



UNIVERSIDADE FEDERAL DE LAVRAS

**AVALIAÇÃO DE MÉTODOS DE ESTIMAÇÃO
DOS PARÂMETROS DA DISTRIBUIÇÃO
BETA-BINOMIAL VIA SIMULAÇÃO MONTE
CARLO**

IVANI DE OLIVEIRA NEGRÃO

2000

IVANI DE OLIVEIRA NEGRÃO

**AVALIAÇÃO DE MÉTODOS DE ESTIMAÇÃO DOS
PARÂMETROS DA DISTRIBUIÇÃO BETA-BINOMIAL VIA
SIMULAÇÃO MONTE CARLO**

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

Orientador

Prof. Luiz Henrique de Aquino

LAVRAS
MINAS GERAIS – BRASIL
2000

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

Negrão, Ivani de Oliveira

Avaliação de métodos de estimação dos parâmetros da distribuição Beta-Binomial via simulação Monte Carlo / Ivani de Oliveira Negrão. – Lavras : UFLA, 2000.

72 p. : il.

Orientador: Luiz Henrique de Aquino.

Dissertação (Mestrado) – UFLA.

Bibliografia.

1. Distribuição Beta-Binomial. 2. Índice de agregação. 3. Simulação Monte Carlo. I. Universidade Federal de Lavras. II. Título.

CDD-519.282

IVANI DE OLIVEIRA NEGRÃO

**AVALIAÇÃO DE MÉTODOS DE ESTIMAÇÃO DOS
PARÂMETROS DA DISTRIBUIÇÃO BETA-BINOMIAL VIA
SIMULAÇÃO MONTE CARLO**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Agronomia, Área de Concentração em Estatística e Experimentação Agropecuária, para obtenção do título de "Mestre".

Aprovada em 15 de fevereiro de 2000.

Prof. Ednaldo Carvalho Guimarães	UFU
Prof. Eduardo Bearzoti	UFLA
Prof. Lucas Monteiro Chaves	UFLA

Luiz Henrique de Aquino
Prof. Luiz Henrique de Aquino
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

À glória de meu Pai Celestial,
Aos meus pais, Guilherme e Vaste,
Aos meus irmãos Elias, Guilherme, Matheus e Otoniel,
Aos meus amigos, Sandro, Mical e Ademir,

DEDICO

AGRADECIMENTOS

A Deus pela força e inspiração.

À Universidade Federal de Lavras (UFLA), em especial ao Departamento de Ciências Exatas (DEX), pelo curso oferecido.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudo concedida.

Aos professores do DEX-UFLA, em especial aos meus orientadores, Luiz Henrique de Aquino e Eduardo Bearzoti, pelas valiosas orientações e amizade.

A todos os funcionários do DEX, na pessoa do Prof. Coordenador Joel Augusto Muniz, pela competência e prestabilidade.

Aos meus colegas de curso Andrea da Cruz Leonel, Carlos dos Santos, Letícia Lima Milani, Marcelo Gomes Judice e Mônica Aparecida Tomé, pelo companheirismo e carinho de verdadeiros irmãos.

Aos colegas da nova turma, em especial a Adriano, Ana Rita, Andrea, Moisés, Nagib e Yara, pela fraternidade que proporcionaram.

Ao colega Fabyano, que tanto contribuiu ao andamento prático da simulação dos dados utilizados neste trabalho.

Aos meus professores de graduação, Cícero, Ednaldo Guimarães, Luiz Alberto, Márcia Petronzio e Rúi Tojeiro, pela lição profissional que cada um me ensinou.

Aos colegas Edmilson e Lucas pela disponibilidade em conseguir e enviar-me os trabalhos, que foram muito valiosos.

A Cilinha, Sebastiana, Cidinha, Vera e Elza por terem sido como mães para mim durante os dois últimos anos.

Aos meus pais, Guilherme e Vaste, pelo carinho, estrutura financeira e emocional fornecidos ao longo de minha vida.

Aos meus irmãos Elias, Guilherme, Matheus e Otoniel por serem meus irmãozinhos queridos.

Aos meus primos Marcos e Rosane, que tanto contribuíram para a minha vida acadêmica, principalmente nos meus primeiros passos em Uberlândia.

Ao meu amigo Sandro Tillmann, pelos anos de amizade tão fundamental que tem me proporcionado, sobretudo por continuar presente mesmo à distância.

À minha amiga Mical Muniz, por ser a irmãzinha que não tive. Sua presença todos esses anos tem feito diferença.

Aos amigos Ademir, Paulo Caixeta, Aurélia Aparecida, Ester Solange, Iêda Vidal e Rose pelo carinho e longos anos de amizade.

SUMÁRIO

	Página
RESUMO.....	i
ABSTRACT.....	ii
1 INTRODUÇÃO.....	1
2 REFERENCIAL TEÓRICO.....	3
2.1 Caracterização e Aplicações da Distribuição Beta-Binomial.....	3
2.2 Ajuste, Reparametrização, Média e Variância da Distribuição Beta-Binomial (DBB).....	7
2.2.1 A Distribuição Beta-Binomial.....	7
2.2.2 Reparametrização da DBB.....	10
2.2.3 Média e Variância da DBB.....	12
2.3 Métodos de Estimação.....	16
2.3.1 Método dos Momentos.....	16
2.3.1.1 Ajuste da DBB pelo Método dos Momentos.....	18
2.3.2 Método da Máxima Verossimilhança.....	19
2.3.2.1 Ajuste da DBB pelo Método da Máxima Verossimilhança.....	20
2.3.3 Método de Médias e Zeros.....	24
2.3.4 Método da Análise de Variância (ANAVA).....	27
2.3.4.1 Ajuste da DBB pelo Método da ANAVA.....	30
2.4 Estatísticas de Avaliação de Estimadores.....	31
2.4.1 Viés.....	31

2.4.2 EQM.....	32
2.4.3 CS.....	33
2.4.4 DP.....	33
2.5 Simulações Prévias Envolvendo Comparação de Estimadores.....	34
3 MATERIAL E MÉTODOS.....	39
3.1 Simulação dos Dados.....	39
3.2 Estimação.....	40
3.3 Avaliação dos Estimadores.....	43
4 RESULTADOS E DISCUSSÃO.....	44
4.1 Estimadores de π	44
4.2 Estimadores de θ	51
5 CONCLUSÕES.....	64
5.1 Estimadores de π	64
5.2 Estimadores de θ	64
REFERÊNCIAS BIBLIOGRÁFICAS.....	66

RESUMO

NEGRÃO, Ivani de Oliveira. Avaliação de métodos de estimação dos parâmetros da distribuição beta-binomial via simulação Monte Carlo. Lavras: UFLA, 2000. 72p. (Dissertação - Mestrado em Agronomia/Área de concentração em Estatística e Experimentação Agropecuária)*

Objetivou-se avaliar os métodos dos momentos, máxima verossimilhança, médias e zeros e análise de variância na estimação dos parâmetros da distribuição beta-binomial, via simulação Monte Carlo. Para tanto, foram implementados programas em SAS®, em que amostras pseudo aleatórias de uma distribuição beta-binomial foram geradas e através de 1500 repetições de diferentes situações foram calculadas as estatísticas avaliadoras de estimadores: viés, erro quadrático médio, consistência simples e desvio padrão. Foi também aplicado o procedimento PROC UNIVARIATE nas sequências das estimativas com o objetivo de verificar a normalidade assintótica da distribuição dos estimadores. Pôde-se concluir que os estimadores da análise de variância e máxima verossimilhança foram os que apresentaram melhores propriedades conjuntamente. O estimador de momentos foi consistente a partir do tamanho de amostras vinte, enquanto os demais foram consistentes para qualquer valor de amostra maior que cinco.

* Comitê Orientador: Luiz Henrique de Aquino - UFLA (Orientador) e Eduardo Bearzoti - UFLA

ABSTRACT

NEGRÃO, Ivani de Oliveira. Evaluation of estimation methods of the beta-binomial distribution parameters via Monte Carlo simulation. Lavras: UFLA, 2000. 72p. (Dissertation - Master in Agronomy/Major in Statistics and Agricultural Experimentation) *

The methods of moments, maximum likelihood, means and zeros and analysis of variance were evaluated in the estimation of the beta-binomial distribution parameters, via Monte Carlo simulation. Programs in SAS[®] were implemented in which pseudo-random samples of a beta-binomial distribution were generated yielding 1,500 simulations of different situations; the evaluating statistics: bias, mean square error, simple consistency and standard deviation were calculated. The PROC UNIVARIATE was also used in the sequences of the estimates with the objective of verifying asymptotic normality of the distribution of estimators. The estimators of analysis of variance and maximum likelihood presented best properties jointly. The moments estimator was consistent for sample sizes greater than 20, while the others were consistent for any sample size greater than five.

* Guidance Committee: Luiz Henrique de Aquino - UFLA (Adviser) and Eduardo Bearzoti - UFLA

1 INTRODUÇÃO

A distribuição beta-binomial (DBB) é uma extensão da distribuição binomial (DB) em que o parâmetro p é permitido variar de grupo a grupo segundo uma distribuição beta. Sua aplicação em Fitossanidade tem sido amplamente empregada, pois dadas as peculiaridades de determinadas espécies de insetos e patógenos, a pressuposição de que p é constante para todo grupo não é válida.

Em termos da reparametrização sugerida por Griffiths (1973), os parâmetros da DBB são π e θ , em que π é a média de p , enquanto θ é uma medida de variação em p . Esses parâmetros possuem interpretações práticas bem objetivas em vários campos de estudo. Em Fitossanidade, por exemplo, π e θ significam índice ou razão de incidência e índice de agregação, respectivamente.

Alguns trabalhos que compararam métodos de estimação para a correlação intra-classe (ρ), que é uma função de θ , têm sido encontrados. No entanto, dada a diversidade das situações encontradas entre as diversas áreas, pode-se notar a preferência de determinados autores da mesma área pelos mesmos métodos. Em Fitossanidade, a maior parte dos trabalhos consultados empregaram o método da máxima verossimilhança, enquanto em Toxicologia e Teratologia, os métodos mais aplicados foram o da análise de variância (ANAVA) e o método dos momentos. Chatfield e Goodhardt (1970) utilizaram o método de médias e zeros na estimação dos parâmetros da DBB, em um estudo sobre o comportamento dos consumidores em compras. Recentemente, Ridout, Demétrio e Firth (1999) compararam, via simulação Monte Carlo, vinte métodos de estimação para a correlação intra-classe, considerando dados gerados pela DBB e duas outras distribuições binomiais mistas. Percebe-se,

assim, que o estudo da DBB e a comparação de métodos de estimação tem preocupado os pesquisadores, o que realça a importância do estudo.

Este trabalho visou investigar, via simulação Monte Carlo, a qualidade de quatro métodos de estimação dos parâmetros π e θ em termos do viés, EQM, desvio padrão e consistência simples, e ainda verificar a normalidade assintótica da distribuição dos estimadores e o seu desempenho.

2 REFERENCIAL TEÓRICO

2.1 Caracterização e Aplicações da Distribuição Beta-Binomial

Seja uma dada característica de uma população que possa ser representada conforme a Figura 1.

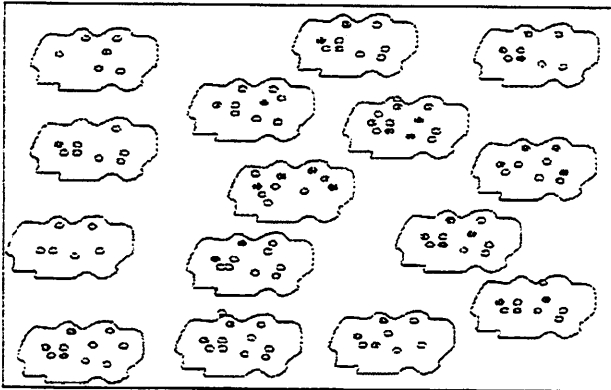


FIGURA 1. Representação de uma amostra aleatória.

A Figura 1 está representando uma amostra aleatória de uma população descrita por uma função de probabilidade $f(\cdot)$, em que foram escolhidos aleatoriamente m grupos (unidades amostrais) com n_i indivíduos cada. Foi considerado, aqui, o caso em que para cada indivíduo observado dentro de cada grupo, a resposta esperada é binária, mais especificamente 0 ou 1.

Estatisticamente, situações em que a uma variável aleatória são atribuídos apenas os valores 0 ou 1 são chamadas de ensaio de Bernoulli (Casela e Berger, 1990). O valor 1 está associado a “sucesso” e 0 a “fracasso”, que podem estar, por exemplo, representando a informação: uma planta observada

(ou parte de uma planta) está ou não atacada por um inseto. Na Figura 1, os indivíduos hachurados representam sucesso e os demais, fracasso.

Seja p a probabilidade de sucesso de um dado ensaio em qualquer grupo. Se n ensaios de Bernoulli são realizados em cada um de m grupos, então o número de sucessos y_i ($y_i = 0, 1, \dots, n_i$), por grupo, é descrito por uma distribuição binomial. No entanto, note que na Figura 1 há uma “região” em que a concentração de indivíduos apresentando sucesso é grande, enquanto em outra, é baixa ou nula. Este tipo de distribuição é classificada como padrão agrupado e ocorre com frequência, por exemplo, em Fitossanidade. Dadas as peculiaridades de determinadas espécies de insetos ou patógenos, a pressuposição de que p é constante para todos os grupos não é válida e logo a aplicação da distribuição binomial torna-se inadequada. Existem diversas distribuições binomiais mistas que visam contornar o problema da variabilidade de p , mas uma mais simples foi proposta por Skellam (1948), em que p é pressuposto variar de grupo a grupo de acordo com uma distribuição beta. A distribuição resultante denomina-se beta-binomial.

Em Fitossanidade, a aplicação da distribuição beta-binomial é muito ampla, pois segundo Madden e Hughes (1995), as variabilidades em p podem estar associadas ao hospedeiro, patógeno, condições ambientais ao longo do campo, ou até mesmo ao fato do processo de multiplicação do patógeno nos grupos ser proporcional à doença inicial. Os autores ainda ressaltam que esta não é uma condição apenas espacial, e consideram o exemplo de examinar sementes quanto à infecção por determinada doença. Neste caso, a probabilidade de y_i em n_i sementes analisadas estarem infectadas pode estar variando de planta a planta, em que cada planta é considerada um grupo. Muitos outros exemplos de aplicação da distribuição beta-binomial em Fitossanidade e na distribuição espacial de incidência de doenças podem ser encontrados em: Madden, Nault e

Murrall (1995), Hoffmann et al. (1996), Nault e Kennedy (1996), Madden, Hughes e Ellis (1995), Hughes, Madden e Munkvold (1996), Madden, Hughes e Munkvold (1996), Hughes et al. (1997) e Miller, Madden e Schmitthenner (1997).

Experimentos que visam testar o efeito de um novo medicamento também podem ser conduzidos como ensaios binários. Dois tratamentos experimentais são designados, avaliando-se a presença ou ausência de anormalidade após a aplicação do medicamento. Em geral, os testes são inicialmente realizados em ratos (ou outros roedores), em que cada leitegada é considerada um grupo, sendo que o surgimento de anormalidades pode ser mais comum em indivíduos de uma leitegada que de outra. Deste modo, a probabilidade p de um dado indivíduo que recebeu o medicamento apresentar anormalidades é descrita por uma distribuição beta (Zucker e Wittes, 1992). Outros exemplos de aplicações da distribuição beta-binomial em Toxicologia podem ser encontrados em: Kupper e Haseman (1978), Haseman e Kupper (1979) e Zucker e Wittes (1992).

Outras aplicações da DBB foram encontradas em Teratologia. Um exemplo dado em Paul (1982) pode ser ligeiramente descrito como segue: dados S_k tratamentos (incluindo um controle), uma pesquisa foi conduzida com o objetivo de verificar se os tratamentos aplicados eram (ou não) responsáveis pelo aparecimento de anormalidades em fetos. Foram designadas m_j fêmeas, cada uma prenha com n_j fetos para cada tratamento, em que a probabilidade p do feto i da fêmea j que recebeu o tratamento k apresentar anormalidade era pressuposto variar de fêmea a fêmea de acordo com uma distribuição beta. Neste caso o número de fêmeas por tratamento a ser testado foi diferente, bem como o número de fetos por fêmea. Traçando-se uma analogia com a Figura 1, seria como se a situação ali representada fosse repetida k vezes. A variável y_i

representou o número de fetos que apresentavam anormalidades em cada fêmea para cada tratamento. Assim, o efeito de cada tratamento foi analisado como se seguisse uma distribuição beta-binomial, e uma vez estimadas as proporções de anormalidades para cada tratamento, Paul (1982) apresentou um pseudo teste-*t* baseado no método de jackknife para testar a igualdade de duas proporções. Mais exemplos de aplicação da distribuição beta-binomial em Teratologia podem ser encontrados em: Elston (1977), Williams (1975), Kupper et al. (1986) e Williams (1988).

O uso da DBB em variáveis binárias correlacionadas como alternativa aos modelos de regressão foi discutido em Ochi e Prentice (1984) e Prentice (1986). Gange et al. (1996) sugeriram a DBB para a descrição de dados, em serviços de saúde, como uma alternativa ao uso de modelos logísticos de regressão. O modelo beta-binomial foi encarado pelos autores como uma extensão do modelo logístico no sentido de aquele incluir um parâmetro que media a correlação entre indivíduos dentro do mesmo grupo. No estudo, foi considerado um exemplo em que foram comparados o modelo logístico tradicional e o beta-binomial na avaliação do reflexo das mudanças na política de período de internação de pacientes em hospitais. Neste exemplo, cada indivíduo foi analisado como um grupo, e considerando que cada paciente permanecia n_i dias internado, a variável y_i representou o número de dias em que o paciente i ficava internado por alguma razão que poderia ter sido evitada. Pressupôs-se que havia uma correlação entre as respostas diárias para o mesmo paciente, e com a DBB representando uma descrição mais cautelosa dos dados, os autores concluíram que resultados incorretos poderiam ter sido indicados caso se tomassem como base apenas as análises através do modelo logístico.

2.2 Ajuste, Reparametrização, Média e Variância da Distribuição Beta-Binomial (DBB)

Este estudo ficou restrito ao caso em que a variável y_i está sendo observada em grupos de mesmo tamanho ($n_i = n$ para todo $i=1, 2, \dots, m$); porém, sua extensão para o caso de diferentes tamanhos de grupos é inteiramente análoga. Por facilidade de notação, entenda-se y por y_i . A apresentação da DBB a seguir foi atribuída a Skellam (1948), embora a notação utilizada esteja baseada em Griffiths (1973) e Casela e Berger (1990).

2.2.1 A Distribuição Beta-Binomial

Suponha que em uma distribuição binomial

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$$

a probabilidade de sucesso p esteja variando de grupo a grupo segundo uma distribuição beta

$$dF(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} dp, \quad 0 < p < 1, \quad \alpha, \beta > 0.$$

A distribuição resultante é chamada beta-binomial e tem por função de probabilidade:

$$f(y) = \binom{n}{y} \frac{B(\alpha+y, n+\beta-y)}{B(\alpha, \beta)}, \quad y=0, 1, \dots, n.$$

A justificativa desse resultado tem como base os conceitos de distribuição conjunta, condicional e marginal: dadas duas variáveis aleatórias y e p com distribuição conjunta $f(y, p)$, a distribuição condicional de y , isto é, a distribuição de y dado um valor fixo de p , é representada por $f(y|p)$ e definida como

$$f(y|p) = \frac{f(y, p)}{f(p)},$$

em que $f(p)$ é a distribuição marginal de p . Portanto, segue-se que

$$f(y, p) = f(y|p)f(p) \quad \text{e} \quad f(y) = \int_0^1 f(y|p)f(p) dp.$$

Logo, assumindo que $f(y|p)$ é uma binomial e que $f(p)$ é uma beta, segue-se que

$$\begin{aligned} f(y) &= \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} dp \\ &= \binom{n}{y} \frac{1}{B(\alpha, \beta)} \int_0^1 p^y (1-p)^{n-y} p^{\alpha-1} (1-p)^{\beta-1} dp \\ &= \binom{n}{y} \frac{1}{B(\alpha, \beta)} \int_0^1 p^{\alpha+y-1} (1-p)^{n+\beta-y-1} dp \end{aligned}$$

$$= \binom{n}{y} \frac{B(\alpha+y, n+\beta-y)}{B(\alpha, \beta)}, \quad y=0, 1, \dots, n.$$

em que

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx,$$

que também pode ser expressa em termos da função gama,

$$\Gamma(m) = \int_0^{\infty} x^{m-1} e^{-x} dx, \quad m > 0,$$

como segue:

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}, \quad m > 0 \text{ e } n > 0$$

A princípio, qualquer outra distribuição de probabilidade contínua com domínio no intervalo $(0, 1)$ poderia ser utilizada para descrever a variabilidade de p , mas a distribuição beta possui um forte apelo algébrico devido à definição da função beta facilitar a obtenção da função de probabilidade (Skellam, 1948). São os parâmetros α e β que determinam as formas da DBB, as quais coincidem com o mesmo padrão observado na distribuição beta, embora a distribuição beta seja contínua e a DBB discreta. Assim, a função de probabilidade pode ser estritamente crescente ($\alpha > 1, \beta = 1$), estritamente decrescente ($\alpha = 1, \beta > 1$), em forma de U ($\alpha < 1, \beta < 1$), ou unimodal ($\alpha > 1, \beta > 1$). O caso $\alpha = \beta$ incorre em uma distribuição simétrica, e $\alpha = \beta = 1$ recai em uma distribuição uniforme discreta (Casela e Berger, 1990).

2.2.2 Reparametrização da DBB

A fórmula obtida para a DBB pela definição é a de representação mais simples, mas Griffiths (1973) sugeriu uma reparametrização, para a mesma, em que os novos parâmetros possuem interpretações próprias nas áreas de aplicação. Os novos parâmetros são:

$$\pi = \frac{\alpha}{\alpha + \beta} \text{ e } \theta = \frac{1}{\alpha + \beta}, \quad 0 < \pi < 1, \quad \theta > 0.$$

Observando-se a média e variância de uma distribuição beta, nota-se que o novo parâmetro π representa exatamente a média de p , enquanto θ é uma medida de variação em p , que é múltipla de sua variância (Griffiths, 1973). Em estudos de pragas e doenças de plantas, π é chamado de índice de incidência e θ índice de agregação (Hughes e Madden, 1993). É importante ressaltar que enquanto medida de variação em p , θ pode ser visto como uma medida de variação entre grupos e, ainda, a DBB pressupõe que indivíduos dentro de grupos possuem mesma probabilidade de sucesso. Neste contexto, o coeficiente da correlação intra-classe ρ é um indicador da similaridade que existe entre indivíduos dentro de um mesmo grupo e pode ser expresso como (Ridout, Demétrio e Firth, 1999):

$$\rho = \frac{\theta}{1 + \theta}.$$

Visto que $\theta > 0$, ρ , para o caso beta-binomial, é sempre positivo, isto é, $0 < \rho < 1$.

Utilizando-se a função beta dada em termos da função gama mencionada anteriormente e o fato de $\Gamma(m) = (m-1)\Gamma(m-1)$, a DBB pode ser reescrita como:

$$\begin{aligned}
 f(y) &= \binom{n}{y} \frac{B(\alpha+y, n+\beta-y)}{B(\alpha, \beta)} \\
 &= \binom{n}{y} \frac{(\alpha+y-1)(\alpha+y-2)\dots\alpha(\beta+n-y-1)(\beta+n-y-2)\dots\beta}{(\alpha+\beta+n-1)(\alpha+\beta+n-2)\dots(\alpha+\beta)} \\
 &= \binom{n}{y} \frac{\prod_{i=0}^{y-1}(\alpha+i) \prod_{i=0}^{n-y-1}(\beta+i)}{\prod_{i=0}^{n-1}(\alpha+\beta+i)}, \quad y = 0, 1, \dots, n.
 \end{aligned}$$

com $\alpha, \beta > 0$.

Portanto, em termos dos parâmetros π e θ , a DBB fica:

$$\begin{aligned}
 f(y) &= \binom{n}{y} \frac{\prod_{i=0}^{y-1} \left(\frac{\pi}{\theta} + i \right) \prod_{i=0}^{n-y-1} \left(\frac{\chi}{\theta} + i \right)}{\prod_{i=0}^{n-1} \left(\frac{1}{\theta} + i \right)} \\
 &= \binom{n}{y} \frac{\frac{1}{\theta^y} \prod_{i=0}^{y-1} (\pi + i\theta) \frac{1}{\theta^{n-y}} \prod_{i=0}^{n-y-1} (\chi + i\theta)}{\frac{1}{\theta^n} \prod_{i=0}^{n-1} (1 + i\theta)}
 \end{aligned}$$

$$= \binom{n}{y} \frac{\prod_{i=0}^{y-1} (\pi + i\theta) \prod_{i=0}^{n-y-1} (\chi + i\theta)}{\prod_{i=0}^{n-1} (1 + i\theta)}, \quad y = 0, 1, \dots, n.$$

em que $0 < \pi < 1$, $\theta > 0$ e $\chi = 1 - \pi$

Além dos aspectos que facilitam a interpretação dos parâmetros, esta reparametrização é particularmente útil na obtenção da função de verossimilhança (Griffiths, 1973).

2.2.3 Média e Variância da DBB

Skellam (1948) deduziu uma fórmula geral para os momentos de uma DBB baseada nos momentos fatoriais em torno da origem das distribuições binomial e beta, dada como:

$$\mu_j = n^j \frac{B(j + \alpha, \beta)}{B(\alpha, \beta)}.$$

No entanto, utilizaram-se, aqui, os conceitos de esperança e variância de distribuições mistas e modelos hierárquicos observados em Casela e Berger (1990), que podem ser expressos no seguinte teorema.

Teorema 1.: Se X e Y são duas variáveis aleatórias, então

$$E(X) = E(E(X/Y))$$

e

$$\text{Var}(X) = E(\text{Var}(X/Y)) + \text{Var}(E(X/Y)),$$

dados que as esperanças existem.

As provas para este teorema podem ser encontradas em Casella e Berger (1990, p.154 e 158).

Considerando $X=y$ e $Y=p$ no Teorema 1, segue-se que a média e variância da DBB são, respectivamente,

$$\begin{aligned}\mu &= E(y) = E(E(y/p)) \\ &= E(np) \\ &= nE(p) \\ &= n \frac{\alpha}{\alpha + \beta}.\end{aligned}$$

e

$$\begin{aligned}\sigma^2 &= \text{Var}(y) = E(\text{Var}(y/p)) + \text{Var}(E(y/p)) \\ &= E(np(1-p)) + \text{Var}(np) \\ &= nE(p(1-p)) + n^2 \text{Var}(p)\end{aligned}$$

Mas

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

e

$$\begin{aligned} E(p(1-p)) &= \frac{1}{B(\alpha,\beta)} \int_0^1 p(1-p)p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 p^\alpha (1-p)^\beta dp \\ &= \frac{B(\alpha+1,\beta+1)}{B(\alpha,\beta)} \\ &= \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} \text{ (Casela e Berguer, 1990).} \end{aligned}$$

Portanto,

$$\begin{aligned} \sigma^2 &= n \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + n^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{n\alpha\beta(\alpha+\beta) + n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

Substituindo α e β pelos seus correspondentes em termos de π e θ , tem-se que:

$$\mu = n \pi$$

e

$$\sigma^2 = n \pi \chi \frac{(1+n\theta)}{(1+\theta)}, \text{ em que } \chi = 1 - \pi.$$

Assim, comparando-se os momentos populacionais média e variância das distribuições binomial e beta-binomial, observa-se que a média de uma DBB é igual à média de uma DB e as variâncias diferem entre si por um termo maior que 1, que é uma função de θ . Escrita como função de ρ , Hughes e Madden (1995) explicaram a variância da DBB como a variância de uma DB multiplicada por um ‘fator de heterogeneidade’ dado por $1 + \rho(n - 1)$. Portanto, o parâmetro θ é um indicador do quanto os dados se afastam de uma DB, sendo que o caso extremo $\theta = 0$ representa a própria DB. Gange et al. (1996) mencionaram que devido ao fator $1 + \rho(n - 1)$ ser maior ou igual a 1, ele modela a superdispersão devida à correlação dentro de grupo. Na prática, dependendo da interpretação que o pesquisador deseja fornecer, a DBB tanto pode ser dada em termos dos parâmetros π e θ como em termos de π e ρ . O primeiro caso geralmente é encontrado em trabalhos que estudaram incidências de doenças ou pragas (por ex., Hoffmann et al., 1996), e o segundo em trabalhos em Toxicologia e Teratologia (por ex., Zucker e Wittes, 1992 e Kupper et al., 1986).

2.3 Métodos de Estimação

2.3.1 Métodos dos Momentos

O método dos momentos talvez seja o mais antigo método de estimação, tendo sido registrado por Karl Pearson no final do século XIX (Casela e Berger, 1990). Este método tem a vantagem de apresentar uma técnica simples, visto que a obtenção dos estimadores consiste basicamente em resolver o sistema obtido ao se igualarem momentos amostrais aos respectivos momentos populacionais.

Seja uma amostra y_1, y_2, \dots, y_m de uma população com função de probabilidade $f(y | \theta_1, \theta_2, \dots, \theta_k)$, ou simplesmente $f(\cdot)$. Os estimadores de momentos são aqueles obtidos ao se resolver, em termos dos parâmetros, o seguinte sistema:

$$\begin{cases} m_1 = \mu_1 \\ m_2 = \mu_2 \\ \vdots \\ m_k = \mu_k \end{cases}$$

em que,

$$m_j = \frac{1}{m} \sum_{i=1}^m y_i^j \quad \text{e} \quad \mu_j = E(y^j),$$

com $j=1,2,\dots,k$.

Portanto, observa-se que m_j é uma função da amostra, enquanto μ_j é uma função dos parâmetros $\theta_1, \theta_2, \dots, \theta_k$, e o momento amostral de ordem j é a média amostral.

Pode-se mostrar facilmente que:

$$E(m_1) = \mu$$

e

$$V(m_1) = \frac{\sigma^2}{m}$$

em que μ e σ^2 representam, respectivamente, a média e a variância populacionais. Portanto, m_1 é um estimador cuja distribuição possui média μ e variância σ^2/m , em que m é o tamanho da amostra.

Na prática, o estimador de momentos mais utilizado é aquele que considera as observações corrigidas por sua média, denominado, na literatura, de momento amostral central e definido como:

$$M_m^j = \frac{1}{m} \sum_{i=1}^m (Y_i - m_1)^j, \quad j=1,2,\dots,k.$$

A partir do momento amostral central de ordem 2, M_m^2 , define-se o chamado estimador da variância amostral, dado por:

$$S^2 = \frac{m}{(m-1)} M_m^2.$$

A motivação para a definição de S^2 vem do fato de que sua esperança matemática representa a variância populacional, sendo inclusive o estimador utilizado pelo PROC MEANS do programa SAS®.

2.3.1.1 Ajuste da DBB pelo Método dos Momentos

Suponha que $f(\cdot)$ seja descrita pela distribuição beta-binomial. Assim os estimadores de momentos para os parâmetros da DBB são obtidos ao se resolver o sistema

$$\begin{cases} m_1 = \mu \\ S^2 = \sigma^2 \end{cases}$$

em que m_1 e S^2 representam, respectivamente, a média e variância amostral

Portanto, de

$$\begin{cases} m_1 = n \pi_m \\ S^2 = n \pi_m \chi_m \frac{(1 + n\theta_m)}{1 + \theta_m} \end{cases}$$

tem-se que:

$$\begin{cases} \pi_m = \frac{m_1}{n} \\ \theta_m = \frac{S^2 - n \pi_m \chi_m}{n \pi_m \chi_m - S^2} \end{cases}$$

que são os estimadores de momentos para π e θ .

2.3.2 Método da Máxima Verossimilhança

O método da máxima verossimilhança tem um apelo intuitivo desejável para todo estimador. Dada uma amostra aleatória de uma população, descrita por uma função de probabilidade $f(\cdot)$, o método da máxima verossimilhança fornece, sempre que possível, estimativas dos parâmetros de $f(\cdot)$, de tal modo que o valor da função de verossimilhança nessas estimativas representa a probabilidade máxima de a amostra ter sido retirada daquela população. Em outras palavras, dada uma amostra aleatória y_1, y_2, \dots, y_m , de uma população descrita por $f(\cdot)$, sempre que existirem os valores de $\theta_{1mv}, \dots, \theta_{kmv}$ que maximizam a função

$$L(\theta | Y) = \prod_{i=1}^m f(y_i | \theta_1, \dots, \theta_k),$$

estes são os estimadores de máxima verossimilhança de $\theta_1, \dots, \theta_k$.

Casela e Berger (1990) ressaltam que na prática existem dois fatos inerentes ao método da máxima verossimilhança. O primeiro está relacionado com o problema de se encontrar o máximo global de uma função, que muitas vezes, dependendo do modelo $f(\cdot)$, exige procedimentos numéricos. O segundo é da sensibilidade numérica da função $L(\theta | Y)$, visto que ela é dada pelo produto das distribuições marginais de cada y . Isto é, quão sensíveis são as estimativas a pequenas mudanças nos dados? Pode ocorrer que uma pequena mudança na amostra gere estimativas de máxima verossimilhança totalmente diferentes, tornando o uso de seus estimadores suspeitos, no caso de “pequenas” amostras.

Para encontrar os máximos da função de verossimilhança, é comum tomar a função auxiliar, ou suporte, $\ln(L(\theta|Y))$, que facilita o tratamento algébrico dos procedimentos de maximização. Para encontrar os estimadores de máxima verossimilhança (EMV), seguem-se os mesmos procedimentos adotados para o cálculo de pontos de máximos de funções, que são, basicamente, i) obter a equação ou o sistema de equações de máxima verossimilhança e resolvê-lo em termos dos parâmetros, determinando, assim, os possíveis pontos críticos; ii) obter a derivada segunda ou o sistema de derivadas segundas e fazer os testes necessários para verificar se os pontos críticos encontrados são pontos de máximo, mínimo ou sela. Aquele que for o máximo global representará a estimativa de máxima verossimilhança (Casela e Berger, 1990).

2.3.2.1 Ajuste da DBB pelo Método da Máxima Verossimilhança

Para a obtenção da função de verossimilhança, pode-se adotar os procedimentos indicados em Griffiths (1973). Supondo que $f(\cdot)$ é uma DBB, o autor utiliza a sua forma reparametrizada, e ainda a definição de frequência observada, para abreviar a representação da função suporte.

Definindo f_y ($y=0, 1, \dots, n$) como as frequências observadas para o valor y , a função de verossimilhança é, então, dada por:

$$L(\pi, \theta) = \prod_{y=0}^n [f(y)]^{f_y} .$$

Seja ainda,

$$S_i = \sum_{y=0}^i f_y \quad (i=0, 1, \dots, n),$$

de tal modo que $S_n = m$ representa o tamanho da amostra. Portanto, com essas notações, a função suporte de verossimilhança é dada por:

$$\ln L = \ln \left\{ \prod_{y=0}^n [f(y)]^{f_y} \right\}$$

$$= \sum_{y=0}^n f_y \ln[f(y)]$$

$$= \sum_{y=0}^n f_y \ln \left[\binom{n}{y} \frac{\prod_{y=0}^{y-1} (\pi+i\theta) \prod_{y=0}^{1-y-1} (\chi+i\theta)}{\prod_{y=0}^{y-1} (1+i\theta)} \right]$$

$$= C - S_n \sum_{i=1}^{n-1} \ln(1+i\theta) + \sum_{i=1}^{n-1} \left\{ (S_n - S_i) \ln(\pi+i\theta) + S_{n-1-i} \ln(\chi+i\theta) \right\}$$

em que

$$C = \sum_{y=0}^n f_y \ln \binom{n}{y} \text{ e } \chi = 1 - \pi$$

Utilizando-se os conceitos de derivadas parciais, tem-se:

$$\begin{aligned} \frac{\partial \ln L}{\partial \pi} &= \sum_{i=0}^{n-1} \frac{\partial}{\partial \pi} \left[C - S_n \sum_{i=1}^{n-1} \ln(1+i\theta) \right] \\ &+ \sum_{i=0}^{n-1} \frac{\partial}{\partial \pi} \left[\sum_{i=1}^{n-1} \left\{ (S_n - S_i) \ln(\pi+i\theta) + S_{n-1-i} \ln(\chi+i\theta) \right\} \right] \\ &= \sum_{i=0}^{n-1} \left[\frac{S_n - S_i}{\pi+i\theta} - \frac{S_{n-1-i}}{\chi+i\theta} \right] \end{aligned}$$

e

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta} &= \sum_{i=0}^{n-1} \frac{\partial}{\partial \theta} \left[C - S_n \sum_{i=1}^{n-1} \ln(1+i\theta) \right] \\ &+ \sum_{i=0}^{n-1} \frac{\partial}{\partial \theta} \left[\sum_{i=1}^{n-1} \left\{ (S_n - S_i) \ln(\pi+i\theta) + S_{n-1-i} \ln(\chi+i\theta) \right\} \right] \\ &= \sum_{i=0}^{n-1} i \left[\frac{S_n - S_i}{\pi+i\theta} + \frac{S_{n-1-i}}{\chi+i\theta} - \frac{S_n}{1+i\theta} \right] \end{aligned}$$

Devido ao fato de o sistema

$$\begin{cases} \frac{\partial \ln L}{\partial \pi} = 0 \\ \frac{\partial \ln L}{\partial \theta} = 0 \end{cases}$$

ser não-linear nos parâmetros π e θ , alguma técnica numérica torna-se necessária para resolvê-lo. Um método numérico frequentemente utilizado para

obtenção de soluções de equações não lineares (ou sistema de equações) é o método de Newton-Raphson, utilizado por diversos autores, como Griffiths (1973), Smith (1983) e Ridout, Demétrio e Firth (1999). No presente caso, trata-se de um sistema não linear de duas equações e duas variáveis (parâmetros) a serem determinadas.

Definindo f_1 como a primeira equação, e f_2 como a segunda, o método de Newton-Raphson fornece a seguinte solução para o sistema:

$$\begin{bmatrix} \pi^{k+1} \\ \theta^{k+1} \end{bmatrix} = \begin{bmatrix} \pi^k \\ \theta^k \end{bmatrix} - J^{-1} \begin{bmatrix} f_1(\pi^k, \theta^k) \\ f_2(\pi^k, \theta^k) \end{bmatrix}$$

em que J^{-1} representa a inversa da matriz jacobiana de f_1 e f_2 . A solução acima não representa necessariamente um ponto de máximo na superfície determinada pela função de verossimilhança, mas caso seja o máximo global, $\pi^k = \pi_{mv}$ e $\theta^k = \theta_{mv}$ são os estimadores de máxima verossimilhança de π e θ .

Uma sub-rotina em Fortran que fornece estimativas de máxima verossimilhança de π e θ foi dada em Smith (1983) e posteriormente modificada (Smith e Ridout, 1995). Já Madden (1994) propôs um programa computacional para o ajuste da DBB a dados de incidência de doenças utilizando, também, o método da máxima verossimilhança.

2.3.3 Método de Médias e Zeros

O método de médias e zeros para a estimação dos parâmetros da DBB foi descrito por Chatfield e Goodhardt (1970). Os autores justificaram seu uso em casos em que a DBB assume forma de U ($\alpha < 1$, $\beta < 1$). Suas justificativas fundamentaram-se nos resultados de Anscombe (1950), que mostrou que o método de médias e zeros é mais eficiente que o método dos momentos no ajuste de distribuições binomiais negativas em forma de J (J -shaped) reversas (alta incidência de zeros na amostra). A conjectura dos autores foi baseada no fato de a distribuição binomial negativa poder ser dada como limite da DBB, quando n e β tendem a infinito.

Em um conjunto de dados em que se observa uma alta incidência de zeros, o método de médias e zeros possui um forte apelo intuitivo, pois seus estimadores são obtidos de modo que as frequências esperada e observada na classe zero se igualem. Isto é, dadas estimativas iniciais dos parâmetros pelo método dos momentos, calcula-se a probabilidade esperada de zeros, comparando-a com a proporção observada de zeros. O método consiste de uma técnica iterativa simples e é apresentado a seguir. Utilizou-se a notação $P(\cdot)$ ao invés de $f(\cdot)$ para evitar confusões de notação de frequência observada e probabilidade teórica.

Sejam m_1 e S^2 , respectivamente, a média e a variância amostral, as estimativas de α e β , pelo método dos momentos, são dadas por:

$$\alpha_1 = \frac{-m_1 [S^2 - m_1(n - m_1)]}{nS^2 - m_1(n - m_1)}$$

$$\beta_1 = \frac{(n - m_1)}{m_1} \alpha_1$$

Calculando a probabilidade teórica de zeros utilizando os parâmetros α_1 e β_1 , tem-se:

$$P_1(0) = \frac{B(\alpha_1, n + \beta_1)}{B(\alpha_1, \beta_1)}$$

Comparando $P_1(0)$ com $\frac{f_0}{m}$, se:

$$P_1(0) > \frac{f_0}{m}$$

então os valores de α e β devem ser aumentados de tal forma que suas razões permaneçam constantes. Os autores sugeriram este acréscimo definindo:

$$\alpha_2 = 1,1\alpha_1 \quad \text{e} \quad \beta_2 = 1,1\beta_1.$$

Assim,

$$m_1 = n \frac{\alpha_1}{\alpha_1 + \beta_1} = n \frac{\alpha_2}{\alpha_2 + \beta_2}.$$

Além disso,

$$\frac{\beta_2 + k}{\alpha_2 + \beta_2 + k} < \frac{\beta_1 + k}{\alpha_1 + \beta_1 + k}, \quad k = 1, 2, \dots, (n-1)$$

e

$$\frac{\beta_2}{\alpha_2 + \beta_2} = \frac{\beta_1}{\alpha_1 + \beta_1}.$$

Portanto,

$$P_2(0) = \frac{B(\alpha_2, n + \beta_2)}{B(\alpha_2, \beta_2)} = \frac{(\beta_2 + n - 1)(\beta_2 + n - 2) \dots \beta_2}{(\alpha_2 + \beta_2 + n - 1)(\alpha_2 + \beta_2 + n - 2) \dots (\alpha_2 + \beta_2)} < P_1(0)$$

e, conseqüentemente, o acréscimo em α e β alterou a probabilidade teórica de zeros na direção desejada; isto é, com o aumento de α e β , a probabilidade teórica tende à proporção observada de zeros.

Caso contrário, se

$$P_1(0) < \frac{f_0}{m},$$

então os valores de α e β devem ser diminuídos, também, de modo que suas razões permaneçam constantes. O fator de redução sugerido foi 0,9. Assim,

$$\alpha_2 = 0,9\alpha_1 \quad \text{e} \quad \beta_2 = 0,9\beta_1.$$

Analogamente, justifica-se que esta redução em α e β faz com que a probabilidade teórica tenda à proporção observada de zeros.

As terceiras estimativas de α e β são obtidas por interpolação linear de modo que, definindo-se

$$z_1 = P_1(0) - \frac{f_0}{m}$$

$$z_2 = P_2(0) - \frac{f_0}{m}$$

e interpolando-se linearmente para $z_3=0$, encontra-se:

$$\alpha_3 = \alpha_1 + \frac{z_1(\alpha_2 - \alpha_1)}{z_1 - z_2}$$

$$\beta_3 = \frac{\alpha_3 \beta_1}{\alpha_1}$$

Portanto, os estimadores de médias e zeros para π e θ são:

$$\pi_{mz} = \frac{\alpha_3}{\alpha_3 + \beta_3} \quad \text{e} \quad \theta_{mz} = \frac{1}{\alpha_3 + \beta_3}$$

Na verdade, pode-se obter tantos acréscimos ou reduções nos valores de α e β quanto se deseje; porém, os autores afirmaram que estas estimativas obtidas são suficientemente acuráticas.

2.3.4 Método da Análise de Variância (ANAVA)

O método da ANAVA é bastante empregado em estudos de genética para estimar componentes de variância em análises realizadas em dados representados por variáveis contínuas.

Apresenta-se, a seguir, um modelo aleatório de análise de variância com um fator, como justificativa teórica do processo de estimação da correlação intra-classe. As abordagens aqui apresentadas estão essencialmente relacionadas com as colocações mais gerais sobre procedimentos de inferências para a correlação intra-classe descritos por Donner (1986).

Supondo que o j -ésimo indivíduo do i -ésimo grupo possa ser matematicamente descrito como

$$y_{ij} = \mu + a_i + e_{ij}, \quad i=1,2,\dots,m, \quad j=1,2,\dots,n$$

em que

μ = média populacional;

a_i = efeito do grupo i . Estes efeitos são tidos como identicamente e independentemente distribuídos com média 0 e variância σ_E^2 ;

e_{ij} = resíduos que são identicamente e independentemente distribuídos com média 0 e variância σ_D^2 ;

a_i e e_{ij} são independentes.

A variância de y_{ij} é dada por

$$\sigma_y^2 = \sigma_E^2 + \sigma_D^2$$

e a correlação intra-classe ρ , definida como

$$\rho = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_D^2}.$$

A estimação de ρ é retratada por Donner (1986) como um “subproduto” do processo de estimação dos parâmetros σ_E^2 e σ_D^2 através da tabela de análise de variância (Tabela 1) associada ao modelo acima descrito. Na prática, σ_E^2 e σ_D^2 são os parâmetros de maior interesse e de interpretação mais simples, enquanto ρ reflete apenas as suas contribuições em relação à variância total da característica sob estudo.

TABELA 1. Análise de variância para um modelo de efeitos aleatórios

CV	GL	SQ	QM	E(QM)
Entre grupos	$m - 1$	$n \sum_i (\bar{y}_i - \bar{y}_{..})^2$	QM_E	$\sigma_D^2 + n \sigma_E^2$
Dentro de grupos	$m(n - 1)$	$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$	QM_D	σ_D^2
Total	$mn - 1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$		

Visto que estimadores não viesados de σ_D^2 e σ_E^2 são dados, respectivamente, por

$$V_D = QM_D \quad \text{e} \quad V_E = \frac{QM_E - QM_D}{n},$$

o estimador de análise de variância, para ρ , é

$$\rho_{av} = \frac{QM_E - QM_D}{QM_E + (n - 1)QM_D}$$

A Tabela 1 apresenta as causas de variação e seus respectivos graus de liberdade, somas de quadrados, quadrados médios entre e dentro de grupos e as esperanças de quadrados médios. É importante observar que até este ponto não se especificou que tipo de distribuição seguem os dados; mas, em particular, se seguem uma distribuição normal, poder-se-ia aplicar o teste F para a verificação da existência da variabilidade entre grupos (Donner, 1986).

2.3.4.1 Ajuste da DBB pelo Método da ANAVA

O uso do método da ANAVA em dados binários foi sugerido por Elston (1977) para estimar a correlação intra-classe em um experimento em Teratologia. Posteriormente, Crowder (1978, 1979) justificou o método da ANAVA na estimação da correlação intra-classe em dados representados como proporções seguindo uma DBB. Desde então, diversos pesquisadores em Teratologia, Toxicologia e outras áreas vêm utilizando esta metodologia.

Supondo agora que a distribuição $f(\cdot)$ em questão seja uma beta-binomial, visto que o parâmetro ρ é uma função de θ , o método da ANAVA pode ser usado para estimar o parâmetro θ , enquanto π , segundo sugestão de Elston (1977), é estimado da mesma forma que no método dos momentos.

Relembrando que para uma beta-binomial, ρ é dado por

$$\rho = \frac{\theta}{1 + \theta},$$

segue-se que

$$\theta = \frac{\rho}{1 - \rho}.$$

Portanto, o estimador de análise de variância de θ é dado por

$$\theta_{av} = \frac{QM_E - QM_D}{n QM_D}.$$

2.4 Estatísticas de Avaliação de Estimadores

A princípio, não se sabe se os métodos apresentados no subtópico 2.3 fornecem os mesmos resultados na estimação dos parâmetros. Existem diversas técnicas que permitem avaliar o desempenho de estimadores e ainda quantificar

as diferenças entre eles, caso existam. Dentre estas técnicas, pode-se destacar as estatísticas viés, erro quadrático médio (EQM), consistência simples (CS) e desvio padrão (DP), cujas definições estão apresentadas a seguir e podem ser consultadas em Casela e Berger (1990).

2.4.1 Viés

O viés de um estimador pontual w , de um parâmetro ω , é a diferença entre o valor esperado de w e ω . Isto é, $\text{Viés}_\omega(w) = E_\omega(w) - \omega$.

Em particular, se o estimador tiver viés nulo, então $E_\omega(w) = \omega$.

2.4.2 EQM

O erro quadrático médio de um estimador pontual w , de um parâmetro ω , é a função definida por $E_\omega(w - \omega)^2$.

Como pode ser observado, o erro quadrático médio mede a diferença ao quadrado média do estimador w e o parâmetro ω . Existem outras funções que podem ser utilizadas para medir a distância entre o estimador e o parâmetro, como, por exemplo, $E_\omega(|w - \omega|)$. Porém, o EQM é preferível por duas razões: uma é o tratamento algébrico mais simples e a outra é o fato de o EQM ter a interpretação

$$E_\omega(w - \omega)^2 = \text{Var}_\omega(w) + [\text{Viés}_\omega(w)]^2.$$

Logo, o EQM incorpora duas componentes, em que a primeira é uma medida da variabilidade do estimador (precisão) e a segunda mede o viés

(acurácia). Portanto, um estimador que tem boas propriedades de EQM, tem variância e viés pequenos. Naturalmente, para encontrar estimadores com essas “boas propriedades” de EQM, é necessário procurar por aqueles que controlam bem tanto o viés quanto a variância. Em particular, se o estimador for não viesado, tem-se:

$$E_{\omega} (w - \omega)^2 = \text{Var}_{\omega} (w).$$

Isto é, se o estimador for não viesado, o EQM do estimador será igual à sua variância.

2.4.3 CS

Diz-se que um estimador w é consistente simples se para todo $\varepsilon > 0$, e todo $\theta \in \Theta$,

$$\lim_{m \rightarrow \infty} P[|w - \omega| < \varepsilon] = 1$$

Se um estimador w satisfaz a terceira definição, pode-se dizer que a probabilidade de a diferença entre w e o parâmetro ω ser tão pequena quanto se deseja é igual a um, desde que o tamanho da amostra seja suficientemente grande. Esta é também uma propriedade desejável para um estimador, visto que é razoável esperar que se o tamanho da amostra cresce, ela representará a população de forma mais fidedigna, no processo de estimação.

2.4.4 DP

O desvio padrão em Casela e Berger (1990) é definido, em termos do momento central de ordem 2 dado, como $\text{Var}(w) = E[w - E(w)]^2$. No entanto, em um processo de simulação, sua definição está baseada no conceito de variância amostral, como já apresentado no subtópico 2.2.1, e define-se desvio padrão de w como a raiz quadrada da variância amostral, dado por:

$$DP_{\omega}(w) = \sqrt{\frac{\sum w^2 - \frac{(\sum w)^2}{n}}{n-1}},$$

após aplicar n vezes o estimador ω .

Casela e Berger (1990) comentam que o DP tem a mesma interpretação que a variância no sentido de que se os valores médios de w forem pequenos, é plausível que w seja igual a $E(w)$ e, contrariamente, para valores maiores de w , espera-se uma maior variabilidade. O DP é, em geral, mais adotado por possuir uma interpretação mais simples, visto que sua unidade de medida coincide com a de w , enquanto a unidade da variância é dada como o quadrado da unidade de w .

Embora as definições dos subtópicos 2.4.1 e 2.4.2 tenham sido dadas em termos da esperança matemática, suas interpretações são entendidas por médias. Isto é, a esperança de um estimador é a média das estimativas obtidas ao se aplicar o referido estimador infinitas vezes, numa mesma situação.

2.5 Simulações Prévias Envolvendo Comparação de Estimadores

Donner e Koval (1980) compararam, via simulação Monte Carlo, o estimador da ANAVA e máxima verossimilhança para a correlação intra-classe utilizando a distribuição normal multivariada como a distribuição de probabilidade implícita. Os autores recomendaram o método da ANAVA em casos em que se tem algum conhecimento a priori se os valores da correlação intra-classe são baixos a moderados ($\rho < 0,5$), enquanto para valores altos de ρ , o método da máxima verossimilhança foi o mais adequado.

Visando estudar o comportamento de um novo estimador de momentos estabilizado (η_{stab} , $\eta = 1/\theta$), Tamura e Young (1987) conduziram um estudo de simulação utilizando a DBB, em que o objetivo principal era verificar a validade da incorporação de informações a priori sobre a proporção de tumores, em roedores, provenientes da aplicação de medicamentos teste. Com base no estudo de Tamura e Young (1986), foram geradas 100 amostras pseudo aleatórias de uma DBB, em que os parâmetros foram: $\pi = 0,15$ combinado com $\eta = 40$ (incidência de tumores moderada), $\pi = 0,02$ combinado com $\eta = 102$ (incidência de tumores baixa). Em ambas as situações foram utilizados $n = 20; 40; 60; 80$ e $m = 10; 20; 40$, e os resultados obtidos pelo estimador η_{stab} foram comparados com os de η_m e η_{mv} . Os autores concluíram que para valores do fator de estabilização entre 0,3 e 0,6, η_{stab} forneceu menor desvio médio absoluto (DMA) que os estimadores de momentos e máxima verossimilhança quando considerada a situação de baixa incidência de tumores e também na situação de incidência moderada para $m \leq 20$.

Feng e Grizzle (1992), motivados pelo problema de estimar o tamanho de amostras em estudos de grupos aleatorizados, compararam o método dos

momentos e da ANAVA na estimação de ρ . Para cada combinação dos parâmetros, π (0,05; 0,20; 0,35), ρ (0,00; 0,03; 0,06), n (30; 90; 270) e m (10; 30; 90), foram geradas 500 amostras pseudo-aleatórias de uma DBB. A avaliação dos estimadores foi feita através do viés, desvio padrão e consistência. Os resultados permitiram concluir que: i) os estimadores foram consistentes para m suficientemente grande ($m > 30$), embora o viés do estimador de momentos tenha sido muito menor que o do estimador da ANAVA para baixos m , ii) as variâncias de ambos foram semelhantes e iii) para n e m fixos, o viés aumentou à medida que ρ e π cresceram.

Yamamoto e Yanagimoto (1992) propuseram um estimador não viesado de momentos (ρ_{ub}) para ρ baseado numa equação de momentos não viesada e, via simulação, o compararam com outros três estimadores de momentos e com o de máxima verossimilhança. Entre os outros três estimadores de momentos está ρ_m , lá chamado estimador de momento central, e ρ_{stab} , estimador de momentos estabilizado de Tamura e Young (1987). Os parâmetros para a geração dos dados foram designados através de experiência prévia com estudos em Teratologia. Foram utilizados dois valores de π (0,05; 0,20), combinados com cinco valores de ρ (0,01; 0,02; 0,04; 0,08; 0,16), considerados baixos, e $\pi=0,40$ combinado com outros cinco valores de ρ (0,20; 0,25; 0,30; 0,35; 0,40), considerados altos. Ambas as situações utilizaram $m=20$ e $n=10$. A partir destes parâmetros, foram geradas 10.000 amostras pseudo-aleatórias de uma DBB e aquelas que conduziam a estimativas indeterminadas ou maiores que um, para ρ , eram excluídas do estudo, e as negativas eram igualadas a zero. Para os quatro estimadores, foram calculados média, viés médio e a razão $E=EQM/(MA)^2$, em que MA representa a média amostral do estimador. Os autores concluíram que as combinações de baixos π ($\pi \leq 0,20$) e ρ ($\rho < 0,08$)

conduziram a muitas estimativas negativas de ρ e, conseqüentemente, a vieses positivos para ρ_{mv} . Nos outros casos, o viés foi negativo para todos estimadores, exceto ρ_{stab} , que foi sempre positivo, e ρ_m , que foi negativo apenas em duas configurações ($\pi=0,05$ combinado com $\rho=0,08$ e $0,16$). O viés de ρ_{ub} foi, em geral, menor que o de ρ_{mv} e a eficiência correspondente, medida por E, foi também maior para ρ_{ub} . Com base nestes resultados, os autores concluíram que seu estimador é “melhor para uso prático”.

Lipsitz, Laird e Brennan (1994) conduziram um estudo de simulação com o objetivo de estudar métodos de estimação para o “coeficiente de concordância” κ , através da correlação intra-classe, em psiquiatria. Neste tipo de pesquisa, deseja-se quantificar a semelhança entre diagnósticos de psiquiatras para pacientes com distúrbios mentais. Os grupos foram constituídos de, no mínimo, dois psiquiatras, que avaliavam o paciente e atribuíam sucesso se o diagnóstico fosse neurose, e fracasso se o distúrbio fosse outro. Para o estudo, foram geradas 1825 amostras pseudo-aleatórias de uma DBB, a partir dos parâmetros $m=26$, $n_i = 3; 4; 5; 6$, $\pi=0,36$ e $\kappa=\rho=0,41$, que foram baseados no exemplo de Fleiss (1971). As estimativas de ρ foram obtidas por ρ_{mv} e outros três estimadores não-iterativos. Para a comparação dos estimadores, foram levados em conta o viés, a eficiência e a porcentagem de intervalos de confiança (95%) que continham o valor real de κ (ρ). Alguns problemas de convergências com o estimador ρ_{mv} foram encontrados em pequenas amostras, embora tenha apresentado a menor variância e 95% de seus intervalos de confiança continham o valor real de κ . Os resultados para os estimadores não-iterativos foram similares aos de máxima verossimilhança.

Kung-Jong, Cumberland e Kuo (1996) propuseram um intervalo de confiança para a correlação intra-classe (ρ) no caso beta-binomial. O intervalo

de confiança foi deduzido a partir dos momentos amostrais m_1 e S^2 e, através de um estudo via simulação Monte Carlo utilizando 10000 amostras (repetições), os autores compararam o intervalo proposto com o obtido pela técnica jackknife. Utilizando os valores paramétricos $m = 20; 50; 100$, $n = 5; 10; 15$, $\pi = 0,10; 0,20; 0,40$ e $\rho = 0,10; 0,20; 0,40$, dados eram simulados através de uma DBB e os intervalos de confiança obtidos, bem como seus respectivos níveis nominais de significância e comprimentos. Os autores concluíram que com $m \geq 50$ o intervalo de confiança proposto obteve um bom desempenho, exceto nas situações em que um alto valor de ρ era combinado com baixos π . Quando $m = 20$, o intervalo proposto ainda apresentou bons resultados nas situações em que ρ era pequeno.

Num amplo estudo, via simulação Monte Carlo, Ridout, Demétrio e Firth (1999) avaliaram vinte estimadores para a correlação intra-classe (ρ) em dados binários. Foram estudados entre vinte estimadores para ρ , os estimadores ρ_{mv} e ρ_{av} e os estimadores de momentos de Tamura e Young (1987) (ρ_{stab}), Kleinman (1973) (ρ_{keq} e ρ_{kpr}) e Yamamoto e Yanagimoto (1992) (ρ_{ub}). Ambos os estimadores ρ_{keq} e ρ_{kpr} utilizam o momento amostral central de ordem 2 como estimador da variância amostral e diferem entre si apenas pelo fato de ρ_{keq} pressupor $n_i = n$, para todo $i = 1, 2, \dots, m$. Os autores justificaram, analiticamente, que: i) para tamanhos de grupos iguais, os estimadores dentro dos conjuntos $\{\rho_{av}, \rho_{ub}\}$ e $\{\rho_{keq}, \rho_{kpr}\}$ são equivalentes e ii) os estimadores ρ_{av} , ρ_{kpr} e ρ_{ub} possuem o mesmo comportamento assintótico, isto é, são também equivalentes à medida que o número de grupos tende a infinito. Para as análises via simulação foram gerados dados pseudo-aleatórios de uma DBB a partir das combinações de três valores de π (0,05; 0,20; 0,50), cinco valores de

ρ (0; 0,05; 0,20; 0,50; 0,80), dois valores de m (10; 50), três distribuições de probabilidade (beta-binomial e duas outras distribuições mistas finitas) e dois arranjos de tamanhos de grupos obtidos dos estudos de Kupper et al. (1986) e Brass (1958), totalizando, assim, 180 configurações. Amostras que satisfaziam pelo menos uma das condições, i) $y_i = 0$ para todo $i=1, \dots, m$, ii) $y_i = n_i$ para todo $i=1, \dots, m$ e iii) $n_i = 1$ para todo $i=1, \dots, m$, eram eliminadas do estudo, e estimativas de ρ fora do espaço paramétrico (0,1) eram substituídas pelo extremo adequado. Este cuidado evitou problemas de estimadores indefinidos e os autores ainda argumentaram que a rejeição deste tipo de dado simulado é razoável, visto que não representa uma situação prática. Para avaliação da qualidade dos estimadores foram obtidas as estatísticas viés, DP e EQM, sendo que os EQM's foram utilizados para calcular a eficiência relativa do estimador em questão, dada pelo quociente $E = EQM(\rho_{mv}) / EQM(\text{estimador})$. Os resultados mostraram que as distribuições dos vieses da maioria dos estimadores foram assimétricas à esquerda, indicando que os estimadores foram negativamente viesados sob algumas condições de simulação e raramente apresentavam vieses positivos. A distribuição do desvio padrão foi assimétrica à direita para todos estimadores e os que apresentaram maior EQM foram os que apresentaram maior viés, sendo ρ_{stab} um deles. Os estimadores assintoticamente equivalentes, ρ_{av} , ρ_{kpr} e ρ_{ub} , comportaram-se de forma similar sob todas as condições de simulação e apresentaram o menor EQM juntamente com ρ_{mv} . Os autores ressaltaram, ainda, que quanto à eficiência, é um tanto difícil comparar os estimadores quando as diferentes distribuições de probabilidades são consideradas, e daí as dificuldades de se tirarem conclusões mais gerais.

3 MATERIAL E MÉTODOS

Os estudos propostos para os métodos de estimação, já descritos em Referencial Teórico, foram conduzidos via simulação Monte Carlo. A técnica em si tem como objetivo a geração de números pseudo-aleatórios segundo uma distribuição de probabilidade pela qual se tenha interesse. Este método tem sido amplamente empregado nas mais diversas áreas da ciência, pois representa, por exemplo, a possibilidade de se testar uma técnica de análise sem a necessidade de se coletarem dados reais (Moore e Tsiatis, 1991). Como vantagens, pode-se apontar uma maior abrangência de situações que podem ser estudadas e a redução nos custos inerentes a uma pesquisa. Embora estudos realizados via simulação permitam conclusões mais abrangentes, deve-se ressaltar, também, que é interessante aplicar as técnicas que foram testadas em dados reais como um comparativo com resultados prévios. Dentre os trabalhos consultados, que utilizaram este procedimento com a distribuição beta-binomial, incluem-se os de Yamamoto e Yanagimoto (1992) e Feng e Grizzle (1992). Neste estudo, foram analisados apenas resultados simulados.

3.1 Simulação dos Dados

Foram implementados quatro programas em SAS[®] com o objetivo de obter dados segundo a distribuição beta-binomial e, em seguida, o ajuste dos parâmetros da distribuição amostral obtida pelos métodos dos momentos, máxima verossimilhança, médias e zeros e ANOVA. Para isso, foram utilizados os valores paramétricos n (5; 10; 20), π (0,05; 0,25; 0,50) e θ (0,01; 0,10; 0,30; 0,50; 1,00; 3,00) e os números de grupos m (5; 10; 20; 80; 100), totalizando 270 configurações. Os valores paramétricos foram escolhidos segundo motivação prática em diversas áreas.

O procedimento de geração dos valores y está baseado no fato de, para um valor x entre zero e um, existir um y tal que $F(x)=y$. Assim, através da função *ranuni* do SAS®, valores r de uma distribuição uniforme eram gerados e comparados com os de uma DBB acumulada x . Se $r > x$, então um sucesso era acrescentado a y e continuava-se o processo até que os n “indivíduos” dentro daquele grupo fossem “observados”. Caso contrário, se $r < x$, o valor 0 era atribuído à variável y e prosseguia-se até que os m grupos fossem gerados. Com base em Yamamoto e Yanagimoto (1992) e Ridout, Demétrio e Firth (1999), amostras que satisfaziam pelo menos uma das condições abaixo eram eliminadas:

- i) $y_i = 0$ para todo $i=1, \dots, m$
- ii) $y_i = n$ para todo $i=1, \dots, m$.

A rejeição deste tipo de amostra justifica-se pelo fato dessas induzirem a estimativas indefinidas, e também pela abrangência prática do estudo. Se para os grupos observados não houver algum sucesso em algum deles, ou contrariamente, se todos os indivíduos dentro dos grupos apresentarem sucesso, não há necessidade prática da pesquisa. O processo de geração das amostras era repetido de tal forma que 1500 amostras úteis eram obtidas para cada combinação dos parâmetros. Estas podem ser entendidas como repetições das amostras geradas sob os mesmos parâmetros e são particularmente úteis por permitirem o cálculo das estatísticas utilizadas na avaliação da qualidade dos estimadores.

3.2 Estimação

O conjunto de valores estimados por cada método era analisado de modo

que estimativas de π e θ fora de seus respectivos espaços paramétrico eram eliminadas.

Os procedimentos do SAS[®] utilizados no processo de estimação foram: PROC SORT, PROC MEANS, PROC FREQ e PROC IML. O primeiro foi utilizado na ordenação dos dados obtidos em 3.1, o segundo no cálculo das somas, médias, variâncias e somas de quadrados não corrigidas, o terceiro na distribuição de frequências de y e o último permitiu a construção das estimativas obtidas para o cálculo do viés, EQM, CS e DP.

Após a ordenação dos dados pelo PROC SORT, eram obtidas, pelo PROC MEANS, as médias e variâncias das 1500 amostras (repetições). As fórmulas da média e variância utilizadas pelo PROC MEANS são, respectivamente, o momento amostral de ordem 1, m_1 , e a variância amostral S^2 , definidas no subtópico 2.3.1. Portanto, as estimativas pelo método dos momentos eram obtidas diretamente através da substituição das médias e variâncias amostrais nas fórmulas π_m e θ_m em 2.3.1.1, para cada uma das 1500 amostras. No caso do método da máxima verossimilhança, estas eram armazenadas em um outro banco de dados para serem posteriormente utilizadas como valores iniciais no processo iterativo.

Para o método da máxima verossimilhança, além das estimativas de momentos a serem utilizadas como valores iniciais no processo iterativo, foram também obtidas, através do PROC FREQ, juntamente com a opção *table y*, as frequências observadas na classe y , f_y ($y=0,1,\dots,n$) para cada amostra. No ambiente de matrizes PROC IML, os valores auxiliares S_i ($i=1,\dots,m$) eram obtidos e uma rotina análoga à encontrada no manual do PROC IML SAS[®], p. 114, foi construída para encontrar a resolução pelo método de Newton-Raphson do sistema de equações de verossimilhança. Foi estabelecido um número máximo de 100 iterações, e o critério de convergência adotado foi

$|w^{k+1} - w^k| < 0,001$, em que w^{k+1} e w^k eram as estimativas na iteração $k+1$ e k , respectivamente.

As estimativas pelo método de médias e zeros foram obtidas conforme o algoritmo em 2.3.4. Tal algoritmo foi implementado dentro de um *data set* específico em que as estimativas iniciais de α e β foram obtidas pelo método dos momentos.

Para o método da ANAVA, além de média e variância, também eram pedidas, no PROC MEANS, as somas e somas de quadrados não corrigidas de y para cada amostra. Através das fórmulas

$$QM_E = \frac{1}{(m-1)} \left[\frac{\sum_{i=1}^m y_i^2}{n} - \frac{\left(\sum_{i=1}^m \sum_{j=1}^n y_{ij} \right)^2}{m n} \right]$$

e

$$QM_D = \frac{1}{m(n-1)} \left[\sum_{i=1}^m \sum_{j=1}^n y_{ij}^2 - \frac{\sum_{i=1}^m y_i^2}{n} \right],$$

os quadrados médios entre e dentro de grupos eram obtidos e substituídos na fórmula de θ_{av} descrita em 2.3.4.1. Como já mencionado em 2.3.4.1, π_{av} foi calculado da mesma maneira que no método dos momentos.

3.3 Avaliação dos Estimadores

Ao final do processo de estimação para cada estimador, foram calculados, no ambiente PROC IML, as médias das estatísticas, viés, erros quadráticos médio (EQM), consistência simples (CS) e desvio padrão (DP), para cada configuração, ao longo das 1500 repetições. As fórmulas utilizadas foram as descritas em 2.4. Relembrando que embora as duas primeiras tenham sido dadas como esperanças, deve se ter em mente que a definição de esperança matemática é uma média ao longo de infinitas repetições da situação. Neste estudo, cada situação foi repetida 1500 vezes. Além disso, foi também aplicado o procedimento PROC UNIVARIATE nas estimativas de π , θ , viés e CS com o objetivo de avaliar suas distribuições, e também verificar a partir de que valor de m os estimadores eram normalmente distribuídos.

As análises através das estatísticas avaliadoras foram conduzidas observando-se seus comportamentos em quatro situações, nas quais foram calculadas as médias de cada uma das estatísticas para cada estimador, considerando-se i) as 270 configurações, ii) cada valor de n , iii) cada valor de m , e iv) cada valor de θ . A primeira situação forneceu um desempenho global dos estimadores e as médias em ii), iii) e iv) permitiram avaliar o desempenho dos estimadores para os diferentes valores de n , m e θ . Tendo-se em mente que se esperavam boas propriedades dos estimadores com o crescimento de m , procurou-se, também, verificar como se comportaram os estimadores com o crescimento de n , dado um valor fixo para m . Quanto ao parâmetro θ , relembrando que este mede o afastamento dos dados de uma distribuição binomial, foi também de interesse prático observar o que ocorre com as estatísticas avaliadoras à medida que θ cresce.

4 RESULTADOS E DISCUSSÃO

4.1 Estimadores de π

As Figuras 2, 3 e 4 mostram o viés, EQM e DP dos quatro estimadores propostos, calculados como a média das 270 configurações. Nota-se claramente que os estimadores π_{mz} e π_{av} são os que apresentaram menor viés, sendo o viés do primeiro negativo e do segundo, positivo. Os vieses de π_m e π_{av} foram, em geral, positivos, e os de π_{mv} , π_{mz} , negativos. Recordando que $\pi_m = \pi_{av}$, é importante ressaltar que embora o viés de π_m tenha sido maior que o de π_{av} , isso é devido apenas ao fato de os dados utilizados para cada método terem sido oriundos de sementes aleatórias. O DP de π_{mv} foi o menor, seguido de π_{mz} , enquanto o EQM de π_{mv} foi o maior, seguindo o mesmo comportamento do viés; isto é, estimador com viés mais alto apresentou maior EQM.

As estimativas médias das distribuições dos estimadores π_m (ou π_{av}), π_{mv} e π_{mz} pertenciam ao intervalo, ora determinado pelos quantis $\xi_{0,25}$ e $\xi_{0,50}$, ora determinados por $\xi_{0,50}$ e $\xi_{0,75}$, sendo que a primeira situação ocorreu com mais frequência quando os valores de θ foram combinados com $\pi=0,50$. Em ambas situações, as médias eram bem próximas da mediana ($\xi_{0,50}$). Alguns casos em que os intervalos foram determinados por $\xi_{0,75}$ e $\xi_{0,95}$ ocorreram para π_{mv} e π_{mz} em situações de baixos m ($m \leq 10$) combinados com $n=5$ e 20 e $\pi=0,05$. As distribuições dos vieses seguiram exatamente o mesmo comportamento da distribuição do estimador, sendo que algumas exceções foram encontradas para os estimadores dos vieses de π_{mv} e π_{mz} , visto que são, em geral, negativos e logo, nesses casos, se os intervalos de

π_{mv} ou π_{mz} fossem determinados por $\xi_{0,50}$ e $\xi_{0,75}$, o dos vieses eram por $\xi_{0,25}$ e $\xi_{0,50}$. A estatística consistência simples era reduzida à medida que m crescia e aumentada nas demais situações. Relembrando que esta estatística fornece a média dos módulos dos vieses, pode-se dizer que as estimativas tendem a se afastar de seu valor real à medida que os parâmetros crescem fixando-se m . Devido ao fato de CS tornar-se tão pequena quanto se deseje, bastando-se, para isso, apenas tomar um número de grupos (m) suficientemente grande, pode-se dizer que todos os estimadores estudados foram consistentes.

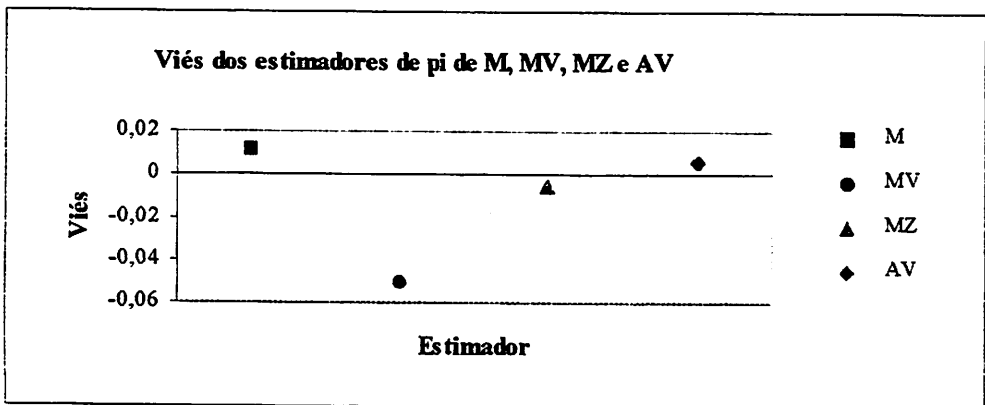


FIGURA 2. Viés médio dos estimadores de momentos (M), máxima verossimilhança (MV), médias e zeros (MZ) e análise de variância (AV) para as 270 configurações dos parâmetros.

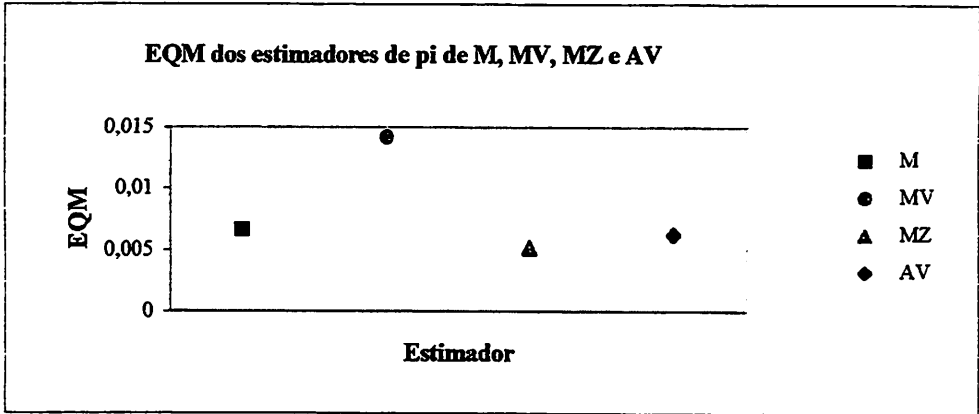


FIGURA 3. EQM médio dos estimadores de momentos (M), máxima verossimilhança (MV), médias e zeros (MZ) e análise de variância (AV) para as 270 configurações dos parâmetros.

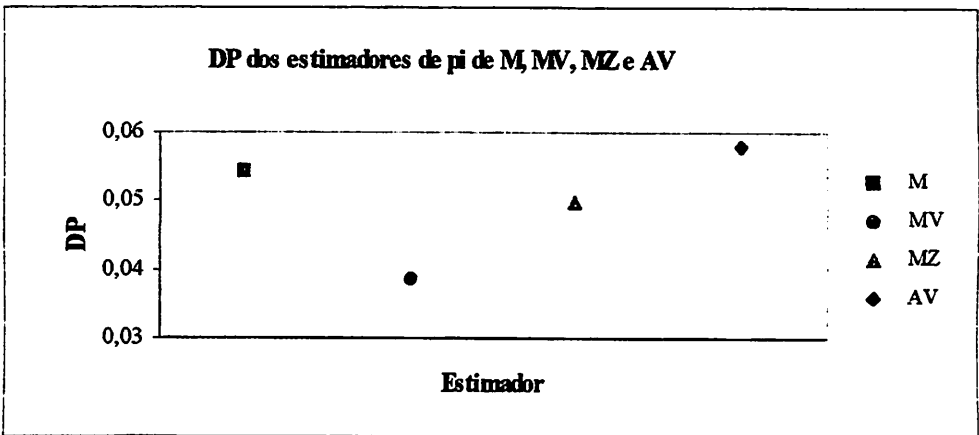


FIGURA 4. DP médio dos estimadores de momentos (M), máxima verossimilhança (MV), médias e zeros (MZ) e análise de variância (AV) para as 270 configurações dos parâmetros.

As Figuras 5, 6 e 7 ilustram o comportamento do viés, EQM e DP dos quatro estimadores de π , à medida que m cresce. Nota-se que para baixos m

($m < 20$), todas essas estatísticas são relativamente altas, tendendo a diminuir a partir de $m = 20$, sendo que, para $m \geq 80$, elas praticamente igualaram-se. Este resultado, em geral, é esperado para qualquer estimador, pois espera-se, com o crescimento de m , que a amostra torne-se mais representativa. Pode-se observar, ainda, que o DP de π_{mv} foi baixo, mesmo para baixos m . Observando a interpretação dada ao EQM de um estimador no subtópico 2.4, nota-se que o EQM de π_{mv} foi o mais alto devido apenas ao seu viés ter sido também o mais alto.

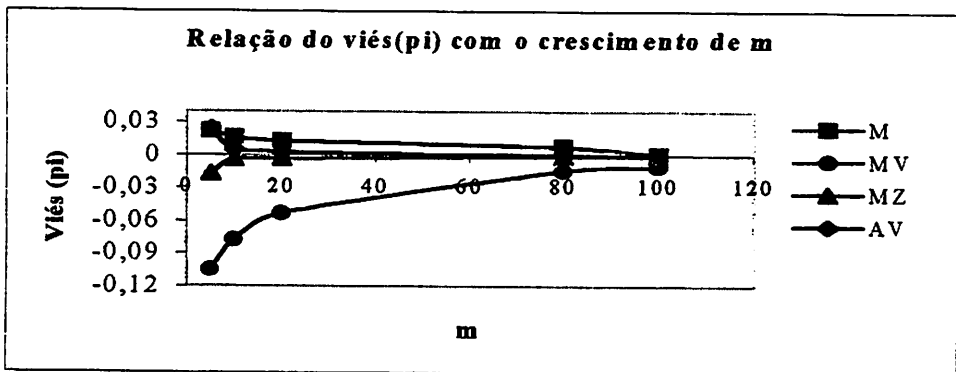


FIGURA 5: Relação do viés dos estimadores de π : π_m , π_{mv} , π_{mz} , π_{av} , com o crescimento de m .

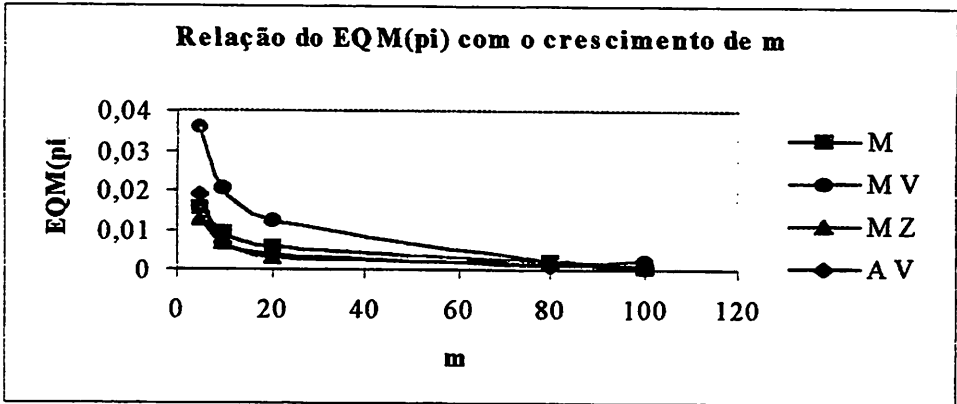


FIGURA 6. Relação EQM dos estimadores de π : π_m , π_{mv} , π_{mz} e π_{av} , com o crescimento de m.

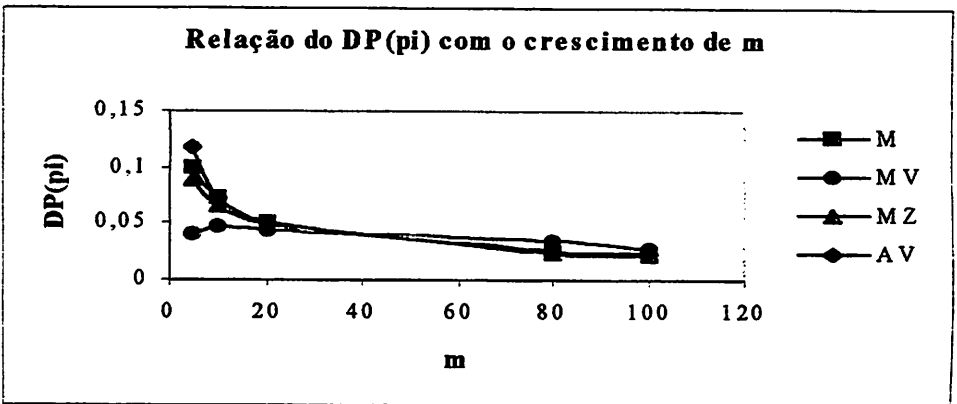


FIGURA 7. Relação DP dos estimadores de π : π_m , π_{mv} , π_{mz} e π_{av} , com o crescimento de m.

O teste de Shapiro-Wilk aplicado às sequências das 1500 estimativas de π_m nas 270 configurações simuladas acusou normalidade nas seguintes situações:

- i) $\theta=0,01$, $\pi=0,25$, $n=20$ e $m \geq 20$; $\pi=0,50$ $n=5$ e $m=100$,
 $n=10$ e 20 e $m \geq 20$.
- ii) $\theta=0,10$, $\pi=0,25$, $n=10$ e 20 e $m \geq 10$; $\pi=0,50$ $n=5$ e $m \geq 80$,
 $n=10$ e 20 e todo m .
- iii) $\theta=0,30$, $\pi=0,25$, $n=10, 20$ e $m \geq 20$; $\pi=0,50$ $n=5$ e 20 e
 $m \geq 20$, $n=10$ e $m \geq 80$.
- iv) $\theta=0,50$, $\pi=0,25$, $n=5, 10$ e $m=100$, $n=10$ e 20 e $m \geq 20$;
 $\pi=0,50$, $n=5, 10$ e 20 e $m \geq 20$
- v) $\theta=1,00$ e $3,00$, $\pi=0,25$, $n=5, 10$, e 20 e $m \geq 80$; $\pi=0,50$, $n=5$
e $m \geq 20$ e $n=10$ e 20 e $m \geq 10$.

Nota-se claramente que a normalidade foi detectada com maior frequência à medida que os parâmetros cresciam e, exceto para $\pi=0,05$, o crescimento de m induziu à normalidade da distribuição de π_m . Este resultado é importante quando se tem interesse na estimação por intervalo do parâmetro π . O teorema central do limite garante que independente da distribuição de uma variável aleatória com média μ e variância σ^2 , sua média terá distribuição normal com média μ e variância σ^2/m , quando m tende a infinito (Casela e Berger, 1990). Assim, supondo que y seja descrita por uma distribuição beta-binomial, $n\pi_m$ será normalmente distribuída com média μ e variância σ^2/m , desde que seja tomado m suficientemente grande. Portanto, pode-se dizer que π_m tem distribuição normal com média μ e variância σ^2/nm . Com base nisso, Madden, Hughes e Munkvold (1996) sugeriram intervalos de confiança

para π , dado por $\hat{\pi}_m \pm z_{1-p/2} S(\hat{\pi}_m)$, em que $z_{1-p/2}$ é o percentil $(1-p/2)100$ de uma distribuição normal padrão e $S(\hat{\pi}_m)$ é o erro padrão. Para pequenas amostras, os autores sugeriram que a distribuição normal fosse substituída pela distribuição t , com $m-1$ graus de liberdade.

O teorema da normalidade assintótica dos estimadores de máxima verossimilhança pode ser usado para construir intervalos de confiança para o estimador π_{mv} , utilizando a matriz de informação de “Fisher” para determinar as variâncias dos estimadores. Os resultados do teste de Shapiro-Wilk aplicado às sequências das 1500 estimativas, obtidas segundo π_{mv} para cada uma das 270 configurações, incorreu em normalidade nos seguintes casos:

- i) $\theta=0,01$, $\pi=0,50$, $n=20$ e $m \geq 20$.
- ii) $\theta=0,10$, $\pi=0,25$, $n=5$ e $m \geq 80$; $n=10$ e $m=100$; $n=20$ e $m \geq 20$; $\pi=0,50$ $n=5$ e $m=100$.
- iii) $\theta=0,30$, $\pi=0,25$ e $0,5$, $n=5$ e 10 e $m \geq 80$; $n=20$ e $m \geq 10$;
- iv) $\theta=0,50$, $\pi=0,25$, $n=5$ e $m=100$, $n=10$ e $m \geq 80$, $n=20$ e $m \geq 10$; $\pi=0,50$, $n=5, 10$ e 20 e $m \geq 20$
- v) $\theta=1,00$ e $3,00$, $\pi=0,05$, $n=10$ e 20 e $m \geq 80$; $\pi=0,25$ $n=10$ e $m \geq 80$, $n=20$ e $m \geq 10$; $\pi=0,50$, $n=10$ e $m \geq 10$, $n=20$ e todo m .

Os resultados obtidos aqui são semelhantes àqueles encontrados para π_m , e também aos fornecidos por π_{mz} . Ou seja, a normalidade assintótica foi detectada com maior frequência em configurações em que os valores reais dos parâmetros cresciam, sendo que, exceto para $\pi=0,05$, o crescimento de n e m

implicou em normalidade, o que era de se esperar. Ressaltando mais uma vez, os resultados para π_{av} não foram comentados porque são exatamente iguais aos de π_m .

As análises considerando o crescimento de n para valores fixos de m implicaram na redução do viés, DP, CS e EQM para o estimador $\pi_m = \pi_{av}$, em todos os casos. As mesmas estatísticas aumentaram nas mesmas situações para o estimador π_{mv} e apenas o viés do estimador π_{mz} . Considerando o crescimento de θ , todas as estatísticas aumentaram em todos os métodos, embora π_{mv} tenha se mostrado menos sensível a este crescimento que os outros estimadores.

4.2 Estimadores de θ

As Figuras 8, 9 e 10 ilustram o viés, EQM e DP dos estimadores θ_m , θ_{mv} , θ_{mz} e θ_{av} calculados como as médias das 270 configurações. Observa-se, através da Figura 8, que θ_{av} e θ_{mv} foram os estimadores que obtiveram o menor viés, sendo que o de θ_{mv} foi negativo, enquanto os dos demais foram positivos. Por meio da Figura 9', nota-se que os estimadores que obtiveram menor EQM foram θ_{mv} e θ_{av} , seguidos tanto para viés quanto DP por θ_{mz} . Observa-se, ainda, que o DP dos estimadores de θ (Figura 10) seguiu a mesma ordem que o EQM, situação contrária à observada para os estimadores de π .

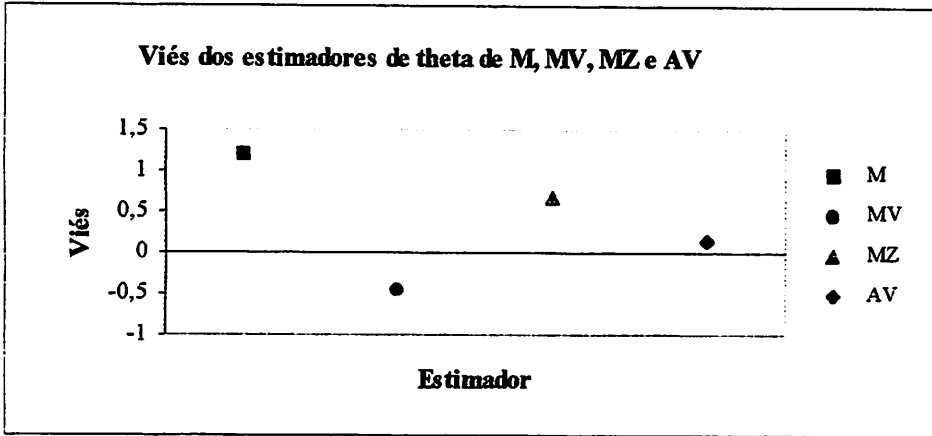


FIGURA 8. Viés médio dos estimadores de θ , θ_m , θ_{mv} , θ_{mz} e θ_{av} , para as 270 configurações dos parâmetros.

Considerando o fato de ρ ser uma função crescente de θ , Yamamoto e Yanagimoto (1992) argumentaram que era de se esperar que resultados obtidos para ρ fossem análogos aos respectivos para θ , ou seja, os estimadores de θ apresentariam o mesmo comportamento que os de ρ mostraram. Lá, a equação de momentos não-viesada que originou ρ_{ub} estava baseada nos mesmos estimadores de momentos aqui utilizados. Como já observado no subtópico 2.5, ρ_{ub} obteve viés e EQM relativamente baixos, e com base nisso esperava-se que θ_m também apresentasse baixos viés e EQM. Na verdade, isso ocorreu quando as configurações dos parâmetros não incluíam $\theta \geq 1,00$ combinados com $m < 80$. Vale ressaltar, ainda, que estimativas negativas neste estudo foram eliminadas, enquanto em Yamamoto e Yanagimoto (1992), as mesmas eram igualadas a zero. Isso pode ter provocado um aumento no viés e EQM globais nas situações em que um grande número de estimativas negativas eram encontradas.

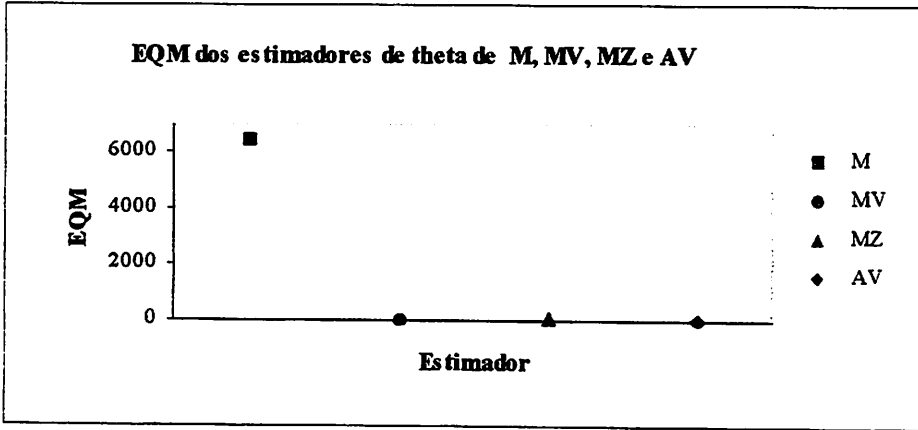


FIGURA 9. EQM médio dos estimadores de θ , θ_m , θ_{mv} , θ_{mz} e θ_{av} , considerando as 270 configurações dos parâmetros.

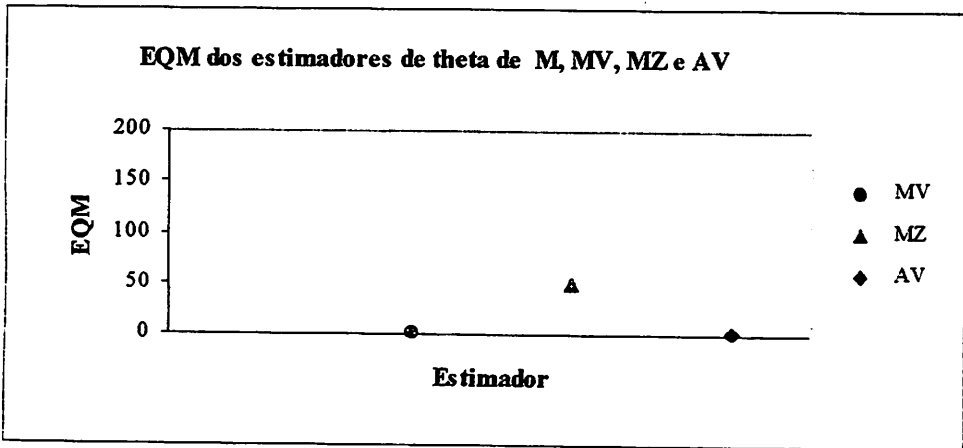


FIGURA 9'. FIGURA 9 ampliada.

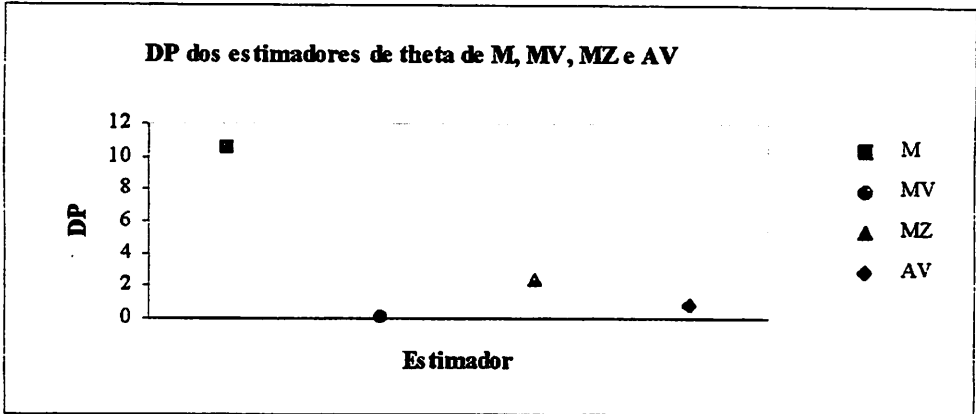


FIGURA 10. DP médio dos estimadores de θ , θ_m , θ_{mv} , θ_{mz} e θ_{av} , considerando as 270 configurações dos parâmetros.

A Figura 11 ilustra o comportamento do viés dos estimadores à medida que θ cresce. Observa-se que o viés de θ_m foi visivelmente baixo para $\theta \leq 0,5$, e a partir desse valor, aumentou consideravelmente. Este resultado condiz com os de Feng e Grizzle (1992), em que os vieses de ρ_m e ρ_{av} aumentaram com o crescimento de ρ . É importante mencionar que tanto o viés quanto o EQM e o DP de θ_m foram baixos, mesmo para $\theta \geq 1,00$, quando $m \geq 80$. Os altos valores observados para as estatísticas de avaliação foram devidos a valores discrepantes que ocorreram quando $m = 10$. Nota-se, ainda, que os vieses de θ_{mv} , θ_{mz} e θ_{av} também aumentaram com o crescimento de θ , mas θ_{mz} e θ_{av} pareceram menos sensíveis a este crescimento.

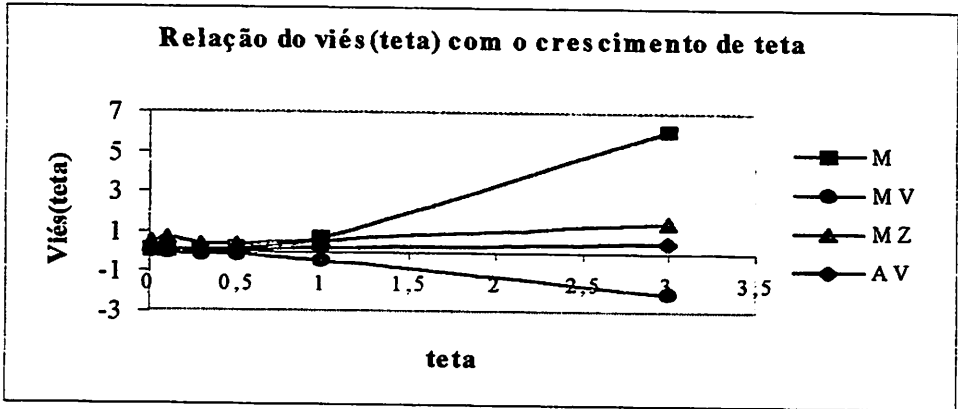


FIGURA 11. Relação do viés dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de θ .

A Figura 12 ilustra o efeito de θ no EQM. Novamente, o EQM de θ_m foi o mais alto, e através da Figura 12' nota-se que θ_{mv} foi o estimador que apresentou menor EQM, seguido dos estimadores θ_{av} e θ_{mz} , quando foi considerado o crescimento de θ . Através da Figura 13, nota-se que o DP de θ_{mv} foi o menor, seguido por θ_{av} e θ_{mz} , e para θ_m observa-se uma forte influência do parâmetro θ ; isto é, o crescimento de θ implicou em crescimento do DP para θ_m .

As Figuras 14, 15 e 16 ilustram a influência dos valores discrepantes nas estatísticas de avaliação para $m=10$, no estimador θ_m . Observa-se que as configurações que mais contribuíram para a “explosão” dos valores do viés, EQM e DP foram aquelas associadas a $m=10$. Na verdade, essa situação ocorreu também com valores de $m=5$ e 20 , mas como as figuras ilustram, os valores mais altos provieram de configurações que incluíam $m=10$.

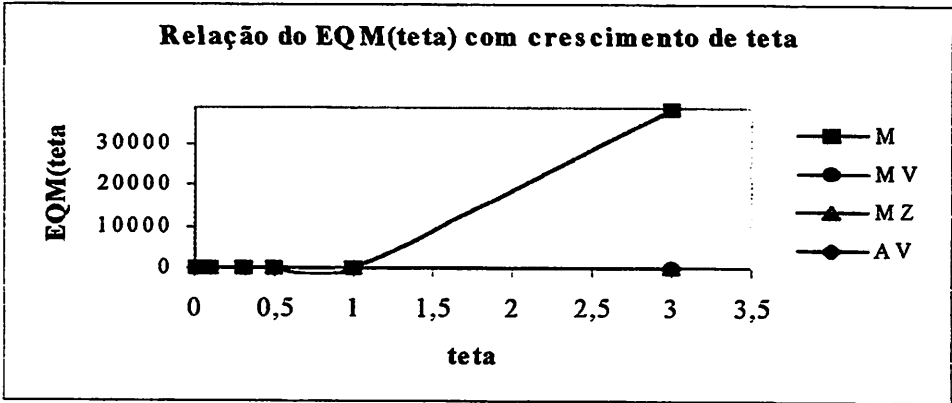


FIGURA 12. Relação do EQM dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de θ .

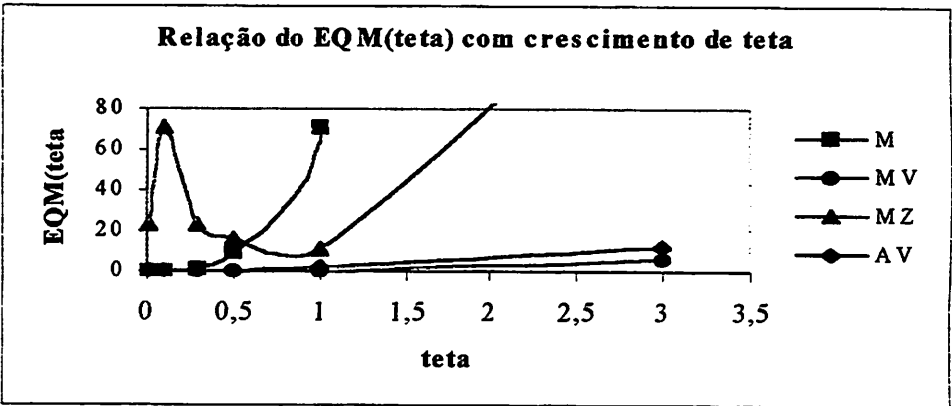


FIGURA 12'. FIGURA 12 ampliada

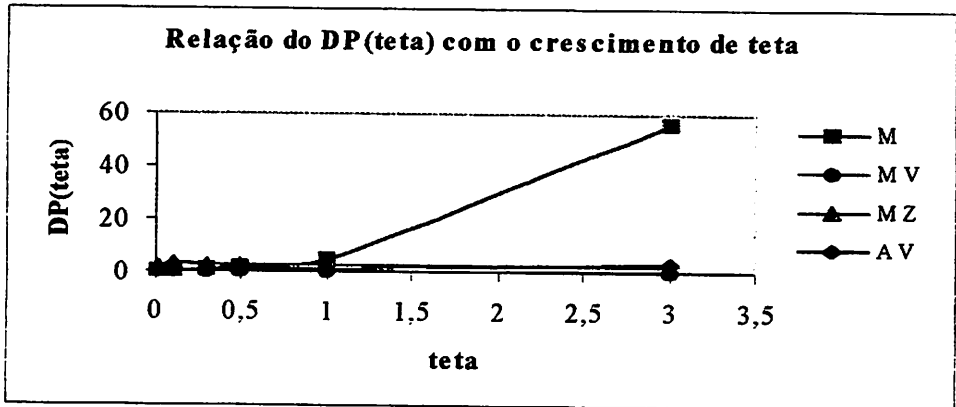


FIGURA 13. Relação do DP dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de θ .

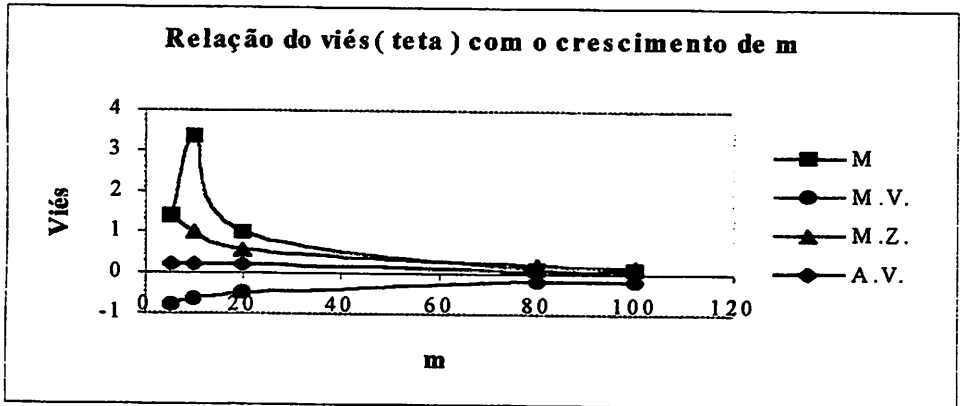


FIGURA 14. Relação do viés dos estimadores de θ , θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de m .

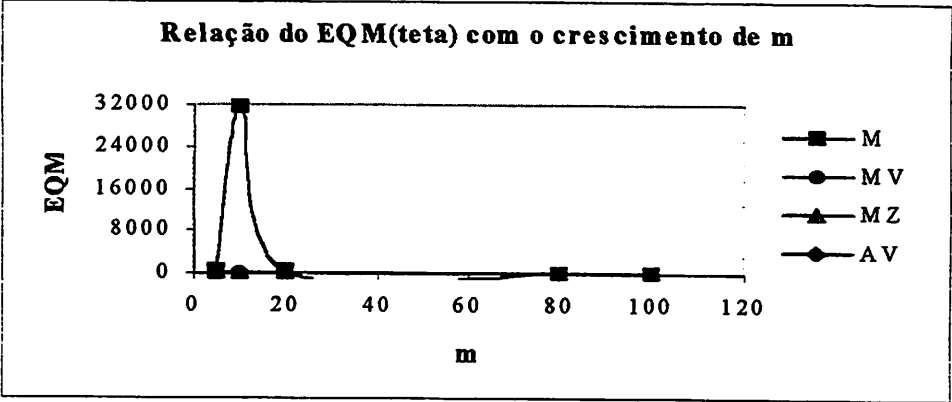


FIGURA 15. Relação do EQM dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de m .

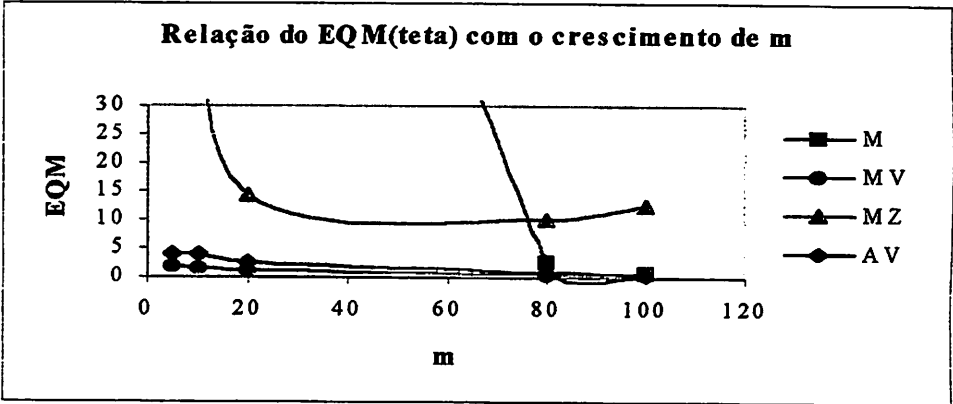


FIGURA 15'. FIGURA 15 ampliada

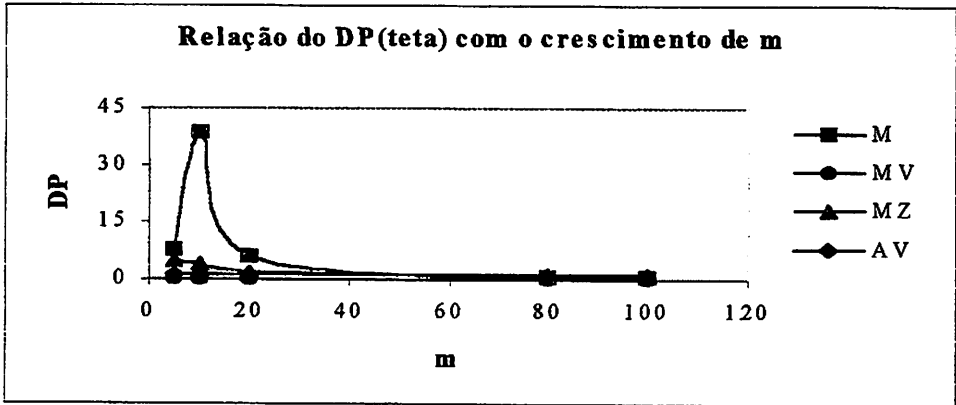


FIGURA 16. Relação do DP dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de m .

A estatística CS também assumiu valores discrepantes quando os valores $\theta \geq 1,00$ foram combinados com $m \leq 20$ e qualquer valor de π e n . O aspecto observado no estimador θ_m na Figura 17 retrata uma situação de inconsistência de θ_m ligada às configurações anteriormente mencionadas. Embora a consistência simples esteja relacionada diretamente apenas com as diferenças em módulo de as estimativas se tornarem tão pequenas quanto se deseje, bastando para isso tomar um número de grupos suficientemente grande, pode-se falar em inconsistência de θ_m naquelas situações. Com o objetivo de sanar as dúvidas inerentes que surgiram, as simulações para as configurações indicadas foram repetidas, e nenhuma diferença significativa ocorreu, o que nos levou a crer que este resultado era consistente.

Um outro problema que deve também ser mencionado é o de estimativas negativas que os estimadores podem fornecer. A eliminação de amostras em que todos os indivíduos dentro de grupo não apresentavam sucesso para todo grupo, ou contrariamente, todos eram sucesso para todos os grupos observados, evitou o

ou contrariamente, todos eram sucesso para todos os grupos observados, evitou o problema de estimadores indefinidos. No entanto, estimativas negativas de θ surgiram para todos os estimadores, e embora tenham sido eliminadas do estudo, essa situação pode ocorrer na prática. Para θ_m , uma maior frequência de estimativas negativas estava associada a amostras que haviam sido geradas a partir de configurações que envolviam baixos m (≤ 20) e $\pi=0,05$ e altos θ ($\geq 1,00$).

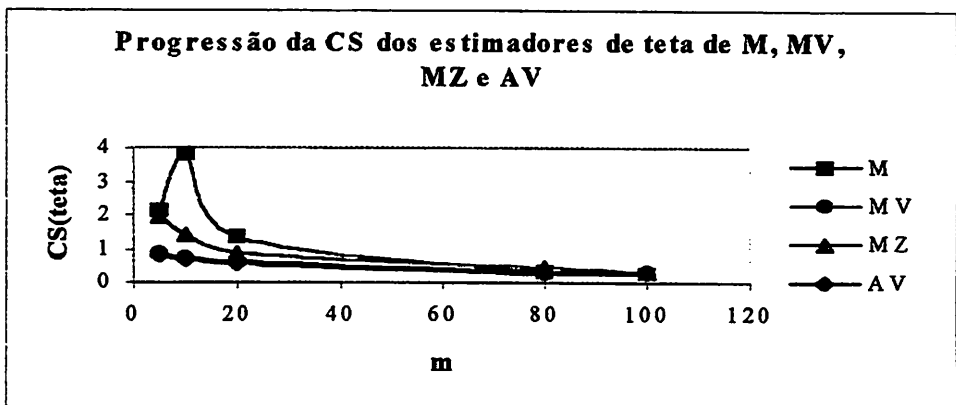


FIGURA 17. Relação da consistência simples dos estimadores de θ : θ_m , θ_{mv} , θ_{mz} e θ_{av} , com o crescimento de m .

Alguns problemas de convergência foram observados nos estimadores de máxima verossimilhança quando $\theta \leq 0,10$ estavam associados com os valores $\pi=0,5$, $n=20$ e $m \leq 10$. Como o método da máxima verossimilhança utilizou, como valores iniciais, as estimativas obtidas pelo método dos momentos, é provável que muitos dos problemas de convergência sejam devidos ao valor inicial e não ao método em si. Vale lembrar que o método de Newton-Raphson garante que se os valores iniciais forem tomados num intervalo adequado, então o

método converge. Porém, estimativas discrepantes foram obtidas pelo método dos momentos quando $\theta \geq 1,00$. Assim, a convergência do método de Newton-Raphson pode ter sido comprometida naqueles casos.

Com as mesmas configurações indicadas no método da máxima verossimilhança, surgiram, também, problemas no método de médias e zeros. Relembrando que os valores iniciais utilizados neste método foram também obtidos pelo método dos momentos, deparou-se também, aqui, com o problema de estimativas discrepantes. Além disso, devido ao fato de este método utilizar o modelo DBB, surgem também as limitações da função gama. Observando a função gama descrita no tópico 2, vemos que ela é definida por uma integral no intervalo de zero a infinito. No SAS®, valores do argumento da função maiores que 171 ou menores que $E-12$, por exemplo, são ignorados, e dados como valores perdidos. Como ilustração, considere o exemplo em que $\pi=0,50$, $\theta=0,01$ e $n=20$. Tomando-se $m=10$, dados foram simulados e as estimativas iniciais (pelo método dos momentos) para π e θ foram determinadas como $\hat{\pi}_m=0,40$ e $\hat{\theta}_m=0,005168$. Substituindo-se estes valores em $P_1(0)$, o argumento para a função gama foi dado como inválido, e logo os valores de $P_1(0)$ e $P_2(0)$ ficaram indefinidos e o método não pode ser continuado. Esse tipo de problema foi mais freqüente em configurações que envolviam baixos θ ($\theta \leq 0,10$) e m ($m \leq 20$) e π ($\pi=0,50$).

Problemas análogos de estimativas negativas com o método da ANAVA surgiram com as mesmas configurações do método de médias e zeros. Aqui, estimativas negativas foram mais frequentes (50% em média) em configurações que envolviam $\theta \leq 0,10$, embora tenham diminuído significativamente mais que nos outros métodos, quando qualquer configuração estava associada a $m \geq 20$.

A distribuição de θ_m foi assimétrica à direita em todos os casos, sendo que a média pertencia ao intervalo determinado pelos quantis $\xi_{0,50}$ e $\xi_{0,75}$ quando $\theta \leq 0,50$, e ao intervalo determinado por $\xi_{0,75}$ e $\xi_{0,95}$ para $\theta \geq 1,00$ e $m \leq 20$. Os outros três estimadores apresentaram distribuições semelhantes à de θ_m , porém, a distribuição de θ_{mv} para configurações envolvendo $m=n=5$ foi assimétrica à esquerda, com a média pertencendo ao intervalo $[\xi_{0,25}, \xi_{0,50}]$. Os vieses de θ_m , θ_{mv} e θ_{mz} foram quase sempre positivos e o de θ_{av} positivo para $\theta \leq 0,10$ e negativo para $\theta \geq 0,30$. Relembrando que o viés é a diferença da estimativa pelo valor real, o fato de o viés ter sido sempre negativo naquelas situações indica que o estimador subestima os valores populacionais (paramétricos) de θ . As distribuições dos vieses de todos os estimadores seguiram exatamente o mesmo padrão das distribuições dos estimadores. O teste de Shapiro-Wilk detectou normalidade apenas em alguns casos para o estimador de máxima verossimilhança. Aqui, assim como para π_{mv} , o teorema da normalidade assintótica dos estimadores de máxima verossimilhança já garante que θ_{mv} tem distribuição normal. Caso se tenha o objetivo de construir intervalos de confiança para θ_{mv} , pode-se usar, também, a variância de θ_{mv} determinada pela matriz de informação de Fisher. Com baixos θ ($\theta \leq 0,50$), o teste detectou normalidade apenas nas configurações determinadas por $\pi=0,50$, $n=20$ e $m \geq 10$. Para $\theta \geq 1,00$, foi detectada normalidade para as sequências de estimativas obtidas a partir de amostras geradas com as configurações envolvendo: i) $\pi=0,05$, $n \geq 10$ e $m \geq 80$, e ii) $\pi \geq 0,25$, $n \geq 10$ e $m \geq 20$. O intervalo de confiança proposto por Kung-Jong, Cumberland e Kuo (1996) para ρ pode, provavelmente, ser usado para construir intervalos de confiança para θ , dada a propriedade ressaltada por Yamamoto e Yanagimoto (1992). No entanto,

como o teste de Shapiro-Wilk não acusou normalidade para os outros estimadores, não é possível afirmar a partir de que valores de m ou n isso poderia ser feito.

As análises em termos do crescimento de n dentro de valores fixos de m mostraram que as estatísticas viés, EQM e CS aumentaram para θ_{mv} , enquanto o DP diminuiu. Para θ_m , todas as estatísticas diminuíram, mas apenas para $m > 20$, sendo que, para valores menores, os resultados foram inconsistentes. O estimador θ_{av} mostrou uma redução das estatísticas avaliadoras somente para $m=100$, enquanto para θ_{mz} , todas aumentaram com o crescimento de n , considerando-se valores fixos para m .

Procurou-se, aqui, dar um enfoque detalhado dos resultados obtidos para os estimadores estudados. Porém, devido a certas dificuldade que podem ser relacionadas com os procedimentos numéricos e limitações inerentes aos métodos não-iterativos (por ex. estimativas negativas), não se pode afirmar que um determinado método é melhor que outro. Neste sentido, um fato importante que deve ser ressaltado é que todos os estimadores apresentaram um comportamento semelhante quando eram tomados $m \geq 80$. Caso o pesquisador tenha noções a priori da faixa dos valores paramétricos da característica sob estudo, acredita-se que este trabalho possa, em geral, sugerir tamanhos amostrais (número de grupos) que forneçam as propriedades desejadas.

5 CONCLUSÕES

5.1 Estimadores de π

Quanto ao viés e EQM, os estimadores de momentos e de médias e zeros do parâmetro π foram os que apresentaram menores valores médios, considerando as 270 configurações. Na mesma condição, o estimador de máxima verossimilhança de π foi o que forneceu maior viés e EQM médios, sendo, em contrapartida, o que obteve o menor DP médio e o menos invariante.

Todos os estimadores obtiveram resultados semelhantes para $m \geq 20$, e foram consistentes para todo $m \geq 5$. Em geral, os estimadores apresentaram normalidade assintótica para todos os valores de θ associados a $\pi \geq 0,25$, $n \geq 10$ e $m \geq 20$.

5.2 Estimadores de θ

Os estimadores θ_{av} e θ_{mv} foram os que apresentaram menores viés e EQM médios, seguidos de θ_{mz} . O estimador θ_{mv} foi, assim como π_{mv} , o que apresentou menor DP médio, seguido de θ_{av} .

O viés de todos os estimadores de θ aumentou com o crescimento de θ , sendo que, para $\theta \leq 0,5$, todos os estimadores apresentaram baixos vieses. No entanto, o estimador θ_m se mostrou o mais “sensível” ao crescimento de θ , quanto ao viés, EQM, DP.

Considerando o comportamento dos estimadores quanto ao crescimento de m , o estimador de momentos foi o que apresentou o pior desempenho quanto

às estatísticas avaliadoras, seguido, em relação ao EQM, por θ_{mz} . Os estimadores θ_{mv} , θ_{mz} e θ_{av} foram consistentes para todos os valores de m , enquanto θ_m foi consistente para $m \geq 20$.

Para $m \geq 80$, todos os estimadores apresentaram comportamento semelhante, e apenas o estimador de máxima verossimilhança de θ apresentou normalidade assintótica em alguns casos ($\theta \leq 0,5$; $\pi = 0,50$; $n = 20$; $m \geq 10$ e $\theta \geq 1,00$; $\pi \geq 0,25$; $n \geq 10$; $m \geq 80$).



REFERÊNCIAS BIBLIOGRÁFICAS

- ANSCOMBE, F.J. Sampling theory of the negative binomial and logarithmic series distributions. **Biometrika**, London, v.37, p.358-382, 1950.
- BRASS, W. Models of birth distributions in human populations. **Bulletin of the International Statistical Institute**, v.36, n.2, p.165-179, 1958.
- CASELA, G.; BERGER, R.L. **Statistical inference**. Belmont: Wadsworth, 1990. 650p.
- CHATFIELD, C.; GOODHARDT, G.J. The beta-binomial model for consumer purchasing behaviour. **Applied Statistics**, London, v.19, p.240-250, 1970.
- CROWDER, M.J. Beta-binomial Anova for proportions. **Applied Statistics**, London, v.27, n.1, p.34-37, 1978.
- CROWDER, M.J. Inference about the intraclass correlation coefficient in the beta-binomial ANOVA for proportions. **Journal Royal Statistical Society**, Cambridge, v.41, n.2, p.230-234, 1979.
- DONNER, A.A. review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. **International Statistical Review**, Netherlands, v.54, n.1, p.67-82, 1986.
- DONNER, A.; KOVAL, J.J. The estimation of intraclass correlation in the analysis of family data. **Biometrics**, Washington, v.36, n.1, p.19-25, Mar. 1980.

- ELSTON, R.C. Response to query: estimating "heritability" of dichotomous trait. **Biometrics**, Washington, v.33, n.2, p.232-233, June 1977.
- FENG, Z.; GRIZZLE, J.E. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. **Statistics in Medicine**, Sussex, v.11, p.1607-1614, 1992.
- FLEISS, J.L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Washington, v.76, p.378-382, 1971.
- GANGE, S.J.; MUÑOZ, A.; SÁEZ, M.; ALONSO, J. Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. **Applied Statistics**, London, v.43, n.3, p.371-382, 1996.
- GRIFFITHS, D.A. Maximum likelihood estimation for the beta-binomial distribution and an application to household distribution of the total number of a disease. **Biometrics**, Washington, v.29, n.4, p.637-648, Dec. 1973.
- HASEMAN, J.K.; KUPPER, L.L. Analysis of dichotomous responses data from certain toxicological experiments. **Biometrics**, Washington, v.35, n.1, p.281-293, Mar. 1979.
- HOFFMANN, J.P.; NYROP, J.P.; KIRKWYLAND, J.J.; RIGGS, D.M.; GILREIN, D.O.; MOYER, D.D. Sequential sampling plan for scheduling control of lepidopteran pests of fresh market sweet corn. **Journal of Economic Entomology**, Lanham, v.89, n.2, p.386-395, Apr. 1996.

- HUGHES, G.; MADDEN, L.V. Some methods allowing for aggregated patterns of disease incidence in the analysis of data from designed experiments. **Plant Pathology**, Oxford, v.44, n.6, p.927-943, Dec. 1995.
- HUGHES, G.; MADDEN, L.V. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. **Phytopathology**, St. Paul, v.83, n.7, p.759-763, July 1993.
- HUGHES, G.; MADDEN, L.V.; MUNKVOLD, G. P. Cluster sampling for disease incidence data. **Phytopathology**, St. Paul, v.86, n.2, p.132-137, Feb. 1996.
- HUGHES, G.; McROBERTS, N.; MADDEN, L.V.; NELSON, S.C. Validating mathematical models of plant-disease progress in space and time. **Journal of Mathematical Applied in Medicine & Biology**, v.14, p.85-112, 1997.
- KLEINMAN, J.C. Proportions with extraneous variance: single and independent samples. **Journal of the American Statistical Association – Serie c**, Alexandria, v.68, n.341, p.46-54, 1973.
- KUNG-JONG, L., CUMBERLAND, W.G.; KUO, L. An interval estimate for the intraclass correlation in beta-binomial sampling. **Biometrics**, Washington, v.52, n.2, p.412-425, June 1996.
- KUPPER, L.L.; HASEMAN, J.K. The use of a correlated binomial model for the analysis of certain toxicological experiments. **Biometrics**, Washington, v.34, n.1, p.69-76, 1978.

- KUPPER, L.L.; PORTIER, C.; HOGAN, M.D.; YAMAMOTO, E. The impact of litter effects on dose-response modeling in teratology. **Biometrics**, Washington, v.42, n.1, p.85-98, Mar. 1986.
- LIPSITZ, S.R.; LAIRD, N.M.; BRENNAN, T.A. Simple moment estimates of the k -coefficient and its variance. **Applied Statistics**, London, v.43, n.2, p.309-323, 1994.
- MADDEN, L.V. BBD – Computer software for fitting the beta-binomial distribution to disease incidence data. **Plant Disease**, St. Paul, v.78, n.5, p.536-540, May 1994.
- MADDEN, L.V.; HUGHES, G. Plant disease incidence: distributions, heterogeneity, and temporal analysis. **Annual Review of Phytopathology**, Palo Alto, v.33, p.529-564, 1995.
- MADDEN, L.V.; HUGHES, G.; MUNKVOLD, G.P. Plant disease incidence: inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. **Crop Protection**, Oxford, v.15, n.7, p.621-632, Nov. 1996.
- MADDEN, L.V.; HUGHES, G.; ELLIS, M.A. Spatial heterogeneity of the incidence of grape downy mildew. **Phytopathology**, St. Paul, v.85, n.3, p.269-275, Mar. 1995.

- MADDEN, L.V., NAULT, L.R., MURRAL, D.J.; APELT, M.R. Spatial pattern analysis of the incidence of aster yellows disease in lettuce. **Researches on Population Ecology**, Kyoto, v.37, n.2, p.297-289, 1995.
- MILLER, S.A.; MADDEN, L.V.; SCHMITTHENNER, A.F. Distribution of *Phitophthora* spp. in field soils determined by immunoassay. **Phytopathology**, St. Paul, v.87, n.1, p.101-107, Jan. 1997.
- MOORE, D.F.; TSIATIS, A. Robust estimation of the variance in moment methods for extra-binomial and extra-poisson variation. **Biometrics**, Washington, v.47, n.2, p.383-401, 1991.
- NAULT, B.A.; KENNEDY, G.G. Sequential sampling plans for use in timing insecticide applications for control of european corn borer (Lepidoptera: pyralidae) in potato. **Journal of Economic Entomology**, Lanham, v.89, n.6, p.1468-1476, 1996.
- OCHI, Y.; PRENTICE, R.L. Likelihood inference in a correlated probit regression model. **Biometrika**, London, v.71, n.3, p.531-543, 1984.
- PAUL, S.R. Analysis of proportions of affected fetuses in teratological experiments. **Biometrics**, Washington, v.38, n.2, p.361-370, June 1982.
- PRENTICE, R.L. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. **Journal of the American Statistical Association - Serie c**, Alexandria, v.81, n.394, p.321-327, Dec. 1986.

RIDOUT, M.S.; DEMÉTRIO, C.G.B; FIRTH, D. Estimating intraclass correlation for binary data. **Biometrics**, Washington, v.55, n.1, p.137-148, Mar. 1999.

STATISTICAL ANALYSIS SYSTEM. SAS/IML[®] Software: usage and reference, version 6. Cary, 1989. 501 pp.

SKELLAM, J.G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. **Journal Royal Statistical Society – serie B**, Cambridge, v.10, p.257-261, 1948.

SMITH, D.M. Algorithm AS189: Maximum likelihood estimation of the parameters of the beta-binomial distribution. **Applied Statistics**, London, v.32, p.196-204, 1983.

SMITH, D.M.; RIDOUT, M.S. Algorithm AS R93: A remark on algorithm AS189: Maximum likelihood estimation of the parameters of the beta-binomial distribution. **Applied Statistics**, London, v.44, p.545-547, 1995.

SOUZA, G. da S. e. **Introdução aos modelos de regressão linear e não-linear**. Brasília: Embrapa-SP/ EMBRAPA-SP/ EMBRAPA-SEA, 1998. 505p.

TAMURA, R.N.; YOUNG, S.S. The incorporation of historical control information in tests of proportions: simulation study of Tarone's procedures. **Biometrics**, Washington, v.42, n.2, p.343-349, June 1986.

- TAMURA, R.N.; YOUNG, S.S. A stabilized moment estimator for the beta-binomial distribution. **Biometrics**, Washington, v.43, n.4, p.813-824, Dec. 1987.
- WILLIAMS, D.A. 394: the analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. **Biometrics**, Washington, v.31, n.4, p.949-952, Dec. 1975.
- WILLIAMS, D.A. Reader reaction: Estimating bias using the beta-binomial distribution in teratology. **Biometrics**, Washington, v.44, n.1, p.305-309, Mar. 1988.
- YAMAMOTO, E.; YANAGIMOTO, T. Moment estimators for the beta-binomial distribution. **Journal of Applied Statistics**, Abingdon, v.19, n.2, p.273-283, 1992.
- ZUCKER, D.; WITTES, J. Testing the effect of treatment in experiments with correlated binary outcomes. **Biometrics**, Washington, v.48, n.3, p.695-710, Mar. 1992.