualquer outro conhecimento da população os valores encontrados em uma amostra aleatória são os melhores guias da distribuição da população. Outra técnica conhecida como *bootstrap* infinito foi introduzida por Conlon e Thomas (1990) e 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 binomiais dois estimadores do parâmetro de interesse podem ser utilizados, quais sejam, o estimador de máxima verossimilhança e o estimador de Pan (2002). Este 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 o erro tipo I por experimento e o poder. Adicionalmente o método de *bootstrap* infinito foi 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 de sucesso da i-ésima população).

2 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 a 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: $p_1 = \ldots = p_k$ e hipóteses H_0 Parcial: $p_1 = \ldots = p_i \neq p_{i+1} = p_{i+2} = \ldots = p_k$. As simulações foram feitas considerando 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 do grupo 1, $p_1 = p_2 = \ldots = p_i$, e do grupo 2, $p_{i+1} = p_{i+2} = \ldots = p_k$, definida por Δ, sendo $\Delta = 0,01;\ 0,05;\ 0,1;\ 0,2;\ 0,3;\ 0,4;\ 0,5;\ 0,6;\ 0,7;\ 0,8\ e\ 0,9$. Para a realização destas simulações, considerou-se que o valor do parâmetro p dentro do grupo 1 foi de 0,01 e do grupo 2 de $p + \Delta$, 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, situações em que os valores de p se aproximavam de 0,5. Nestes casos, considerou-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. Nesta última situação procurou-se avaliar a *performance* dos testes em situações em que ambos os grupos tivessem valores de p próximos de 1/2, situação em que os testes binomiais possuem melhores desempenhos.

Na segunda etapa, o mesmo procedimento de simulação foi realizado para medir o poder sob a hipótese H_0 parcial, e sob a hipótese alternativa H_1 ($p_1 \neq p_2 \neq \ldots \neq p_k$). Em cada simulação realizada para avaliar o poder sob a hipótese alternativa H_1, foi feita a combinação entre os mesmos números de populações binomiais (k), tamanhos amostrais (n), diferença entre p_k e p_1 dada 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 distintos entre si, 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. O número de comparações, nestas situações específicas, foram dadas 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.

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 proporção de sucesso 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

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

A estatística (4) é a amplitude estudentizada para cada par de populações binomiais (i, h) e é a base do teste que é descrito na seqüência, sendo o seu valor obtido em cada simulação realizada. 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:

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

sendo 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 valores-p denominados ajustados (Thorpe e Holland, 2000) 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 valores-p de cada par de populações foram comparados com os valores-p nominais (α) de 0,01 e 0,05. Para os casos em que o valor-p 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 realizado 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.

3 Resultados e Discussão

3.1 Erro tipo I

Foram considerados a seguir a discussão sobre o erro tipo em duas situações distintas. Na primeira, o erro tipo I foi avaliado sob a hipótese nula completa, e na segunda, sob a hipótese nula parcial.

3.1.1 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 e 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). 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%. Houve tendências de melhores resultados nos testes, 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) e diferentes tamanhos de amostras (n) para os estimadores de Pan (Pan, 2002) e Máxima Verossimilhança (MV) ao valor nominal de 5%

		$p = 0,1$		$p = 0,5$		$p = 0,9$	
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 $0,5$, os testes tenham apresentado um 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 tamanhos não diferentes 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 de $0,5$ 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 (resultados não apresentados).

3.1.2 Erro tipo I sob H_0 parcial

Novamente, ao avaliar as taxas de erro tipo I sob H_0 parcial, observou-se o mesmo comportamento geral dos testes para o valor nominal de significância de 1% e 5%. É conveniente salientar que para garantir que o espaço paramétrico dos $p_i's$ 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 grande ($\Delta = 0,9$) 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, conforme a Figura 3.

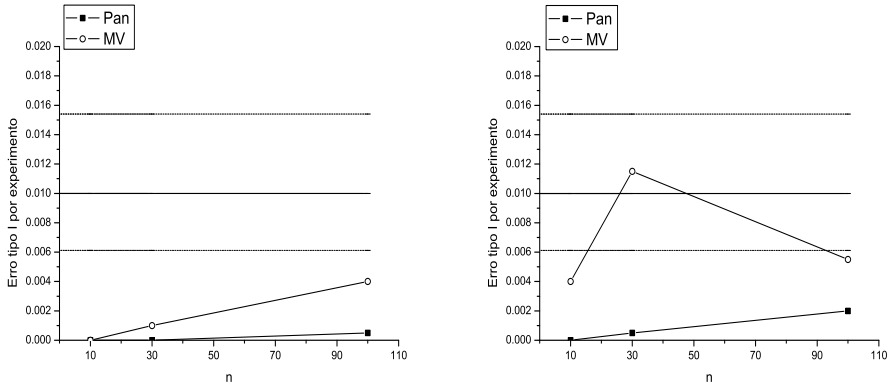
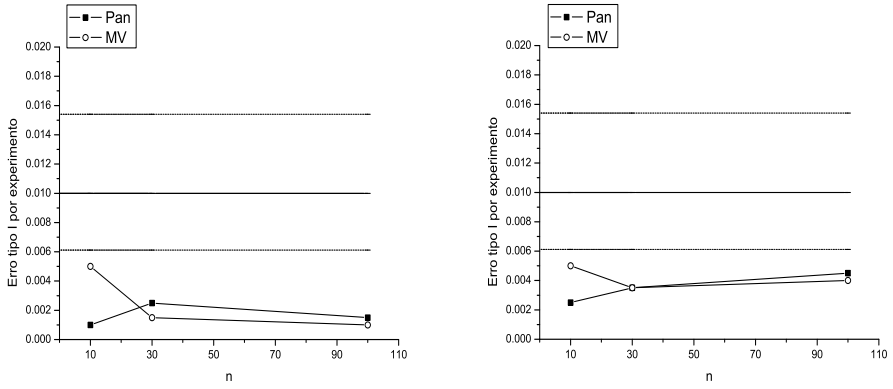


Figura 1: Taxas de erro tipo I por experimento dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 5$ e (b) $k = 10$, com $\Delta = 0,05$ e para $\alpha = 1\%$ considerando a hipótese H_0 parcial.

Nas Figuras 1 (a) e (b) são 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 e $\alpha = 1\%$. No teste de *bootstrap* MV, observou-se *performance* um pouco superior, pois os níveis de significância foram superiores aos do teste de Pan, mas inferiores (conservativo) ou iguais (ideal) ao nível nominal. 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).

Nas Figuras 2 (a) e (b) são apresentadas as taxas de erro 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 tenderam a se igualar com relação as taxas observadas de erro tipo I por experimento.



(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) $k = 5$ e (b) $k = 10$, com $\Delta = 0,5$ e para $\alpha = 1\%$ considerando a hipótese H_0 parcial.

Nas Figuras 3 (a) e (b) estão apresentadas as taxas de erro tipo I por experimento sob H_0 parcial para os testes *bootstrap* de Pan e MV para $\Delta = 0,9$ e com $k = 5$ e 10. Novamente em todas as situações de n os testes de bootstrap foram 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 possuem piores desempenhos. O teste bootstrap MV foi um pouco superior ao teste de *bootstrap* de Pan, embora de forma bem inexpressiva. O mesmo comportamento foi observado para o nível nominal de 5%.

Procurando avaliar situações em que os valores de p se aproximavam de $0,5$, ainda sob H_0 parcial, foram feitas simulações adicionais onde se avaliou o erro tipo I

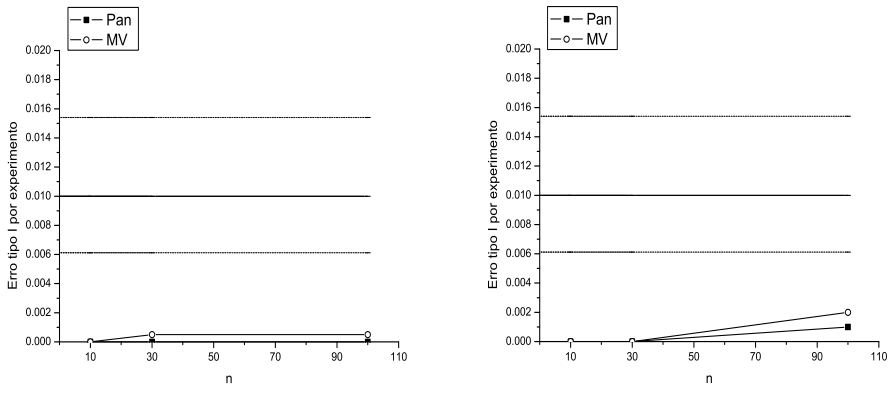


Figura 3: Taxas de erro tipo I por experimento dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n) e número de populações iguais (a) $k = 5$ e (b) $k = 10$, com $\Delta = 0,9$ e para $\alpha = 1\%$ considerando a hipótese H_0 parcial.

por experimento considerando os dois níveis nominais de significância de 1% e 5%.

As taxas de erro tipo I por experimento sob H_0 parcial dos dois testes de *bootstrap* 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, são apresentadas nas Figuras 4 (a) e (b). 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 4 (b)) os resultados foram todos conservativos, embora menos conservativos do que os casos semelhantes sob H_0 parcial com um dos grupos tendo valores de p afastados de $0,5$. Houve uma tendência das taxas aproximarem-se do valor nominal 5% com o aumento de n. Os testes MV e Pan tiveram resultados bastante parecidos.

3.2 Poder do teste

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 possuíam 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 3.2.1 e 3.2.2.

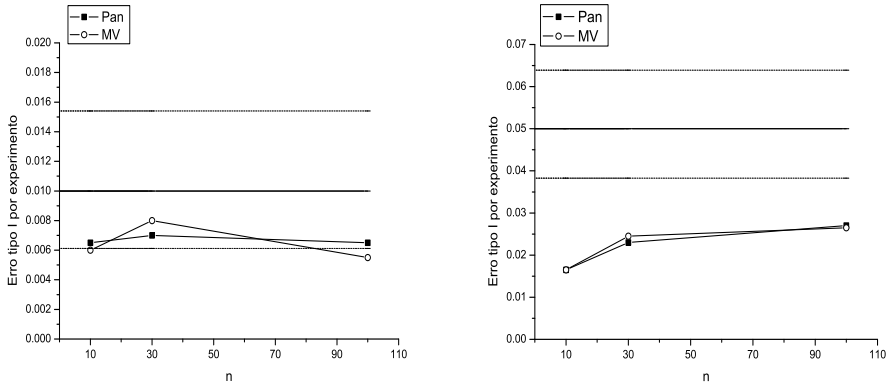


Figura 4: Taxas de erro tipo I por experimento dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n), com $k = 10$, $\Delta = 0,01$; $p^{(1)} = 0,50$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$ considerando a hipótese H_0 parcial.

3.2.1 Poder do teste sob H_1

Na Tabela 2 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 entre a maior e a menor proporção binomial das k populações. Para valores pequenos de Δ ($\Delta = 0,1$), o poder é inferior ao valor nominal $\alpha = 0,05$ quando n é pequeno ($n = 10$). Isso aconteceu com todos os valores de k dos testes estudados. Observou-se um acréscimo considerável no poder dos dois testes com o aumento de n de 10 para 30 ou para 100. Pode-se observar também, 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.

O teste *bootstrap* de MV foi quase sempre superior ao teste *bootstrap* de Pan em relação ao poder. Eles tendem a igualar seus desempenhos quando n aumenta. Quando os Δ eram grandes ($\Delta \geq 0,5$), em geral, os testes também tenderam a apresentar desempenhos iguais. No entanto, quando as diferenças foram muito grandes ($\Delta = 0,9$), em que 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 ocorreu devido ao fato apontado por outros pesquisadores (Agresti e Coull, 1998; Pan, 2002), de que o estimador das proporções add-4, que 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, tem 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.

Tabela 2 – Poder (%) sob H_1 para diferentes números de populações (k) e diferentes tamanhos de amostras (n) para os estimadores de Pan (Pan, 2002) e Máxima Verossimilhança (MV) ao nível nominal de 5%

		$\Delta = 0,1$		$\Delta = 0,5$		$\Delta = 0,9$	
k	n	Pan	MV	Pan	MV	Pan	MV
2	10	1,25	1,25	81,45	81,45	100,00	64,35
2	30	19,80	35,30	100,00	100,00	100,00	95,45
2	100	88,15	89,70	100,00	100,00	100,00	100,00
5	10	0,00	0,70	37,50	39,15	99,50	65,65
5	30	2,85	8,30	99,25	99,85	100,00	95,65
5	100	61,70	66,85	100,00	100,00	100,00	100,00
10	10	0,00	0,55	17,80	19,35	97,00	59,80
10	30	0,90	4,05	97,55	98,40	100,00	95,85
10	100	42,80	50,80	100,00	100,00	100,00	100,00

Uma situação retratada para $\alpha = 1\%$ refere-se ao poder dos testes expresso em função da diferença Δ. Nas Figuras 5 (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 o aumento de Δ, o que é esperado pela teoria (Mood; Graybill; Boes, 1974).

Uma redução no poder foi observada na Figura 5 (a) a partir de $\Delta = 0,8$ para o teste MV com $n = 10$, 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. No teste Pan nesta mesma situação foi observada uma curva de poder estimada condizente com o esperado, ou seja, monótona crescente. Para amostras grandes (Figura 5 (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 das Figuras 5 (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 o problema de queda de poder eliminado com $n = 100$. Diante disso, observa-se que o tamanho de amostra tem um importante papel na qualidade do estimador e portanto na *performance* do teste.

Nas Figuras 6 (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 $k = 2$ (Figuras 5 (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 atingiram 100% e esse efeito deixou de existir. Observou-se novamente com $n = 10$ e $\Delta > 0,7$, que o teste MV tem as mesmas deficiências relatadas anteriormente, sendo superado pelo teste de Pan.

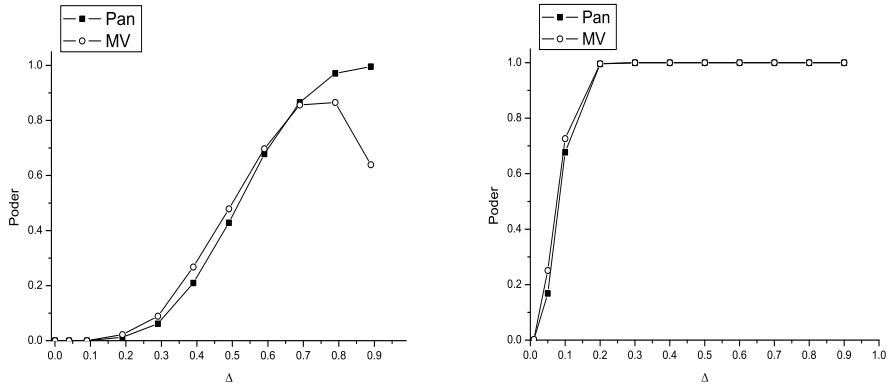


Figura 5: Poder sob H_1 dos testes de *bootstrap* Pan e MV em função da diferença Δ, com $k = 2$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$, para $\alpha = 1\%$.

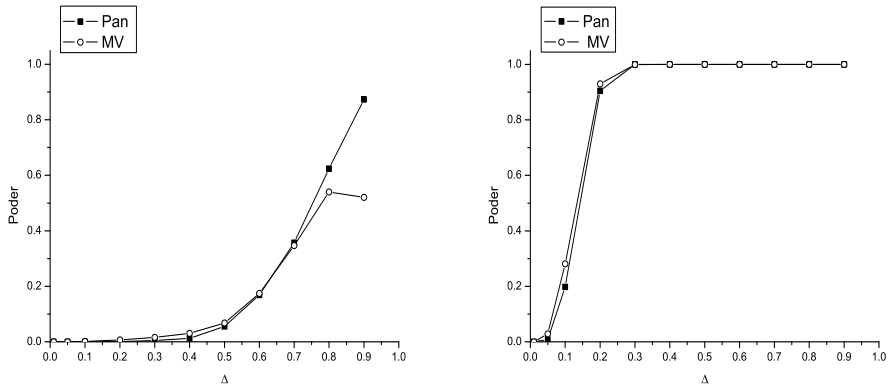


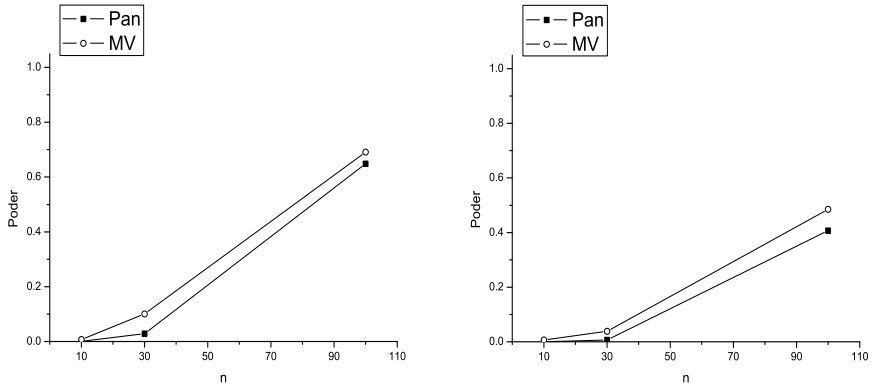
Figura 6: Poder sob H_1 dos testes de *bootstrap* Pan e MV em função da diferença Δ, com $k = 10$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$, para $\alpha = 1\%$.

3.2.2 Poder do teste sob H_0 parcial

Nas Figuras 7 (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 $\alpha = 1\%$. Pode-se observar nas Figuras 7 (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, independentemente 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, sendo que este crescimento foi maior para $k = 5$.

Ao comparar os valores de poder dos testes Pan e MV nas Figuras 7 (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 7 (a)) para $k = 10$ (Figura 7 (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.



(a) (b)

Figura 7: 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 $\alpha = 1\%$.

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 foram apresentados nas Figuras 8 (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 8 (a)) para $k = 10$ (Figura 8 (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 $n \geq 30$.

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 baseado neste estimador pudessem ser resultantes

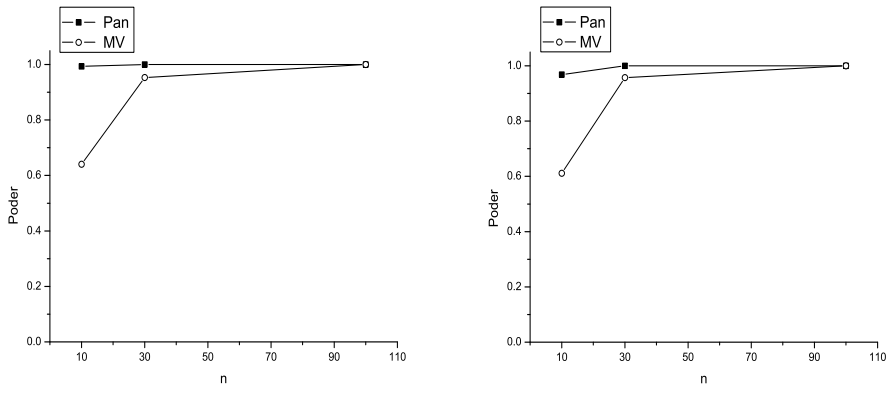


Figura 8: Poder sob H_0 parcial dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n), diferença $\Delta = 0,9$ e número de populações iguais (a) $k = 5$ e (b) $k = 10$, para $\alpha = 1\%$.

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 $0,5$.

Nas Figuras 9 (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 tiveram 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 tiveram um pequeno acréscimo. Se forem comparados estes resultados com aqueles da Figura 7 (b), pode-se observar que houve uma drástica redução de poder se for fixado o mesmo valor de n.

Nas Figuras 10 (a) e (b) foram apresentados os valores de poder para $\alpha = 1$ e 5%, respectivamente, em função de n para $\Delta = 0,4$. Novamente observou-se para todos os valores de n, que os dois testes tiveram *performance* coincidente 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 $n \geq 30$.

Nas Figuras 11 (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 tiveram 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 11 (b), foi 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 11 (a)), se for fixado um teste e um valor de Δ. Além disso, para grandes amostras

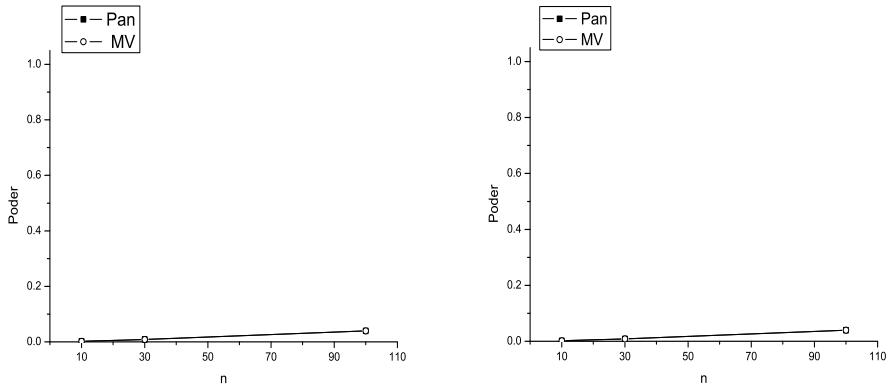


Figura 9: Poder sob H_0 parcial dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n), $k = 10$, $\Delta = 0,1$; $p^{(1)} = 0,45$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$.

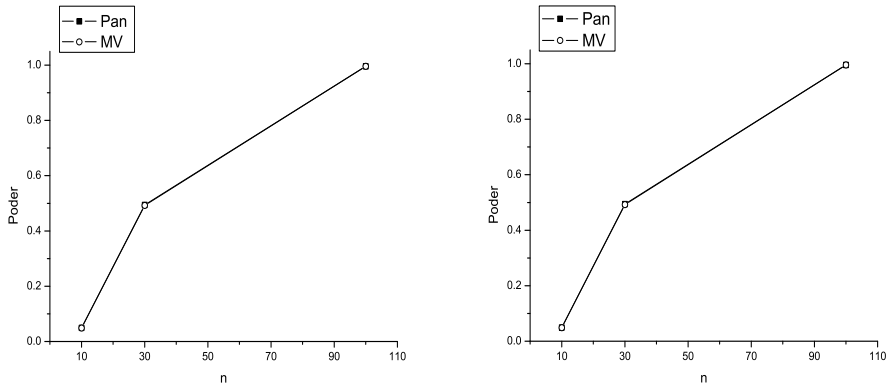


Figura 10: Poder sob H_0 parcial dos testes de *bootstrap* Pan e MV em função dos tamanhos amostrais (n), $k = 10$, $\Delta = 0,4$; $p^{(1)} = 0,30$ e valores nominais de significância iguais (a) $\alpha = 1\%$ e (b) $\alpha = 5\%$.

o teste de MV não teve redução do poder e as *performances* de ambos os testes aproximaram 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 foi refletida no poder do teste associado, como já havia sido preconizado.

Finalmente para $k = 10$, nas Figuras 12 (a) e (b) são apresentados os valores de poder para $\alpha = 1\%$ em função de Δ e com $n = 10$ e 100, respectivamente. Foram

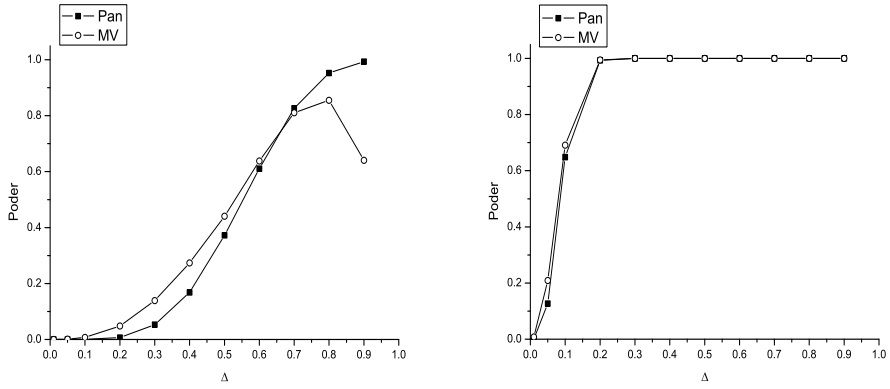


Figura 11: Poder sob H_0 parcial dos testes de *bootstrap* Pan e MV em função da diferença Δ, com $k = 5$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$ para $\alpha = 1\%$.

observados nos testes curvas de poder estimadas bem parecidas, se comparadas as Figuras 11 (a) e (b) com as Figuras 12 (a) e (b). Estes resultados também foram condizentes com os apresentados nas Figuras 6 (a) e (b) sob a hipótese H_1.

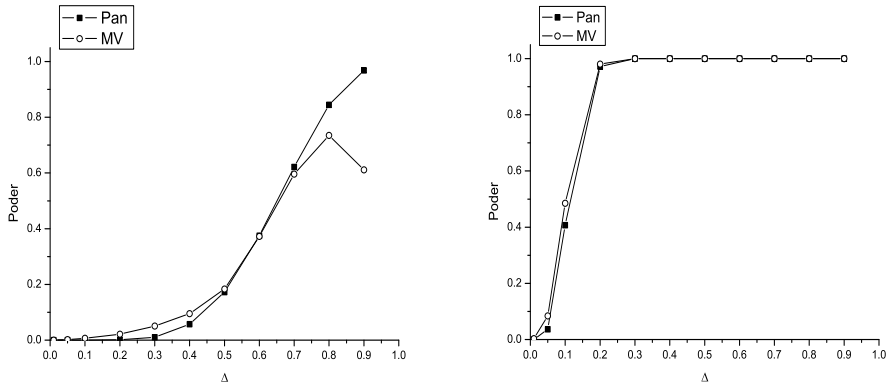


Figura 12: Poder sob H_0 parcial dos testes de *bootstrap* Pan e MV em função da diferença Δ, com $k = 10$ e tamanhos amostrais iguais (a) $n = 10$ e (b) $n = 100$ para $\alpha = 1\%$.

De maneira geral, no teste *bootstrap* de MV observou-se 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 $n = 10$, no poder dos testes Pan observou-se uma 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.

Conclusões

Os testes de comparações múltiplas em populações binomiais tiveram 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, dentro das condições definidas na abrangência desse estudo.

Nas diferentes configurações avaliadas recomenda-se a utilização do teste *bootstrap* de Pan em consequência 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 ($n \leq 10$).

BIASE, N. G.; FERREIRA, D. F. Multiple comparison for binomial proportions using *bootstrap*. *Rev. Mat. Estat.*, São Paulo, v.24, n.1, p.95-112, 2006.

▪ *ABSTRACT: 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 trials. However, 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 power. These two infinite* bootstrap *tests are distinguished by the estimators of the parameter p_i. One of these tests considered the maximum likelihood estimator (ML) and the other took into account the Pan's estimator. Both tests were evaluated in different configurations considering the number of populations and the parameter values resulting from 2,000 Monte Carlo simulations. Pan´s and ML* bootstrap *tests had excellent performances, controlling the experimentwise type-I error rates at the same levels or at lower levels than those of significance nominal values in addition to elevated powers. Pan's* bootstrap *tests is preferable due to the better performance in situations where the binomial proportions are distant from 1/2 and sample sizes are small ($n \leq 10$).*

▪ *KEYWORDS: Bernoulli trials; Monte Carlo method; Maximum likelihood and Pan's estimators.*

Referências

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

CONLON, M.; THOMAS, R. G. A new confidence interval for the difference of two binomial proportions. *Comput. Stat. Data Anal.*, Amsterdam, v.9, n.2, p.237-241, 1990.

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

LINDSEY, J. K. *Applying generalized linear models.* New York: Springer-Verlag, 1997. 257p.

MANLY, B. F. J. *Randomization,* bootstrap *and Monte Carlo methods in biology.* 2nd. ed. London: Chapman-Hall, 1998. 399p.

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

PAN, W. Approximate confidence intervals for one proportion and two proportions. *Comput. Stat. Data Anal.*, Amsterdam, v.40, n.1, p.143-157, 2002.

THORPE, D. P.; HOLLAND, B. Some multiple comparison procedures for variance from non-normal populations. *Comput. Stat. Data Anal.*, Amsterdam, v.35, p.171-199, 2000.

Recebido em 10.03.2006.

Aprovado após revisão em 15.05.2006.