Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Modelos para análise de dados superdispersos de indução de
haploidia em milho

Andreza Jardelino da Silva

Dissertação apresentada para obtenção do título
de Mestra em Ciências. Área de concentração:
Estatística e Experimentação Agronômica

Piracicaba
2017

Andreza Jardelino da Silva
Bacharel em Estatística

**Modelos para análise de dados superdispersos de indução de haploidia em
milho**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:

Prof^ª Dr^ª **CLARICE GARCIA BORGES DEMÉTRIO**

Dissertação apresentada para obtenção do título de Mestra
em Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Silva, Andreza Jardelino da
Modelos para análise de dados superdispersos de indução de haploidia em milho / Andreza Jardelino da Silva. – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – Piracicaba, 2017.
66 p.

Dissertação (Mestrado) – USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Indução de haploidia 2. Modelo misto 3. Regressão logística 4. Superdispersão. I. Título.

DEDICATÓRIA

*Ao meu companheiro Ranieri, sem
o qual minha vida seria um ponto final.*

AGRADECIMENTOS

Agradeço a Deus, por sua misericórdia na minha vida, pela força e coerência nas minhas atitudes, pela conquista dos meus sonhos.

Aos meus pais, Materna e Cleodon, e minha irmã Aline, pelo carinho e compreensão ao longo destes anos, sei o quanto foi difícil devido à distância.

Ao meu companheiro, Ranieri da Silva Nascimento, por manter minha vida tranquila, por me dar apoio em todos os momentos difíceis, por ser meu amigo, e acima de tudo, por me amar e manter viva em mim a esperança.

A todos os professores do Departamento de Ciências Exatas da ESALQ/USP, em especial, à professora Renata, por me ajudar quando mais precisei, à professora Roseli, pelo carinho e atenção que sempre teve comigo, aos professores Silvio e Afrânio, pelas contribuições dadas no exame de qualificação.

À minha orientadora, professora Clarice Garcia Borges Demétrio, por se importar com o meu desenvolvimento, pela amizade, carinho e disposição neste trabalho, além da excelente companhia.

Aos amigos que conquistei na minha vida, em especial, Andréia, Reginaldo, Mirian, Cristian, Tiago, Patrícia, Sueme, Flaviane, Ítalo, Carmem e Altemir.

Ao meu amigo Filipe Inácio Matias, meu “filho”, pelas felicidades e risadas que vivenciamos ao longo destes anos, momentos estes que serão eternos em minhas memórias.

À minha amiga Giovana Fumes, por todas as horas de felicidades que passamos juntas, pelo apoio e ajuda na construção desta dissertação.

À toda equipe do departamento de Ciências Exatas da ESALQ/USP, em particular, à Solange, que sempre me confortou com suas palavras.

Ao professor Roberto e sua aluna Evellyn, do Departamento de Genética da ESALQ/USP, por concederem o conjunto de dados e pelo auxílio nas dúvidas que surgiram.

Ao CNPq, pelo auxílio financeiro concedido nestes dois anos de mestrado.

A todos, que contribuíram de forma direta ou indiretamente para a realização deste trabalho, o meu muito obrigada.

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	10
1 INTRODUÇÃO	11
2 REVISÃO BIBLIOGRÁFICA	13
2.1 Modelos Lineares Generalizados (MLGs)	13
2.1.1 Estrutura dos MLGs	13
2.1.2 Método de Estimação	14
2.1.3 Função Desvio (<i>deviance</i>)	16
2.1.4 Superdispersão e Subdispersão	17
2.1.5 Intervalo de confiança	19
2.1.6 Critérios de seleção nos MLGs	19
2.1.7 Técnicas de diagnóstico	20
2.2 Modelos Lineares Generalizados Mistos (MLGMs)	23
2.2.1 Estrutura dos MLGMs	23
2.2.2 Método de Estimação	24
2.2.3 Técnicas de diagnóstico para os MLGMs	27
2.2.4 Modelo misto para ensaios casualizados em blocos com subamostras	27
2.3 Fundamento Genético	28
3 MATERIAL E MÉTODOS	31
3.1 Material	31
3.2 Métodos	32
3.2.1 Análise descritiva e modelos	32
4 RESULTADOS E DISCUSSÃO	39
5 CONCLUSÕES E CONSIDERAÇÕES FINAIS	53
REFERÊNCIAS	55
APÊNDICE	59
ANEXO	66

RESUMO

Modelos para análise de dados superdispersos de indução de haploidia em milho

O milho é uma espécie alógama cujo produto comercial são os híbridos, os quais originam-se do cruzamento de duas linhagens endogâmicas. Uma forma para obtenção de tais linhagens é por meio das técnicas de indução de haploidia e posterior obtenção dos duplo-haploides, permitindo até 100% de homozigose. Essas técnicas retornam resultados importantes no melhoramento de milho. Uma variável de interesse importante, obtida a partir dessas técnicas é a taxa de indução de haploidia, a qual trata-se de uma proporção entre o número de sementes haploides e o número total de sementes. O conjunto de dados foi obtido pelo cruzamento da linhagem indutora LI- ESALQ, com cinco genótipos comerciais de milho (2B587PW, 30F53H, BM820, DKB390 e STATUS VIPTERA), em duas gerações F1 e F2, por meio de um delineamento em blocos ao acaso, na área experimental do Departamento de Genética da ESALQ/USP. A teoria dos modelos lineares generalizados (MLGs) possibilita mais opções para a distribuição da variável resposta, exigindo somente que a mesma pertença à família exponencial sob a forma canônica. Tal classe de distribuições pode ser ainda expandida para modelos que permitem efeitos aleatórios no preditor linear, caracterizando a classe dos modelos lineares generalizados mistos (MLGMs). O objetivo deste trabalho foi analisar a taxa de indução de haploidia em milho tropical, utilizando um modelo binomial misto, com efeito aleatório em nível de indivíduo. O método de estimação foi o de máxima verossimilhança. Com base em tal modelagem, verificou-se que o genótipo 30F53H, destacou-se em relação aos demais quanto à eficiência da taxa de indução de haploidia. Todas as análises foram implementadas no *software* R.

Palavras-chave: Indução de haploidia; Modelo misto; Regressão logística; Superdispersão

ABSTRACT

Models for the analysis of overdispersed haploid induction data in maize

The maize is an allogeneic species whose commercial product are the hybrids, which are generated by the crossing of two endogenous lines. An alternative to obtain these lines is using the haploid induction techniques and subsequent doubled haploid production, that allows up to 100% homozygous. Artificial production of doubled haploids is important in plant breeding. An important variable, that results from these techniques, is the haploid induction rate, which is a proportion between the number of haploid seeds and the total number of seeds. The data set was obtained by crossing the inductive line LI-ESALQ, with five commercial genotypes of corn (2B587PW, 30F53H, BM820, DKB390 and STATUS VIPTERA), in two generations F1 e F2, in a randomized block design, in the experimental area of Department of Genetics, ESALQ/USP. The generalized linear models (GLMs) allow more options for the variable response distribution, requiring only that it belongs to the exponential family in canonical form. The GLM class can be expanded to models that allow random effects in the linear predictor, the mixed generalized linear models (MGLM) class. This work aimed to analyse the haploid induction rate in the tropical maize. The binomial mixed model, that included random effects in individual level, was proposed. The maximum likelihood method was used to estimate the parameters. The result revealed that the genotype 30F53H stands out in relation to the others regarding the efficiency in the haploid induction rate. All the analyzes were implemented in the software R.

Keywords: Haploid induction; Logistic regression; Mixed model; Overdispersion

LISTA DE FIGURAS

Figura 1 - Diagrama de Hasse para o experimento casualizado em blocos com subamostras	28
Figura 2 - Croqui do experimento conduzido	31
Figura 3 - Esquema do experimento conduzido	32
Figura 4 - Tipos de sementes: À esquerda, tipos de coloração do embrião. À direita, na parte superior, semente haploide (embrião branco) e, na parte inferior, semente diploide (embrião roxo) (MELCHINGER et al., 2013)	32
Figura 5 - Diagrama de Hasse para o experimento casualizado em blocos com subamostras para análise individual	33
Figura 6 - Diagrama de Hasse para o experimento em estudo casualizado em blocos com subamostras para análise conjunta	36
Figura 7 - Dados de indução de haploidia: <i>Box Plots</i> para números totais de sementes (m_{ijk}) e sementes haploides (y_{ijk}); e proporções de sementes haploides (p_{ijk}), por genótipo, para as gerações F1 e F2	40
Figura 8 - Dados de indução de haploidia: Gráficos de interação entre os números médios de sementes ($\bar{m}_{ij}$) e de sementes haploides ($\bar{y}_{ij}$); e proporções médias de sementes haploides ($\bar{p}_{ij}$) versus genótipos, por blocos, para as gerações F1 e F2	41
Figura 9 - Dados de indução de haploidia: Gráfico de dispersão para proporção de sementes haploides (p_{ijk}) versus número de sementes por espiga (m_{ijk}), para as gerações F1 e F2	42
Figura 10 - Dados de indução de haploidia: Gráficos semi-normais de probabilidade com envelopes de simulação para os Modelos 1, 2 e 3, para as gerações F1 e F2	44
Figura 11 - Dados de indução de haploidia: Gráficos dos componentes do desvio e resíduos de Pearson, versus índices, para as gerações F1 e F2 considerando o modelo binomial-logístico-normal com efeito aleatório de indivíduo . . .	45
Figura 12 - Dados de indução de haploidia: Intervalo de confiança para o preditor linear para as gerações F1 e F2, considerando o modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 2)	46

Figura 13 -Dados de indução de haploidia: Gráficos de interação entre bloco versus a taxa média de indução ($\bar{p}_{ijk}$), por genótipo; interação entre bloco versus a taxa média de indução ($\bar{p}_{ijk}$), por geração; interação de geração versus a taxa média de indução ($\bar{p}_{ijk}$), por genótipo, respectivamente	47
Figura 14 -Dados de indução de haploidia: Gráficos semi-normais de probabilidade com envelope de simulação para os Modelos 1, 2 e 3	49
Figura 15 -Dados de indução de haploidia: Gráfico dos componentes do desvio e dos resíduos de Pearson para o modelo binomial-logístico-normal com efeito aleatório de indivíduo	50
Figura 16 -Dados de indução de haploidia: Gráfico das proporções observadas versus estimadas de indução de haploidia para o modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)	52
Figura 17 -Esquema do banco de dados	66

LISTA DE TABELAS

Tabela 1 - Análise de variância do modelo misto para experimento conduzido em blocos com subamostras	28
Tabela 2 - Esquema da análise do desvio, considerando-se o diagrama de Hasse da Figura 5	34
Tabela 3 - Esquema da análise do desvio, considerando-se o diagrama de Hasse da Figura 6	37
Tabela 4 - Valores referentes ao número de espigas (k_{ij}), números médios de sementes ($\bar{m}_{ij}$) e sementes haploides ($\bar{y}_{ij}$); e proporções médias de sementes haploides ($\bar{p}_{ij}$), por parcela	39
Tabela 5 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados para geração F1	43
Tabela 6 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados para geração F2	43
Tabela 7 - Estimativas dos parâmetros para as gerações F1 e F2 para modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 2)	46
Tabela 8 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados - sem efeito de interação geração $\times$ genótipo	48
Tabela 9 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados - com efeito de interação geração $\times$ genótipo	48
Tabela 10 - Estimativas dos parâmetros do modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)	51
Tabela 11 - Proporções médias para o ajuste do modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)	51
Tabela 12 - Distribuições de probabilidade pertencentes à família exponencial	60

1 INTRODUÇÃO

De acordo com Conab (2016), o Brasil encontra-se entre os três maiores produtores de milho no mundo; somente nos anos de 2015/2016 foram produzidos, aproximadamente, 80 mil toneladas de grãos. Esse avanço na produção é devido ao desenvolvimento de técnicas de manejo mais eficientes e ao melhoramento genético de plantas, como a utilização de genótipos mais adaptados de híbridos superiores.

O milho é uma espécie alógama, cujo produto comercial são os híbridos. Tais híbridos são oriundos do cruzamento de duas linhagens endogâmicas, que podem ser desenvolvidas de forma rápida e eficiente, por meio de técnicas de indução de haploidia e posterior duplicação cromossômica para obtenção de linhagens duplo-haploides, as quais permitem até 100% de homozigose (BORÉM; MIRANDA, 2013). Essas técnicas têm suscitado grande interesse no meio científico.

Um indivíduo haploide é aquele cujo núcleo celular possui apenas uma única repetição cromossômica (DE LAAT; GOHDE; VOGELZAKG, 1987). Em melhoramento genético, a utilização de haploides pode ser uma ferramenta importante para o desenvolvimento de linhagens homozigotas.

A técnica de duplo-haploide retorna resultados importantes na genética do melhoramento de milho, porém, a taxa de indução de haploidia é baixa, variando de 0 a 10%, sendo, geralmente, menor em condições de clima tropical (PRASANNA; CHAIKAM; MAHUKU, 2012).

O interesse de estudo dos pesquisadores é a taxa de indução de haploidia, a qual trata-se da razão entre o número de sementes haploides e o número total de sementes em cada espiga. Estatisticamente, a utilização dos modelos lineares generalizados (MLGs) permite fazer uma análise para dados cujas observações são proporções.

A teoria dos MLGs, proposta por Nelder e Wedderburn (1972), possibilita o uso de várias distribuições para a variável resposta, com a única condição de que tais distribuições pertençam à família exponencial sob a forma canônica. Além disso, a modelagem por meio da classe MLGs proporciona uma maior flexibilidade entre a relação da média da variável resposta univariada com o preditor linear, obtida por meio de uma função monótona e diferenciável, denominada função de ligação. Se o modelo apresentar efeitos fixos e aleatórios no preditor linear, têm-se os chamados modelos lineares generalizados mistos

(MLGMs).

O objetivo desta dissertação foi identificar um modelo adequado para analisar dados de taxa de indução de haploidia em experimentos oriundos da técnica de duplo-haploide. Para tal finalidade, utilizou-se a metodologia dos MLGMs, com ênfase em técnicas para dados com superdispersão.

A estrutura deste trabalho está organizada em cinco capítulos, distribuídos da forma que se segue. No capítulo 2, é apresentada uma revisão bibliográfica sobre os modelos lineares generalizados e extensões, bem como técnicas de diagnóstico, além de alguns fundamentos genéticos. No capítulo 3, tem-se a descrição do conjunto de dados que motivou esta pesquisa e o desenvolvimento dos métodos a serem utilizados na análise. No capítulo 4, são mostrados os resultados e as discussões. Finalmente, no capítulo 5, são destacadas as contribuições deste trabalho para a análise dos dados de indução de haploidia em milho e propostas para trabalhos futuros. Alguns detalhes estão disponíveis nos apêndices e no anexo.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo apresentar uma breve revisão de literatura sobre a classe dos MLGs, MLGMs e alguns fundamentos genéticos utilizados no desenvolvimento deste trabalho.

2.1 Modelos Lineares Generalizados (MLGs)

2.1.1 Estrutura dos MLGs

Sejam n variáveis aleatórias independentes, $Y_1, \dots, Y_n$, cada uma com função densidade ou função de probabilidade na família exponencial expressa por

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \quad i = 1, \dots, n, \quad (1)$$

em que $E(Y_i) = \mu_i = b'(\theta_i)$ e $\text{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i$, sendo $V_i = V(\mu_i) = d\mu_i/d\theta_i$ a função de variância, $\phi^{-1} > 0$, o parâmetro de dispersão e θ , o parâmetro canônico (CORDEIRO; DEMÉTRIO, 2013). Uma vez que se conhece a função de variância especificada como $V(\mu_i) = b''(\theta_i)$, é possível determinar qual será a classe de distribuições correspondente na família exponencial. Segundo Fisher (1934), a família exponencial de distribuição garante aos parâmetros estatísticas suficientes.

Os MLGs estão definidos como uma teoria unificada da modelagem estatística, proposta por Nelder e Wedderburn (1972), e envolvem três componentes:

- i) Um componente aleatório, representado pelas variáveis aleatórias independentes, as quais são as variáveis respostas, $Y_1, \dots, Y_n$, que possuem a mesma distribuição de probabilidade pertencente à família exponencial na forma canônica, dada pela expressão (1), com médias μ_i e parâmetro de escala constante $\phi > 0$.
- ii) Um componente sistemático, representado pelas covariáveis, $X_1, \dots, X_n$, que constituem a estrutura linear do modelo, isto é,

$$\eta_i = \sum_{j=1}^P X_{ij}\beta_j \quad i = 1, \dots, n,$$

ou, matricialmente,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

em que $\boldsymbol{\eta}$ é o vetor dos preditores lineares, $\mathbf{X}$ a matriz de delineamento do modelo $(n \times p)$ e $\boldsymbol{\beta}$ é o vetor $(p \times 1)$ de parâmetros a serem estimados.

- iii) Uma função $g(\cdot)$, monótona e diferenciável, denominada função de ligação, que permite fazer a ligação entre a média μ_i e o preditor linear η_i , isto é,

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

Algumas distribuições que pertencem à família exponencial estão explicitadas no Apêndice A.

2.1.2 Método de Estimação

Alguns métodos de estimação podem ser utilizados para a obtenção do vetor de parâmetros $\boldsymbol{\beta}$, como mínimos quadrados, Bayesiano e da máxima verossimilhança (CORDEIRO; DEMÉTRIO, 2013). Neste trabalho, será descrito o método da máxima verossimilhança.

Assim, usando-se a expressão dada em (1), considere o logaritmo da função de máxima verossimilhança na forma

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}_i, \phi; \mathbf{y}_i) = \phi^{-1} \sum_{i=1}^n [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n c(y_i, \phi). \quad (2)$$

Para estimação do vetor de parâmetros $\boldsymbol{\beta}$ obtém-se a derivada de primeira ordem da expressão dada em (2), isto é, a função score, e para obtenção da informação de Fisher, calcula-se o valor esperado da derivada de segunda ordem da função. Ambas as expressões podem ser representadas, respectivamente, por

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \phi \mathbf{X}^T \mathbf{W}^{1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu}),$$

e

$$\mathbf{K}(\boldsymbol{\beta}) = \mathbf{E} \left(-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) = \phi \mathbf{X}^T \mathbf{W} \mathbf{X},$$

em que $\mathbf{X}$ é uma matriz $(n \times p)$ de posto completo, cujas linhas são $\mathbf{x}_i^T$, com $i = 1, \dots, n$,

$\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ sendo a matriz dos pesos, com

$$w_i = \frac{1}{V(\mu_i)[g'(\mu_i)]^2},$$

$\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ a matriz da função de variância, $\mathbf{y} = (y_1, \dots, y_n)^T$ e $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$.

Para iniciar o processo via algoritmo iterativo de Newton-Raphson, define-se um valor inicial $\boldsymbol{\beta}^{(0)}$ para $\boldsymbol{\beta}$, e uma expansão da função escore $\mathbf{U}(\boldsymbol{\beta})$ é realizada, baseada numa aproximação de Taylor de primeira ordem em torno da vizinhança do ponto $\boldsymbol{\beta}^{(0)}$, da forma

$$\mathbf{U}(\boldsymbol{\beta}) \cong \mathbf{U}(\boldsymbol{\beta}^{(0)}) + \mathbf{U}'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}), \quad (3)$$

em que $\mathbf{U}'(\boldsymbol{\beta})$ é a derivada primeira da função escore em relação ao vetor de parâmetro $\boldsymbol{\beta}$, avaliada no ponto $\boldsymbol{\beta}^{(0)}$. A partir de (3), pode-se, explicitar um algoritmo de estimação para o processo iterativo de Newton-Raphson, dado por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [-\mathbf{U}'(\boldsymbol{\beta}^{(m)})]^{-1}\mathbf{U}(\boldsymbol{\beta}^{(m)}), \quad \text{com } m = 0, 1, \dots$$

Nem sempre a matriz $-\mathbf{U}'(\boldsymbol{\beta})$ é positiva definida. Comumente, realiza-se uma substituição pelo valor esperado de $-\mathbf{U}'(\boldsymbol{\beta})$, denotado por $\mathbf{K}(\boldsymbol{\beta})$ (informação de Fisher dada em (2.1.2)). Deste modo, a equação para o processo iterativo é expressa por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathbf{K}(\boldsymbol{\beta}^{(m)})]^{-1}\mathbf{U}(\boldsymbol{\beta}^{(m)}), \quad \text{com } m = 0, 1, \dots \quad (4)$$

Com base em (4), obtém-se o algoritmo de Mínimos Quadrados Ponderados Iterativos (MQPI), descrito na forma

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)}, \quad \text{com } m = 0, 1, \dots, \quad (5)$$

em que m é a m -ésima iteração; $\boldsymbol{\beta}^{(m+1)}$ é o vetor de parâmetros na iteração no passo $m+1$; $\mathbf{X}$ é a matriz de delineamento; $\mathbf{W}$ é a matriz dos pesos, com $\mathbf{W} = \text{diag}\{w_i\}$ e expressa por

$$w_i^{(m)} = \frac{1}{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2},$$

e $\mathbf{z}$ é o vetor da variável dependente ajustada na m -ésima iteração, e pode ser escrito como

$$\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu}).$$

A cada iteração, o vetor das estimativas dos parâmetros, $\widehat{\boldsymbol{\beta}}^{(m+1)}$, é atualizado com base nas estimativas anteriores $\widehat{\boldsymbol{\beta}}^{(m)}$. A convergência de (5) ocorre num número finito de passos, independentemente dos valores iniciais atribuídos, entretanto, para que ocorra de forma mais rápida, é usual utilizar uma estimativa inicial dada por $\boldsymbol{\eta}^{(0)} = g(\mathbf{y})$.

Entre vários critérios para se avaliar a convergência tem-se

$$\sum_{j=1}^p \left(\frac{\beta_j^{(m)} - \beta_j^{(m+1)}}{\beta_j^{(m)}} \right)^2 < \xi,$$

para um ξ suficientemente pequeno (CORDEIRO; DEMÉTRIO, 2013).

2.1.3 Função Desvio (*deviance*)

Nos modelos lineares generalizados, a qualidade do ajuste é avaliada por meio da função desvio, ou comumente chamada de *deviance*. Essa função é a distância entre o logaritmo da função de verossimilhança dos modelos saturado e sob pesquisa. Considerando-se a equação em (2) como função apenas do vetor de parâmetros $\boldsymbol{\beta}$, tem-se o modelo saturado, em que $p = n$, expresso por

$$\ell(\mathbf{y}; \mathbf{y}) = \phi^{-1} \sum_{i=1}^n [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)] + \sum_{i=1}^n c(y_i, \phi),$$

(neste caso, a estimativa de máxima verossimilhança de μ_i , é dada por $\tilde{\mu}_i = y_i$) e o modelo sob pesquisa, em que $p < n$, é da forma

$$\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) = \phi^{-1} \sum_{i=1}^n [y_i \hat{\theta}_i - b(\hat{\theta}_i)] + \sum_{i=1}^n c(y_i, \phi),$$

em que $\tilde{\theta}_i$ e $\hat{\theta}_i$ são as estimativas de máxima verossimilhança do parâmetro canônico sob os modelos saturado e reduzido, respectivamente (CORDEIRO; DEMÉTRIO, 2013).

A função desvio escalonado é definida por

$$S_p = S(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}; \mathbf{y})\}. \quad (6)$$

Reescrevendo (6) tem-se que

$$S(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi^{-1} 2 \sum_{i=1}^n [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)] = \phi^{-1} D_p$$

em que D_p é a função desvio. Ao obter um valor significativamente pequeno do desvio, têm-se evidências de que, para uma pequena quantidade de parâmetros obtém-se um ajuste tão bom quanto o ajuste do modelo saturado (CORDEIRO; DEMÉTRIO, 2013).

No Apêndice B, são apresentadas as funções desvio para alguns modelos.

2.1.4 Superdispersão e Subdispersão

Na análise de dados de proporções, o modelo binomial tem como pressuposições a independência entre os eventos de Bernoulli e a probabilidade de sucesso constante. Assim, supondo $Y_i \sim \text{Bin}(n_i, \pi_i)$ tem-se que

$$E(Y_i) = \mu_i = n_i \pi_i \quad \text{e} \quad \text{Var}(Y_i) = n_i \pi_i (1 - \pi_i) = \mu_i \left(1 - \frac{\mu_i}{n_i}\right).$$

Na prática, entretanto, a maioria dos dados são de natureza heterogênea, levando a uma relação média-variância diferente da especificada pelo modelo binomial.

A falha nas pressuposições, faz com que o modelo binomial não se ajuste às observações, ocasionando uma variância observada maior (superdispersão) ou menor (subdispersão) do que a esperada (HINDE; DEMÉTRIO, 1998; RIDOUT; BESBEAS, 2004).

A superdispersão pode ocorrer devido à variabilidade do material experimental, correlação entre respostas individuais, amostragem em conglomerados, variáveis não observadas, excessos de zeros nas observações (HINDE; DEMÉTRIO, 1998), falha nas pressuposições do modelo (KOKONENDJI, 2014) e autocorrelação (CHRIÉL; STRYHN; DAUPHIN, 1999).

Os modelos para análise de dados superdispersos podem ser divididos em duas categorias (HINDE; DEMÉTRIO, 1998). A primeira leva em consideração uma forma geral para a função de variância com adição de um parâmetro de dispersão ϕ , permitindo maior flexibilidade para a variância. No caso de modelagem de proporção, seja o modelo quase-binomial, com

$$E(Y_i) = \mu_i = n_i \pi_i \quad \text{e} \quad \text{Var}(Y_i) = \phi n_i \pi_i (1 - \pi_i) = \phi \mu_i \left(1 - \frac{\mu_i}{n_i}\right),$$

em que $\phi > 0$ é fator de heterogeneidade (FINNEY, 1971). O estimador de ϕ pode ser expresso por

$$\hat{\phi} = \frac{X^2}{n-p},$$

sendo X^2 a estatística de Pearson para um modelo que apresenta p parâmetros e n observações.

A segunda forma de modelar a superdispersão considera modelos de dois estágios. Se a variável resposta segue a distribuição binomial e o parâmetro (P_i) do modelo segue a distribuição beta, isto é, $Y_i|P_i \sim \text{Bin}(n_i, P_i)$ e $P_i \sim \text{Beta}(\alpha_i, \beta_i)$, tem-se como resultado o modelo beta-binomial com

$$E(Y_i) = \mu_i = n_i\pi_i \quad \text{e} \quad \text{Var}(Y_i) = n_i\pi_i(1 - \pi_i)[1 + \phi(n_i + 1)] = \phi_i\mu_i \left(1 - \frac{\mu_i}{n_i}\right),$$

em que $\phi = 1/(\alpha_i + \beta_i + 1)$, considerando que $\alpha_i + \beta_i$ seja aproximadamente constante (GRIFFITHS, 1973). É importante ressaltar que essa distribuição não pertence à família exponencial canônica.

Pode-se assumir ainda, para modelos com dois estágios para dados de proporção, que $Y_i|P_i \sim \text{Bin}(n_i, P_i)$ tem um efeito aleatório no preditor linear, isto é, $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\sigma} Z_i$ com $Z_i \sim N(0, 1)$, e $\text{logit}(P_i) = \log(P_i/(1 - P_i))$, chamado de modelo logístico-normal, e sendo o exemplo mais simples de modelos lineares generalizados mistos (detalhes sobre esta classe de modelos serão vistos na Seção 2.2). Desse modo,

$$E(Y_i) \approx n_i\pi_i \quad \text{e} \quad \text{Var}(Y_i) \approx n_i\pi_i(1 - \pi_i)[1 + \phi(n_i - 1)\pi_i(1 - \pi_i)],$$

também conhecido como modelo III de Williams (WILLIAMS, 1982).

Outra forma de modelar a superdispersão para dados de proporção, é considerar que as respostas Y_i são resultantes de um somatório das respostas R_{ij} de n_i ensaios de Bernoulli, que não são independentes (HINDE; DEMÉTRIO, 1998; CORDEIRO; DEMÉTRIO, 2013). Deste modo, $Y_i|R_{ij} \sim \text{Bin}(n_i, \pi_i)$; $R_{ij} \sim \text{Bernoulli}(\pi_i)$ e $\text{Cov}(R_{ij}, R_{ik}) = \rho$. Então,

$$E(Y_i) = n_i\pi_i \quad \text{e} \quad \text{Var}(Y_i) = n_i\pi_i(1 - \pi_i)[1 + \rho(n_i - 1)].$$

É importante ressaltar que a restrição $\text{Var}(Y_i) > 0$ faz com que $-1/(n_i - 1) < \rho < 1$, sendo ρ a correlação entre as observações (CORDEIRO; DEMÉTRIO, 2013).

Para a análise de dados na forma de contagens com subdispersão, têm-se diferentes propostas, como os modelos de Poisson-binomial (KOKONENDJI, 2014), Poisson

duplo e COM-Poisson (RIDOUT; BESBEAS, 2004). Brooks, James e Gray (1991) apresentaram ainda uma análise de dados das frequências de leitegadas de porcos que apresentam subdispersão considerando efeito aleatório no tempo.

De forma geral, o problema de dados com superdispersão ou subdispersão não está na estimação dos parâmetros relacionados à média, mas sim, na estimativa do erro padrão, ou seja, na matriz de variâncias e covariâncias. Isso implica que os testes de hipóteses serão afetados, podendo induzir a escolhas inadequadas de modelos.

2.1.5 Intervalo de confiança

Assintoticamente, tem-se que $\hat{\beta} \sim N_p(\beta, \phi^{-1}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})$, em que $\mathbf{K}^{-1} = \phi^{-1}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ é o inverso da matriz de informação de Fisher, $\mathbf{X}$ é a matriz de delineamento com efeito fixo, e $\mathbf{W}$ é a matriz de pesos. Desta forma, a construção do intervalo de confiança assintótico para um valor específico do vetor de parâmetros β com coeficiente $1 - \alpha$, é expresso da forma

$$[\hat{\beta}_r - a_{\alpha/2}(\hat{k}^{ii})^{1/2}; \hat{\beta}_r + a_{\alpha/2}(\hat{k}^{ii})^{1/2}],$$

sendo $\hat{k}^{ii}$ o elemento (i, i) do inverso da matriz estimada da informação de Fisher ($\hat{\mathbf{K}}^{-1}$), e $\Phi(-a_{\alpha/2}) = \alpha/2$, em que $\Phi(\cdot)$ é a função densidade acumulada (f.d.a.) da distribuição normal padrão (CORDEIRO; DEMÉTRIO, 2013).

2.1.6 Critérios de seleção nos MLGs

A ideia fundamental na seleção de modelos é encontrar um que esteja bem ajustado aos dados, e contenha o menor número possível de parâmetros, isto é, um modelo parcimonioso, princípio fundamental na modelagem estatística.

Deste modo, têm-se os critérios da informação de Akaike (AIC) (AKAIKE, 1973) e de informação Bayesiano (BIC) (SCHWARZ, 1978), definidos respectivamente, por

$$\begin{aligned} \text{AIC} &= -2\hat{\ell}_p + 2p, \\ \text{BIC} &= -2\hat{\ell}_p + p \log n, \end{aligned}$$

em que p é o número de parâmetros do modelo sob pesquisa, $\hat{\ell}_p$ é o logaritmo da função

de máxima verossimilhança para o modelo sob pesquisa, e n é o número de observações do conjunto de dados.

Tais critérios visam a minimização de um modelo sem que se utilizem testes estatísticos (PAULA, 2013). De acordo com Cordeiro e Demétrio (2013), as expressões do AIC e BIC para os MLGs são dadas, respectivamente, pelas equações

$$\begin{aligned} \text{AIC}^* &= D_p - 2\hat{\ell}_p + 2p, \\ \text{BIC}^* &= D_p - 2\hat{\ell}_p + p \log(n), \end{aligned}$$

em que D_p é a função desvio do modelo sob pesquisa.

Considera-se o melhor modelo aquele que apresentar o menor valor de AIC^* . Essa regra de decisão também é válida para o BIC^* (CORDEIRO; DEMÉTRIO, 2013).

2.1.7 Técnicas de diagnóstico

Nesta seção, serão descritas algumas técnicas de diagnósticos utilizadas para verificação de ajuste nos MLGs.

2.1.7.1 Resíduos

Os resíduos no contexto dos MLGs são comumente usados para explorar a adequação do modelo ajustado no que diz respeito à escolha da distribuição proposta para a variável resposta (CORDEIRO G.M. E LIMA NETO, 2006). Isso é explicado devido à preocupação da ocorrência de desvios sistemáticos, ocasionados por uma escolha inadequada da função de ligação e/ou da função de variância, pela presença de pontos discrepantes que podem ter sido ocasionados devido a pontos que estão nos extremos da amplitude da variável independente, ou que foram erroneamente coletados, ou ainda, por algum motivo não controlado que influenciou no momento da obtenção dos dados.

Em modelos gaussianos, a análise de diagnóstico consiste em técnicas para analisar os resíduos, a fim de verificar homocedasticidade, normalidade e independência. Essas verificações são de suma importância, uma vez que, a violação de um desses pressupostos pode ocasionar resultados não confiáveis comprometendo as estimativas e previsões acerca do modelo (MYERS; MONTGOMERY D. C. E VINING, 2010). Nos MLGs, os resíduos mais usados são

(i) Resíduos ordinários:

$$r_i = y_i - \hat{\mu}_i,$$

em que y_i representa a variável resposta e $\hat{\mu}_i$ sua respectiva média estimada.

(ii) Resíduo de Pearson Generalizado:

$$r_P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}, \quad (7)$$

em que $V(\hat{\mu}_i)$ é a função de variância de acordo com a distribuição utilizada.

(iii) Resíduo de Pearson generalizado estudentizado:

$$r_{PE} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}V(\hat{\mu}_i)(1 - \hat{h}_i)}},$$

em que $\hat{\phi}$ é uma estimativa do parâmetro de dispersão ϕ , $\hat{h}_i$ são os elementos da diagonal da matriz de projeção $\widehat{\mathbf{H}}$, expressa por

$$\widehat{\mathbf{H}} = \widehat{\mathbf{W}}^{1/2} \mathbf{X} (\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{W}}^{1/2}.$$

(iv) Componentes do desvio:

$$d_r = \text{sinal}(y_i - \hat{\mu}_i) \times \sqrt{d_i}, \quad (8)$$

em que $d_i = 2\{y_i[\tilde{\theta} - \hat{\theta}] - b(\tilde{\theta}) + b(\hat{\theta})\}$. Esse resíduo será a raiz quadrada de cada elemento do desvio com o sinal igual ao de $y_i - \hat{\mu}_i$.

(v) Componentes do desvio estudentizado:

$$d_e = \frac{\text{sinal}(y_i - \hat{\mu}_i) \times \sqrt{d_i}}{\sqrt{\hat{\phi}(1 - \hat{h}_i)}},$$

sendo que a distribuição de d_e está mais próxima da normalidade. Além disso, para grandes amostras são considerados pontos discrepantes os $|d_e| > 2$ (WILLIAMS, 1984).

2.1.7.2 Gráfico semi-normal de probabilidades

A estimação do parâmetro ϕ faz com que as estatísticas desvio residual e X^2 de Pearson se aproximem do número de graus de liberdade, portanto, essas estatísticas não podem mais ser usadas para a verificação do ajuste de um modelo com superdispersão (CORDEIRO; DEMÉTRIO, 2013).

Análises gráficas por meio dos resíduos, podem ser utilizadas para avaliar se o preditor linear e a função de ligação são os mais adequados, e verificar possíveis pontos discrepantes. Para a identificação da especificidade de $V(\mu)$, pode-se usar um gráfico de resíduos versus valores ajustados, entretanto, para escolhas de modelos superdispersos que envolvem o parâmetro de escala ϕ esse gráfico não será tão útil. Liang e McCullagh (1993) usam, para dados binários, gráficos dos resíduos versus tamanhos amostrais m_i para propor um modelo adequado. Ganio e Schafer (1992) consideram para modelos de superdispersão, o gráfico da variável adicionada a fim de comparar funções de variância.

Uma outra abordagem para diagnóstico é utilizar o gráfico semi-normal de probabilidades com envelope simulado. Este gráfico, comumente conhecido por *half normal plot* (hnp), é muito utilizado para a verificação da qualidade do ajuste do modelo proposto (HINDE; DEMÉTRIO, 1998; VIEIRA, 1998).

Para a obtenção do hnp é necessário plotar de forma ordenada os valores absolutos de uma medida de diagnóstico (sendo mais comumente usados nos MLGs, os componentes do desvio estudentizado) versus os respectivos valores esperados das estatísticas de ordem da distribuição semi-normal, calculados por

$$\Phi^{-1} \left(\frac{i + n - \frac{1}{8}}{2n + \frac{1}{2}} \right),$$

em que $\Phi^{-1}(\cdot)$ é a função acumulada da distribuição normal padrão e n refere-se ao tamanho da amostra obtida, com $i = 1, \dots, n$.

Uma proposta para ajudar na análise gráfica é a construção de um envelope simulado. No caso do modelo estar bem ajustado, o envelope simulado conterá os valores observados (ATKINSON, 1985). Na sequência, os passos para a construção de um envelope simulado são apresentados.

- (i) Após ajustar um modelo, obter q_i , os valores de uma quantidade diagnóstica escolhida em valor absoluto e colocá-los em ordem crescente;

- (ii) Gerar 99 simulações para a variável resposta do modelo ajustado com os mesmos valores para as variáveis regressoras;
- (iii) Realizar o ajuste do modelo para as 99 amostras e, para cada ajuste, obter valores da quantidade diagnóstica escolhida, $q_{j(i)}^*$, $j = 1, \dots, 99$, $i = 1, \dots, n$, em valor absoluto e ordem crescente;
- (iv) Para cada modelo ajustado calcular os percentis 25%, 50% e 97,5%;
- (v) Gerar os gráficos desses percentis e dos q_i observados versus as estatísticas de ordem da distribuição semi-normal.

No *software* R, este tipo de gráfico está implementado por meio da função `hnp` (MORAL; HINDE; DEMÉTRIO, 2016) para algumas distribuições de probabilidades, diversas funções de ligações e diferentes tipos de resíduos.

2.2 Modelos Lineares Generalizados Mistos (MLGMs)

A classe de MLGs considera covariáveis de efeitos fixos no preditor linear. Entretanto, existem experimentos nos quais tem-se a necessidade de se incluírem também efeitos aleatórios no componente sistemático, definindo os chamados Modelos Lineares Generalizados Mistos (MLGMs) (HINDE; DEMÉTRIO, 1998).

2.2.1 Estrutura dos MLGMs

Considere que $\mathbf{Y}|\mathbf{u}$ tem uma distribuição condicional, tal que, $\mathbf{Y}$ é um vetor com n respostas independentes, e que esta distribuição condicional pertença à família exponencial na forma canônica expressa em (1), com

$$E[Y_i|u] = \mu_i \quad \text{e} \quad \text{Var}[Y_i|u] = \phi V(\mu_i),$$

e um preditor linear

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u},$$

em que $\mathbf{x}_i^T$ é a i -ésima linha da matriz de delineamento do experimento de dimensão $(n \times p)$ para os efeitos fixos, $\boldsymbol{\beta}$ é o vetor de parâmetros de efeito fixo, $\mathbf{z}_i^T$ é a i -ésima linha da matriz de delineamento para os efeitos aleatórios e $\mathbf{u}$ é o vetor de efeitos aleatórios, e função de ligação $\eta_i = g(\mu_i)$. Para especificar o modelo, assume-se que $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

Portanto, tem-se que

$$E(Y_i) = E[E(Y_i|\mathbf{u})] = E(\mu_i) = E[g^{-1}(\eta_i)] = E[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u})],$$

e

$$\begin{aligned} \text{Var}(Y_i) &= \text{Var}[E(Y_i|\mathbf{u})] + E[\text{Var}(Y_i|\mathbf{u})] \\ &= \text{Var}(\mu_i) + E[\phi V(\mu_i)] \\ &= \text{Var}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u})] + E\{\phi[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u})]\}, \end{aligned}$$

sendo $g^{-1}(\cdot)$ a inversa da função de ligação, que pode apresentar funções não-lineares. Deste modo, o modelo não poderá ser escrito de forma mais resumida sem que se façam suposições específicas, bem como a distribuição condicional de Y_i .

Ao incluir efeitos aleatórios no preditor linear, induz-se uma correlação entre as observações, sendo a covariância dada por

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}[E(Y_i|\mathbf{u}), E(Y_j|\mathbf{u})] + E[\text{Cov}(Y_i, Y_j|\mathbf{u})] \\ &= \text{Cov}(\mu_i, \mu_j) + E(0) \\ &= \text{Cov}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u}), g^{-1}(\mathbf{x}_j^T \boldsymbol{\beta} + \mathbf{z}_j^T \mathbf{u})], \end{aligned}$$

assumindo independência na distribuição condicional de Y_i dado $\mathbf{u}$.

2.2.2 Método de Estimação

O MLGM apresenta efeitos aleatórios no preditor linear e permite que a variável resposta assuma uma distribuição que pertença à família exponencial na forma canônica como mostrado na Seção 2.2.1.

Seja $\mathbf{Y}$ o vetor da variável resposta independente e $\mathbf{u}$ o vetor q -dimensional de efeitos aleatórios, sendo $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Sob independência condicional, segue-se a função densidade sob a forma

$$f(Y_i|\mathbf{u}, \boldsymbol{\beta}, \phi),$$

em que $\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u}$, para qualquer função de ligação $g(\cdot)$ conhecida, com os vetores de covariáveis $\mathbf{x}_i$ e $\mathbf{z}_i$ de dimensões p e q , respectivamente, $\boldsymbol{\beta}$ um vetor de efeitos

fixos p -dimensional, e ϕ um parâmetro de dispersão.

A estimação dos parâmetros no modelo pode ser feita por meio da maximização da função de verossimilhança marginal, integrando em relação aos efeitos aleatórios. Assim, a verossimilhança é dada por

$$f_i(Y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \phi) = \int \prod_{j=1}^{n_i} f_i(Y_i|\mathbf{u}, \boldsymbol{\beta}, \phi) f(u_j|\boldsymbol{\Sigma}) du_j,$$

sendo a função de verossimilhança completa

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \phi) = \prod_{i=1}^N f_i(Y_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \phi). \quad (9)$$

Assumindo independência condicional, tem-se que

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \phi) = \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_i(Y_i|\mathbf{u}, \boldsymbol{\beta}, \phi) f(u_j|\boldsymbol{\Sigma}) du_j.$$

Há dificuldades em maximizar (9) devido ao número de N integrais sobre os efeitos aleatórios. Estas integrais, em geral, só são resolvidas analiticamente quando a variável resposta segue distribuição gaussiana. Quando a variável resposta segue outra distribuição de probabilidade, o resultado deve ser obtido por métodos numéricos, uma vez que não se consegue obter uma solução analítica. Neste contexto, os métodos numéricos utilizados são as quadraturas de Gauss-Hermite e adaptativa de Gauss-Hermite, os métodos de Laplace e Monte Carlo, sendo, neste trabalho, abordados os dois primeiros métodos.

Método da quadratura de Gauss-Hermite

O método de Gauss-Hermite é utilizado para resolver integrais da forma

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx, \quad (10)$$

em que $f(x)$ é uma função definida no intervalo $[a, b]$ dos reais. A integral definida em (10) é aproximada por meio de uma soma ponderada, entre a função $f(x)$, avaliada nos chamados pontos de Gauss, e w_i , pesos de integração (GIVENS; HOETING, 2005). Dessa

forma, tem-se que

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^n w_i f(x_i), \quad (11)$$

em que n é o número de pontos utilizados para a aproximação dado em (11), x_i , são as raízes de um polinômio de Hermite $H_n(x)$ com $(i = 1 < 2, \dots, n)$, e os pesos w_i são dados por

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2}.$$

Uma desvantagem desse método, é que dependendo do suporte da função $f(x)$, os pontos selecionados podem ou não estar dentro da área de interesse. De acordo com Givens e Hoeting (2005), uma forma de corrigir esse problema, é reescalando tais pontos, de maneira que os mesmos passam a estar dentro da área com maior densidade da função de $f(x)$, chamado método da Quadratura Adaptativa de Gauss-Hermite abordado a seguir.

Método da quadratura adaptativa de Gauss-Hermite

Segundo Molenberghs e Verbeke (2005), este método faz com que os pontos da quadratura sejam centrados e escalonados, assim a função $e^{-x^2} f(x)$ assume uma distribuição gaussiana. A média desta distribuição é igual à moda $\hat{x}$ de $\log[e^{-x^2} f(x)]$ e

$$\left[-\frac{\partial^2}{\partial z^2} \log[f(z)w(z)]|_{z=\hat{z}} \right]^{-1}.$$

Logo, a quadratura adaptada será

$$z_q^+ = \hat{z} + \left[-\frac{\partial^2}{\partial z^2} \log[f(z)w(z)]|_{z=\hat{z}} \right]^{-1/2} z_q,$$

com peso sob a forma

$$w_q^+ = \left[-\frac{\partial^2}{\partial z^2} \log[f(z)w(z)]|_{z=\hat{z}} \right]^{-1/2} \frac{w(z_q^+)}{w(z_q)} w_q.$$

Assim, a integral da quadratura de Gauss-Hermite adaptada é uma aproximação da soma ponderada expressa por

$$\int f(z)w(z)dz \approx \sum_{q=1}^Q w_q^+ f(z_q^+).$$

Ao assumir $Q = 1$, isto é, apenas um ponto da quadratura, o método da Quadratura Adaptativa é reduzido ao método de Laplace (MOLENBERGHS; VERBEKE, 2005).

2.2.3 Técnicas de diagnóstico para os MLGMs

Os resíduos utilizados para verificação do ajuste nos MLGMs são os apresentados em (7) e (8) na seção 2.1.7. Entretanto, outros tipos de resíduos podem ser encontrados em Urbano (2012).

2.2.4 Modelo misto para ensaios casualizados em blocos com subamostras

Para um melhor entendimento dos termos a serem incluídos em um modelo, um exemplo da análise de variância quando a variável resposta tem distribuição gaussiana será apresentado nesta seção.

Considere um experimento casualizado em blocos com I tratamentos, J blocos e K indivíduos por parcela, em que a variável resposta é contínua, Y_{ijk} com $i = 1, \dots, I; j = 1, \dots, J$ e $k = 1, \dots, K$. Pode-se supor inicialmente que $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$.

Um modelo estatístico para Y_{ijk} pode ser dado por

$$Y_{ijk} = \beta_0 + \beta_j + \tau_i + \varepsilon_{ij} + \xi_{k(ij)},$$

em que β_0 é uma constante, β_j é o efeito fixo de bloco j , τ_i é o efeito fixo de tratamento i , ε_{ij} é o efeito aleatório de parcela com $\varepsilon_{ij} \sim N(0, \sigma_1^2)$ e $\xi_{k(ij)}$ é o efeito aleatório de indivíduo dentro da parcela com $\xi_{k(ij)} \sim N(0, \sigma_2^2)$.

Pelo diagrama de Hasse (BRIEN, 2007) apresentado na Figura 1, pode-se obter o quadro da análise de variância com as esperanças dos quadrados médios, apresentado na Tabela 1. A estatística F para testar igualdade entre médias de tratamentos, isto é, para testar a hipótese $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$, é dada por

$$F = \frac{\text{QM Tratamento}}{\text{QM Resíduo}} \sim F_{I-1, (I-1)(J-1)}.$$

Figura 1 - Diagrama de Hasse para o experimento casualizado em blocos com subamostras

Tabela 1 - Análise de variância do modelo misto para experimento conduzido em blocos com subamostras

Causas de Variação	GL	SQ	E[QM]
Bloco	$J - 1$	$Y'Q_BY$	$\sigma_2^2 + k\sigma_1^2 + q_B(\Psi)$
Entre Parcela[Bloco]	$J(I - 1)$	$Y'Q_{PB}Y$	
Genótipo	$I - 1$	$Y'Q_TY$	$\sigma_2^2 + k\sigma_1^2 + q_T(\Psi)$
Resíduo	$(J - 1)(I - 1)$	$Y'Q_{Res}Y$	$\sigma_2^2 + k\sigma_1^2$
Entre indivíduos [P^B]	$IJ(K - 1)$	$Y'Q_{SPB}Y$	σ_2^2

$$Q_{res} = M_{pb} - M_b - M_g + M_G, \quad M_G = X_G(X'_G X_G)^{-1} X'_G, \quad M_b = X_b(X'_b X_b)^{-1} X'_b, \quad M_T = X_T(X'_T X_T)^{-1} X'_T, \\ M_{pb} = X_{pb}(X'_{pb} X_{pb})^{-1} X'_{pb}, \quad M_{spb} = X_{spb}(X'_{spb} X_{spb})^{-1} X'_{spb}$$

2.3 Fundamento Genético

Em avaliações da proporção (taxa) de indução de haploidia no melhoramento da cultura do milho, a técnica de duplo-haploides para obtenção das linhagens permite ganhos econômicos de até 100% de homozigose (BORÉM; MIRANDA, 2013), blocos gênicos

completos, menores gastos com mão de obra e menor tempo de obtenção de linhagens.

A obtenção de linhagens por meio da metodologia duplo-haploide tem sido largamente utilizada nos últimos cinco anos, abrangendo grande número de espécies, incluindo milho, trigo, cevada e arroz (LI et al., 2013). As técnicas de indução de haploides podem ser *in vivo* ou *in vitro*. Neste trabalho, a técnica de manejo utilizada é a *in vivo*, a qual, baseia-se no uso de linhagens indutoras de haploidia, que carregam o gene R-navajo e expressam a cor roxa devido à formação da proteína antocianina. Tal pigmentação constitui importante marcador fenotípico, pois, as sementes haploides apresentam coloração roxa apenas no endosperma e o embrião permanece na cor branca (CHASE; NANDA, 1965).

Ao ocorrer dupla fecundação normal para formação da semente, são necessários dois núcleos germinativos paternos, um para fecundar a oosfera feminina formando o embrião diploide (2N) e outro para fecundar os dois núcleos polares femininos formando o endosperma triploide (3N). Na formação de uma semente haploide, os núcleos polares são fecundados formando um endosperma roxo (3N), porém, a oosfera sofre perda da fertilização (CHALYK et al., 2003; SARKAR; COE JR, 1966) ou eliminação cromossômica após a fertilização (FISCHER, 2004; ZHANG et al., 2008), dando origem a um embrião branco haploide (N).

O sistema de coloração relatado não é perfeito, uma vez que inúmeras interações gênicas pleiotrópicas ou epigenéticas podem interferir na expressão do gene R-navajo de tal forma que a cor roxa pode não ser expressa, levando à seleção de falso-positivos e falso-negativos. Para evitar isso, uma segunda etapa conduzida em campo é necessária. Após a duplicação cromossômica com colchicina e formação de mudas DH0 (primeira planta duplicada e provável duplo-haploide), essas devem ser levadas ao campo e autofecundadas formando o DH1 (primeira geração após duplicação cromossômica). O DH1 será levado ao campo e cada espiga semeada em uma fileira para observação das plantas, pois se o DH0 já for uma linhagem, ao plantar o DH1 em fileira na geração seguinte, essa deverá ser homogênea confirmando a predominância de locos homozigóticos e formação do duplo-haploide. Se DH1 apresentar segregação, ou seja, presença de heterozigosidade, trata-se de um falso-positivo (não houve indução de haploidia) (DEIMLING; RÖBER; GEIGER, 1997; PRASANNA; CHAIKAM; MAHUKU, 2012).

A taxa de indução de haploidia, geralmente, é muito baixa, possivelmente por ineficiência do indutor tropical. Os indutores de regiões temperadas, geralmente, são mais

eficientes como os genótipos comerciais MHI (6-8%), RWS (8-10%) e PHI (12-14%). As primeiras linhagens descobertas foram W23 (1,5%) e Stock 6 (3%) (KERMICLE, 1969; COE, 1959), que deram origem a grande parte das linhagens indutoras conhecidas atualmente.

3 MATERIAL E MÉTODOS

3.1 Material

O conjunto de dados que motivou este trabalho é proveniente de um experimento conduzido nos anos de 2015/2016 no Laboratório de Melhoramento de Plantas Alógamas sob a orientação do Dr. Roberto Frieth-Neto do Departamento de Genética da USP/ ESALQ, Piracicaba-SP, localizada à 22°42'30" sul e 47°38'30" oeste.

O estudo consistiu na avaliação da taxa de indução de haploidia em milho. Para isso, o indutor gimnogenético tropicalizado de haploidia, LI- ESALQ, foi utilizado como genitor masculino em cruzamentos para formação de haploides com cinco híbridos tropicais comerciais divergentes fenotipicamente (Genótipo 1 - 2B587PW, Genótipo 2 - 30F53H, Genótipo 3 - BM820, Genótipo 4 - DKB390 e Genótipo 5 - STATUS VIPTERA). Além disso, os genitores femininos foram avaliados em duas gerações: F1, composta pelo genótipo comercial e a F2, a qual consiste na autofecundação da geração F1, gerando uma máxima variabilidade devido à segregação genética (BERNARDO, 2009)).

Estes genótipos obtidos das gerações F1 e F2 foram semeados, respectivamente, em áreas contíguas (Locais 1 e 2) em linhas de 0,85 m, intercalados com o indutor para que houvesse condições ideais para a realização da polinização ao acaso, em um campo isolado. Foi utilizado o delineamento causalizado em blocos, com três repetições (Figuras 2 e 3), com número variável de subamostras nas parcelas. Cada parcela foi constituída de famílias de irmãos germanos, gerando um total de 900 observações (a geração F1 com 532 observações e a geração F2 com 368 observações).

Figura 2 - Croqui do experimento conduzido

Figura 3 - Esquema do experimento conduzido

Para identificar as sementes haploides, as espigas de cada planta do genitor feminino foram colhidas e debulhadas. A identificação foi realizada visualmente, por meio da coloração do endosperma e do embrião pelo marcador R-navajo (CHASE; NANDA, 1965), como mostra a Figura 4. Os números de sementes identificadas como haploides foram utilizadas para determinar a taxa de indução de haploidia para cada cruzamento.

Figura 4 - Tipos de sementes: À esquerda, tipos de coloração do embrião. À direita, na parte superior, semente haploide (embrião branco) e, na parte inferior, semente diploide (embrião roxo) (MELCHINGER et al., 2013)

O objetivo do experimento foi verificar se existiam diferenças em relação à taxa de indução de haploidia para os diferentes cruzamentos propostos entre o indutor LI-ESALQ e os cinco híbridos comerciais nas duas gerações, F1 e F2.

O esquema do banco de dados pode ser visualizado no anexo deste trabalho.

3.2 Métodos

3.2.1 Análise descritiva e modelos

De acordo com a literatura, dependendo do tipo de população, as gerações F1 e F2 apresentam variabilidades distintas (FALCONER; MACKAY, 1995) conforme descrito na Seção 3.1. Neste trabalho, inicialmente, foram feitas análises descritivas para identificar

o comportamento da distribuição dos dados, em cada geração, e, posteriormente, propor modelos adequados.

Análises individuais

Seja Y_{ijk} o número de sementes haploides, de um total de m_{ijk} sementes, obtido na k -ésima espiga ($k = 1, \dots, k_{ij}$) do i -ésimo genótipo ($i = 1, \dots, 5$) no j -ésimo bloco ($j = 1, 2, 3$). Para a análise descritiva dos dados, por geração, foram obtidos os valores médios por casela para o número de sementes haploides ($\bar{y}_{ij}$), para o número de sementes por espiga ($\bar{m}_{ij}$) e para proporção de sementes haploides ($\bar{p}_{ij} = \sum p_{ijk}/k_{ij}$). Adicionalmente, foram feitos *Box Plots*, para y_{ijk} , m_{ijk} e p_{ijk} e gráficos de dispersão para $\bar{y}_{ij}$, $\bar{m}_{ij}$ e $\bar{p}_{ij}$ versus genótipos por blocos, para as gerações F1 e F2, além de gráficos de dispersão para p_{ijk} versus m_{ijk} .

A fim de se entender a relação entre fatores utilizados e confundimentos para a geração dos dados apresenta-se, na Figura 5, o diagrama de Hasse, de forma semelhante ao apresentado na Seção 2.2.4.

Figura 5 - Diagrama de Hasse para o experimento casualizado em blocos com subamostras para análise individual

A partir do diagrama de Hasse apresentado na Figura 5, tem-se o esquema da análise do desvio resultante para as análises individuais das gerações F1 e F2 como mostra a Tabela 2, a qual leva em consideração a forma de casualização dos tratamentos.

Tabela 2 - Esquema da análise do desvio, considerando-se o diagrama de Hasse da Figura 5

Causas de Variação	Graus de liberdade	
	F1	F2
Bloco	2	2
Entre Parcela[Bloco]	12	12
Genótipo	4	4
Resíduo	8	8
Entre Espigas [B^P]	517	353

Neste experimento, a probabilidade de sucesso é a semente ser haploide, fato este que não pode ser considerado homogêneo, pois, no processo de pulverização, a espiga de cima recebe maior quantidade de pólen do que a espiga da parte inferior na mesma planta. Outro fator importante a ser considerado, é a falta de conhecimento sobre de qual planta cada espiga foi gerada, pois há uma dependência entre as espigas de mesma planta. Devido a essa falta de informação, tem-se um efeito aleatório a nível de indivíduo (espiga) e outro em nível de parcela.

A distribuição Binomial tem a pressuposição de independência entre os eventos de Bernoulli e probabilidade de sucesso constante (MYERS; MONTGOMERY D. C. E VINING, 2010). Sendo assim, pode ocorrer uma variação maior do que a esperada pelo modelo, sendo este fato, um indicativo para o fenômeno de superdispersão, o qual fora discutido na Seção 2.1.4. Assim, alguns modelos para tal análise foram propostos.

Inicialmente, admite-se que o número de sementes haploides Y_{ijk} tem distribuição binomial, com índice m_{ijk} e probabilidade de sucesso P_{ijk} , isto é, $Y_{ijk}|P_{ijk} \sim \text{Bin}(m_{ijk}, P_{ijk})$, e considera-se a função de ligação logística e o preditor linear expresso por

$$\eta_{ijk} = \log \left(\frac{P_{ijk}}{1 - P_{ijk}} \right) = \beta_0 + \beta_j + \tau_i + \varepsilon_{ij} \quad (\text{Modelo 1}) \quad (12)$$

em que β_0 é uma constante, β_j é o efeito fixo de bloco j , τ_i é o efeito fixo do genótipo i ,

ε_{ij} é um efeito aleatório de parcela, sendo $\varepsilon_{ij} \sim N(0, \sigma_p^2)$. Na realidade, esse é um modelo binomial-logístico-normal, isto é, um caso particular do MLGM.

Foram considerados ainda, modelos alternativos em que o preditor linear assumiu as expressões

$$\eta_{ijk} = \log \left(\frac{P_{ijk}}{1 - P_{ijk}} \right) = \beta_0 + \beta_j + \tau_i + \xi_{k(ij)} \quad (\text{Modelo 2}) \quad (13)$$

e

$$\eta_{ijk} = \log \left(\frac{P_{ijk}}{1 - P_{ijk}} \right) = \beta_0 + \beta_j + \tau_i + \varepsilon_{ij} + \xi_{k(ij)} \quad (\text{Modelo 3}) \quad (14)$$

em que ε_{ij} é um efeito aleatório de parcela e $\xi_{k(ij)}$ é um efeito aleatório de espigas, sendo $\xi_{k(ij)} \sim N(0, \sigma_e^2)$.

Para seleção dos modelos apresentados, inicialmente, realizou-se uma visualização gráfica com a intenção de identificar o “melhor” ajuste dentro de cada modelo proposto (1, 2 e 3). Para tal, foram construídos gráficos semi-normais de probabilidade com envelope simulado para os componentes do desvio, usando-se o pacote **hnp** (MORAL; HINDE; DEMÉTRIO, 2016), com nível de confiança aproximado de 0,95.

Sequencialmente, a fim de comparar os diferentes modelos, foram utilizados os critérios de *Akaike* (AIC) e Bayesiano (BIC). Além disso, para o modelo selecionado, foram feitos gráficos de resíduos dos componentes do desvio e dos resíduos de Pearson, versus índice.

Após encontrar o modelo que se ajustou melhor às observações, obtiveram-se as estimativas por máxima verossimilhança para os parâmetros utilizando a função **glmer** do pacote **lme4** (BATES et al., 2015) e construídos os intervalos de confiança assintóticos para o preditor linear por meio do pacote **coefplot2** (BOLKER; SU, 2011). Todas as análises foram realizadas utilizando-se o *software* R (R Core Team, 2015).

Análise conjunta

Para análise conjunta das observações nas duas gerações, seja Y_{ijks} o número de sementes haploides, de um total de m_{ijks} sementes, obtido na k -ésima espiga ($k = 1, \dots, k_{ij}$) do i -ésimo genótipo na s -ésima geração ($s = 1, 2$) do j -ésimo bloco ($j = 1, 2, 3$).

Na análise descritiva dos dados, foram feitos gráficos de interação para blocos com genótipo versus $\bar{p}_{ijk}$, gráficos de interação para blocos com geração versus $\bar{p}_{ijk}$ e gráficos

de interação para geração com genótipo versus $\bar{p}_{ijk}$.

A fim de se entender a relação entre fatores utilizados e confundimentos para a geração dos dados, apresenta-se na Figura 6, o diagrama de Hasse, de forma semelhante ao apresentado no item 2.2.4.

Figura 6 - Diagrama de Hasse para o experimento em estudo casualizado em blocos com subamostras para análise conjunta

A partir do diagrama de Hasse apresentado na Figura 6, tem-se a análise do desvio resultante para a análise conjunta das gerações como mostra a Tabela 3, a qual leva em consideração a forma de casualização dos tratamentos. Nota-se que o efeito de geração está totalmente confundido com o efeito de local.

Tabela 3 - Esquema da análise do desvio, considerando-se o diagrama de Hasse da Figura 6

Causas de Variação	Graus de liberdade
Local	1
Geração	1
Locais[Blocos]	4
Parcela[Locais^Blocos]	24
Genótipo	4
Geração#Genótipo	4
Resíduo	16
Espigas[Locais^Blocos ^Parcela]	870

Com base na função de ligação logística, em que o preditor linear apresenta a estrutura de um delineamento casualizado em blocos, e $Y_{ijks}|P_{ijks} \sim \text{Bin}(m_{ijks}, P_{ijks})$, o primeiro modelo proposto para a análise conjunta das observações das duas gerações é

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \xi_{k(ijs)} \quad (\text{Modelo 1}) \quad (15)$$

em que β_0 é uma constante de efeito fixo, γ_s é o efeito fixo da geração s , $\beta_{j(s)}$ é o efeito fixo de bloco j dentro do local s , τ_i é o efeito fixo do genótipo i , $\xi_{k(ijs)}$ é o efeito aleatório em nível de indivíduo (k -ésima espiga), sendo $\xi_{k(ijs)} \sim N(0, \sigma_e^2)$.

Um segundo modelo foi construído apenas com o efeito aleatório de parcela, isto é,

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \varepsilon_{ij(s)} \quad (\text{Modelo 2}) \quad (16)$$

em que β_0 é uma constante de efeito fixo, γ_s é o efeito fixo da geração s , $\beta_{j(s)}$ é o efeito fixo de bloco j dentro do local s , τ_i é o efeito fixo do genótipo i , $\varepsilon_{ij(s)}$ é um efeito aleatório de parcela dentro do local s , sendo $\varepsilon_{ij(s)} \sim N(0, \sigma_p^2)$.

Para a construção do próximo modelo, assumiram-se dois efeitos aleatórios, um para espigas e outro para parcelas, isto é,

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \varepsilon_{ij(s)} + \xi_{k(ijs)} \quad (\text{Modelo 3}) \quad (17)$$

em que β_0 é uma constante de efeito fixo, γ_s é o efeito fixo da geração s , $\beta_{j(s)}$ é o efeito fixo de bloco j dentro do local s , τ_i é o efeito fixo do genótipo i , $\varepsilon_{ij(s)}$ é um efeito aleatório de parcelas dentro do local s com sendo $\varepsilon_{ij(s)} \sim N(0, \sigma_p^2)$, e $\xi_{k(ijs)}$ é o efeito aleatório em nível de indivíduo (k -ésima espiga) com $\xi_{k(ijs)} \sim N(0, \sigma_e^2)$.

Os modelos que se seguem, levam em consideração o possível efeito de interação entre geração e genótipo em cada situação mostrada nos Modelos 1, 2 e 3. Deste modo, têm-se

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \theta_{is} + \xi_{k(ijs)} \quad (\text{Modelo 4}) \quad (18)$$

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \theta_{is} + \varepsilon_{ij(s)} \quad (\text{Modelo 5}) \quad (19)$$

e

$$\eta_{ijks} = \log \left(\frac{P_{ijks}}{1 - P_{ijks}} \right) = \beta_0 + \gamma_s + \beta_{j(s)} + \tau_i + \theta_{is} + \varepsilon_{ij(s)} + \xi_{k(ijs)} \quad (\text{Modelo 6}) \quad (20)$$

em que β_0 é uma constante de efeito fixo, γ_s é o efeito fixo da geração s , $\beta_{j(s)}$ é o efeito fixo de bloco j dentro do local s , τ_i é o efeito fixo do genótipo i , θ_{is} é o efeito de interação do genótipo i com a geração s , $\varepsilon_{ij(s)}$ é um efeito aleatório de parcela dentro do local s com $\varepsilon_{ij(s)} \sim N(0, \sigma_p^2)$, $\xi_{k(ijs)}$ é o efeito aleatório em nível de indivíduo (k -ésima espiga) com $\xi_{k(ijs)} \sim N(0, \sigma_e^2)$.

Os Modelos 1, 2, 3, 4, 5 e 6 foram comparados e selecionados, inicialmente, por meio dos critérios AIC e BIC. Além disso, foram construídos os gráficos semi-normais de probabilidades com envelopes simulados para os resíduos componentes dos desvios, com um nível de confiança de 0,95. Adicionalmente, foram feitos os gráficos dos componentes do desvio e resíduos de Pearson, versus os índices das observações.

Após encontrar o modelo que se ajustou melhor as observações, construiu-se o gráfico das proporções estimadas versus observadas. Obtiveram-se para tal modelo as estimativas por máxima verossimilhança para os parâmetros por meio da função **glmer** do pacote **lme4** (BATES et al., 2015). Todas as análises foram realizadas utilizando-se o *software* R (R Core Team, 2015).

4 RESULTADOS E DISCUSSÃO

Análises individuais

Pela análise descritiva apresentada na Tabela 4, observa-se que houve uma variabilidade grande quanto ao números de espigas (k_{ij}) obtidos por parcela (26 a 52 para a geração F1 e 5 a 43 para a geração F2), números médios de sementes haploides por espiga ($\bar{m}_{ij}$), números médios de sementes haploides por espiga ($\bar{y}_{ij}$) e proporções médias de sementes haploides ($\bar{p}_{ij}$).

Tabela 4 - Valores referentes ao número de espigas (k_{ij}), números médios de sementes ($\bar{m}_{ij}$) e sementes haploides ($\bar{y}_{ij}$); e proporções médias de sementes haploides ($\bar{p}_{ij}$), por parcela

Geração	Bloco		Genótipo 1	Genótipo 2	Genótipo 3	Genótipo 4	Genótipo 5
F1	Bloco 1	k	33	37	26	32	42
		$\bar{m}$	498,1515	535,4324	518,9231	530,5938	479,8571
		$\bar{y}$	3,7879	15,9459	2,9231	3,0313	2,9524
		$\bar{p}$	0,00787	0,02933	0,00612	0,00593	0,00651
	Bloco 2	k	39	36	35	45	52
		$\bar{m}$	474,3590	584,5833	600,4571	511,6000	446,1923
		$\bar{y}$	2,7436	14,4444	4,4000	5,6667	4,3846
		$\bar{p}$	0,00591	0,02686	0,00795	0,01135	0,01178
	Bloco 3	k	28	36	29	26	36
		$\bar{m}$	507,5357	550,5000	537,7241	593,0385	496,8333
		$\bar{y}$	5,3929	13,5556	5,9310	8,8462	6,0000
		$\bar{p}$	0,01133	0,02519	0,01281	0,01579	0,01232
Total		100	109	90	103	130	
F2	Bloco 1	k	24	43	26	21	39
		$\bar{m}$	456,7500	403,2093	497,5385	389,5714	385,7436
		$\bar{y}$	3,5417	11,6512	4,6538	5,7619	5,8205
		$\bar{p}$	0,00799	0,03165	0,01048	0,01478	0,01493
	Bloco 2	k	20	14	17	28	33
		$\bar{m}$	439,3500	342,0714	315,8235	368,9286	321,1818
		$\bar{y}$	3,1000	10,5714	5,2353	6,2143	4,6970
		$\bar{p}$	0,00734	0,03482	0,01921	0,01604	0,01571
	Bloco 3	k	12	19	5	27	40
		$\bar{m}$	264,0833	360,7368	246,6000	350,3333	308,6750
		$\bar{y}$	3,2500	21,3158	3,0000	5,1852	4,6750
		$\bar{p}$	0,01409	0,06155	0,01706	0,01355	0,01472
Total		56	76	48	76	112	

Por meio da Figura 7, observa-se que os números de sementes por espiga (m_{ijk}), os números de sementes haploides (y_{ijk}) e as proporções de sementes haploides variam bastante. Tem-se que o genótipo 2 (30F53H) destaca-se em relação aos demais, quanto aos números e proporções de sementes haploides, em ambas as gerações.

Figura 7 - Dados de indução de haploidia: *Box Plots* para números totais de sementes (m_{ijk}) e sementes haploides (y_{ijk}); e proporções de sementes haploides (p_{ijk}), por genótipo, para as gerações F1 e F2

Figura 8 - Dados de indução de haploidia: Gráficos de interação entre os números médios de sementes ($\bar{m}_{ij}$) e de sementes haploides ($\bar{y}_{ij}$); e proporções médias de sementes haploides ($\bar{p}_{ij}$) versus genótipos, por blocos, para as gerações F1 e F2

Na Figura 8, foram apresentados os gráficos dos números médios de sementes ($\bar{m}_{ij}$), números médios de sementes haploides ($\bar{y}_{ij}$) e proporções médias de sementes haploides ($\bar{p}_{ij}$) por parcela, considerando-se os diferentes genótipos e diferentes blocos para cada uma das gerações. Existem evidências de interação entre blocos e genótipos, apenas, para o número médio de sementes por parcela mas não para a proporção de sementes haploides que é a variável resposta. Adicionalmente, pela Figura 9, pode-se notar que, aparentemente, não existe uma relação entre proporção de sementes haploides e número de sementes por espiga, ou seja, a proporção de sementes haploides não depende do tamanho da espiga.

Figura 9 - Dados de indução de haploidia: Gráfico de dispersão para proporção de sementes haploides (p_{ijk}) versus número de sementes por espiga (m_{ijk}), para as gerações F1 e F2

Ajustando-se os Modelos 1, 2 e 3, com preditores lineares dados por (12), (13) e (14), como definidos na Seção 3.2, foram obtidos os resultados apresentados na Tabela 5 para a geração F1 e a Tabela 6 para a geração F2. Observou-se que em ambas as tabelas, há presença de superdispersão, uma vez que, o desvio residual é muito maior que os graus de liberdade. Os Modelos 2 e 3 apresentaram os menores valores para o critério AIC e BIC, porém, não houve convergência para o Modelo 3. Assim, o Modelo 2 foi o selecionado como o melhor modelo. Este resultado foi confirmado pelos gráficos semi-normais de probabilidade com envelope de simulação apresentados na Figura 10, e pelos gráficos dos componentes do desvio e resíduos de Pearson versus índice, para as gerações F1 e F2, apresentados na

Figura 11.

Tabela 5 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados para geração F1

Fonte de variação	Modelo 1		Modelo 2		Modelo 3	
	GL	Desvio	GL	Desvio	GL	Desvio
Bloco	2	13,44	2	26,62	2	18,48
Genótipo	4	90,23	4	155,20	4	90,48
Espiga[bloco^parcela]	524	3920,70	524	2962,60	524	2960,00
σ_p^2	0,0328				0,0136	
σ_e^2			0,5020		0,4864	
AIC	3936,7		2978,6		2978,0	
BIC	3971,0		3012,8		3016,5	

Tabela 6 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados para geração F2

Fonte de variação	Modelo 1		Modelo 2		Modelo 3	
	GL	Desvio	GL	Desvio	GL	Desvio
Bloco	2	11,88	2	5,21	2	15,20
Genótipo	4	115,42	4	69,28	4	69,20
Espiga[bloco^parcela]	360	3461,40	360	2081,10	359	2081,10
σ_p^2	0,02398				$2,265 \times 10^{-07}$	
σ_e^2			0,8137		0,8136	
AIC	3477,4		2097,1		2099,1	
BIC	3508,7		2128,3		2134,2

Figura 10 - Dados de indução de haploidia: Gráficos semi-normais de probabilidade com envelopes de simulação para os Modelos 1, 2 e 3, para as gerações F1 e F2

Figura 11 - Dados de indução de haploidia: Gráficos dos componentes do desvio e resíduos de Pearson, versus índices, para as gerações F1 e F2 considerando o modelo binomial-logístico-normal com efeito aleatório de indivíduo

Na Tabela 7, apresentam-se as estimativas para os parâmetros do Modelo 2, com seus respectivos erros-padrão e componentes de variância. Nota-se que, tanto os componentes de variância quanto as estimativas de erro padrão foram maiores na geração F2 quando comparados com a geração F1, confirmando deste modo, que tal geração tem uma maior variabilidade do material genético.

Tabela 7 - Estimativas dos parâmetros para as gerações F1 e F2 para modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 2)

Parâmetros	Estimativas (Erros-padrão)	
	F1	F2
β_0	-5,30 (0,11)	-5,15 (0,16)
β_2	0,18 (0,09)	0,10 (0,13)
β_3	0,49 (0,10)	0,30 (0,14)
τ_2	1,21 (0,12)	1,38 (0,19)
τ_3	0,07 (0,13)	0,38 (0,21)
τ_4	0,35 (0,12)	0,48 (0,19)
τ_5	0,20 (0,12)	0,43 (0,18)
σ_e^2	0,5020	0,8137

Na Figura 12, foram construídos intervalos de confiança aproximados para as estimativas dos preditores lineares dos genótipos, sem considerar os efeitos dos blocos e do intercepto. Nota-se que o genótipo 2 apresentou maior indução de sementes haploides, pois se destaca dos demais genótipos apresentando estimativa superior aos demais, além do fato, que é o único cujo intervalo de confiança que não se sobrepõe aos demais. Ainda é notável, que os genótipos 1, 3, 4 e 5 apresentam algum efeito de haploidia, porém em menor escala, pois o zero não está contido em nenhum destes intervalos calculados. Vale ressaltar que estes intervalos aproximados foram construídos com 95% de confiança.

Figura 12 - Dados de indução de haploidia: Intervalo de confiança para o preditor linear para as gerações F1 e F2, considerando o modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 2)

Análise conjunta

Os gráficos da Figura 13 mostraram que existem evidências que, provavelmente, não há presença de interação significativa entre blocos com genótipos, blocos com geração e geração com genótipos.

Figura 13 - Dados de indução de haploidia: Gráficos de interação entre bloco versus a taxa média de indução ($\bar{p}_{ijk}$), por genótipo; interação entre bloco versus a taxa média de indução ($\bar{p}_{ijk}$), por geração; interação de geração versus a taxa média de indução ($\bar{p}_{ijk}$), por genótipo, respectivamente

Ajustando-se os modelos com preditores lineares expressos em (15), (16), (17), (18), (19) e (20), e que são descritos na Seção 3.2, foram obtidos os resultados apresentados nas Tabelas 8 e 9.

Tabela 8 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados - sem efeito de interação geração $\times$ genótipo

	Modelo 1		Modelo 2		Modelo 3	
Fonte de variação	GL	Desvio	GL	Desvio	GL	Desvio
Bloco[Local]	4	31,02	4	116,62	4	30,98
Geração	1	18,64	1	166,15	1	18,79
Genótipo	4	212,20	4	1723,72	4	212,08
Geração $\times$ Genótipo						
Espiga[Geração $\wedge$ Bloco $\wedge$ Parcela]	889	5057,9	889	7526,3	888	5057,9
σ_p^2			$1,5 \times 10^{-16}$		$3,7 \times 10^{-08}$	
σ_e^2	0,6325				0,6324	
AIC	5079,9		7548,3		5081,9	
BIC	5132,7		7601,2		5139,5	

Tabela 9 - Análise de desvio e estimativas dos componentes de variância para os modelos ajustados - com efeito de interação geração $\times$ genótipo

	Modelo 4		Modelo 5		Modelo 6	
Fonte de variação	GL	Desvio	GL	Desvio	GL	Desvio
Bloco[Local]	4	31,25	4	116,44	4	31,04
Geração	1	18,92	1	174,74	1	18,94
Genótipo	4	212,65	4	1713,96	4	212,35
Geração $\times$ Genótipo	4	2,06	4	14,03	4	2,09
Espiga[Geração $\wedge$ Bloco $\wedge$ Parcela]	885	5055,9	885	7512	884	5055,9
σ_p^2			$3,1 \times 10^{-18}$		$2,4 \times 10^{-07}$	
σ_e^2	0,6297				0,6292	
AIC	5085,9		7542		5087,9	
BIC	5157,9		7614		5164,7	

Muito embora foram apresentados os modelos descritos anteriormente, os Modelos 4, 5 e 6 apresentaram problemas de convergência, entretanto, os resultados foram semelhantes aos Modelos 1, 2 e 3 sem considerar o efeito de interação. Dentre os modelos que convergiram, o modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1) apresentou os menores valores de AIC e BIC, e ainda pelo princípio da parcimônia, escolhe-se tal modelo como o mais adequado ao conjunto de dados. Na Figura 14, são apresentados os gráficos semi-normais de probabilidade com envelope de simulação e na Figura 15, têm-se os gráficos dos componentes do desvio e resíduos de Pearson versus índices, os quais corroboram a escolha do modelo.

Figura 14 - Dados de indução de haploidia: Gráficos semi-normais de probabilidade com envelope de simulação para os Modelos 1, 2 e 3

Figura 15 - Dados de indução de haploidia: Gráfico dos componentes do desvio e dos resíduos de Pearson para o modelo binomial-logístico-normal com efeito aleatório de indivíduo

Uma possível explicação para os pontos discrepantes presentes nas Figuras 14 e 15, é que, em geral, as coletas de dados no campo são realizadas por um grande número de avaliadores, e embora exista uma padronização estabelecida nos protocolos de fenotipagem, sempre existe uma diferença de experiência de cada avaliador. Além disso, a verificação da taxa de indução de haploidia é realizada por meio da proporção da contagem do número de sementes haploides em relação à contagem de todas as sementes de uma espiga. A definição se a semente é haploide, ou não, é visual e ainda pode ser mais difícil por não existir um padrão de expressão do gene R-navajo (CHAIKAM et al., 2015).

Outro fator que também deve-se incluir no erro é a polinização das espigas, e, conseqüentemente, a indução de sementes haploides. Devido às condições do experimento em estudo, sabe-se que a polinização foi ao acaso, induzido pelo vento, em campo isolado com apenas uma única linhagem indutora. A direção ou o sentido do vento é imprevisível, podendo favorecer, ou não, a polinização de plantas devido à localização no campo. Assim, a direção do vento em campo, não pode ser controlada, sendo uma das variantes que elevam o resíduo.

A Tabela 10 mostra as estimativas dos parâmetros para o modelo selecionado.

Tabela 10 - Estimativas dos parâmetros do modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)

Parâmetros	Estimativas (Erros-padrão)
β_0	-5,39 (0,10)
γ_{22}	0,38(0,11)
$\beta_{2(1)}$	0,18 (0,10)
$\beta_{3(1)}$	0,51 (0,10)
$\beta_{2(2)}$	0,11 (0,12)
$\beta_{3(2)}$	0,30 (0,12)
τ_2	1,28 (0,10)
τ_3	0,18 (0,11)
τ_4	0,40 (0,11)
τ_5	0,30 (0,10)
σ_e^2	0,6325

Na Tabela 11, têm-se as proporções médias da taxa de indução de haploidia considerando o modelo binomial-logístico-normal com efeito aleatório de indivíduo. Observa-se nesta tabela que, as médias das proporções provenientes do G2 apresentaram maior taxa de indução nas duas gerações quando comparadas aos demais genótipos, sendo maior na geração F2. Além disso, na geração F2, existem plantas que provavelmente apresentam melhor desempenho na taxa de indução devido à variabilidade existente nessa geração. Entretanto, deve ser lembrado que o efeito de gerações está confundido com efeitos de locais.

Tabela 11 - Proporções médias para o ajuste do modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)

Gerações	Genótipos					Médias de Gerações
	G1	G2	G3	G4	G5	
F1	0,0076	0,0270	0,0085	0,0105	0,0097	0,0127
F2	0,0088	0,0392	0,0129	0,0147	0,0146	0,0180
Médias de Genótipos	0,0164	0,0662	0,0214	0,0252	0,0243	

Na Figura 16, apresenta-se o gráfico das proporções observadas versus valores estimados para o modelo binomial-logístico-normal com efeito aleatório de indivíduo, o que confirma novamente a escolha do modelo.

Figura 16 - Dados de indução de haploidia: Gráfico das proporções observadas versus estimadas de indução de haploidia para o modelo binomial-logístico-normal com efeito aleatório de indivíduo (Modelo 1)

5 CONCLUSÕES E CONSIDERAÇÕES FINAIS

As proporções de sementes haploides obtidas no experimento planejado para a comparação de cinco genótipos de milho tropical, apresentaram uma dispersão maior (superdispersão) do que aquela prevista pelo modelo binomial, para ambas as gerações F1 e F2. Essa variabilidade extra, foi, provavelmente, induzida por vários fatores aleatórios não controlados.

Dentre os modelos propostos para a análise, o que melhor se ajustou aos dados foi o binomial-logístico-normal, considerando-se efeito aleatório de espigas, tanto para as análises individuais como para a análise conjunta de ambas as gerações.

Com base nos intervalos de confiança aproximados das análises individuais, o genótipo 30F53H destacou-se em relação aos demais, apresentando a maior taxa de indução de haploidia, em ambas as gerações. Estudos adicionais precisam ser feitos para obtenção dos intervalos de confiança para as médias marginais.

Em trabalhos futuros sugere-se um maior controle em relação às origens das espigas, uma vez que, há uma dependência entre as mesmas. Recomenda-se também, a utilização de avaliadores treinados para que não haja diferenças entre estes na identificação e seleção de sementes haploides, e que os mesmos possam ser identificados. Além disso, para que a comparação entre gerações fique livre de confundimentos há necessidade de um planejamento mais adequado do experimento.

REFERÊNCIAS

- AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: AKADEMINAI KIADO. **Second International Symposium on Information Theory**. Budapest, 1973. p. 267–281.
- ATKINSON, A C. **Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis**. Oxford: Clarendon, 1985. 282 p.
- BATES, Douglas; MÄCHLER, Martin; BOLKER, Ben; WALKER, Steve. **Fitting Linear Mixed-Effects Models Using lme4**. 2015. R package version 3.3.1. Disponível em: <<http://R-Forge.R-project.org>>. Acesso em: 16 Dezembro de 2016.
- BERNARDO, Rex. Should maize doubled haploids be induced among f1 or f2 plants? **Theoretical and applied genetics**, Springer, Berlin, v. 119, n. 2, p. 255–262, 2009.
- BOLKER, Ben; SU, Yu-Sung. **coefplot2: Coefficient plots**. 2011. R package version 0.1.3.2. Disponível em: <<http://R-Forge.R-project.org>>. Acesso em: 16 Dezembro de 2016.
- BORÉM, A.; MIRANDA, G. V. **Melhoramento de Plantas**. 6. ed. Viçosa: UFV, 2013. 523 p.
- BRIEN, C J. Determining the analysis of variance table. **Satatistical Modelling.**, 2007. Disponível em: <<http://chris.brien.name/ee2/course/SM06.pdf>>. Acesso em: 20 Abr. 2016.
- BROOKS, R.J.; JAMES, W.H.; GRAY, E. Modelling Sub-Binomial Variation in the Frequency of Sex Combinations in Litters of Pigs. **Biometrics**, JSTOR, Boston, v. 47, p. 403–417, 1991.
- CHAIKAM, Vijay; NAIR, Sudha K; BABU, Raman; MARTINEZ, Leocadio; TEJOMURTULA, Jyothsna; BODDUPALLI, Prasanna M. Analysis of effectiveness of r1-nj anthocyanin marker for in vivo haploid identification in maize and molecular markers for predicting the inhibition of r1-nj expression. **Theoretical and Applied Genetics**, Springer, New York, v. 128, n. 1, p. 159–171, 2015.
- CHALYK, S; BAUMANN, A; DANIEL, G; EDER, J. Aneuploidy as a possible cause of haploid-induction in maize. **Maize Genetics Cooperation Newsletter**, Iowa, v. 77, p. 29, 2003.
- CHASE, Sherret S; NANDA, Devender K. Comparison of variability in inbred lines and monoploid-derived lines of maize (zea mays l.). **Crop Science**, Madison, v. 5, n. 3, p. 275–276, 1965.
- CHRIÉL, Mariann; STRYHN, Henrik; DAUPHIN, Gwenaëlle. Generalised linear mixed models analysis of risk factors for contamination of danish broiler flocks with salmonella typhimurium. **Preventive veterinary medicine**, Elsevier, Fort Collins, v. 40, n. 1, p. 1–17, 1999.
- COE, E H. A line of maize with high haploid frequency. **The American Naturalist**, JSTOR, Chicago, v. 93, n. 873, p. 381–382, 1959.

CONAB, COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento de Safra Brasileira: grãos - 8 levantamento, safra 2015/2016**, Brasília, 2016.

Disponível em:

<http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_06_10_14_42_08_boletim_graos_maio_2016_-_final.pdf>. Acesso em: 30 Abr. 2016.

CORDEIRO, Gauss Moutinho; DEMÉTRIO, Clarice G. B. **Modelos Lineares Generalizados e Extensões**. São Paulo, 2013. Disponível em:

<<http://www.lce.esalq.usp.br/arquivos/aulas/2013/LCE5868/livro.pdf>>. Acesso em: 02 jul. 2015.

CORDEIRO G.M. E LIMA NETO, E.A. **Modelos Paramétricos**. 2006. Disponível em: <http://www.ufjf.br/clecio_ferreira/files/2013/05/Livro-Gauss-e-Eufrasio.pdf>. Acesso em: 02 de março de 2016.

DE LAAT, AMM; GOHDE, W; VOGELZAKG, MJDC. Determination of ploidy of single plants and plant populations by flow cytometry. **Plant breeding**, Wiley Online Library, Malden, v. 99, n. 4, p. 303–307, 1987.

DEIMLING, S; RÖBER, F; GEIGER, HH. Methodik und genetik der in-vivo-haploideninduktion bei mais. **Vortr Pflanzenzüchtg**, Germany, v. 38, p. 203–224, 1997.

FALCONER, DS; MACKAY, TFC. Introduction to quantitative genetics. **Longman**, American Genetic Association, Newport, v. 19, n. 8, p. 1, 1995.

FINNEY, David John. **Probit Analysis**. 3. ed. England: Cambridge University Press, 1971. 390 p.

FISCHER, E. **Molecular genetic studies on the occurrence of paternal DNA transmission during in vivo haploid induction in maize (Zea mays)**. 2004. Tese (Doutorado) — Dissertation, University of Hohenheim, Stuttgart, Germany, 2004.

FISHER, R.A. Two new properties of mathematical likelihood. **Philosophical Transactions of the Royal Society A**, Nottingham, v. 144, p. 285–307, 1934.

GANIO, Lisa M; SCHAFER, Daniel W. Diagnostics for overdispersion. **Journal of the American Statistical Association**, Taylor & Francis, Boston, v. 87, n. 419, p. 795–804, 1992.

GIVENS, G. H; HOETING, J. A. **Computational statistics**. Jonh Wiley and sons, Hoboken, 2005.

GRIFFITHS, DA. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. **Biometrics**, JSTOR, Boston, p. 637–648, 1973.

HINDE, J.; DEMÉTRIO, Clarice G. B. Overdispersion: Models and Estimation. **Computational Statistics & Data Analysis**, Amsterdam, v. 27, p. 151–170, 1998.

KERMICLE, JL. Androgenesis conditioned by a mutation in maize. **Science**, Baltimore, v. 166, n. 3911, p. 1422–1424, 1969.

KOKONENDJI, Célestin C. Over-and underdispersion models. **Methods and Applications of Statistics in Clinical Trials: Planning, Analysis, and Inferential Methods**, John Wiley & Sons, Inc., Hoboken, v. 2, p. 506–526, 2014.

LI, Huihui; SINGH, Ravi P; BRAUN, Hans-Joachim; PFEIFFER, Wolfgang H; WANG, Jiankang. Doubled haploids versus conventional breeding in CIMMYT wheat breeding programs. **Crop Science**, The Crop Science Society of America, Inc., Madison, v. 53, n. 1, p. 74–83, 2013.

LIANG, Kung-Yee; MCCULLAGH, Peter. Case studies in binary dispersion. **Biometrics**, Boston, p. 623–630, 1993.

MCCULLAGH P. E NELDER, J. A. **Generalized Linear Models**. 2. ed. London: Chapman and Hall, 1989. 511 p.

MELCHINGER, Albrecht E; SCHIPPRACK, Wolfgang; WÜRSCHUM, Tobias; CHEN, Shaojiang; TECHNOW, Frank. Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. **Scientific Reports**, Nature Publishing Group, London, v. 3, 2013.

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. Springer, New York, 2005.

MORAL, Rafael A; HINDE, John; DEMÉTRIO, Clarice GB. **hnp: Half-Normal Plots with Simulation Envelopes**. 2016. R package version 1.0. Disponível em: <<https://CRAN.R-project.org/packages=hnp>>. Acesso em: 30 jul. 2016.

MYERS, R.H.; MONTGOMERY D. C. E VINING, G. G. **Generalized Linear Models: With Applications in Engineering and the Sciences**. 2. ed. New Jersey: John Wiley, 2010. 496 p.

NELDER, J.A; WEDDERBURN, R.W.N. Generalized Linear Models. **Journal of the Royal Statistical Society A**, Hoboken, v. 135, p. 370–384, 1972.

PAULA, Gilberto Alvarenga de. **Modelos de Regressão com apoio computacional**. São Paulo, 2013. Disponível em: <<http://www.ime.usp.br/~giapaula/livro.pdf>>. Acesso em: 14 Ago. 2015.

PRASANNA, BM; CHAIKAM, Vijay; MAHUKU, George. **Doubled haploid technology in maize breeding: theory and practice**. México: CIMMYT, 2012. 51 p.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <<http://www.R-project.org/>>. Acesso em: 20 Nov. 2016.

RIDOUT, Martin S; BESBEAS, Panagiotis. An empirical model for underdispersed count data. **Statistical Modelling**, SAGE Publications, New York, v. 4, n. 1, p. 77–89, 2004.

SARKAR, KR; COE JR, E H. A genetic analysis of the origin of maternal haploids in maize. **Genetics**, Genetics Society of America, Bethesda, v. 54, n. 2, p. 453, 1966.

SCHWARZ, Gideon. Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Disponível em: <[http://dx. doi. org/10.1214/aos/1176344136](http://dx.doi.org/10.1214/aos/1176344136)>. Acesso em: 26 Nov. 2016.

URBANO, Mariana Regassi. **Modelagem de mortalidade natural e superdispersão em dados entomológicos**. 2012. 117 p. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

VIEIRA, Afranio Marcio Correa. **Modelos para dados de proporções com superdispersão aplicados ao controle biológico**. 1998. 61 p. Dissertação (Mestrado) — Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1998.

WILLIAMS, David A. Extra-binomial variation in logistic linear models. **Applied statistics**, Hoboken, v. 31, n. 2, p. 144–148, 1982.

WILLIAMS, D. A. Residuals in generalized linear models. **International Biometrics Conference**, Tokyo, p. 59–68, 1984.

ZHANG, Zili; QIU, Fazhan; LIU, Yongzhong; MA, Kejun; LI, Zaiyun; XU, Shangzhong. Chromosome elimination and in vivo haploid production induced by stock 6-derived inducer line in maize (*zea mays* l.). **Plant cell reports**, Springer, Knoxville, v. 27, n. 12, p. 1851–1860, 2008.

APÊNDICE

Apêndice A

Algumas distribuições na família exponencial.

Normal

Se uma variável aleatória $Y \sim N(\mu, \sigma^2)$, sua função densidade será escrita por

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-(y - \mu)^2}{2\sigma^2} \right], \quad \mu \in \mathbb{R}, \ y \in \mathbb{R}, \ \sigma > 0,$$

em que $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \theta^2/2 = \mu^2/2$ e $c(y, \phi) = 1/2\{\log(2\pi\sigma^2) + y^2/\sigma^2\}$. Tem-se, ainda, que a função de variância $V(\mu) = 1$,

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \phi V(\mu) = \sigma^2.$$

Poisson

Sendo uma variável aleatória $Y \sim \text{Poisson}(\mu)$, sua função densidade ficará escrita da forma

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}, \quad \mu > 0, \ y \in \{0, 1, \dots\},$$

em que $\theta = \log(\mu)$, $\phi = 1$, $b(\theta) = e^\theta = \mu$ e $c(y, \phi) = -\log(y!)$. Tem-se, ainda, que a função de variância $V(\mu) = \mu$,

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \phi V(\mu) = \mu.$$

Binomial

Assumindo que $Y \sim \text{Bin}(n, \pi)$, em que Y é o número de sucessos em n ensaios independentes de Bernoulli, sendo cada um com probabilidade de sucesso π , a função de probabilidade de Y pode ser escrita sob a forma

$$f(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad \pi \in [0, 1], \ y \in \{0, 1, \dots, n\},$$

em que $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = 1$, $b(\theta) = -n \log(1 - \pi)$ e $c(y, \phi) = \log\left(\binom{n}{y}\right)$. Tem-se, ainda, que a função de variância $V(\mu) = n\pi(1 - \pi) = \frac{\mu}{n}(n - \mu)$,

$$E(Y) = n\pi = \mu \quad \text{e} \quad \text{Var}(Y) = \phi V(\mu) = \frac{\mu}{n}(n - \mu).$$

Gama

Se $Y \sim \text{Gama}(\alpha, \frac{\alpha}{\mu})$, em que Y é uma variável aleatória, então a função densidade é

$$f(y; \alpha; \mu) = \frac{\left(\frac{\alpha}{\mu}\right)^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\left\{\frac{-\alpha y}{\mu}\right\}, \quad \alpha > 0, \mu > 0, y > 0,$$

em que $\theta = -1/\mu$, $\phi = 1/\alpha$, $b(\theta) = -\log(-\theta)$ e $c(y, \phi) = \alpha \log(\alpha y) - \log \Gamma(\alpha) - \log y$. Tem-se, ainda, que a função de variância $V(\mu) = \mu^2$,

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \phi V(\mu) = \frac{\mu^2}{\alpha}.$$

Normal Inversa

Se uma variável aleatória $Y \sim \text{NI}(\mu, \phi)$, sua função densidade é expressa como

$$f(y; \mu, \phi) = \frac{1}{\sqrt{2\pi\phi y^3}} \exp\left[\frac{-(y - \mu)^2}{2\phi y \mu^2}\right], \quad \mu \in \mathbb{R}, y > 0, \phi > 0,$$

em que $\theta = -1/2\mu^2$, $\phi = \sigma^2$, $b(\theta) = -(-2\theta)^{1/2}$ e $c(y, \phi) = 1/2[\log(2\pi\phi y^3) - 1/(\phi y)]$. Tem-se, ainda, que a função de variância $V(\mu) = \mu^3$,

$$E(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \phi V(\mu) = \mu^3 \sigma^2.$$

Na Tabela 12, encontra-se um resumo das distribuições apresentadas.

Tabela 12 - Distribuições de probabilidade pertencentes à família exponencial

Distribuição	$b(\theta)$	θ	ϕ	$V(\mu)$
Normal	$\theta^2/2$	μ	σ^2	1
Poisson	e^θ	$\log(\mu)$	1	μ
Binomial	$n \log(1 + e^\theta)$	$\log\{\mu/(n - \mu)\}$	1	$\frac{\mu}{n}(n - \mu)$
Gama	$-\log(-\theta)$	$-1/\mu$	$1/\alpha$	μ^2
Normal Inversa	$-(-2\theta)^{1/2}$	$-1/2\mu^2$	σ^2	μ^3

Apêndice B

Função desvio para algumas distribuições de probabilidade (PAULA, 2013).

Normal

Sendo $\theta_i = \mu_i$, tem-se que $\tilde{\theta}_i = y_i$ e $\hat{\theta}_i = \hat{\mu}_i$, e o desvio será dado por

$$D_p = 2 \sum_{i=1}^n [y_i(y_i - \hat{\mu}_i) + \hat{\mu}_i^2/2 - y_i^2/2] = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2,$$

que, naturalmente, coincide com a soma de quadrados de resíduos do modelo de regressão linear ordinário.

Poisson

Aqui, $\theta_i = \log(\mu_i)$, tem-se que $\tilde{\theta}_i = \log(y_i)$ para a $y_i > 0$ e $\hat{\theta}_i = \log(\hat{\mu}_i)$, e o desvio será dado por

$$D_p = 2 \sum_{i=1}^n [y_i \log(y_i/\hat{\mu}_i) + (\hat{\mu}_i - y_i)],$$

Se $y_i = 0$, o i -ésimo termo de $D(\mathbf{y}; \hat{\mu})$ vale $2\hat{\mu}_i$, e assim, para o modelo de Poisson, segue como

$$d^2(y_i; \hat{\mu}_i) = \begin{cases} 2\{y_i \log(y_i/\hat{\mu}_i) + (\hat{\mu}_i - y_i)\} & \text{se } y_i > 0, \\ 2\hat{\mu}_i & \text{se } y_i = 0, \end{cases}$$

em que $D_p = \sum_{i=1}^2 d_i^2$.

Binomial

Assumindo que $Y_i \sim \text{Binomial}(n_i, \mu_i)$, $i = 1, \dots, k$, tem-se que $\tilde{\theta}_i = \log\{y_i/(n_i - y_i)\}$ para $0 < y_i < n_i$, e $\hat{\theta}_i = \log\{\hat{\mu}_i/(n_i - \hat{\mu}_i)\}$, então o desvio é expresso por

$$D_p = 2 \sum_{i=1}^n [y_i \log(y_i/\hat{\mu}_i) + (n_i - y_i) \log\{(n_i - y_i)/(n_i - \hat{\mu}_i)\}].$$

Porém, quando $y_i = 0$ e $y_i = n_i$, o i -ésimo termo de $D(\mathbf{y}; \hat{\mu})$ será $2n_i \log[n_i/(n_i - \hat{\mu}_i)]$ e $2n_i \log(n_i/\hat{\mu}_i)$, respectivamente. Reescrevendo, para a binomial, cada desvio será expresso por

$$d^2(y_i; \hat{\mu}_i) = \begin{cases} y_i \log(y_i/\hat{\mu}_i) + (n_i - y_i) \log\{(n_i - y_i)/(n_i - \hat{\mu}_i)\} & \text{se } 0 < y_i < n_i, \\ 2n_i \log[n_i/(n_i - \hat{\mu}_i)] & \text{se } y_i = 0, \\ 2n_i \log(n_i/\hat{\mu}_i) & \text{se } y_i = n_i, \end{cases}$$

em que $D_p = \sum_{i=1}^3 d_i^2$.

Gama

Sendo $\tilde{\theta}_i = -1/y_i$ e $\hat{\theta}_i = -1/\hat{\mu}_i$, o desvio para valores positivos será escrito na forma

$$D_p = 2 \sum_{i=1}^n [\log(\hat{\mu}_i/y_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i].$$

Se algum componente de y_i for igual a zero, o desvio será indeterminado. McCullagh P. e Nelder (1989) sugerem substituir $D(\mathbf{y}; \hat{\mu})$ da forma

$$D_p^* = 2\phi C(\mathbf{y}) + 2\phi \sum_{i=1}^n \log \hat{\mu}_i + 2\phi \sum_{i=1}^n y_i/\hat{\mu}_i,$$

em que, $C(\mathbf{y})$ será uma função arbitrária, todavia limitada. Pode-se usar por exemplo, $C(\mathbf{y}) = \sum_{i=1}^n y_i/(1 + y_i)$.

Normal Inversa

Nesse caso, $\tilde{\theta}_i = -1/2y_i^2$ e $\hat{\theta}_i = -1/2\hat{\mu}_i^2$, o desvio será expresso por

$$D_p = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / (y_i \hat{\mu}_i^2).$$

Apêndice C - Programação para ajuste do modelo

```

#####
## Modelos propostos para geração F1
#####

f1<-subset(dados.ind,GERACAO=="1")

BLOC01<-f1$BLOCO
GENOTIPO1<-f1$GENOTIPO
GERACAO1<-f1$GERACAO

resp1 <- with(f1, cbind(HAPLOIDE, TOTAL-HAPLOIDE))
ind1 <- factor(1:dim(resp1)[1])

#####
model1 <- glmer(resp1 ~ BLOC01 + GENOTIPO1 + (1|GENOTIPO1:BLOC01), family =
binomial, data=f1)
summary(model1)
anova(model1)

model2 <- glmer(resp1 ~ BLOC01 + GENOTIPO1 + (1|ind1) , family = binomial,
data=f1)
summary(model2)
anova(model2)

model3 <- glmer(resp1 ~ BLOC01 + GENOTIPO1 + (1|GENOTIPO1:BLOC01) + (1|ind1),
family = binomial, data=f1)
summary(model3)
anova(model3)

#####

#####
## Modelos propostos para geração F2
#####

f2<-subset(dados.ind,GERACAO=="2")

```


64

```
BLOC02<-f2$BLOC0
```

```
GENOTIPO2<-f2$GENOTIPO
```

```
GERACA02<-f2$GERACA0
```

```
resp2 <- with(f2, cbind(HAPLOIDE, TOTAL-HAPLOIDE))
```

```
ind2 <- factor(1:dim(resp2)[1])
```

```
#@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

```
model1 <- glmer(resp2 ~ BLOC02 + GENOTIPO2 + (1|GENOTIPO2:BLOC02), family =  
binomial, data=f2)
```

```
summary(model1)
```

```
anova(model1)
```

```
model2 <- glmer(resp2 ~ BLOC02 + GENOTIPO2 + (1|ind2) , family = binomial,  
data=f2)
```

```
summary(model2)
```

```
anova(model2)
```

```
model3 <- glmer(resp2 ~ BLOC02 + GENOTIPO2 + (1|GENOTIPO2:BLOC02) + (1|ind2),  
family = binomial, data=f2)
```

```
summary(model3)
```

```
anova(model3)
```

```
#@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

```
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

```
## Modelos propostos para Análise conjunta das gerações
```

```
$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
```

```
HAPLOIDE <- dados.ind$HAPLOIDE
```

```
TOTAL <- dados.ind$TOTAL
```

```
GENOTIPO<-dados.ind$GENOTIPO
```

```
EPOCA<-dados.ind$EPOCA
```

```
GERACA0<-dados.ind$GERACA0
```

```
resp <- with(dados.ind, cbind(HAPLOIDE, TOTAL-HAPLOIDE))
```

```
ind <- factor(1:dim(resp)[1])
```

[illegible]

```
model1<- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +
(1|ind), family=binomial, data = dados.ind)
summary(model1)
anova(model1)
```

```
model2<- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +  
(1|PARCE:GERACAO), family=binomial, data = dados.ind)  
summary(model2)  
anova(model2)
```

```
model3 <- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +
(1|PARCE:GERACAO) + (1|ind), family=binomial, data = dados.ind)
summary(model3)
anova(model3)
```

```
model4<- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +
GERACAO:GENOTIPO + (1|ind), family=binomial, data = dados.ind)
summary(model4)
anova(model4)
```

```
model5<- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +  
GERACAO:GENOTIPO + (1|PARCE:GERACAO), family=binomial, data = dados.ind)  
summary(model5)  
anova(model5)
```

```
model6<- glmer(resp ~ BLOCO:GERACAO + GERACAO + GENOTIPO +  
GERACAO:GENOTIPO + (1|PARCE:GERACAO) + (1|ind), family=binomial,  
data = dados.ind)  
summary(model6)  
anova(model6)
```

[illegible]

ANEXO

Esquema do banco de dados

	GENOTIPO	GERACAO	EPOCA	ESPIGA	TOTAL	HAPLOIDE	PERC.INDU
1	30F53H	1	1	1	467	7	1,4989293
2	30F53H	1	1	2	681	7	1,0279001
...	...	...	...	...	...	...	...
37	30F53H	1	1	37	309	15	4,8543689
38	30F53H	1	2	1	730	4	0,5479452
...	...	...	...	...	...	...	...
73	30F53H	1	2	36	318	4	1,2578616
74	30F53H	1	3	1	521	19	3,6468330
...	...	...	...	...	...	...	...
109	30F53H	1	3	36	571	8	1,4010508
110	2B587PW	1	1	1	561	1	0,1782531
...	...	...	...	...	...	...	...
209	2B587PW	1	3	28	415	1	0,2409639
210	STATUS	1	1	1	553	8	1,4466546
...	...	...	...	...	...	...	...
209	2B587PW	1	3	28	415	1	0,2409639
210	STATUS	1	1	1	553	8	1,4466546
...	...	...	...	...	...	...	...
429	BM820	1	3	29	287	10	3,4843206
430	DKB390	1	1	1	682	6	0,8797654
...	...	...	...	...	...	...	...
532	DKB390	1	3	26	611	21	3,4369885
533	30F53H	2	1	1	425	11	2,5882353
...	...	...	...	...	...	...	...
608	30F53H	2	3	19	308	4	1,2987013
609	2B587PW	2	1	1	555	4	0,7207207
...	...	...	...	...	...	...	...
664	2B587PW	2	3	12	190	6	3,1578947
665	STATUS	2	1	1	322	1	0,3105590
...	...	...	...	...	...	...	...
776	STATUS	2	3	40	143	1	0,6993007
777	BM820	2	1	1	318	0	0,0000000
...	...	...	...	...	...	...	...
824	BM820	2	3	5	249	6	2,4096386
825	DKB390	2	1	1	482	7	1,4522822
...	...	...	...	...	...	...	...
900	DKB390	2	3	27	178	0	0,0000000

Figura 17 - Esquema do banco de dados