CLEIDE SILVEIRA BRASIL PEIXOTO

PROPOSIÇÃO DE ESTIMADORES PARA FUNÇÕES LINEARES BINOMIAIS COM AMOSTRAS INFLACIONADAS DE ZERO

LAVRAS – MG

2013

CLEIDE SILVEIRA BRASIL PEIXOTO

PROPOSIÇÃO DE ESTIMADORES PARA FUNÇÕES LINEARES BINOMIAIS COM AMOSTRAS INFLACIONADAS DE ZERO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador

Dr. Marcelo Angelo Cirillo

LAVRAS – MG

2013

Ficha Catalográfica Elaborada pela Coordenadoria de Produtos e Serviços da Biblioteca Universitária da UFLA

Peixoto, Cleide Silveira Brasil.
Proposição de estimadores para funções lineares binomiais com amostras inflacionadas de zero / Cleide Silveira Brasil Peixoto. – Lavras : UFLA, 2013.
77 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2013.
Orientador: Marcelo Angelo Cirillo.
Bibliografia.

1. Funções lineares. 2. Estimador ZIB. 3. Wald. 4. Wilson. 5. Zeros. I. Universidade Federal de Lavras. II. Título.

CDD – 519.544

CLEIDE SILVEIRA BRASIL PEIXOTO

PROPOSIÇÃO DE ESTIMADORES PARA FUNÇÕES LINEARES BINOMIAIS COM AMOSTRAS INFLACIONADAS DE ZERO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em, 15 de julho de 2013

Dr. Ricardo Tavares	UFOP
Dr. Fernando Luiz Pereira de Oliveira	UFOP
Dr. Márcio Balestre	UFLA
Dr. Renato Ribeiro de Lima	UFLA

Dr. Marcelo Angelo Cirillo
Orientador

LAVRAS – MG

2013

Ao Senhor Deus,
por ter me sustentado dia após dia;
À minha família;
Aos meus filhos Eduardo e Rachel;
Ao meu esposo Josaphat.

DEDICO

AGRADECIMENTOS

A Deus, Senhor da minha vida;

À Universidade Federal de Lavras (UFLA), ao Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela oportunidade de concretizar esse trabalho.

Ao professor Dr. Marcelo Angelo Cirillo, pela orientação, competência, apoio, incentivo e amizade.

Aos professores do Departamento de Ciências Exatas da Universidade Federal de Lavras (DEX-UFLA), pela competência, ensinamentos e comprometimento.

Ao professor Dr. Lurimar Smera Batista, do Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), pelo apoio e comprometimento.

Aos colegas e amigos da Pós-Graduação: Ângela, Azly, Edmary, Jailson, Jaime, José Otaviano (*in memorian*), Luiz Vásquez, Marcio, Nelson, Norma, Regilson, Tânia, Isabel e Walter pelo companheirismo e convivência durante todo o curso, principalmente nos momentos de estudo.

Em especial, ao amigo Augusto, pelas horas de estudo em que compartilhou seus conhecimentos com competência e dedicação.

Aos meus pais; Josias *(in memoriam*) e Rosália *(in memoriam*) e minha tia Eulina *(in memoriam*), pelo amor, compreensão e educação; aos meus irmãos; Roberto, Leny, Eloisa e Alberto; aos meus sogros; Josaphat e Jair; a minha tia Raquel e minha cunhada Ana, pelo apoio, incentivo e carinho.

Aos meus filhos; Eduardo e Rachel, pelo afeto, confiança, incentivo e momentos descontraídos e alegres.

Ao meu esposo, amigo e companheiro, Josaphat, pelo amor, compreensão, apoio, estímulo e confiança em todos esses anos.

Meus sinceros agradecimentos a todos que, de alguma forma, contribuíram para elaboração desse trabalho.

RESUMO

Com o propósito de obter inferências estatísticas nas mais diversas áreas de pesquisa, objetivou-se, neste trabalho propor um aprimoramento, por meio da incorporação de estimadores robustos, que substituam o estimador de máxima verossimilhança, em amostras binomiais com excessos de zeros, nos métodos Wald e Wilson, usualmente utilizados para estimação de funções lineares binomiais. Para a validação dos métodos, utilizou-se simulação Monte Carlo, combinando diferentes cenários envolvendo valores paramétricos, números de ensaios de Bernoulli, tamanho amostral e diferentes porcentagens de valores nulos contidos na amostra. Recomenda-se a utilização do método de Wald nas situações cujas estimativas das proporções binomiais robustas a excesso de zero, que maximizam a variância das funções lineares binomiais, utilizando a componente sistemática ρ_1 e coeficientes ortogonais. Em se tratando do método de Wilson, por apresentar probabilidades de cobertura incoerentes ao nível nominal de confiança, não é recomendável seu uso.

Palavras-chave: Funções Lineares. Estimador ZIB. Wald. Wilson. Zeros.

ABSTRACT

With the purpose of obtaining statistical inferences in different research areas, this paper aims to propose an improvement by incorporating robust estimators, which substitute the maximum likelihood estimator of binomial samples with excess zeros on the Wald and Wilson methods, usually used to estimate binomial linear functions. In order to validate the methods, we used the Monte Carlo simulation combining different scenarios involving parametric values, the number of Bernoulli trials, sample size and different percentages of null values in the sample. We recommend the use of the Wald method in situations in which the estimates of the robust binomial proportions present excess of zeros, which maximize the variance of the binomial linear functions using the $\rho_1$sistematic component and orthogonal coefficients. Regarding the Wilson method, we do not recommend its use due to presenting the possibility of incoherent coverage at the nominal level of confidence.

Keywords: Linear functions. ZIB estimator. Wald. Wilson. Zeros.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A distribuição binomial tem sido amplamente utilizada na análise de dados de proporção, e isso tem despertado interesse na comunidade científica quanto à pesquisa estatística, no que tange à formulação e melhoramento de procedimentos inferenciais que proporcionam resultados estatísticos mais acurados e/ou precisos. Entre as metodologias estatísticas propostas na literatura para análise de dados na escala de proporção, pode ser mencionada a construção de intervalos de confiança para funções lineares de proporções binomiais, utilizando-se os métodos de Wald e Wald ajustado (PRICE; BONETT, 2004) e Wilson (TEBBS; ROTHS, 2008).

Uma dificuldade encontrada ao utilizar a distribuição binomial é o fato de não se tratar de uma distribuição contínua e os estudos citados tomam por base os pressupostos de uma distribuição normal. Isso gera problemas de natureza estatística, devido à aproximação assintótica e imprecisão na construção de intervalos de confiança assimétricos, aplicados em amostras binomiais.

Outra dificuldade na aplicação de métodos clássicos conhecidos na literatura como Wald e Wilson para a estimação intervalar de duas ou mais proporções binomiais, deve-se ao fato de que tais métodos não foram elaborados para situações em que as amostras apresentam excesso de observações nulas. Da mesma forma, estimativas intervalares para funções lineares são “frágeis” para amostras com essas características.

Tendo por base o problema de amostras que apresentam excesso de observações nulas, este projeto tem por justificativa propor uma modificação nos métodos intervalares de Wald e Wilson, incorporando estimadores robustos a amostras binomiais inflacionadas de zeros, utilizadas na composição de funções lineares binomiais. Copas (1988) advertiu que, mesmo assumindo o modelo

adequado, alguns zeros podem ser considerados *outliers* e diferentes métodos de estimação são sensíveis a essa discrepância. Nesse caso é preciso encontrar métodos robustos de estimação, que sejam capazes de considerar a presença de dados discrepantes e fornecer uma estimativa coerente do parâmetro que se deseja obter.

Contudo, para que estes procedimentos sejam viáveis de serem aplicados, torna-se necessário fazer uma avaliação, utilizando-se simulações com o método Monte Carlo, de suas principais propriedades: probabilidade de cobertura e amplitude média intervalar. Cirillo, Ferreira e Safádi (2009) mencionam que métodos de computação intensiva são convenientes de serem aplicados, para que a probabilidade de cobertura seja melhorada em relação à aproximação do verdadeiro nível de confiança, nas inferências realizadas específicas a famílias binomiais.

Assim, conhecendo-se estas propriedades, tais métodos podem ser empregados em situações reais específicas de experimentos com respostas referentes a dados de proporção. Em potencial, podemos exemplificar, proporção de contaminação de fungos filamentosos em diferentes amostras de café, contagem de insetos vivos e/ou mortos (dependendo do interesse) em parcelas experimentais com diferentes densidades populacionais, proporção de plantas que germinaram com diferentes dosagens de promotores de crescimento, etc.

Em virtude do que foi mencionado, objetivou-se,neste trabalho, propor soluções para realização de inferências em amostras binomiais contaminadas de observações nulas, através de um aprimoramento pelo uso de estimadores robustos, nos métodos de Wald e Wilson utilizados na construção de famílias binomiais.

2 REFERENCIAL TEÓRICO

2.1 Modelo binomial inflacionado de zeros

Na maioria das aplicações é factível se encontrem amostras com valores observados iguais a zero. Frente a essa situação, dada as grandes quantidades denomina-se a amostra como inflacionada de zeros e para acomodar esse efeito surge então uma classe especial de distribuição estatística. Em particular para distribuições discretas, Kemp e Kemp (1988) afirmam que o estimador de máxima verossimilhança clássico de distribuições discretas tem muitas características desejáveis, mas no caso de uma distribuição binomial, deve-se atentar para casos que não sigam exatamente essa distribuição, como por exemplo, a ocorrência de superdispersão nos dados. Mesmo quando um método leva a estimadores explícitos para determinadas distribuições, esses podem não ser desejáveis.

Estudos teóricos comprovam que um método pode ser eficiente em determinada região do espaço paramétrico e ter comportamento contrário em outras regiões.

De acordo com Ruckstuhl e Welsh (2001), em casos de dados de proporção, um modelo a ser utilizado é o modelo binomial inflacionado de zeros (ZIB). Esse modelo pode ser representado como uma distribuição de mistura de dois componentes: uma componente supõe que a ocorrência de zeros γ, enquanto outra componente é caracterizada por uma distribuição binomial, com probabilidade $(1-\gamma)$.

Assim, o modelo ZIB tem função massa de probabilidade definida por

$$P(Y=y)=\begin{cases}\gamma+(1-\gamma)(1-\pi)^{m}, & \text{se} \quad y=0\\ (1-\gamma)\binom{m}{y}\pi^{y}(1-\pi)^{m-y}, & \text{se } y=1,2,...,m,\end{cases} \tag{1}$$

com $E(Y)=(1-\gamma)m\pi$ e $Var(Y)=[(1-\gamma)m\pi][(1-\pi)(1-\gamma m)]$, em que o parâmetro γ, conforme dito anteriormente, representa a proporção de ocorrência de zeros, existindo a restrição $0\le\gamma<1$. No caso de $\gamma>0$, observa-se uma inflação na $Var(Y)$, ocorrendo assim superdispersão, devido ao excesso de zeros. Por outro lado, se $\gamma=0$, então $Var(Y)=m\pi(1-\pi)$, será exatamente a mesma de um modelo binomial.

Existem diversas soluções na literatura para estimação robusta em dados discretos, como é o modelo binomial. Ruckstuhl e Welsh (2001) discutiram o desempenho dos estimadores de mínima disparidade e estimadores E, que serão apresentados a seguir para melhor compreensão deste trabalho.

2.2 Estimadores de mínima disparidade

Devido aos questionamentos de que alguns métodos de estimação poderiam ser robustos, mas não eficientes, como por exemplo, o estimador de mínima distância de Hellinger (SIMPSON, 1987), surgiu a motivação para o estudo de outros métodos de distâncias mínimas.

No método de Hellinger são atribuídos pesos menores às observações mais discrepantes no modelo, tornando-o equivalente ao estimador de máxima verossimilhança em modelos conhecidos. Em contrapartida, Lindsay (1994) mostrou que esse estimador possui uma curva de influência que não contempla algumas situações de robustez enganosas.

Com isso, o autor propôs estudar outros métodos de distâncias mínimas para investigar esses fenômenos, utilizando a chamada “função de ajuste residual”, cuja configuração determina completamente a robustez e a eficiência de determinado estimador. A partir daí, Lindsay (1994) apresentou o estimador de mínima disparidade, que é outra classe de estimadores que representa bem um modelo binomial, sob uma ou mais condições adversas.

2.2.1 Estimadores-E

Os estimadores-E se enquadram na classe de estimadores de mínima disparidade e têm por idéia central o desenvolvimento de uma metodologia capaz de manipular uma “contaminação grosseira”, no sentido de não contemplar observações outliers, na obtenção das estimativas dos parâmetros. Em outras palavras, entende-se a obtenção de estimativas robustas. Nesse contexto, uma alternativa é modificar a log verossimilhança para reduzir o efeito das observações na cauda das distribuições (RUCKSTUHL; WELSH, 2001). Contudo, também podem haver efeitos de uma contaminação não restrita apenas à calda e, nesses casos, tende a deflacionar ou inflacionar várias classes de modo arbitrário. Isso sugere que é mais plausível trabalhar na frequência de escala do que na escala de observações individuais.

O modo de decidir que classes controlar é relacionar as frequências relativas com a função de probabilidade do modelo binomial, um conhecimento que nos leva à direção da estimação de mínima distância na escala de frequência relativa (RUCKSTUHL; WELSH, 2001).

A escolha da função de disparidade é arbitrária e será confirmada pelas propriedades e desempenho do estimador. Faz sentido trabalhar com a disparidade que produza o estimador de máxima verossimilhança, quando sob influência do modelo binomial.

A disparidade da verossimilhança $H(\pi, f_n)$ é dada por:

$$H(\pi, f_n) = \sum_{y=0}^{m} \rho(x) p_\pi(y), \tag{2}$$

em que $x = \frac{f_n(y)}{p_\pi(y)}$ e $f_n(y) = n^{-1} \sum_{y=0}^{m} I(Y_i = y)$, *y = 0, ..., m,* é computada como a proporção das observações iguais a *y* em uma amostra de tamanho *n*, sendo $p_\pi(y)$ a probabilidade de ocorrência de y considerando a proporção π e $\rho(x) = x \ln(x)$.

A disparidade da verossimilhança é minimizada pelo estimador de verossimilhança $(\hat{\pi}_{emv})$ para o modelo binomial é mencionada em Ruckstuhl e Welsh (2001), por meio de uma transformação da função $\rho(x)$ em uma função linear $\rho_1(x)$, que tem o efeito de reduzir as taxas em que $\rho(x)$ tende ao infinito quando x é grande, ou a zero quando x é pequeno. Impondo que a nova função $\rho_1(x)$ tenha derivação contínua, tem-se:

$$\rho_1(x) = \begin{cases} [\ln(c_1) + 1]x - c_1, & \text{se } x < c_1; \\ x \ln(x), & \text{se } c_1 \le x \le c_2; \\ [\ln(c_2) + 1]x - c_2, & \text{se } x > c_2; \end{cases} \tag{3}$$

em que $\rho_1(x)$ é garantida sobre determinados valores assumidos pelas constantes c_1 e c_2. Por recomendação de Ruckstuhl e Welsh (2001), utiliza-se $c_1 < c_2 = 1$. O estimador de $\hat{\pi}$ que minimiza $H(\pi, f_n)$ é dado por:

$$\hat{\pi} = \arg\min_{\pi} \{\pi, f\}. \quad (4)$$

2.2.2 Estimadores para a proporção do modelo binomial robusto a excesso de zeros utilizando a função de mínima disparidade

Mantendo a mesma filosofia dos estimadores-E, Silva (2009) propôs um estimador para a proporção de sucessos em uma amostra proveniente de uma população binomial inflacionada de zeros, dado o fato de que o estimador de máxima verossimilhança, quando aplicado nessa situação, apresenta "erros grosseiros" já mencionados anteriormente.

Em contrapartida, os estimadores robustos pertencentes à classe dos estimadores-E apresentam uma sensibilidade em relação às constantes de afinidade assumidas c_1 e c_2. Além do mais, torna-se inviável, no que tange à sua praticidade de aplicação em muitas situações, uma vez que o argumento desse estimador, que minimiza a função de disparidade,requer que todo o domínio paramétrico $(0 \leq \pi \leq 1)$ seja "percorrido".

Com essa motivação Silva (2009) propôs o estimador $\hat{\pi}_{zib}$ em duas abordagens distintas. Uma primeira abordagem, considera $\rho_1(x)$, definido em (3), e define $\hat{\pi}_{zib}$ como

$$\hat{\pi}_{zib} = \sum_{y=0}^{m} \rho_1(x)\hat{\pi}_{emv}. \quad (5)$$

Com a finalidade de diminuir a rapidez com que a função $\rho(x) = x\ln(x)$ tende ao infinito com o crescimento de x, na proposta do

estimador $\hat{\pi}_{zib}$, segundo Silva (2009), a função $\rho_1(x)$ é modificada com o propósito de diminuir essa taxa de crescimento.

Note que em (5) considera-se a estimativa de máxima verossimilhança $\hat{\pi}_{emv}$ entretanto, considera-se uma componente descrita pela função $\rho_1(x)$ definido em (3) que acomoda o efeito das observações presentes na cauda da distribuição. Com esse propósito, segue uma definição da função Holder contínua definida por (6)

$$\rho_2(x)=\begin{cases}\left\{[\ln(c_1)+\dfrac{(1-u)\ln(c_1)+1}{u}]c_1^{1-u}\right\}x^u-[(1-u)\ln(c_1)+1]\dfrac{c_1}{u}, & x<c_1\\ x\ln(x), \text{ se } c_1\le x\le c_2 & \\ \left\{[\ln(c_2)+\dfrac{(1-u)\ln(c_2)+1}{u}]c_2^{1-u}\right\}x^u-[(1-u)\ln(c_2)+1]\dfrac{c_2}{u}, & x>c_2\end{cases} \quad (6)$$

Em se tratando da definição (5) além das constantes de afinidade c_1 e c_2 tem-se uma nova constante definida em u cuja finalidade é reduzir a velocidade na qual a função ρ(x)=xlnx $\to \infty$, quando x $\to \infty$, uma vez que essa função é aprimorada conforme descrita em (3) e (6). Naturalmente, para $u=1$, $\rho_2(x)$ é equivalente a $\rho_1(x)$.

Ressalta-se que a especificação destes valores estão relacionados a acurácia das estimativas, portanto, torna-se necessário determinar os valores adequados utilizando-se um procedimento de simulação em caráter exploratório, no sentido de averiguar quais valores proporcionam melhor acurácia.

Especificada essas constantes, procede-se com a incorporaçao das função (6) em (5). Convém ressaltar que as funções $\rho_2(x)$ e $\rho_1(x)$ são

contínuas e diferenciáveis em R^{+}. Logo o estimador $\hat{\pi}_{zib}$ na segunda abordagem é dado por

$$\hat{\pi}_{zib} = \sum_{y=0}^{m} \rho_2(x)\hat{\pi}_{emv}. \tag{7}$$

Para melhores esclarecimentos, segue um exemplo (SILVA, 2009), considerando a função $\rho_2(x)$ em uma amostra de tamanho n, na qual cada unidade amostral foi considerada independente e identicamente distribuida por uma binomial (m, π). Dessa forma, uma amostra de tamanho 20, foi simulada pelo método Monte Carlo, assumindo um modelo binomial inflacionado de zeros, com 100 ensaios de Bernoulli, probabilidade de sucessos $\pi = 0,3$ e proporção de valores nulos $\gamma = 0,2$.

Tabela 1 Valores assumidos para ilustrar a estimação de π utilizando $\hat{\pi}_{zib}$

Unidades amostrais
25, 30, 31, 25, 24, 27, 0, 26, 31, 32, 26, 28, 21, 27, 29, 30, 0, 0, 28, 32

Importante ressaltar que pelo fato de utilizar a função $\rho_2(x)$, fixou-se $c_1 = 0,1$ e $c_2 = 1$, o valor de $u = 0,26$ foi pesquisado, através de uma rotina computacional descrita no Anexo A, para obter uma estimativa de π_{zib} próxima do valor de $\pi = 0,3$. A seguir são apresentados os passos para obtenção do estimador $\hat{\pi}_{zib}$.

(1) Obtenção das estimativas de máxima verossimilhança

$$\hat{\pi}_{emv} = \frac{1}{m}\sum_{i=1}^{n} y_i \, f_n(y) \tag{8}$$

em que,

$$f_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y = y_i)\,,\ y = 0,...,m\,, \tag{9}$$

$$f_n(0) = \frac{I(Y=0) + I(Y=0) + ... + I(Y=0)}{20} = \frac{3}{20} = 0{,}15,$$
$$\vdots$$
$$f_n(32) = \frac{I(Y=32) + I(Y=32) + ... + I(Y=32)}{20} = \frac{2}{20} = 0{,}10.$$

Vale lembrar que o mesmo procedimento deve ser feito para cada y, observando que, para o caso em questão, tem-se apenas valores não nulos quando y for igual às unidades amostrais. Com isso, pode-se calcular o valor de $\hat{\pi}_{emv}$.

$$\hat{\pi}_{emv} = \frac{0\,f_n(0) + 1\,f_n(1) + ... + 100\,f_n(100)}{100} = 0,236\,.$$

Com base nessa estimativa foi consultada a tabela de estimadores para $\gamma = 0,2$ (SILVA, 2009), para se obter um valor próximo ao encontrado.

(2) Cálculo das probabilidades considerando-se a estimativa de máxima verossimilhança dada por 0,236.

$$p\hat{\pi}_{emv}(0) = \binom{100}{0} 0,236^{0}(1-0,236)^{100} = 2,04\times10^{-12}$$

$$p\hat{\pi}_{emv}(1) = \binom{100}{1} 0,236^{1}(1-0,236)^{99} = 6,30\times10^{-11}$$

$$p\hat{\pi}_{emv}(2) = \binom{100}{2} 0,236^{2}(1-0,236)^{98} = 9,63\times10^{-10}$$

$$\vdots$$

$$p\hat{\pi}_{emv}(100) = \binom{100}{100} 0,236^{100}(1-0,236)^{0} = 1,96\times10^{-63}$$

(3) Cálculo do estimador $\hat{\pi}_{zib}$ utilizando $\rho_2(x)$ conforme definições dadas em (6) e (7) .

Na tabela 2, a seguir, estão relacionados todos os valores de $p\hat{\pi}_{emv}$, x e $\rho_2(x)$, para cada um dos 100 ensaios de Bernoulli, envolvidos no cálculo de $\hat{\pi}_{zib}$. Para efeito de ilustração, uma vez que cada ensaio bernoulli é considerado, apresentaram -se os resultados para os primeiros valores de m, alguns valores intermediários, finalizando com os valores finais, m=100.

Tabela 2 Valores calculados do $\hat{\pi}_{zib}$ sob $\rho_2(x)$

m	$p\hat{\pi}_{emv}$	x	$\rho_2(x)$	$\rho_2(x)\,p\hat{\pi}_{emv}$
0	$2{,}0386\text{x}10^{-12}$	$7{,}3579\text{x}10^{10}$	2568,8177	$5{,}2368\text{x}10^{09}$
1	$6{,}2970\text{x}10^{-11}$	0	0,2707	$1{,}7049\text{x}10^{11}$
2	$9{,}6280\text{x}10^{-10}$	0	0,2707	$2{,}6069\ \text{x}10^{10}$
3	$9{,}7163\text{x}10^{-09}$	0	0,2707	$2{,}6305\text{x}10^{09}$
4	$7{,}2783\text{x}10^{-08}$	0	0,2707	$1{,}9705\text{x}10^{08}$
5	$4{,}3167\text{x}10^{-07}$	0	0,2707	$1{,}1686\ \text{x}10^{07}$
⋮	⋮	⋮	⋮	⋮
21	0,0804	0,6214	-0,2956	-0,0237
⋮	⋮	⋮	⋮	⋮
100	$1{,}9552\text{x}10^{-63}$	0		

$$\hat{\pi}_{zib} = \sum_{y=0}^{m} \rho_2(x)\hat{\pi}_{emv} = 0{,}2900$$

Em se tratando do $\hat{\pi}_{zib}$ considerando $\rho_2(x)$ e $u=1$ o procedimento é análogo. Entretanto, o valor de c_1 deverá ser escolhido de forma apropriada através de um estudo exploratório, fixando na rotina computacional (Anexo A) um conjunto de valores para c_1.

Comparando a estimativa de $\hat{\pi}_{zib}$, com o valor paramétrico $(\pi = 0{,}3)$, nota-se que o valor de u utilizado resultou em uma estimativa acurada, pois utilizando o critério estabelecido de $\left|\hat{\pi}_{zib} - \hat{\pi}_{emv}\right| \leq t$, onde arbitrariamente t foi especificado em 0,15, pode-se verificar que o valor de u é adequado para realização dessa inferência.

Reportamos às considerações de que o a estimativa de máxima verossimilhança, com incorporação de amostras binomiais inflacionadas de zeros foi comparada com a estimativa de máxima verossimilhança através de margem de erro definida arbitrariamente, assim validou-se a acurácia dessa da estimativa obtida por (7) em relação a (8).

2.3 Funções lineares de proporções binomiais

Seja $Y_i\left(i=1,2,...,q\right)$ uma variável aleatória binomial independente com parâmetros n_i e π_i, em que i representa a população binomial e $\hat{\pi}_i = Y_i/n_i$ a proporção da amostra e seja $z_{\alpha/2}$ o quantil superior da distribuição normal padrão.

O parâmetro que representa uma função linear de proporções, segundo mencionam Price e Bonett (2004), pode ser definido por $\psi = \sum_{i=1}^{q} \delta_i \pi_i$, em que δ_i é um coeficiente conhecido e π_i é o valor paramétrico referente a *i*-ésima população binomial. Tendo por base essa definição, os métodos de estimação intervalar são descritos a seguir.

2.3.1 Intervalo de confiança para os métodos Wald e Wald ajustado

Um intervalo de confiança de aproximadamente $100(1-\alpha)\%$ Wald para ψ é apresentado por Price e Bonett (2004) por

$$\sum_{i=1}^{q} \delta_i \hat{\pi}_i \pm z_{\alpha/2} \sqrt{\sum_{i=1}^{q} \frac{\delta_i^2 \hat{\pi}_i \left(1-\hat{\pi}_i\right)}{n_i}}, \qquad (10)$$

sendo n_i o tamanho amostral referente a i-ésima população binomial, $\hat{\pi}_i = Y_i/n_i$ e δ_i é um coeficiente conhecido.

Dada a inferência dos intervalos de confiança para a diferença de duas proporções binomiais, especialmente ao considerar amostras pequenas, em que

as probabilidades de cobertura das estimativas intervalares não são condizentes com o nível de confiança desejado, confirmado por Price e Bonett (2004) em estudos de simulação, considerando diferentes cenários de valores paramétricos, nos quais os resultados obtidos para essas probabilidades foram dadas entre 50% e 80%. Em função dessa deficiência, surge então o intervalo de confiança de Wald ajustado, proposto por Agresti e Coull (1998), inicialmente para comparação de duas proporções binomiais, de tal forma que, para cada proporção os autores propuseram a adição de quatro pseudo observações, sendo dois sucessos e dois fracassos.

Posteriormente, Agresti e Caffo (2000) propuseram a expansão do método de Wald ajustado para diferença entre duas proporções binomiais. Uma generalização para q proporções binomais, é representada pela expressão (11).

$$\sum_{i=1}^{q} \delta_i \hat{\pi}_i \pm z_{\alpha/2} \sqrt{\sum_{i=1}^{q} \delta_i^2 \hat{\pi}_i \left(1-\hat{\pi}_i\right) \Big/ \left(n_i + 4/k\right)}\,, \qquad (11)$$

em que $\hat{\pi}_i = \left(Y_i + 2/k\right)/\left(n_i + 4/k\right)$ e k é o número de coeficientes δ_i diferentes de zero em ψ. Nota-se que (10) se reduz aos casos estudados por Agresti e Coull (1998), quando $q = 1$ e $\delta = 1$, e por Agresti e Caffo (2000), quando $q = 2$, $\delta_1 = 1$ e $\delta_2 = -1$.

Price e Bonnet (2004) ressaltaram que a generalização dos métodos de Wald e Wald ajustado, através da aplicação de funções lineares, apresenta como vantagem situações de estudos de comparações que envolvam contrastes complexos, componentes de tendência, efeitos principais e efeitos de interações. Ainda, de acordo com esses autores, em geral o método Wald ajustado apresenta melhores resultados de probabilidades de cobertura que o método Wald.

2.3.2 Intervalo de confiança do método de Wilson

Tendo por base uma inversão nos valores dos erros padrões, na obtenção dos limites inferiores e superiores obtidos via uma aproximação normal, Wilson (1927) propôs uma classe de métodos de estimação, denominada por MOVER. Tal sigla provém de "Method of Variance Estimates Recovery" conforme descreve Zou, Huang e Zhan (2009). As expressões analíticas para a obtenção de um intervalo pertencente a essa classe, doravante menciondo nesse trabalho por método de Wilson, é descrito a seguir.

Supondo que deseja-se construir a $100(1-\alpha)\%$ de aproximação o intervalo de confiança para $\pi_1 + \pi_2$, em que as estimativas de π_1 e π_2 são consideradas independentes. Pelo teorema central do limite (TCL), o limite inferior (*A*) é dado por

$$A = \hat{\pi}_1 + \hat{\pi}_2 - z_{\alpha/2}\sqrt{\hat{\text{var}}(\hat{\pi}_1) + \hat{\text{var}}(\hat{\pi}_2)}. \qquad (12)$$

Supondo que os $100(1-\alpha)\%$ intervalos de confiança (a_i, b_i) para cada parâmetro $\hat{\pi}_i$, $i = 1, 2$ estão disponíveis, nota-se que não há necessidade de especificar as abordagens adotadas para obter (a_i, b_i). Entre todos os valores dos parâmetros plausíveis de $\hat{\pi}_1$ fornecidos pelo (a_1, b_1) e que de $\hat{\pi}_2$ fornecidos pelo (a_2, b_2), $a_1 + a_2$ é geralmente mais próximo de A que $\hat{\pi}_1 + \hat{\pi}_2$. Como resultado, para A, pode-se estimar $\hat{\text{var}}(\hat{\pi}_1)$ em $\hat{\pi}_1 = a_1$ e $\hat{\text{var}}(\hat{\pi}_2)$ em $\hat{\pi}_2 = a_2$.

Além disso, pode-se reparar as estimativas da variância estimadas de $\hat{\pi}_i\left(a_i, b_i\right)$, $i=1, 2$. Pelo TCL e usando $z_{\alpha/2}$, o quantil da distribuição normal padrão, tem-se:

$$a_i = \hat{\pi}_i - z_{\alpha/2}\sqrt{\hat{\text{var}}\left(\hat{\pi}_i\right)}, \tag{13}$$

o que fornece uma estimativa de variância em $\hat{\pi}_i = a_i$ como $\hat{\text{var}}_a\left(\hat{\pi}_i\right) = \left(\hat{\pi}_i - a_i\right)^2 \big/ z_{\alpha/2}^2$ e

$$b_i = \hat{\pi}_i + z_{\alpha/2}\sqrt{\hat{\text{var}}\left(\hat{\pi}_i\right)}, \tag{14}$$

o que fornece uma estimativa de variância em $\hat{\pi}_i = b_i$ como $\hat{\text{var}}_b\left(\hat{\pi}_i\right) = \left(b_i - \hat{\pi}_i\right)^2 \big/ z_{\alpha/2}^2$.

Note que as variâncias estimadas $\hat{\text{var}}_a\left(\hat{\pi}_i\right)$ e $\hat{\text{var}}_b\left(\hat{\pi}_i\right)$ são diferentes, exceto quando o intervalo $\left(a_i, b_i\right)$ é simétrico em relação a $\hat{\pi}_i$.

Conectando o estimador de variância obtida na Eq. (12) resulta em

$$\begin{aligned} A &= \hat{\pi}_1 + \hat{\pi}_2 - z_{\alpha/2}\sqrt{\hat{\text{var}}\left(\hat{\pi}_1\right) + \hat{\text{var}}\left(\hat{\pi}_2\right)} \\ &= \hat{\pi}_1 + \hat{\pi}_2 - z_{\alpha/2}\sqrt{\left(\hat{\pi}_1 - a_1\right)^2 \big/ z_{\alpha/2}^2 + \left(\hat{\pi}_2 - a_2\right)^2 \big/ z_{\alpha/2}^2} \\ &= \hat{\pi}_1 + \hat{\pi}_2 - \sqrt{\left(\hat{\pi}_1 - a_1\right)^2 + \left(\hat{\pi}_2 - a_2\right)^2}. \end{aligned} \tag{15}$$

Passos análogos com a noção de que $\hat{\pi}_1 + \hat{\pi}_2$ está na vizinhança de B produzem o limite superior de B como

$$B = \hat{\pi}_1 + \hat{\pi}_2 + \sqrt{(b_1 - \hat{\pi}_1)^2 + (b_2 - \hat{\pi}_2)^2} \,. \tag{16}$$

Reescrevendo $\hat{\pi}_1 - \hat{\pi}_2$ como $\hat{\pi}_1 + (-\hat{\pi}_2)$ e notando que os limites de confiança para $-\hat{\pi}_2$ são dadas por $(-a_2, -b_2)$, obtemos limites de confiança para $\hat{\pi}_1 - \hat{\pi}_2$ como

$$A = \hat{\pi}_1 - \hat{\pi}_2 - \sqrt{(\hat{\pi}_1 - a_1)^2 + (b_2 - \hat{\pi}_2)^2} \tag{17}$$

e

$$B = \hat{\pi}_1 - \hat{\pi}_2 + \sqrt{(b_1 - \hat{\pi}_1)^2 + (\hat{\pi}_2 - a_2)^2} \,. \tag{18}$$

Zou e Donner (2008) apresentaram uma justificativa analítica para a aplicabilidade geral desse intervalo de confiança, reescrevendo *A* e *B* em relação a $\hat{\pi}_1 + \hat{\pi}_2$ e a $\hat{\pi}_1 - \hat{\pi}_2$ como $\delta_1\hat{\pi}_1 + \delta_2\hat{\pi}_2$, onde δ_1 e δ_2 são constantes, pode-se reescrever os intervalos como

$$A = \delta_1\hat{\pi}_1 + \delta_2\hat{\pi}_2 - \sqrt{[\delta_1\hat{\pi}_1 - \min\{\delta_1 a_1, \delta_1 b_1\}]^2 + [\delta_2\hat{\pi}_2 - \min\{\delta_2 a_2, \delta_2 b_2\}]^2} \tag{19}$$

e

$$B = \delta_1\hat{\pi}_1 + \delta_2\hat{\pi}_2 + \sqrt{\left[\delta_1\hat{\pi}_1 - \max\{\delta_1 a_1, \delta_1 b_1\}\right]^2 + \left[\delta_2\hat{\pi}_2 - \max\{\delta_2 a_2, \delta_2 b_2\}\right]^2}. \quad (20)$$

Supondo $100(1-\alpha)\%$ intervalo de confiança para $\sum_{i=1}^{q}\delta_i\pi_i$, em que $k > 2$, uma aplicação de resultados de indução matemática em

$$\begin{cases} A = \sum_{i=1}^{q}\delta_i\hat{\pi}_i - \sqrt{\sum_{i=1}^{q}\left[\delta_i\hat{\pi}_i - \min\{\delta_i a_i, \delta_i b_i\}\right]^2} \\ B = \sum_{i=1}^{q}\delta_i\hat{\pi}_i + \sqrt{\sum_{i=1}^{q}\left[\delta_i\hat{\pi}_i - \max\{\delta_i a_i, \delta_i b_i\}\right]^2} \end{cases}. \quad (21)$$

Pode-se agora aplicar o intervalo de confiança em (21) para as funções lineares de proporções binomiais. Tem-se, pelo menos, três intervalos de uma proporção única, ou seja, Wald, Wald ajustado e Wilson. Logo tem-se três procedimentos para as funções lineares de proporções binomiais.

Especificamente, seja $Y_i\,(i = 1, 2, ..., q)$ uma variável binomial independente com parâmetros (n_i, π_i), $\hat{\pi}_i = Y_i/n_i$ são as estimativas da amostra para π_i. Uma função linear de proporções binomiais pode ser definida como $\sum_{i=1}^{q}\delta_i\pi_i$, em que os δ_i são constantes conhecidas. Utilizando-se as equações definidas em (20), os limites inferior (a_i) e superior (b_i) do intervalo de confiança de $100(1-\alpha)\%$ Wald podem ser obtidos, fixando $\hat{\pi}_i = Y_i/n_i$, por

$$a_i = \hat{\pi}_i - z_{\alpha/2}\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)/n_i} \quad (22)$$

e

$$b_i = \hat{\pi}_i + z_{\alpha/2}\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)/n_i}\ . \tag{23}$$

O intervalo de Wilson para $\sum_{i=1}^{q}\delta_i\pi_i$ pode ser obtido fixando $\hat{\pi}_i = Y_i/n_i$, obtendo-se

$$\left[\frac{\left(\hat{\pi}_i + z_{\alpha/2}^2/\left(2n_i\right) \mp z_{\alpha/2}\sqrt{\left[\hat{\pi}_i\left(1-\hat{\pi}_i\right) + z_{\alpha/2}^2/\left(4n_i\right)\right]/n_i}\right)}{\left(1+z_{\alpha/2}^2/n_i\right)}\right]. \tag{24}$$

O intervalo de Wald ajustado para $\sum_{i=1}^{q}\delta_i\pi_i$ (PRICE; BONETT, 2004) pode ser obtido, fixando $\hat{\pi}_i = \left(Y_i + 2/k\right)/\left(n_i + 4/k\right)$ (*k* é o número de coeficientes não nulos de δ_i), por

$$a_i = \hat{\pi}_i - z_{\alpha/2}\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)/n_i} \tag{25}$$

e

$$b_i = \hat{\pi}_i + z_{\alpha/2}\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)/n_i}\ . \tag{26}$$

Note-se que o método de Wald ajustado para uma proporção é uma aproximação do método de escore Wilson para o intervalo de 95%.

3 METODOLOGIA

Em função dos objetivos propostos, a metodologia utilizada para obtenção dos resultados segue estruturada nos seguintes tópicos: 3.1 – Simulação de amostras binomiais inflacionadas de zeros; 3.2 – Estimação das probabilidades de sucessos (π) do modelo ZIB; 3.3 – Definição e estimação das funções lineares de proporções binomiais.

3.1 Simulação de amostras binomiais inflacionadas de zeros

Utilizando-se técnicas de simulação Monte Carlo, as amostras binomiais inflacionadas de zeros foram geradas considerando o modelo ZIB, Ruckstuhl e Welsh (2001), caracterizado pela mistura de duas componentes de tal forma que uma componente supõe que a ocorrência de zeros seja dada por uma probabilidade γ, enquanto que a outra representa uma distribuição binomial com probabilidade $(1 - \gamma)$. Dessa forma, o modelo ZIB é dado pela expressão a seguir (1), conforme descrito na Seção 2.1.

Assim, os valores paramétricos assumidos no processo de simulação foram definidos para duas etapas independentes de simulações distintas que permitem estudos em diferentes cenários de tamanhos amostrais.

Na primeira etapa: tamanho amostral (n= 30, 40, 50, 60, 70, 80 e 90), probabilidade de sucesso (π=0,5 e 0,7), proporção de valores nulos esperada nas amostras geradas (γ=0,2 e 0,3) e mantendo fixado o número de experimentos Bernoulli em m=100.

Na segunda etapa: tamanho amostral (n= 8 e 12), probabilidade de sucesso (π=0,5 e 0,75), proporção de valores nulos esperada nas amostras geradas (γ=0,25) e o número de experimentos Bernoulli (m=30, 40 e, 50).

Dada a combinação desses fatores em cada etapa, a obtenção das estimativas de π, será dada conforme descrito no item 3.2.

3.2 Estimação das probabilidades de sucessos (π) do modelo ZIB

Com as especificações paramétricas descritas em 3.1, para cada configuração simulada, as estimativas de π_{zib} serão obtidas utilizando-se o estimador proposto por Silva (2009), definido por:

$$\hat{\pi}_{zib} = \sum_{y=0}^{m} \rho_i(x)\hat{\pi}_{emv}\,, \tag{27}$$

em que $\hat{\pi}_{emv}$, é o estimador de máxima verossimilhança de π e definida por

$$\hat{\pi}_{emv} = \frac{1}{m}\sum_{y=0}^{m} y f_n(y), \tag{28}$$

sendo

$$f_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y = y_i) \text{ e} \tag{29}$$

$$\rho_i(x)=\begin{cases}\left\{[\ln(c_1)+\dfrac{(1-u)\ln(c_1)+1}{u}]c_1^{1-u}\right\}x^u-[(1-u)\ln(c_1)+1]\dfrac{c_1}{u}, & x<c_1\\ x\ln(x), \text{ se } c_1\le x\le c_2\\ \left\{[\ln(c_2)+\dfrac{(1-u)\ln(c_2)+1}{u}]c_2^{1-u}\right\}x^u-[(1-u)\ln(c_2)+1]\dfrac{c_2}{u}, & x>c_2\end{cases} \quad (30),$$

com $x=\dfrac{f_n(y)}{p_n(y)}$ fixado, os valores de $i=1$ e 2. O estimador dado em (30) é construído em duas abordagens, que diferem entre si na base dos valores a serem assumidos pelas constantes c_1, c_2 e u, pesquisadas através da rotina descrita no ANEXO A, para obter uma estimativa de π_{zib} próxima do valor de π estabelecido. Essas duas abordagens são:

a) Abordagem 1 consiste em fixar os valores de $u=c_2=1$ e pesquisar um valor de $c_1<c_2=1$. Assim $\rho_1(x)$ tende a um crescimento maior quando $x\to\infty$. Mantendo a especificação ($c_1<c_2=1$), segundo Ruckstuhl e Welsh (2001) os estimadores de $\hat{\pi}_{emv}$ tendem a ser mais robustos.

b) Abordagem 2 consiste em fixar $c_1=0{,}1$, mantendo a restrição $(c_1<c_2=1)$, e pesquisar o valor de u, de modo a reduzir o crescimento de $\rho_2(x)$ quando $x\to\infty$.

Convém salientar que a acurácia e precisão desse estimador dependem dos valores das constantes de afinidades c_1 e c_2 que torne robusta a proporção de valores nulos esperadas, sendo essa representada por γ.

Para efeito de comparação, estimou-se o viés relativo para as estimativas de $\hat{\pi}_{emv}$ e $\hat{\pi}_{zib}$, de acordo com as expressões

$$v_{emv} = \frac{\hat{\pi}_{emv} - \pi}{\pi} \; ; \; v_e = \frac{\hat{\pi}_e - \pi}{\pi}. \tag{31}$$

3.3 Definição e estimação das funções lineares de proporções binomiais

Seguindo os procedimentos para geração das amostras binomiais inflacionadas de zeros, vistos na Seção 3.1, e a metodologia a ser empregada para estimação dos parâmetros π_i, *i=1;2;...;q* , vistos na Seção 3.2, utilizados na estimação das funções lineares binomiais definidas por (32)

$$\psi = \sum_{i=1}^{q} \delta_i \pi_i , \tag{32}$$

sendo *q* o número total de populações binomiais, o i-ésimo coeficiente associado à proporção de sucessos referente a i-ésima população binomial, é denotado por δ_i , seguindo as especificações definidas na tabela 3.

Tabela 3 Valores dos coeficientes utilizados na especificação das funções lineares binomiais

F=*q*	vetor de coeficientes utilizado na composição de ψ

F1=3	$\underset{\sim}{\delta}_1 = (2,-1,-1)$
F2=5	$\underset{\sim}{\delta}_2 = (4,-1,-1,-1,-1)$
F3=7	$\underset{\sim}{\delta}_3 = (6,-1,-1,-1,-1,-1,-1)$
F4=10	$\underset{\sim}{\delta}_4 = (9,-1,-1,-1,-1,-1,-1,-1,-1,-1)$

F=q notação utilizada para representar a família binomial fixado *q* proporções

Para cada função linear foram computadas as estimativas intervalares para ψ, considerando os intervalos de confiança de Wald e Wilson conforme as seções 2.3.1 e 2.3.2.

Por fim, em função dos valores paramétricos, e cenários de avaliação, os intervalos adaptados para proporções infladas de zeros foram comparados utilizando a probabilidade de cobertura e amplitude média intervalar. Com essas medidas, a precisão das estimativas intervalares e a acurácia, em relação ao nível de confiança nominal, fixado em 95%. Foram realizadas 2000 simulações Monte Carlo. A obtenção desses resultados foi feita através da elaboração de um programa em R (ANEXO B), em que a probabilidade de cobertura é medida pela porcentagem de intervalos que contém o parâmetro ψ, fixado o coeficiente de confiança mencionado, e o comprimento médio, dado pela média dos comprimentos dos intervalos em cada amostra simulada.

4 RESULTADOS E DISCUSSÃO

Considerando-se os cenários de avaliação, mencionados na seção 3.3, em uma primeira etapa foi fixado o número de ensaios de Bernoulli *m*=100, na obtenção das amostras para estudo dos métodos propostos.

Em se tratando da escolha das constantes a serem utilizadas na obtenção das estimativas de π_E Silva e Cirillo (2010), apresentaram estudos relacionados a inferência de um modelo binomial contaminado pela mistura de populações binomiais, sendo as amostras obtidas através de simulações Monte Carlo. Foram considerados diversos valores da constante de afinidade c_1, envolvida na função $\rho(x)$, sendo $0,1 \leq c_1 \leq 0,9$ e tamanhos de amostras iguais a 10, 50 e 80, além das taxas de misturas iguais a 0,20 e 0,40.

Para estes cenários, o estimador π_E apresentou desempenho diferenciado, de acordo com o grau de contaminação definida pela probabilidade da mistura a qual as populações simuladas foram submetidas. Dessa forma, as estimativas obtidas indicaram que a robustez desse estimador é mais pronunciada para determinados valores da constante de afinidade c_1 . Em função dos resultados obtidos, optou-se na escolha das proporçoes de valores nulos avaliados nesse trabalho representados por γ=0,2 e 0,3.

Os resultados descritos nas Tabelas 4 – 7 corresponderam às estimativas do parâmetro π utilizando-se os estimadores de máxima verossimilhança (π_{emv}) e o estimador robusto a excesso de zeros (π_{zib}).

Em síntese, os resultados evidenciaram que nas amostras binomiais contaminadas com excessos de zeros, as estimativas de máxima verossimilhança não são acuradas, uma vez que os vieses relativos estimados foram consideravelmente elevados. Porém, ao considerar as estimativas $\hat{\pi}_{zib}$,

observou-se que, para todos os tamanhos amostrais e valores de γ os vieses relativos apresentaram resultados inferiores a 0,01 para os valores paramétricos especificados em π=0,5 (Tabelas 4 e 5) e π=0,7 (Tabelas 6 e 7), excluindo-se, evidentemente, pequenas flutuações associadas ao erro Monte Carlo.

Tabela 4 Resultados comparativos dos estimadores π_{emv} e π_{zib}, considerando o valor paramétrico $\pi = 0,5$, com restrição $c_2 = u = 1$, caracterizando a componente sistemática ρ_1

n	γ	c_1	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	0,2	0,2900	0,3995	-0,2010	0,5000	-0,0001
30	0,3	0,4300	0,3504	-0,2992	0,5018	0,0036
40	0,2	0,2600	0,4006	-0,1988	0,5007	0,0014
40	0,3	0,4300	0,3499	-0,3002	0,5001	0,0003
50	0,2	0,2600	0,4002	-0,1996	0,4941	-0,0119
50	0,3	0,4300	0,3508	-0,2984	0,4986	-0,0027
60	0,2	0,2500	0,3985	-0,2030	0,4983	-0,0034
60	0,3	0,4300	0,3506	-0,2988	0,4983	-0,0034
70	0,2	0,2500	0,3998	-0,2024	0,4913	-0,0173
70	0,3	0,4300	0,3504	-0,2992	0,4974	-0,0052
80	0,2	0,2400	0,3991	-0,2018	0,4962	-0,0076
80	0,3	0,4300	0,3497	-0,3006	0,4982	-0,0037
90	0,2	0,2400	0,4000	-0,2000	0,4914	-0,0171
90	0,3	0,4300	0,3499	-0,3002	0,4975	-0,0049

Tabela 5 Resultados comparativos dos estimadores π_{emv} e π_{zib}, considerando o valor paramétrico $\pi = 0,5$, com restrição $c_1 = 0,1$ e $c_2 = 1$, caracterizando a componente sistemática ρ_2.

n	γ	*u*	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	0,2	0,1540	0,3995	-0,2010	0,5102	0,0203
30	0,3	0,1800	0,3514	-0,2972	0,4930	-0,0141
40	0,2	0,1400	0,4000	-0,2000	0,5026	0,0052
40	0,3	0,1730	0,3499	-0,3002	0,5020	0,0041
50	0,2	0,1310	0,4003	-0,1994	0,4997	-0,0006
50	0,3	0,1700	0,4002	-0,1995	0,4941	-0,0119
60	0,2	0,1240	0,4002	-0,1995	0,4952	-0,0096
60	0,3	0,1660	0,3507	-0,2986	0,4996	-0,0008
70	0,2	0,1170	0,3993	-0,2014	0,5039	0,0078
70	0,3	0,1640	0,3509	-0,2982	0,4970	-0,0061
80	0,2	0,1130	0,4003	-0,1994	0,4900	-0,0201
80	0,3	0,1610	0,3500	-0,3000	0,4982	-0,0037
90	0,2	0,1060	0,3999	-0,2002	0,5102	-0,0204
90	0,3	0,1590	0,3491	-0,3018	0,5000	-0,0010

Tabela 6 Resultados comparativos dos estimadores π_{emv} e π_{zib}, considerando o valor paramétrico $\pi = 0,7$, com restrição $c_2 = u = 1$, caracterizando a componente sistemática ρ_1.

n	γ	c_1	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	0,2	0,0001	0,5580	-0,2029	0,8322	0,1888
30	0,3	0,2700	0,4889	-0,3016	0,7011	0,0015
40	0,2	0,1500	0,5611	-0,1984	0,6977	-0,0033
40	0,3	0,2700	0,4924	-0,2966	0,6994	-0,0008
50	0,2	0,1500	0,5610	-0,1996	0,6961	-0,0056
50	0,3	0,2700	0,4907	-0,1986	0,7051	0,0073
60	0,2	0,1500	0,5591	-0,2990	0,6979	-0,0030
60	0,3	0,2700	0,4899	-0,3001	0,7050	0,0072
70	0,2	0,1500	0,5597	-0,2004	0,6961	-0,0056
70	0,3	0,2800	0,4902	-0,2997	0,6951	-0,0069
80	0,2	0,1500	0,5605	-0,1990	0,6944	-0,0080
80	0,3	0,2800	0,4904	-0,2994	0,6978	-0,0012
90	0,2	0,1500	0,5597	-0,2004	0,6976	-0,0035
90	0,3	0,2800	0,4921	-0,2970	0,6949	-0,0073

Tabela 7 Resultados comparativos dos estimadores π_{emv} e π_{zib}, considerando o valor paramétrico $\pi = 0{,}7$, com restrição $c_1 = 0{,}1$ e $c_2 = 1$, caracterizando a componente sistemática ρ_2.

n	γ	u	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	0,2	0,1300	0,5589	-0,2820	0,7078	0,0134
30	0,3	0,1400	0,4904	-0,2994	0,7025	0,0040
40	0,2	0,1270	0,5582	-0,2026	0,7020	-0,0019
40	0,3	0,1400	0,4920	-0,2971	0,6979	-0,0017
50	0,2	0,1240	0,5604	-0,1986	0,7001	0,0044
50	0,3	0,1400	0,4908	-0,2990	0,6972	-0,0014
60	0,2	0,1220	0,5605	-0,2013	0,6967	0,0017
60	0,3	0,1390	0,4895	-0,2986	0,7025	0,0028
70	0,2	0,1190	0,5604	-0,3001	0,7008	-0,0072
70	0,3	0,1390	0,4884	-0,2997	0,7015	0,0030
80	0,2	0,1180	0,5590	-0,1993	0,7002	-0,0008
80	0,3	0,1390	0,4899	-0,2994	0,6953	-0,0068
90	0,2	0,1160	0,5592	-0,2004	0,7048	-0,0038
90	0,3	0,1390	0,4897	-0,2970	0,6983	-0,0052

Garantida a acurácia das estimativas de proporção binomial para as configurações paramétricas descritas na seção 3.3, por meio da tabela 3, procedeu-se com a composição das funções lineares binomiais para o método de Wald. Computou-se a probabilidade média de cobertura média dada as 2000 realizações Monte Carlo. Os resultados estão apresentados nas Figuras 1 – 8.

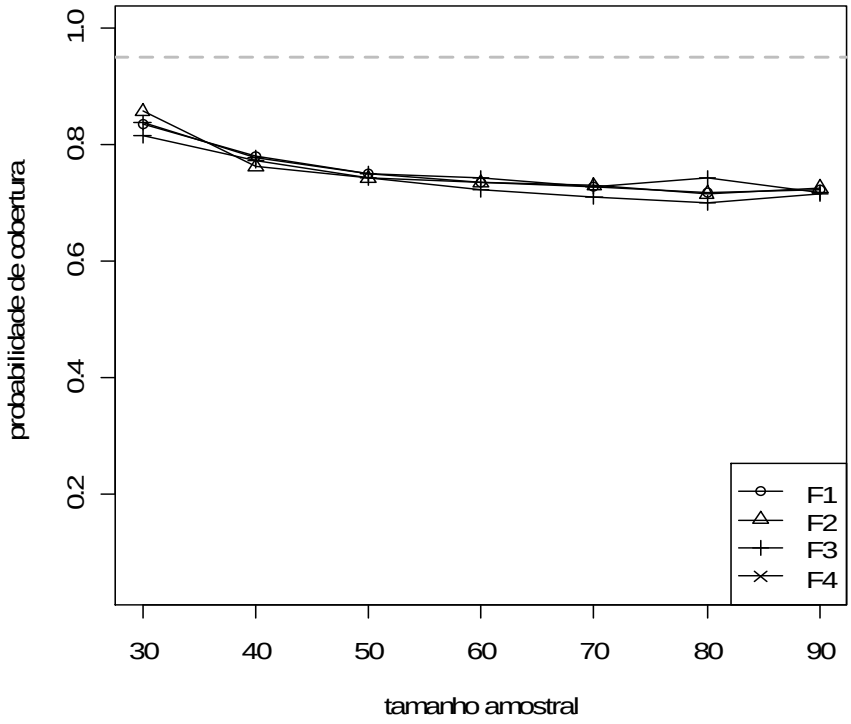

Figura 1 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,2$ e $\pi = 0,5$ e a componente sistemática ρ_1

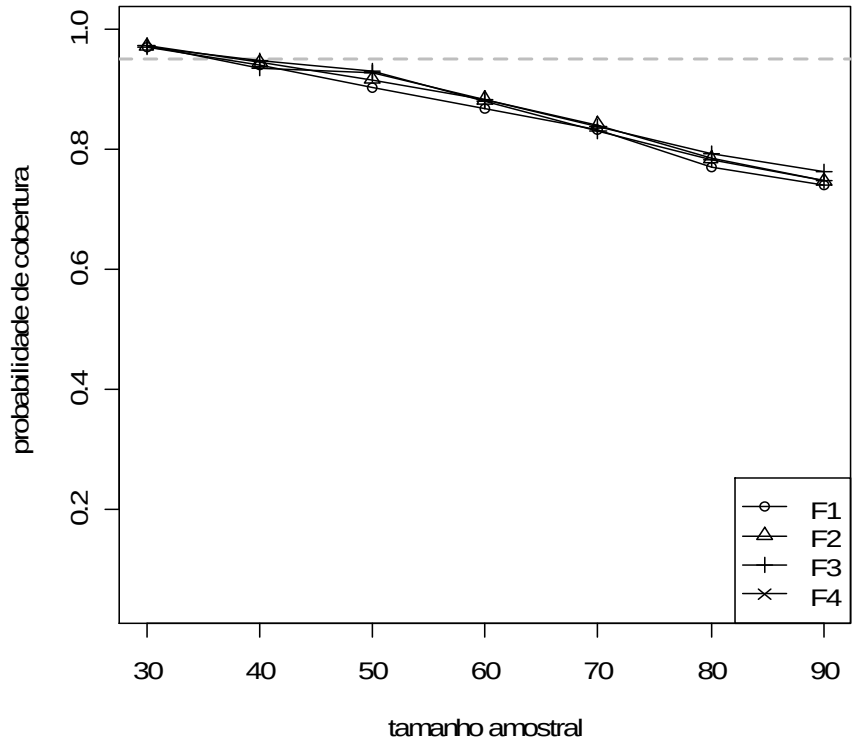

Figura 2 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,2$ e $\pi = 0{,}5$ e a componente sistemática ρ_2

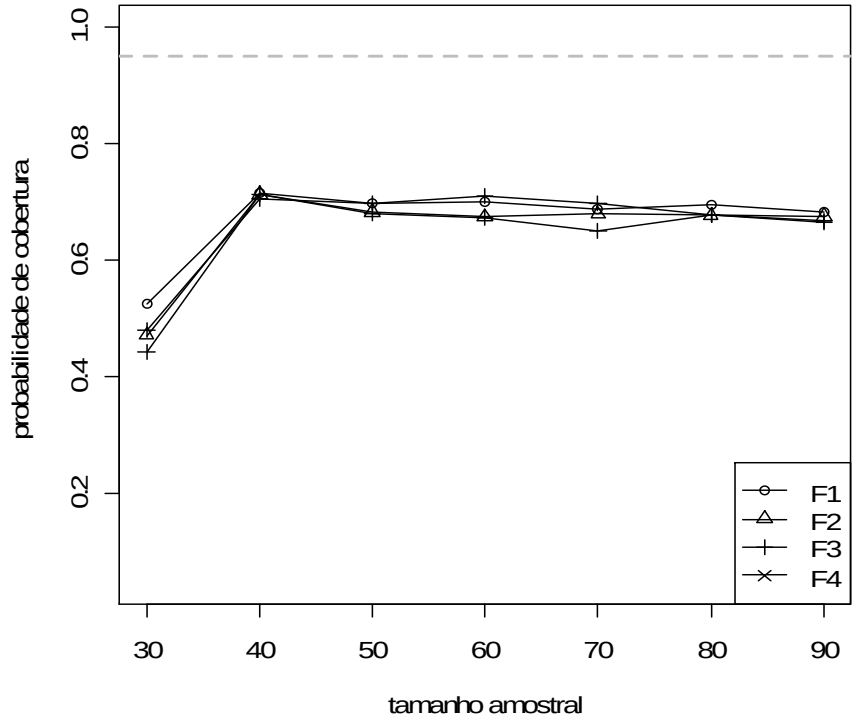

Figura 3 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,2$ e $\pi = 0,7$ e a componente sistemática ρ_1

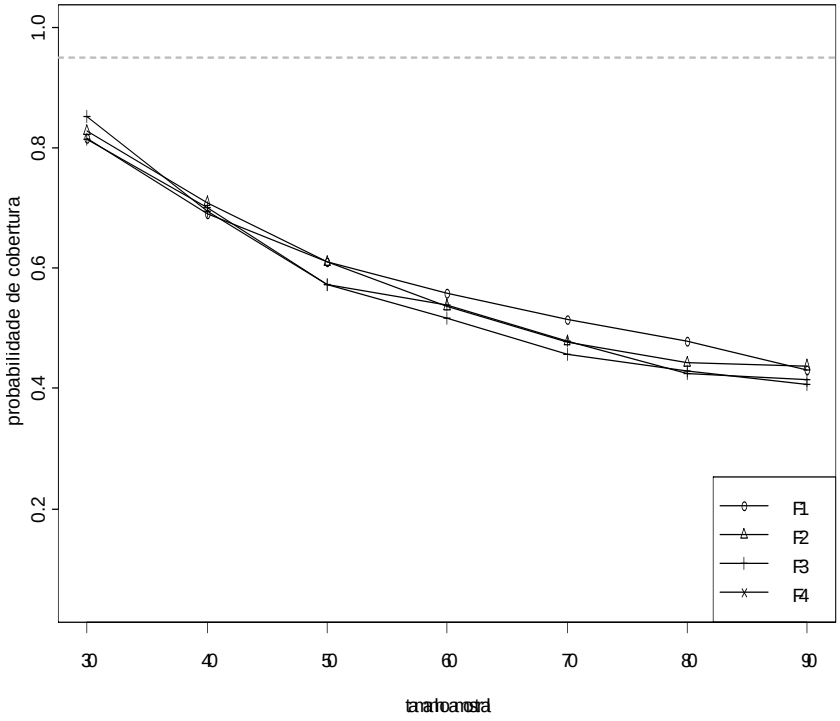

Figura 4 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,2$ e $\pi = 0,7$ e a componente sistemática ρ_2

Conforme os resultados ilustrados nas Figuras 1 – 4, nota-se que, de um modo geral, o aumento do tamanho amostral resultou em uma redução da probabilidade de cobertura, em contradição ao nível nominal de confiança fixado em 95%.

Tal fato, sugere que o desempenho dos métodos de estimação propostos nesse trabalho são diferenciados em relação às estimativas das proporções binomiais serem próximas a 0,5 sendo esse valor que maximiza a variância da

distribuição binomial pertencente a cada população envolvida na composição da família binomial.

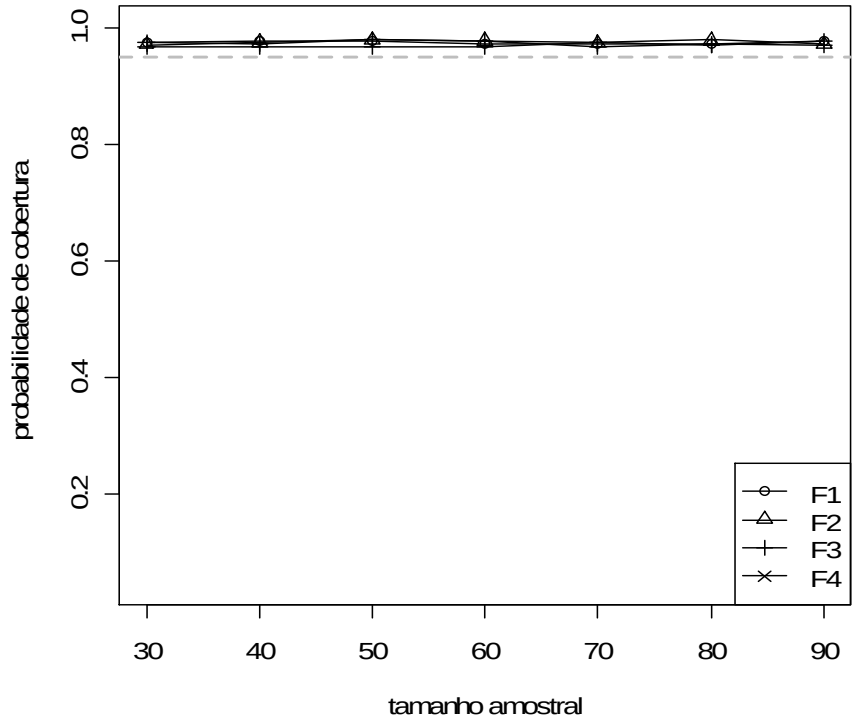

Figura 5 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,3$ e $\pi = 0,5$ e a componente sistemática ρ_1

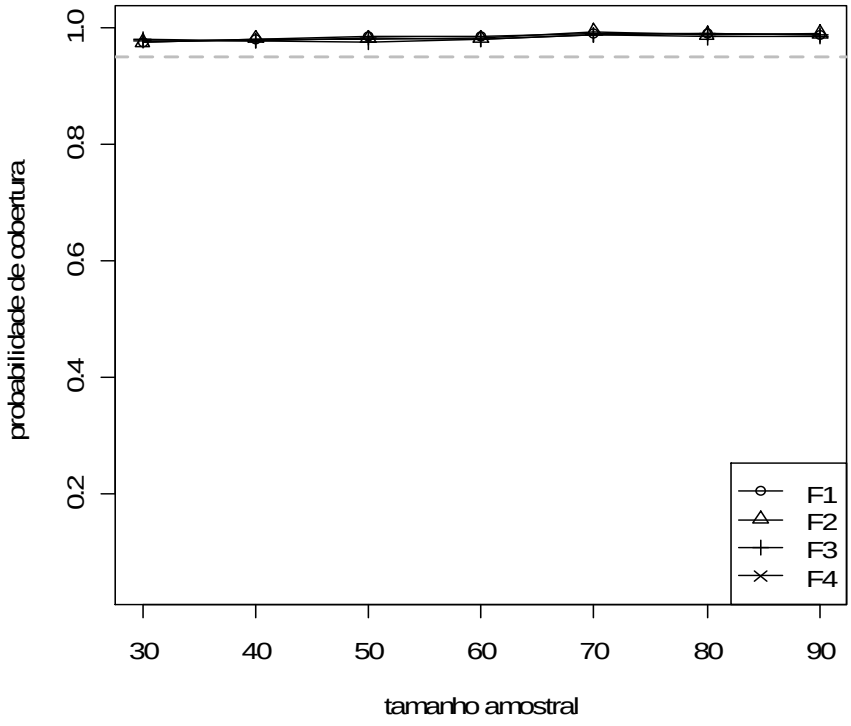

Figura 6 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,3$ e $\pi = 0,7$ e a componente sistemática ρ_1

Com as mesmas configurações, porém aumentando a proporção, de zeros ($\gamma = 0,3$) observou-se que, ao utilizar a componente ρ_1 o método apresentou probabilidades de cobertura superior ao nível nominal de confiança de 95%. Tal fato notório ocorreu para todos os tamanhos amostrais e, em todas as famílias binomiais (Figuras 5 e 6).

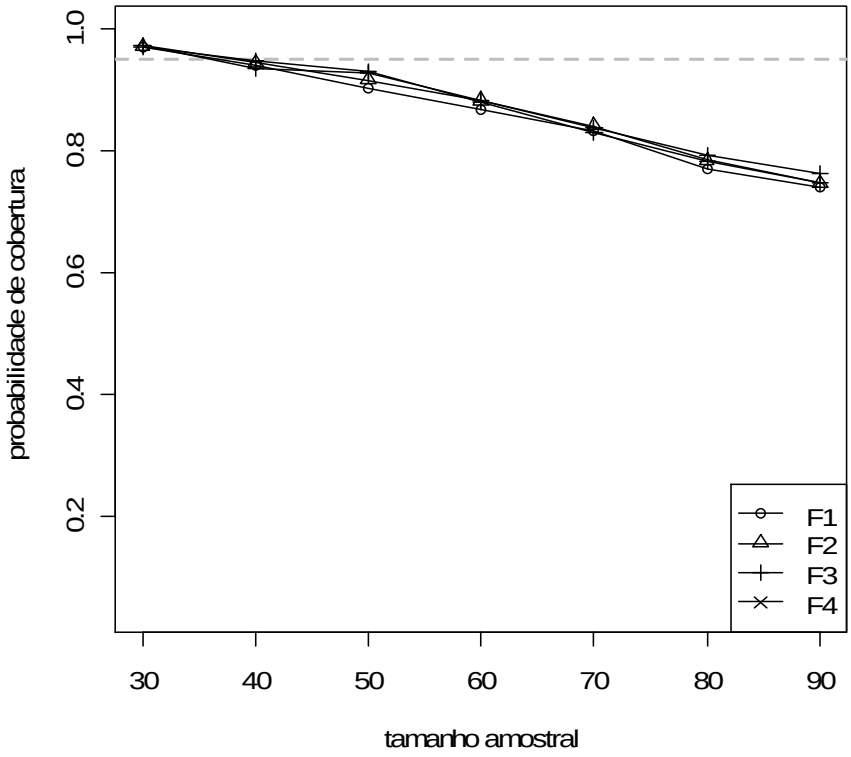

Figura 7 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,3$ e $\pi = 0,5$ e a componente sistemática ρ_2

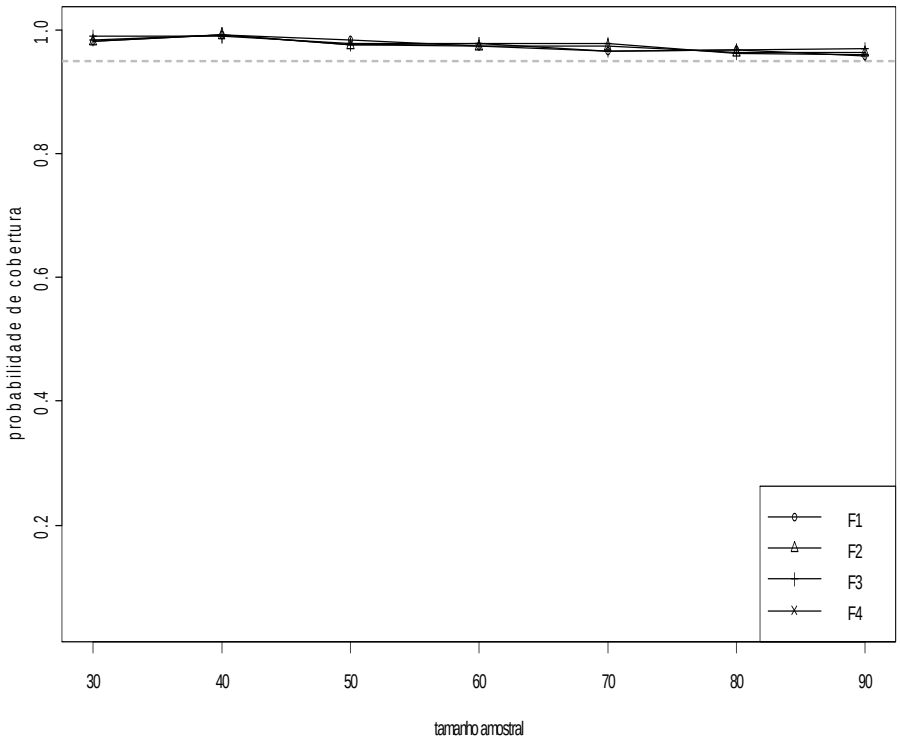

Figura 8 Probabilidade de cobertura assumindo os parâmetros $\gamma = 0,3$ e $\pi = 0,7$ e a componente sistemática ρ_2

Ressalta-se que apenas a composição das famílias binomiais ,utilizando os valores paramétricos $\pi = 0,7$ e a componente sistemática ρ_2, apresentou probabilidades condizentes ao nível nominal de confiança especificado em 95%. Ao utilizar a descrição que caracteriza a componente ρ_2 (Figuras 5 e 6), tendo por base esses resultados, há evidências estatísticas para afirmar que assumindo baixas quantidades de valores nulos, o método de Wald com a incorporação das estimativas robustas apresentou resultados incoerentes com o nível de confiança

nominal. Portanto, não recomenda-se esse método na prática em situações similares aos cenários avaliados.

Outra abordagem de estudo em relação às propriedades do método de Wald aplicado a funções lineares binomiais foi dado por Cirillo, Ferreira e Safádi (2009), em relação ao emprego do método *bootstrap* infinito (COULON; THOMAS, 1990), estendido à diferença de *q* populações binomiais referenciadas por um único parâmetro, mantendo a distribuição de probabilidade conjunta.

Neste contexto, o *bootstrap* não apresentou resultados que possam justificar seu uso e a expansão do método para um número maior de populações, representadas consideradas nas funções lineares, também resultou em probabilidades de cobertura incoerentes com resultados semelhantes aos observados neste trabalho.

Em função dos resultados obtidos apenas para m=100, com o intuito de investigar o desempenho dos métodos propostos e dada as recomendações de Ruckstuhl e Welsh (2001) em relação às constantes c_1 e c_2 em função do grau de contaminação. Dessa forma, mantendo os cenários a serem avaliados nas amostras obtidas por simulações Monte Carlo, porém assumindo a proporção de valores nulos em γ=0,25 com os números de ensaios de Bernoulli *m*=30, 40 e 50 a acurácia e precisão dos estimadores foram avaliados pela comparação dos respectivos vieses calculados para as estimativas de π_{emv} e π_{zib}.

Para estes cenários, os resultados descritos nas Tabelas 8 e 9 foram obtidos com restrição $c_2 = u = 1$, caracterizando a componente sistemática ρ_1.

Tabela 8 Resultados comparativos dos estimadores π_{emv} e π_{zib} para o valor paramétrico $\pi = 0,5$ e componente sistemática ρ_1.

m	*n*	c_1	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	8	0,2800	0,3721	-0,2558	0,5000	0,0000
40	8	0,3300	0,3751	-0,2498	0,4957	-0,0086
50	8	0,3600	0,3747	-0,2505	0,5041	0,0082
30	12	0,2000	0,3734	-0,2533	0,4952	-0,0096
40	12	0,2700	0,3757	-0,2487	0,4999	-0,0002
50	12	03100	0,3738	-0,2524	0,5068	0,0137

Tabela 9 Resultados comparativos dos estimadores π_{emv} e π_{zib} para o valor paramétrico $\pi = 0,75$ e componente sistemática ρ_1.

m	*n*	c_1	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	8	0,0400	0,5610	-0,2519	0,7484	-0,0021
40	8	0,0800	0,5654	-0,2461	0,7543	0,0058
50	8	0,1200	0,5616	-0,2512	0,7535	0,0047
30	12	0,0010	0,5629	-0,2494	0,7427	-0,0097
40	12	0,0500	0,5661	-0,2466	0,7464	-0,0048
50	12	0,1000	0,5615	-0,2513	0,7551	0,0069

Nota-se novamente que a escolha dos valores para c_1 em ambos os valores paramétricos avaliados, as estimativas produzidas para π_{zib} foram acuradas.

Em se tratando da componente ρ_2, os resultados descritos nas Tabelas 10 e 11 correspondem às estimativas π_{emv} e π_{zib} com a mesma especificação $\gamma = 0,25$, porém mantendo a restrição $c_1 = 0,1$ e $c_2 = 1$. Ressalta-se que novamente o estimador $\hat{\pi}_{zib}$ resultou em estimativas acuradas.

Tabela 10 Resultados comparativos dos estimadores π_{emv} e π_{zib} para o valor paramétrico $\pi = 0{,}5$ e componente sistemática ρ_2.

m	*n*	*u*	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	8	0,1860	0,3768	-0,2465	0,4970	-0,0059
40	8	0,1960	0,3774	-0,2452	0,5036	0,0072
50	8	0,2050	0,3745	-0,2511	0,4975	-0,0049
30	12	0,1500	0,3767	-0,2466	0,5028	0,0057
40	12	0,1690	0,3756	-0,2488	0,4975	-0,0050
50	12	0,1800	0,3765	-0,2470	0,5001	0,0002

Tabela 11 Resultados comparativos dos estimadores π_{emv} e π_{zib} para o valor paramétrico $\pi = 0{,}75$ e componente sistemática ρ_2.

m	*n*	*u*	$\hat{\pi}_{emv}$	*viés*	$\hat{\pi}_{zib}$	*viés*
30	8	0,1270	0,5652	-0,2464	0,7487	-0,0046
40	8	0,1960	0,5597	-0,2537	0,7452	-0,0064
50	8	0,1350	0,5645	-0,2473	0,7487	-0,0018
30	12	0,1170	0,5646	-0,2472	0,7489	-0,0015
40	12	0,1250	0,5627	-0,2497	0,7525	0,0033
50	12	0,1290	0,5609	-0,2521	0,7528	0,0038

Confirmada a acurácia do estimador $\hat{\pi}_{zib}$ em função das escolhas apropriadas para as constantes c_1, c_2 e u, justificadas pela acurácia das estimativas, retratadas pelos vieses inferiores a 0,01, procedeu-se com a proposição dos estimadores para famílias binomiais, substituindo os estimadores de máxima verossimilhança $\hat{\pi}_{emv}$ por $\hat{\pi}_{zib}$.

Com essa modificação, os resultados aprimorados com a incorporação dos estimadores π_{zib} descritos na Tabelas 12-14, correspondem às probabilidades de cobertura obtidas para o método de Wald, considerando a

quantidade de valores nulos $\gamma = 0,25$ e os valores dos coeficientes F1, F2, F3 e F4 (Tabela 3), utilizados na composição das funções lineares binomiais com uma proporção de valores nulos fixada em $\gamma = 0,25$ com restrição $c_2 = u = 1$, caracterizando a componente sistemática ρ_1.

Tabela 12 Probabilidades de cobertura de 95% de confiança e amplitude intervalar para o método Wald, para $\pi = 0,5$ e $\pi = 0,75$, utilizando a componente sistemática ρ_1.

			π = 0,5							
m	*n*	c_1	F1	$A\psi$	F2	$A\psi$	F3	$A\psi$	F4	$A\psi$
30	8	0,280	0,989	1,634	0,987	2,987	0,985	4,320	0,988	6,323
40	8	0,330	0,996	1,650	0,997	3,010	0,996	4,367	0,996	6,390
50	8	0,360	0,998	1,659	0,998	3,028	0,999	4,391	0,999	6,429
30	12	0,200	0,913	1,319	0,904	2,408	0,909	3,491	0,911	5,122
40	12	0,270	0,958	1,333	0,946	2,430	0,952	3,524	0,962	5,160
50	12	0,310	0,968	1,342	0,969	2,450	0,973	3,543	0,975	5,203
			π = 0,75							
m	*n*	c_1	F1	$A\psi$	F2	$A\psi$	F3	$A\psi$	F4	$A\psi$
30	8	0,280	0,809	1,288	0,770	2,318	0,755	2,864	0,758	5,011
40	8	0,330	0,883	1,310	0,888	2,368	0,876	3,466	0,886	5,038
50	8	0,360	0,938	1,357	0,940	2,490	0,940	3,570	0,951	5,252
30	12	0,200	0,668	1,035	0,613	1,871	0,602	2,705	0,582	3,935
40	12	0,270	0,727	1,043	0,697	1,904	0,684	2,720	0,677	4,049
50	12	0,310	0,771	0,985	0,746	1,781	0,738	2,574	0,699	3,761

$A\psi$: amplitude intervalar e c_1: constante de afinidade

Os resultados descritos na tabela 12 evidenciam que o método de Wald aprimorado, apresentou uma sensibilidade em relação ao valor paramétrico, de modo que, ao apresentar valores distantes de $\pi = 0,5$, os resultados referentes à

probabilidade de cobertura tendem a apresentar valores inferiores ao nível nominal de confiança. Entretanto, ao assumir a componente sistemática ρ_2 (Tabela 13) nota-se que, em ambos os valores paramétricos, na maioria das situações apresentaram probabilidades de cobertura superiores ou próximas ao nível nominal de confiança. Portanto, recomenda-se o método de Wald aprimorado com o uso da componente sistemática ρ_2.

Tabela 13 Probabilidades de cobertura de 95% de confiança e amplitude intervalar para o método Wald, para $\pi = 0,5$ e $\pi = 0,75$, utilizando a componente sistemática ρ_2.

			π = 0,5							
m	*n*	*u*	F1	*Aψ*	F2	*Aψ*	F3	*Aψ*	F4	*Aψ*
30	8	0,186	0,999	1,672	0,999	3,054	0,999	4,427	0,999	6,479
40	8	0,196	1,000	4,455	1,000	3,074	1,000	4,455	1,000	6,522
50	8	0,205	1,000	1,687	1,000	3,079	1,000	4,462	1,000	6,531
30	12	0,150	0,930	1,323	0,919	2,417	0,946	3,519	0,929	5,129
40	12	0,169	0,987	1,353	0,985	2,471	0,984	3,581	0,978	5,240
50	12	0,180	0,994	1,367	0,996	2,495	0,996	3,616	0,992	5,293
			π = 0,75							
m	*n*	*u*	F1	*Aψ*	F2	*Aψ*	F3	*Aψ*	F4	*Aψ*
30	8	0,127	0,965	1,288	0,976	2,318	0,970	2,864	0,962	5,011
40	8	0,132	0,991	1,429	0,993	2,611	0,992	3,778	0,989	5,522
50	8	0,135	0,999	1,449	0,998	2,651	0,997	3,837	0,996	5,604
30	12	0,117	0,729	1,045	0,742	1,915	0,724	2,766	0,718	4,061
40	12	0,125	0,897	1,125	0,908	2,045	0,906	2,959	0,908	4,363
50	12	0,129	0,954	1,155	0,946	2,109	0,949	3,045	0,964	4,476

$A\psi$: amplitude intervalar ; u: constante utilizada na componente sistemática ρ_2

Mantendo o mesmo enfoque na avaliação do aprimoramento do método de Wilson, com a incorporação do estimador π_{zib}, os resultados descritos na Tabela 14, apresentam em maior parte, probabilidade de cobertura inferior ao nível nominal de confiança em ambos os valores paramétricos. Portanto, pode-se

afirmar que o uso dos estimadores π_{zib} não proporcionou resultados estatísticos que asseguram a inferência em amostras binomiais inflacionadas de zeros, por meio da composição dos contrastes ortogonais previamente definidos na composição das famílias binomiais, tornando o método não recomendável na prática, frente a situações semelhantes aos cenários simulados.

Tabela 14 Probabilidades de cobertura de 95% de confiança e amplitude intervalar para o método Wilson, considerando os valores paramétricos $\pi = 0,5$ e $\pi = 0,75$

			π=0,5							
m	*n*	c_1	F1	$A\psi$	F2	$A\psi$	F3	$A\psi$	F4	$A\psi$
30	8	0,280	0,782	3,206	0,789	5,470	0,845	7,729	0,988	11,14
40	8	0,330	0,750	3,370	0,701	5,623	0,700	7,875	0,701	11,31
50	8	0,360	0,751	3,141	0,699	5,244	0,713	7,358	0,727	10,59
30	12	0,200	0,720	3,122	0,707	5,379	0,701	7,621	0,729	11,00
40	12	0,270	0,720	3,125	0,704	5,000	0,717	7,015	0,720	10,08
50	12	0,310	0,747	3,127	0,645	5,245	0,655	7,281	0,655	10,53
					π=0,75					
m	*n*	c_1	F1	$A\psi$	F2	$A\psi$	F3	$A\psi$	F4	$A\psi$
30	8	0,280	0,763	4,159	0,739	6,702	0,687	9,171	0,691	12,90
40	8	0,330	0,792	4,234	0,742	6,747	0,737	9,957	0,886	12,91
50	8	0,360	0,844	4,290	0,780	6,789	0,752	9,294	0,700	12,97
30	12	0,200	0,771	4,144	0,694	6,654	0,672	9,051	0,634	12,76
40	12	0,270	0,807	4,278	0,766	6,825	0,695	9,250	0,656	12,80
50	12	0,310	0,823	4,279	0,754	6,797	0,709	9,171	0,673	12,77

$A\psi$: amplitude intervalar e c_1: constante de afinidade

Por meio dos resultados descritos nas Tabelas, o método Wilson não apresentou probabilidades de cobertura satisfatórias para as famílias binomiais analisadas com as proporções $\pi = 0,5$ e $\pi = 0,75$, conforme a Tabela 14. Entretanto, convém ressaltar que , para todas as funções lineares binomiais avaliadas, o método de Wilson mostrou-se preciso, uma vez que os valores das

amplitudes intervalar para cada família apresentou diferença na primeira ou casa decimal.

Em relação à probabilidade de cobertura média Zou, Huang e Zhan (2009) em amostras binomiais sem excesso de zero, avaliaram funções lineares com diferentes proporções binômias e coeficientes não ortogonais e concluíram que a cobertura mínima para o método de Wilson é próxima a 81,21% fixado o nível nominal de 90% de confiança e 86,22% para o nível nominal de 95%.

Com base nesses resultados, nota-se que o método de Wilson não é adequado em algumas situações mencionadas por Zou, Huang e Zhan (2009) para avaliar as funções lineares binomiais. Em concordância com esses resultados, por meio da Tabela 14, entende-se que as proposições apresentadas nessa tese, para amostras binomiais inflacionadas de zeros, não são recomendáveis de serem incorporadas para esse método, uma vez que os resultados foram similares aos resultados obtidos por Zou, Huang e Zhan (2009).

5 CONCLUSÃO

Em consonância com os objetivos propostos, conclui-se que o aprimoramento dado pela incorporação de estimadores robustos a excesso de zeros é recomendável para o método de Wald, em situações reais similares aos cenários simulados enfatizando as seguintes situações.

a) Recomenda-se o método de Wald considerando as estimativas das proporções binomais robustas a excessos de zeros que maximizam a variância das funções lineares binomiais $(\hat{\pi}_{zib} \approx 0,5)$, com a incorporação da componente $\rho_{1.}$e coeficientes ortogonais.
b) O Método de Wilson por não apresentar probabilidades de cobertura coerentes ao nível nominal de confiança não apresentou um desempenho que justificasse a incorporação de estimativas robustas a excesso de zeros, na composição das famílias binomiais com coeficientes ortogonais.

Para trabalhos futuros, como instrumento de estudo pretende-se continuar a proposição de métodos alternativos à estimação de funções binomiais incluindo coeficientes quadráticos e não ortogonais, baixas e elevadas proporções de observações nulas e correções no método de Wald nas situações avaliadas cujos resultados foram prejorativos, em relação ao nível nominal de confiança.

REFERÊNCIAS

AGRESTI, A.; CAFFO, B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. **American Statistician**, Alexandria, v. 54, n. 4, p. 280-288, Nov. 2000.

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

CIRILLO, M. A.; FERREIRA, D. F.; SAFÁDI, T. S. Avaliação de métodos de estimação intervalar para funções lineares binomiais via Bootstrap infinito. **Ciência e Agrotecnologia**, Lavras, v. 33, p. 1741-1746, 2009. Edição especial.

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, 1990.

COPAS, J. B. Binary regression models for contaminated data. **Journal of the Royal Statistics Society: Series B, Methodological**, London, v. 50, n. 2, p. 225-265, Apr. 1988.

KEMP, C. D.; KEMP, A. W. Rapid estimation for discrete distributions. **The Statistician**, London, v. 37, n. 2, p. 243-255, June 1988.

LINDSAY, B. G. Efficiency versus robustness: the case for minimum hellinger distance and related methods. **The Annals of Statistic**, Harward, v. 22, n. 4, p. 1081-1114, Sept. 1994.

PRICE, M. R.; BONETT, D. G. An improved confidence interval for a linear function of binomial proportions. **Computational Statistics & Data Analysis**, Amsterdam, v. 45, n. 3, p. 449-456, Apr. 2004.

RUCKSTUHL, A. F.; WELSH, A. H. Robust fitting of the binomial model. **The Annals of Statistic**, Harward, v. 29, n. 4, p. 1117-1136, 2001.

SILVA, A. M. **Proposta e avaliação de um estimador robusto em dados binomiais inflacionados de zeros**. 2009. 47 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2009.

SILVA, A. M.; CIRILLO, M. A. C. Estudo por simulação Monte Carlo de um estimador robusto utilizado na inferência de um modelo binomial contaminado. **Acta Scientiarum Technology**, Maringá, v. 32, n. 3, p. 303-307, 2010.

SIMPSON, D. G. Minimum Hellinger distance estimation for the analisys of count data. **Journal of the American Statistical Association**, New York, v. 82, n. 399, p. 802-807, Sept. 1987.

TEBBS, J. M.; ROTHS, S. A. New large-sample binomial confidence interval for a linear combination of binomial proportions. **Journal of Statistical Planning and Inference**, Amsterdam, v. 138, n. 6, p. 1884-1893, July 2008.

WILSON, E. B. Probable inference, the Law of sucession, and statistical inference. **Journal of the American Statistical Association**, New York, v. 22, p. 209-212, 1927.

ZOU, G. Y.; DONNER, A. Construction of confidence limits about effects measures: a general approach. **Statistics in Medicine**, New York, v. 27, n. 10, p. 1693-1702, May 2008.

ZOU, G. Y.; HUANG, W.; ZHAN, X. A note on confidence interval estimation for a linear function of binomial proportions. **Computational Statistics & Data Analysis**, Amsterdam, v. 53, n. 4, p. 1080-1085, Feb. 2009.

ANEXOS

ANEXO A - Programa utilizado para pesquisar valores dos coeficientes c_1, sob as restrições $c_2 = u = 1$, e u, sob as restrições $c_1 = 0{,}1$ e $c_2 = 1$

```
library(VGAM) ## pacote requerido ##

# ###################### Parâmetros de simulação ######################

m=100          # número de ensaios de Bernoulli

n=90           # número de populações

pr=0.2         # (pr = pi) parâmetro de proporção da binomial

gama=0.30      # proporção de zeros

u=0.13         # u =alfa (coeficiente pesquisado na abordagem 2)
               # u = alfa =1 (fixado na abordagem 1)

nsim=2000      # número de simulações via Monte Carlo

# ###################### Constantes de afinidade ###################### #

c1=0.1         # c1= sigma = 0.1 (fixado na abordagem 2) #
               # c1= sigma (coeficiente pesquisado na abordagem 1) #

c2=1           # c2 = 1 (fixado nas 2 abordagens, c1<c2=1) #

# ###################### Definição das funções ###################### #
```

```
vcont=c(rep(0,m+1))

y=seq(0,m)

rho=c(rep(0,m+1))

estimadores=matrix(0,nsim,3,dimnames=list
(c(rep("Simulação",nsim)),
c("j", "EMV", "Pzib")))


######################## Calculo do fn(y) ############################

fny=function(m,n,dados,vet)

{

   for (a in 1:(m))

  {

       prop=0

       aux=vet[a]

   for (b in 1:n)

   {

       if (aux==dados[b]) prop=prop+1

    }

       vcont[a]=(prop)/n

    }
```

```
        return(vcont)

 }

estimaPzib=function(x,p,c1,c2,alfa)

{

   estPzib=0

   for (b in 1:length(x))

  {

   if (x[b]>=c1 && x[b]<=c2)            rho[b]=x[b]*log(x[b])

   if(x[b]<c1)              rho[b]=((c1^(1-u)*log(c1)+((1-u)*log(c1)+1)
                            *(c1^(1-u)/u))*x[b]^u)-(((1-u)*log(c1)+1)*c1/u)

if(x[b]>c2)                 rho[b]=((c2^(1-u)*log(c2)+((1-u)*log(c2)+1)
                            *(c2^(1-u)/u))*x[b]^u)-(((1-u)*log(c2)+1)*c2/u)

        auxPzib=rho[b]*p[b]

        estPzib=auxPzib + estPzib

}

        return (estPzib)

}

for (j in 1:nsim)

{
```

```
    amos1=rzibinom(n, m, pr, phi = gama)

    soma=sum(amos1)
    while (soma==0)

       {
           amos1=rzibinom(n, m, pr, phi = gama)

       }


  resfny=fny(m,n,amos1,y)

  estmv=sum(y*resfny)/m  #### Estima fn(y) ####

  prob=dbinom(y,m,estmv)

  x=(resfny/prob)

  Pzib=estimaPzib(x,prob,c1,c2,u)

  estimadores[j,1]=j

  estimadores[j,2]=estmv

  estimadores[j,3]=Pzib

}

  m1= mean(estimadores[,2])

  m2= mean(estimadores[,3])

vies=(m2-pr)/pr

m1 #estimador de máxima verossimilhança#
```

```
m2 #estimador robusto#

vies
####################### Fim do programa ##########################
##################################################################
```

ANEXO B - Programa para estimar as funções lineares binomiais, em amostras inflacionadas de zeros

```
library(VGAM)

############## parâmetros de simulação a serem alterados ###############

pr=c(0.7,0.7,0.7,0.7)          # parâmetros da familia binomial

gama=0.30                      # porcentagem de excesso de zeros

npop=10                        # número de populações

coef=c(2,-1,-1)                # coeficientes

n=c(70,70,70)                  # tamanho amostral para cada população

nfixo=70                       # nfixo = n

c1=0.1  ; c2=1                 # constantes de afinidade
                               # (c1=0.1, fixado na abordagem 2)
                               # (c1 é pesquisado na rotina 1, para a abord. 1)

alfa2=abs(qnorm(0.025))        # nível de significancia

u=c(0.139,0.139,0.139)         # constantes de afinidade para o método
                               # de estimação holder

#####################################################################
#################### parâmetros de simulação fixos ##################

   nsim=2000
```

```
   m=100
   psi= coef%*%pr
#####################################################################
# ##################### Definição das funções ######################### #

vcont=c(rep(0,m+1))

k=seq(0,m)

rho=c(rep(0,m+1))

contaw=0

resul=matrix(0,nsim,5)

 ########################## Calculo do fn(k) ########################## #

fnk=function(m,n,dados,vet)

{

 for (a in 1:(m))

 {

   prop=0

   aux=vet[a]

 for (b in 1:n)

 {

    if (aux==dados[b]) prop=prop+1

 }
```

```
    vcont[a]=(prop)/n
  }
    return(vcont)

}

estimaPzib=function(x,p,c1,c2,u)

{

        estPzib=0

for (b in 1:length(x))

  {

if (x[b]>=c1 && x[b]<=c2)      rho[b]=x[b]*log(x[b])

if(x[b]<c1)                    rho[b]=((c1^(1-u)*log(c1)+((1-u)*log(c1)+1)
                               *(c1^(1-u)/u))*x[b]^u)-(((1-u)*log(c1)+1)*c1/u)

if(x[b]>c2)                    rho[b]=((c2^(1-u)*log(c2)+((1-u)*log(c2)+1)
                               *(c2^(1-u)/u))*x[b]^u)-(((1-u)*log(c2)+1)*c2/u)

    auxPzib=rho[b]*p[b]

    estPzib=auxPzib + estPzib

  }

   return (estPzib)

}

for (j in 1:nsim)
```

```
{
 estpsi=0

 estmvp=as.double(c(seq(npop)))  ;  est=c(seq(npop))

  for (a in 1:npop)

  {

     auxsoma=0 ; auxfn=0 ; auxxa=0 ; auxest=0

     parpop=pr[a]

     amos=rzibinom(nfixo,m,parpop,phi=gama)

     auxsoma=sum(amos)

     while (auxsoma==0)  amos=rzibinom(n,m,parpop,phi=gama)

     resfnk=fnk(m,nfixo,amos,k)

     auxest=sum(k*resfnk)/m

     estmvp[a]=auxest              #### Estima fn(k) da k-ésima pop ####

     prob=dbinom(k,m,auxest)

     xa=resfnk/prob

     s=u[a]

     est[a]= estimaPzib(xa,prob,c1,c2,s)

  }

# ################ Estimação das funções lineares (psi) ############### #
```

```
    estpsi=coef%*%est
# ############################ IC de wald ######################### #
    qtde=sum((coef^2*est*(1-est)/n))
    ls=estpsi+(alfa2*sqrt(sum(qtde)))
    li=estpsi-(alfa2*sqrt(sum(qtde)))
    amp=ls-li
    resul[j,5]=amp        # armazena amplitude
    resul[j,2]=ls         # armazena limite superior
    resul[j,3]=li         # armazena limite inferior
    resul[j,4]=estpsi     # armazena estimativa de psie
    resul[j,1]=j          # armazena o número da simulação
  if (psi>=li && psi<=ls) contaw=contaw+1
  }
  pc=contaw/nsim  ; mediapsi=mean(resul[,4])  ; amp_psi=mean(resul[,5])

# ################## Resultados a serem observados ################## #
  pc                    # probabilidade de cobertura da abordagem
  mediapsi              # estimativa média de psi
  amp_psi               # amplitude dos intervalos
```

```
########################## Fim do programa ########################
```

ANEXO C - Programa para estimar as funções lineares binomiais, em amostras inflacionadas de zeros, para o método Wilson

```
library(VGAM)


# ############## parâmetros de simulação a serem alterados ###############

  pr=c(0.3,0.3,0.3,0.3)          # parâmetros da familia binomial

  gama=0.20                      # porcentagem de excesso de zeros

  npop=4                         # número de populações

  coef=c(3,-1,-1,-1)             # coeficientes

  n=c(30,30,30,30)               # tamanho amostral para cada população

  nfixo=30                       # obs: valor igual ao n

  c1=0.44 ; c2=1                 # constantes de afinidade

  alfa2=abs(qnorm(0.025))     # nível de significância

  u=c(1, 1, 1, 1)      # constantes de afinidade para o método de estimação holder

# ###################################################################
```

```
# ################### parâmetros de simulação fixos ################## #
   nsim=200

   m=100

   psi= coef%*%pr

# ################################################################### #

# ################### DEFINIÇÃO DAS FUNÇÕES ################# #

vcont=c(rep(0,m+1))

k=seq(0,m)

rho=c(rep(0,m+1))

contaw=0

resul=matrix(0,nsim,5)


# ############# Calculo do fn(k) ############### #

fnk=function(m,n,dados,vet)
{
  for (a in 1:(m))
  {
```

```
    prop=0
    aux=vet[a]

 for (b in 1:n)

 {
    if (aux==dados[b]) prop=prop+1
 }
    vcont[a]=(prop)/n

 }

     return(vcont)
}

estimaPzib=function(x,p,c1,c2,u)

{
          estPzib=0

 for (b in 1:length(x))

 {

if (x[b]>=c1 && x[b]<=c2) rho[b]=x[b]*log(x[b])
if(x[b]<c1)                    rho[b]=((c1^(1-u)*log(c1)+((1-u)*log(c1)+1)*(c1^(1-
u)/u))*x[b]^u)-(((1-u)*log(c1)+1)*c1/u)
```

```
if(x[b]>c2)                    rho[b]=((c2^(1-u)*log(c2)+((1-u)*log(c2)+1)*(c2^(1-
u)/u))*x[b]^u)-(((1-u)*log(c2)+1)*c2/u)

      auxPzib=rho[b]*p[b]

      estPzib=auxPzib + estPzib

  }

     return (estPzib)

}

for (j in 1:nsim)

{

 estpsi=0

 estmvp=as.double(c(seq(npop)))   ;   est=c(seq(npop)) ; li_inf=c(seq(npop)) ;
li_sup=c(seq(npop))
  for (a in 1:npop)

  {

    auxsoma=0 ; auxfn=0 ; auxxa=0 ; auxest=0
       parpop=pr[a]
```

```
    amos=rzibinom(nfixo,m,parpop,pstr0=gama)

    auxsoma=sum(amos)

    while (auxsoma==0)  amos=rzibinom(n,m,parpop,pstr0=gama)

    resfnk=fnk(m,nfixo,amos,k)

    auxest=sum(k*resfnk)/m

    estmvp[a]=auxest              #### Estima fn(k) da k-ésima pop ####

    prob=dbinom(k,m,auxest)

    xa=resfnk/prob

    s=u[a]

    est[a]= estimaPzib(xa,prob,c1,c2,s)

    li_sup[a]=est[a]+1.96*sqrt((est[a]*(1-est[a])/n[a]))
    li_inf[a]=est[a]-1.96*sqrt((est[a]*(1-est[a])/n[a]))

}


  # ########### Estimação das funções lineares (psi) ############# #
```

```
estpsi=coef%*%est

par_int=cbind(li_inf,li_sup)

menor=min(par_int[,1])

maior=max(par_int[,2])

aux_estpsi=coef*est

L=estpsi-sqrt(sum(aux_estpsi-menor)^2)

U=estpsi+sqrt(sum(aux_estpsi-maior)^2)

amp=U-L

resul[j,5]=amp      # armazena amplitude

resul[j,2]=U       # armazena limite superior

resul[j,3]=L       # armazena limite inferior

resul[j,4]=estpsi   # armazena estimativa de psie

resul[j,1]=j        # armazena o número da simulação

if (psi>=L && psi<=U) contaw=contaw+1
```

```
  }

  pc=contaw/nsim  ; mediapsi=mean(resul[,4])  ; amp_psi=mean(resul[,5])

### Resultados a serem observados

 ### Prob. de cobertura de holder

   pc

### Estimativa media de psi

   mediapsi

## Amplitude dos intervalos

  amp_psi

########################## Fim do programa ########################
#################################################################
```