CAPÍTULO 2

ESTIMAÇÃO E PREDIÇÃO SOB MODELO LINEAR MISTO COM ÊNFASE NA ORDENAÇÃO DE MÉDIAS DE TRATAMENTOS GENÉTICOS

RESUMO

O presente artigo propôs-se a refletir teoricamente o processo de estimação/ predição de médias de tratamentos, nos delineamentos em blocos, com ênfase nas suas aplicações em testes de genótipos, no melhoramento vegetal. Neste sentido, procurou-se comparar as análises baseadas no modelo linear fixo (análise intrablocos) e no modelo linear misto com genótipos aleatórios (análise recuperando informação intertratamentos), buscando identificar os fatores que podem determinar diferentes classificações genotípicas. A análise teórica permitiu constatar que a abordagem de modelo misto (com tratamentos aleatórios), comparativamente às análises tradicionais (médias marginais e análise intrablocos), em geral, leva a: *i)* maior homogeneidade das médias de tratamentos; e *ii)* seleção de diferentes tratamentos genéticos, quando a variância genotípica for baixa em relação à variância do erro e os ensaios forem não ortogonais e desbalanceados. Ademais, se os tratamentos forem oriundos de várias populações, a predição *BLUP* poderá determinar diferente classificação das médias de tratamentos, em relação à análise intrablocos, mesmo sob ortogonalidade e balanceamento.

Palavras-chave*: modelo linear misto, delineamentos em blocos, predição de médias, médias BLUP, ordenação de tratamentos, selecionando genótipos.*

ESTIMATION AND PREDICTION UNDER LINEAR MIXED MODELS EMPHASIZING THE RANKING OF MEANS OF GENETIC TREATMENTS

ABSTRACT

This study was aimed at reviewing the theory of estimation/prediction of treatment means, in randomized block designs, emphasizing aspects of interest to plant breeders. Comparisons were made between analyses based on fixed (intrablock) and mixed (with random treatments effects - recovering intergenotypic information) linear models for identifying determining factors that may affect the classification of genotypes. It could be verified that the mixed model approach, in comparison with the traditional analyses (marginal means and intrablock analysis) in general leads to: *i)* more uniformly distributed treatment means; and *ii)* selection of different genetic treatments when the genetic variance is small, relative to the environmental variance, and designs are non-orthogonal and unbalanced. In addition, if treatments of distinct reference populations are evaluated in the same experiment, *BLUP* prediction can lead to different ranking of means, in comparison with the intrablock analysis, even if designs are balanced and orthogonal.

Key words*: linear mixed models, analysis of block designs, estimation and prediction of means, BLUP means, ranking of treatments, ordering and selecting genotypes.*

1. INTRODUÇÃO

No melhoramento de plantas ainda é comum o uso de análise baseada em modelo fixo para a estimação de médias de tratamentos (ex: genótipos), mesmo quando estes foram obtidos por amostragem numa população. Isto é, em situações em que o modelo é tipicamente misto, pois inclui, além de efeitos fixos, os efeitos aleatórios dos genótipos. Em boa parte dos casos, a modelagem mista é utilizada, com o rigor da suposição, apenas para a estimação de componentes de variância e para a construção de testes "F" apropriados na análise da variância.

Entre as razões que levam os melhoristas práticos a não utilizarem predições baseadas em modelos mistos estão a falta de vivência com estes métodos e a sua pequena divulgação (Bueno Filho, 1997). Acrescenta-se que os efeitos prejudiciais da abordagem tradicional normalmente são tidos como mínimos, a ponto de não recompensar os esforços com a adoção da nova metodologia. A ordem de classificação dos genótipos, em geral, não se altera no caso de ensaios que seguem delineamentos ortogonais e balanceados. Assim, na prática, a estimação de médias admitindo-se modelo fixo (ex: análise intrablocos) quando, na verdade, é misto, não modificaria o resultado final da seleção.

Por outro lado, a ocorrência de desbalanceamento não planejado, decorrente da perda de parcelas, é um fato normal nesse tipo de experimentação. Ademais, nas fases preliminares do processo seletivo, quando os genótipos são numerosos e ainda possuem natureza aleatória (Piepho, 1994), é comum o uso de delineamentos não ortogonais como *BIB* (*blocos incompletos balanceados*) e *PBIB* (*blocos incompletos parcialmente balanceados*). Também têm ganhado aplicação crescente os delineamentos aumentados (Federer, 1956; 1958), os quais, por construção, são desbalanceados e não ortogonais. Nestes casos, a possibilidade de classificações genotípicas diferenciadas entre as duas abordagens analíticas é uma realidade. Assim, optar-se pela conveniência da suposição de um fator como fixo ou aleatório pode estar longe de ser prática inofensiva (Bueno Filho, 1997).

Atualmente, a metodologia de modelos mistos tem-se tornado mais acessível aos usuários graças à sua implementação em sistemas estatístico-computacionais de ampla divulgação como o *SAS*® (*Statistical Analysis System*). Logo, a sua rigorosa aplicação é perfeitamente exeqüível sempre que o modelo subjacente aos dados for de tal natureza, o que deverá garantir maior confiabilidade às estimativas. Neste caso, covariâncias biologicamente comprovadas passam a ser levadas em conta não só nos testes estatísticos, mas também na estimação e predição de efeitos de implicação

direta no ordenamento dos genótipos e, por conseguinte, no resultado da seleção. Os estimadores correspondentes, em geral, têm variância menor do que os de modelo fixo, sendo por isso mais eficientes (Henderson, 1975; Verbeke & Molenberghs, 1997; Federer, 1998).

O propósito deste artigo é apontar, por meio de explicitações teóricas, os fatores que podem determinar diferenças na classificação das médias genotípicas (de tratamentos), quando estas forem obtidas por modelos fixo ou misto, de análise. O desenvolvimento centra-se na abordagem de modelos lineares mistos, por razões de generalidade e de divulgação. A ênfase principal está num modelo de delineamento em blocos, admitindo os efeitos de blocos como fixos e os de tratamentos como aleatórios. Entretanto, procurou-se também avaliar a extensão das constatações obtidas em algumas variações deste modelo.

2. UM MODELO DE DELINEAMENTO EM BLOCOS

Considere-se, para fins de ilustração, um delineamento experimental em blocos, com a tratamentos (genótipos) de efeitos g_i ($i=1,2,...,a$) e b blocos (completos ou incompletos) de efeitos b_j ($j=1,2,...,b$). Com o propósito de generalização, faz-se n_{ij} ser o número de vezes que o tratamento i aparece no bloco j ($n_{ij}=0,1,2,...$). Portanto: $\Sigma_i\Sigma_j n_{ij} = n$ (número de observações); $\Sigma_i n_{ij}=n_{.j}=k_j$ (tamanho ou número de parcelas do bloco j); e $\Sigma_j n_{ij}=n_{i.}=n_i$ (número de repetições do tratamento i); além de que: $\Sigma_i n_i=\Sigma_j k_j=n$. Denota-se ainda por Y_{ijr} a observação num caráter ou variável aleatória Y (observável), relativa à *r-ésima* parcela ($r=1,2,...,n_i$) que recebeu o tratamento i, identificada também pelo bloco j. Um modelo linear que caracteriza esse conjunto de dados pode ser:

$$Y_{ijr} = m + b_j + g_i + e_{ijr}\ ; \qquad \text{com:} \quad e_{ijr} \sim N(0, \sigma_e^2);$$
$$g_i \sim N(0, \sigma_g^2);$$
$$E(Y_{ijr}) = m + b_j\ ; \text{ e } \quad Var(Y_{ijk}) = \sigma_g^2 + \sigma_e^2.$$

Neste modelo, o efeito de bloco (b_j) é assumido como fixo e o de tratamento (g_i), como aleatório. A constante m é de natureza sempre fixa e e_{ijr} é uma variável aleatória não observável. Isso caracteriza o que se conhece na literatura por um *modelo misto*, pois incorpora uma mistura de tipos de efeitos, fixos e aleatórios (Searle, 1987). Dessa forma, os tratamentos testados representam uma amostra de uma população de genótipos, cujas respostas são distribuídas *normalmente*, em torno de uma média comum ($\mu_P = m + \bar{b}$) e com variância σ_g^2; ou seja, os tratamentos são realizações (valores) de variáveis

aleatórias não observáveis, as quais correspondem aos efeitos g_i's (desvios genotípicos aleatórios em relação à média μ_P). O tratamento estatístico desse tipo de modelo, no campo do melhoramento genético vegetal, tem recebido a denominação de *análise com recuperação da informação intervarietal* ou *intergenotípica* (Wolfinger *et al.*,1997; Federer, 1998).

Matricialmente, a expressão que generaliza essa e outras modelagens mistas alternativas pode ser escrita a partir do vetor $\mathbf{y}_{(nx1)}$ de observações, na forma do chamado *modelo linear misto geral*:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon \;\; ; \;\; \text{com:} \;\; \varepsilon \sim N(\phi,\mathbf{R}); \quad \gamma \sim N(\phi,\mathbf{G}); \quad E(\mathbf{y})=\mathbf{X}\beta; \;\; \text{e} \;\; Var(\mathbf{y})= \mathbf{V}_{(n)} = \mathbf{ZGZ'} + \mathbf{R} \; .$$

Neste caso, tem-se: todos os efeitos fixos reunidos no vetor paramétrico $\beta_{(px1)}$; os efeitos aleatórios no vetor paramétrico $\gamma_{(qx1)}$, exceto os erros que compõem o vetor $\varepsilon_{(nx1)}$; $\mathbf{X}_{(nxp)}$ e $\mathbf{Z}_{(nxq)}$ representam as matrizes de incidências dos efeitos contidos em β e γ, respectivamente; $\mathbf{G}_{(q)}$ e $\mathbf{R}_{(n)}$ são as matrizes de variâncias-covariâncias dos vetores aleatórios γ e ε, respectivamente; e as covariâncias entre vetores diferentes são assumidas nulas (Henderson, 1984). Aqui, por simplificação, adotar-se-á: $\mathbf{G}=\mathbf{I}_{(a)}\sigma_g^2$ e $\mathbf{R}=\mathbf{I}_{(n)}\sigma_e^2$, em que $\mathbf{I}_{(.)}$ denota uma matriz identidade e $a=q$ (número de níveis do fator aleatório).

3. ESTIMAÇÃO E PREDIÇÃO NUM MODELO LINEAR MISTO

Sabe-se que, sob as condições definidas anteriormente, o *método de quadrados mínimos ordinário* (*OLS*) não é mais um bom procedimento de estimação, pois assume a simples estrutura $\mathbf{V}=\mathbf{I}\sigma_e^2$, minimizando: $(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)$. A recomendação para esse caso recai sobre o *método de quadrados mínimos generalizado* (*GLS*), o qual contempla qualquer estrutura não singular de $\mathbf{V}$ (matriz de variâncias e covariâncias das observações), o que leva a minimizar a expressão mais genérica: $(\mathbf{y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta)$. Todavia, para isso é necessário conhecer a matriz $\mathbf{V}$, através de $\mathbf{G}$ e $\mathbf{R}$, ou, alternativamente, inserir alguma estimativa de $\mathbf{V}$ no problema de minimização *GLS*. Nesta última situação, o que deve ser feito é encontrar razoáveis estimativas de $\mathbf{G}$ e de $\mathbf{R}$ por meio de algum método de estimação. Entre estes, destacam-se pelo volume de aplicações, os

procedimentos *ANOVA* baseados no método dos momentos (Fisher, 1918; Henderson, 1953) e os métodos de máxima verossimilhança, *ML* (Hartley & Rao, 1967) e *REML* (Patterson & Thompson, 1971).

Em várias situações, a preferência tem sido dada aos métodos baseados em *verossimilhança*, os quais exploram a suposição de que γ e ε têm distribuição normal (Littell *et al.*, 1996; Verneque, 1994). Todavia, no caso de modelos mistos, não existe consenso sobre a melhor forma de estimar componentes de variância (Christensen *et al.*, 1992). Optando-se por *ML* ou *REML* é necessário, então, construir uma função objetivo e maximizá-la em relação a todos os parâmetros desconhecidos. Segundo a abordagem do SAS Institute (1997), com alguns cálculos é possível reduzir o problema de maximização apenas aos parâmetros em $\mathbf{G}$ e $\mathbf{R}$. Assim, os correspondentes logaritmos da função de verossimilhança (l_{ML}) e da função de verossimilhança restrita/residual (l_{REML}) são:

$$l_{ML}(\mathbf{G},\mathbf{R}) = -(1/2)\, log|\mathbf{V}| - (n/2)\, log(\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}) - (n/2)\,[1+ log(2\pi/n)]\,; \quad \text{e}$$

$$l_{REML}(\mathbf{G},\mathbf{R}) = -(1/2)\, log|\mathbf{V}| - (1/2)\, log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - [(n-p)/2]\, log(\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}) - [(n-p)/2]\{1+log[2\pi/(n-p)\}.$$

em que: $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$; p, aqui, é o posto (*rank*) de $\mathbf{X}$; e, dada uma matriz $\mathbf{A}$ qualquer, $\mathbf{A}^{-}$ denota uma inversa generalizada de $\mathbf{A}$ (tal que $\mathbf{A}\mathbf{A}^{-}\mathbf{A}=\mathbf{A}$).

Do processamento numérico de uma destas expressões, através de algoritmos iterativos como Newton-Raphson (implementado no *PROC MIXED* do sistema *SAS*) ou *EM* (*Expectation Maximization*), pode-se obter as estimativas *ML* ou *REML* de interesse ($\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$). Nos outros métodos, a estimação fundamenta-se na construção de formas quadráticas do tipo $\mathbf{y}'\mathbf{P}_t\mathbf{y}$ ($t=1,2,...,s$; sendo s o número de parâmetros σ_t^2 a serem estimados), as quais são equacionadas com suas esperanças matemáticas, $E(\mathbf{y}'\mathbf{P}_t\mathbf{y})$. As formas quadráticas equivalem às somas de quadrados obtidas na correspondente análise de variância e $E(\mathbf{y}'\mathbf{P}_t\mathbf{y})$ é uma função dos parâmetros σ_t^2. Descrições detalhadas são encontradas em Valério Filho (1991), Searle *et al.* (1992), Lopes *et al.* (1993), entre outros.

De posse dos valores paramétricos dos componentes de variância ($\mathbf{G}$ e $\mathbf{R}$) ou de suas estimativas ($\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$), passa-se, então, aos problemas de estimar o vetor de efeitos fixos β (ou uma função a ele associada) e de predizer o vetor de efeitos aleatórios γ (ou também alguma função de γ). Ambos os problemas podem ser resolvidos, simul-

taneamente, através das chamadas *equações de modelo misto* (EMM), propostas por Henderson em 1948 (Littell *et al*., 1996; Henderson, 1984):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \beta^0 \\ \tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}.$$

Não se conhecendo as matrizes **G** e **R**, simplesmente substituem-nas por $\hat{\mathbf{G}}$ e $\hat{\mathbf{R}}$. Manipulações de álgebra matricial levam, por conseguinte, às soluções do sistema:

$$\beta^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \text{ ; e}$$
$$\tilde{\gamma} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta^0) = \mathbf{C}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta^0)$$

em que: $\mathbf{C}=\mathbf{G}\mathbf{Z}'$ é a matriz de covariâncias entre **y** e γ (covariância entre observações fenotípicas e valores genotípicos verdadeiros).

A obtenção das soluções do sistema por meio destas expressões pode, entretanto, ter um alto custo computacional, haja vista a dimensão *nxn* da matriz **V** a ser invertida. O uso de uma inversa generalizada da matriz de coeficientes em EMM representa uma opção de menor esforço computacional, uma vez que esta matriz tem dimensão *(p+q)x(p+q)*, inferior a *nxn* (McLean *et al*., 1991). André (1999) deduz por absorção e substituição nas EMM, uma outra alternativa neste sentido.

É notório que, se $\mathbf{G}^{-1}$ tende para a matriz nula (ex: $\sigma_g^2 \to \infty$, no caso particular $\mathbf{G}=\mathbf{I}\sigma_g^2$), as EMM tendem para as equações de *GLS* para estimar β e γ, quando os componentes de γ são considerados fixos (Robinson, 1991). Por outro lado, quando $\mathbf{G}^{-1}$ domina as EMM (ex: $\sigma_g^2 \to 0$, sob $\mathbf{G}=\mathbf{I}\sigma_g^2$), $\tilde{\gamma}$ tende para zero. Nos casos intermediários, $\mathbf{G}^{-1}$ opera reduzindo (*shrinking*, em inglês) a magnitude das estimativas de γ supostamente fixo, até zero (SAS Institute, 1997).

Se **G** e **R** são conhecidas, β^0 (ou, mais provavelmente, alguma função estimável $\mathbf{L}'\beta^0$) é chamado *melhor estimador linear não viesado* (*BLUE - best linear unbiased estimator*) de β (ou de $\mathbf{L}'\beta$), e $\tilde{\gamma}$ é denominado *melhor preditor linear não viesado* (*BLUP - best linear unbiased predictor*) de γ. O uso do termo *preditor* tem apenas o propósito de distinguir estimadores de efeitos aleatórios daqueles de efeitos fixos (Robinson, 1991). Porém, como já mencionado, **G** e **R** geralmente são desconhecidas,

dispondo-se apenas de estimativas obtidas por algum método. Neste caso, os termos *BLUE* e *BLUP* não mais se aplicam, sendo apropriado substituí-los por *EBLUE* (*empirical best linear unbiased estimator*) e *EBLUP* (*empirical best linear unbiased predictor*), respectivamente (SAS Institute, 1997; Littell *et al.*, 1996). O termo *empírico* é adicionado, portanto, para indicar esse tipo de aproximação.

A correspondente matriz de variâncias-covariâncias dos parâmetros, $\mathbf{C}_{\beta^0,\tilde{\gamma}}$ ou $\hat{\mathbf{C}}_{\beta^0,\tilde{\gamma}}$, é dada por:

$$\mathbf{C}_{\beta^0,\tilde{\gamma}} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \quad \text{ou} \quad \hat{\mathbf{C}}_{\beta^0,\tilde{\gamma}} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z}+\hat{\mathbf{G}}^{-1} \end{bmatrix}^{-}.$$

Dado que os resultados de partição destas matrizes são gerais, conhecendo-se ou não as matrizes paramétricas **G** e **R**, pode-se simplesmente escrever:

$$\mathbf{C}_{\beta^0,\tilde{\gamma}} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}'_{21} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \quad ; \quad \text{com:} \quad \begin{cases} \mathbf{C}_{11} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \; ; \; \mathbf{C}_{21} = -\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\mathbf{C}_{11}; \text{ e} \\ \mathbf{C}_{22} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1} - \mathbf{C}_{21}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \end{cases}.$$

Note-se que $\mathbf{C}_{11}$ é a fórmula familiar da matriz de variâncias-covariâncias de β^0, solução de *quadrados mínimos generalizados*. Assim, entre outras propriedades, tem-se (Henderson, 1984; Searle *et al.*, 1992):

$$Var\,(\mathbf{L}'\beta^0) = \mathbf{L}'\mathbf{C}_{11}\mathbf{L};$$
$$Var\,(\tilde{\gamma}) = \mathbf{C}\mathbf{V}^{-1}\mathbf{C}' - \mathbf{C}\mathbf{V}^{-1}\mathbf{X}\mathbf{C}_{11}\mathbf{X}'\mathbf{V}^{-1}\mathbf{C} = \mathbf{G} - \mathbf{C}_{22}; \text{ e}$$
$$Var\,(\mathbf{L}'\beta^0 + \tilde{\gamma}) = Var\,(\mathbf{L}'\beta^0) + Var\,(\tilde{\gamma}).$$

4. EXPLICITAÇÃO DO *BLUP* DE γ

Para entender as conseqüências da suposição de aleatoriedade dos efeitos de tratamentos sobre suas estimativas de médias é necessário primeiramente analisar o preditor $\tilde{\gamma}$. É conveniente, portanto, derivar a expressão de componentes individuais do vetor $\tilde{\gamma}$, ou seja, de cada $\tilde{g}_i$, o *BLUP* de g_i ($i=1,2,...,a$). Sem perda de generalidade, omite-se, neste momento, o índice j (de blocos), mantendo-se apenas i e r, relativos ao tratamento e a sua repetição. Sob a estrutura de componentes de variância, $\mathbf{G}=\mathbf{I}\sigma_g^2$ e $\mathbf{R}=\mathbf{I}\sigma_e^2$, tem-se $\mathbf{V} = \bigoplus_{i=1}^{a}\mathbf{B}_i$ (sendo: $\mathbf{B}_i = \sigma_g^2\mathbf{J}_{(n_i)} + \sigma_e^2\mathbf{I}_{(n_i)}$; e $\bigoplus_{i=1}^{a}$, a operação matricial soma direta, isto é, a obtenção de uma matriz bloco diagonal com as a matrizes $\mathbf{B}_i$), logo:

$$\widetilde{\gamma} = \begin{bmatrix} \widetilde{g}_1 \\ \widetilde{g}_2 \\ \vdots \\ \widetilde{g}_a \end{bmatrix} = \mathbf{C}\mathbf{V}^{-1} \begin{bmatrix} Y_{11} - \hat{Y}_{11} \\ Y_{12} - \hat{Y}_{12} \\ \vdots \\ Y_{1n_1} - \hat{Y}_{1n_1} \\ Y_{21} - \hat{Y}_{21} \\ Y_{22} - \hat{Y}_{22} \\ \vdots \\ Y_{an_a} - \hat{Y}_{an_a} \end{bmatrix} ; \quad \text{com:} \quad \begin{cases} \mathrm{C} = \bigoplus_{i=1}^{a} \mathbf{1}'_{(n_i)} \sigma_g^2 \\ \mathrm{V}^{-1} = \bigoplus_{i=1}^{a} \mathrm{B}_i^{-1} \\ \mathrm{B}_i^{-1} = \dfrac{1}{\sigma_e^2} \left[\mathrm{I}_{(n_i)} - \lambda_i \mathrm{J}_{(n_i)} \right] \\ \lambda_i = \dfrac{\sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2} \end{cases}$$

onde: $\mathbf{1}'_{(n_i)}$, $\mathbf{I}_{(n_i)}$ *e* $\mathbf{J}_{(n_i)}$ são, respectivamente, um vetor linha de *uns*, uma matriz identidade e uma matriz quadrada com todos os elementos iguais a *um*, todos de ordem n_i.

Portanto, a matriz $\mathbf{CV^{-1}}$, de ordem *axn*, assume o formato:

$$\mathbf{C}\mathbf{V}^{-1} = \frac{\sigma_g^2}{\sigma_e^2} \begin{bmatrix} \overbrace{1-n_1\lambda_1 \quad 1-n_1\lambda_1 \quad \cdots \quad 1-n_1\lambda_1}^{n_1 \text{ colunas}} & \overbrace{0 \quad 0 \quad \cdots \quad 0}^{n_2 \text{ colunas}} & \cdots & \overbrace{0 \quad 0 \quad \cdots \quad 0}^{n_a \text{ colunas}} \\ 0 \quad 0 \quad \cdots \quad 0 & 1-n_2\lambda_2 \quad 1-n_2\lambda_2 \quad \cdots \quad 1-n_2\lambda_2 & \cdots & 0 \quad 0 \quad \cdots \quad 0 \\ \vdots \quad \vdots \quad \vdots \quad \vdots & \vdots \quad \vdots \quad \vdots \quad \vdots & \vdots & \vdots \quad \vdots \quad \vdots \quad \vdots \\ 0 \quad 0 \quad \cdots \quad 0 & 0 \quad 0 \quad \cdots \quad 0 & \cdots & 1-n_a\lambda_a \quad 1-n_a\lambda_a \quad \cdots \quad 1-n_a\lambda_a \end{bmatrix}.$$

Por conseguinte, como ilustram Searle *et al.* (1992), o preditor do efeito genotípico de um tratamento *i*, ou seja, o *BLUP*(g_i), fica determinado por:

$$BLUP(g_i) = \widetilde{g}_i = \frac{n_i \sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2} (\overline{Y}_i - \mu^0) = n_i \lambda_i (\overline{Y}_i - \mu^0) .$$

Isto resulta da multiplicação da *i-ésima* linha de $\mathbf{CV^{-1}}$ pelo vetor $(\mathbf{y} - \hat{\mathbf{y}})$ ou $(\mathbf{y} - \mathbf{X}\beta^0)$:

$$\begin{aligned} \widetilde{g}_i &= \frac{\sigma_g^2}{\sigma_e^2} \sum_{r=1}^{n_i} (1 - n_i \lambda_i)(Y_{ir} - \hat{Y}_{ir}) = \frac{\sigma_g^2}{\sigma_e^2} (1 - n_i \lambda_i) \sum_{r=1}^{n_i} (Y_{ir} - \hat{Y}_{ir}) = \frac{\sigma_g^2}{\sigma_e^2} \left(1 - n_i \frac{\sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2}\right) \sum_{r=1}^{n_i} (Y_{ir} - \hat{Y}_{ir}) \Rightarrow \\ \widetilde{g}_i &= \frac{\sigma_g^2}{\sigma_e^2} \left(\frac{\sigma_e^2 + n_i \sigma_g^2 - n_i \sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2}\right) \sum_{r=1}^{n_i} (Y_{ir} - \hat{Y}_{ir}) = \frac{\sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2} \sum_{r=1}^{n_i} (Y_{ir} - \hat{Y}_{ir}) = \lambda_i \sum_{r=1}^{n_i} (Y_{ir} - \hat{Y}_{ir}) = \lambda_i (Y_{i.} - \hat{Y}_{i.}) . \end{aligned} \tag{1}$$

Sabendo-se que: $Y_{i.} = n_i \overline{Y}_i$ e $\hat{Y}_{i.} = n_i \overline{\hat{Y}}_i$, tem-se, finalmente:

$$\widetilde{g}_i = n_i \lambda_i (\overline{Y}_i - \overline{\hat{Y}}_i) = \frac{n_i \sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2} (\overline{Y}_i - \mu^{0_i}) = n_i \lambda_i (\overline{Y}_i - \mu^{0_i}) \tag{2}$$

em que: $\overline{\hat{Y}}_i = \mu^{0_i}$, representa a média dos valores ajustados (para os efeitos fixos) nas parcelas que receberam o tratamento *i*, ou seja, a média ambiental esperada naquelas parcelas.

Nos delineamentos em blocos, μ^{0_i} representa a soma: (média geral) + (efeito médio dos blocos que receberam o tratamento *i*). Assim, reintroduzindo-se o índice de blocos (*j*) tem-se:

$$\mu^{0_i} = \frac{1}{n_i}\left(\sum_{j=1}^{b} n_{ij} m^0 + \sum_{j=1}^{b} n_{ij} b_j^0\right) = \frac{1}{n_i}\left(n_i m^0 + \sum_{j=1}^{b} n_{ij} b_j^0\right) = m^0 + \frac{1}{n_i}\sum_{j=1}^{b} n_{ij} b_j^0 = m^0 + \bar{b}^{0_i};$$

em que: m^0 e b_j^0 ($j=1,2,...,b$) são os elementos do vetor solução β^0; e $\bar{b}^{0_i}$ denota o efeito médio dos blocos que receberam o tratamento *i*.

Desse modo, o preditor do valor genotípico do tratamento *i*, o *BLUP*(g_i) ou simplesmente $\tilde{g}_i$, pode ainda ser escrito como:

$$\tilde{g}_i = \frac{n_i \sigma_g^2}{\sigma_e^2 + n_i \sigma_g^2}\left[\bar{Y}_i - \left(m^0 + \frac{1}{n_i}\sum_{j=1}^{b} n_{ij} b_j^0\right)\right] = n_i \lambda_i [\bar{Y}_i - (m^0 + \bar{b}^{0_i})] \qquad (3)$$

Este desenvolvimento mostra que o uso de uma constante μ^0, comum para todo *i*, como apresentam Searle *et al.* (1992, p. 271), não se aplica a todo tipo de situação e delineamentos. Por isso, deu-se aqui preferência à notação μ^{0_i}. Entretanto, $\bar{Y}_i$ também estima tudo isso ($m + \bar{b}^i$) mais o efeito do tratamento (g_i). Logo, a diferença ($\bar{Y}_i - \mu^{0_i}$) contém, realmente, só a informação relativa ao efeito aleatório do tratamento *i*, ou seja, a média de seus efeitos genotípicos estimados por parcela. É também oportuno observar que o termo $n_i\lambda_i$ é equivalente à *herdabilidade de médias* de tratamentos ($h^2_{\bar{Y}_i} = \sigma_g^2 \big/ \sigma_g^2 + \sigma_e^2/n_i$), conceito de ampla aplicação em genética.

Sob o ponto de vista experimental, é notória a importância do uso de repetições e casualização para uma predição imparcial de g_i através de $\tilde{g}_i$, ressaltando-se que isto também é fundamental no caso de modelos fixos e seus respectivos *BLUE*'s. O raciocínio teórico a seguir procurará, então, ilustrar a relevância destes princípios experimentais.

Para contornar as dificuldades de um tratamento matemático através de esperanças condicionadas, considere-se apenas as observações relacionadas a um dado genótipo *i*, cujo valor genotípico tenha sido predito por uma das expressões de $\tilde{g}_i$ (1, 2 ou 3). Para simplificação, considere-se ainda um delineamento binário, com $n_{ij}=0$ ou $n_{ij}=1$. Assim, pode-se afirmar que, sob as condições da predição, o valor esperado para cada uma das correspondentes unidades experimentais é dado por: $E(\hat{Y}_{ijr}) = m + b_j$. Contudo, para esse

conjunto particular de dados (relativos ao genótipo i), g_i é uma constante, desconhecida, mas não uma variável aleatória. Logo, $E(g_i)=g_i$. Ademais, desconhecendo-se o número de repetições e as regras de alocação dos tratamentos às parcelas, não se pode, ainda, assumir: $E(e_{ijr})=0$; mas, sim: $E(e_{ijr})=e_{ijr}$. Neste caso, reintroduzindo-se o índice j (de blocos) numa das expressões preliminares de $\widetilde{g}_i$ em (1), tem-se:

$$E(\widetilde{g}_i) = E[\lambda_i \sum_{r=1}^{n_i}(Y_{ijr} - \hat{Y}_{ijr})] \;=\; \lambda_i \sum_{r=1}^{n_i} E(m + b_j + g_i + e_{ijr} - \hat{Y}_{ijr})$$

$$E(\widetilde{g}_i) = \lambda_i \sum_{r=1}^{n_i}(m + b_j + g_i + e_{ijr} - m - b_j) = \lambda_i \sum_{r=1}^{n_i}(g_i + e_{ijr}) = n_i\lambda_i g_i + \lambda_i \sum_{r=1}^{n_i} e_{ijr}.$$

Observa-se que, sob n_i muito baixo e na ausência de casualização, $E(\widetilde{g}_i)$ carrega um termo $\lambda_i \sum_{r=1}^{n_i} e_{ijr} \neq 0$. Isto prejudica a qualidade do preditor $\widetilde{g}_i$ que terá um viés no sentido do efeito ambiental médio incidente nas parcelas que receberam o genótipo i. Todavia, sob casualização e n_i grande, pode-se assumir, tranqüilamente, que $\sum_{r=1}^{n_i} e_{ijr} = 0$ e, nestas condições, $\widetilde{g}_i$ resulta em predições não tendenciosas de g_i. Isto porque, garantidas as condições que tornam $E(e_{ijr})=0$, tem-se: $E(\widetilde{g}_i) = n_i\lambda_i g_i$; e, sob n_i grande, $n_i\lambda_i \to 1$ (herdabilidade máxima), o que implica em: $E(\widetilde{g}_i) \to g_i$. Neste caso, a informação dos outros genótipos torna-se irrelevante para a predição do valor genético do tratamento i (suposição inerente à modelagem fixa). Mas, considerando-se $0 \leq n_i\lambda_i \leq 1$, à medida que a herdabilidade diminui ($n_i\lambda_i \to 0$), o valor absoluto de $\widetilde{g}_i$ reduz-se, proporcionalmente, no sentido do valor esperado populacional, $E(g_i)=0$. Isto revela o aumento da importância do relacionamento entre os genótipos, na predição do valor genético de cada um.

Em síntese, $\widetilde{g}_i \to g_i$ quando a herdabilidade tende para *um*, senão, $\widetilde{g}_i \to E(g_i)$ à medida que a herdabilidade tende para *zero*. De fato, assumir $E(\widetilde{g}_i)=g_i$ somente é desejável se o experimento der condições para tal, isto é, se for capaz de fornecer informações individuais em número suficientemente grande ($h^2_{\bar{Y}_i} \to 1$). Senão, à medida que a informação individual diminui ($h^2_{\bar{Y}_i} \to 0$), é preferível predizer a performance de cada genótipo atribuindo-se um peso crescente às informações de seus "parentes", ou seja, fica cada vez mais seguro admitir $E(\widetilde{g}_i)=E(g_i)$. Assumir $E(\widetilde{g}_i)=g_i$ (suposição do modelo fixo), nestes casos, implica num risco crescente de $\widetilde{g}_i$ produzir estimativas muito pobres do verdadeiro efeito genotípico. Enfim, a abordagem de modelos mistos usufrui da flexibilidade de ponderar a informação individual, em detrimento daquela dos genótipos

aparentados, conforme a repetibilidade associada à primeira (Piepho, 1994). Já a metodologia de modelos fixos (*OLS*) não dispõe desta prudência.

5. O EFEITO "SHRINKAGE" NAS MÉDIAS *BLUP*

Antes de analisar a influência da suposição dos efeitos de tratamentos sobre a estimação ou predição das médias genotípicas, convém introduzir o conceito de médias *BLUP*. Como o vetor γ contém apenas os desvios genotípicos associados aos tratamentos, para predizer, por exemplo, a resposta fenotípica de um genótipo *i* num bloco (ambiente) *j* é necessário construir uma função linear de parâmetros fixos e aleatórios: $\widetilde{\mathbf{y}} = \mathbf{X}\beta^0 + \mathbf{Z}\widetilde{\gamma} \Rightarrow \widetilde{Y}_{ij} = (m^0 + b_j^0) + \widetilde{g}_i$. Tal expressão representa a soma: (valor médio do ambiente particular *j*) + (efeito do genótipo *i*). Contudo, se o pesquisador estiver interessado não apenas na informação dos efeitos genotípicos $\widetilde{g}_i$ (suficientes para o ordenamento e a seleção de genótipos), nem em predições de parcelas individuais ($\widetilde{Y}_{ij}$), mas na resposta média de cada genótipo, a expressão anterior também não responderá ao seu questionamento.

A nova função deve levar em conta, não o efeito individual de bloco, mas, o efeito médio de blocos, assumido comum para todos os genótipos sob comparação. Isto corresponde às chamadas médias de tratamentos ajustadas para os efeitos fixos do modelo. A expressão que atende a esse interesse é aqui denominada *BLUP*(μ_P+g_i) ou médias genotípicas *BLUP*, sendo dada por: $\widetilde{\overline{Y}}_i = (m^0 + \overline{b}^0) + \widetilde{g}_i$. Esta, sim, determina o ajuste das médias de cada tratamento para um mesmo referencial, a constante ($m^0 + \overline{b}^0$). Computacionalmente, este termo da expressão é obtido construindo-se uma função linear dos efeitos fixos, $\mathbf{L}'\beta$, comum para todos os tratamentos. A matriz $\mathbf{L}'_{[a\times(1+b)]}$ pode ter suas linhas todas iguais a: [*1* k_1/n k_2/n ... k_b/n], o que gera uma média ponderada dos efeitos de blocos pelos seus respectivos tamanhos, $\overline{b}^0$, à qual é adicionada a constante m^0. Acrescentando-se o preditor $\widetilde{g}_i$ tem-se, então, a média de interesse.

Searle *et al.* (1992) trata do problema de "estimar" ou "predizer" uma função linear do tipo: $\mathbf{w} = \mathbf{L}'\beta + \gamma$. Os autores comentam que, para $\mathbf{L}'\beta$ estimável, $\widetilde{\mathbf{w}} = \mathbf{L}'\beta^0 + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta^0)$ tem propriedades de melhor preditor linear não viesado (erro médio quadrático mínimo, linearidade em relação a **y** e não tendenciosidade), sendo, por

isso, chamado de *BLUP*(**w**): $\tilde{\mathbf{w}} = \mathbf{L}'\beta^0 + \tilde{\gamma}$. No presente caso, tem-se: *BLUP*(**w**)= $BLUP(\mu_P + g_i) = \hat{\mu}_P + \tilde{g}_i = (m^0 + \overline{b}^0) + \tilde{g}_i$, o que corresponde à média *BLUP* do genótipo *i*. Littell *et al.* (1996) ilustram o tratamento desse tipo de problema através do sistema *SAS*. Os autores referem-se a tais combinações lineares como *funções predizíveis*, para diferenciá-las das *funções estimáveis* que combinam apenas efeitos fixos.

Definido o significado das médias *BLUP*, reconsidere-se agora a expressão de $\tilde{g}_i$. O termo $n_i\lambda_i$ (a herdabilidade de médias) representa um peso aplicado ao mais simples estimador de g_i, o desvio ($\overline{Y}_i - \mu^{0_i}$). Vale esclarecer que, se o componente σ_g^2 não puder ser isolado, este ponderador embora presente em expressões do *BLUP*(g_i) não deve ser referido como herdabilidade, mas como *repetibilidade* do efeito genotípico estimado (Piepho, 1994). Mas, no presente caso, $n_i\lambda_i$ é de fato uma função da variância genética na população da qual os tratamentos foram amostrados (σ_g^2), da variância ambiental influenciando a resposta do caráter (σ_e^2) e do número de repetições do tratamento *i* (n_i). Assim, a questão de interesse imediato é avaliar a sua influência sobre o desvio genotípico médio ($\overline{Y}_i - \mu^{0_i}$).

Fazendo-se $\phi_g = \sigma_g^2 / \sigma_e^2$, relação que reflete a herdabilidade em nível de indivíduos ou parcelas ($h^2_{Y_{ij}} = {\sigma_g^2}/{\sigma_e^2+\sigma_g^2} = {\phi_g}/{1+\phi_g}$), tem-se:

$$n_i\lambda_i = \frac{n_i\sigma_g^2}{\sigma_e^2 + n_i\sigma_g^2} = \frac{n_i\phi_g\sigma_e^2}{\sigma_e^2 + n_i\phi_g\sigma_e^2} = \frac{\sigma_e^2 n_i\phi_g}{\sigma_e^2(1 + n_i\phi_g)} = \frac{n_i\phi_g}{1 + n_i\phi_g}.$$

Numa situação de variabilidade genética muito superior à ambiental ($h^2_{Y_{ij}}$ de valor bastante elevado), tem-se: $\phi_g \to \infty$ e $n_i\lambda_i \to 1$. Isto significa que a diferença ($\overline{Y}_i - \mu^{0_i}$) reflete, integralmente, o valor genotípico do tratamento *i* em relação à média μ_P da população, estimada por: $\hat{\mu}_P = m^0 + \overline{b}^0$. Nesta situação, a resposta média esperada de um genótipo *i*, o $BLUP(\mu_P + g_i)$, tende para: $\hat{\mu}_P + (\overline{Y}_i - \mu^{0_i})$. Em blocos completos balanceados, $\mu^{0_i} = \mu^0 = \mu_P$; então, sob $n_i\lambda_i \to 1$, o $BLUP(\mu_P + g_i)$ se reduz a $\overline{Y}_i$. Neste caso, as respostas genotípicas obtidas pelo preditor $BLUP(\mu_P + g_i)$ dispersam ao máximo

entre si, igualmente às respectivas médias marginais simples não ajustadas ($\overline{Y}_i = \sum_{k=1}^{n_i} Y_{ij} / n_i$).

Por outro lado, quando essa relação de variâncias for muito baixa ($\phi_g \rightarrow 0$), o referido peso também diminui ($n_i\lambda_i \rightarrow 0$) e a diferença ($\overline{Y}_i - \mu^{0_i}$) pouco ou nada informará sobre o valor genotípico individual do tratamento *i*. Seja porque os tratamentos não diferem substancialmente entre si ($\sigma_g^2 \cong 0$), seja por erro experimental muito elevado ($\sigma_e^2 \rightarrow \infty$). Neste caso, a resposta média esperada de um genótipo *i*, o $BLUP(\mu_P + g_i)$, tende para $\hat{\mu}_P$, pois, $\tilde{g}_i \rightarrow 0$. Ou seja, todos os tratamentos terão respostas preditas idênticas ($\hat{\mu}_P$). Desse modo, variações fenotípicas observadas entre genótipos não são mais do que flutuações erráticas em torno da média populacional μ_P; pois, suas propriedades genéticas individuais não são significativamente importantes ou, pelo menos, não puderam ser discriminadas pelo experimento em questão.

Sob tais circunstâncias, não haverá dispersão alguma entre as respostas genotípicas médias preditas. Com efeito, não seria coerente qualquer variação entre estas numa situação de ausência de variabilidade genética. Portanto, a abordagem de modelos mistos mostra-se coerente com a realidade e, por isso, é tida como conceitualmente mais completa (Resende *et al.,* 1996a; Bueno Filho, 1997).

Na maioria das situações práticas, entretanto, $n_i\lambda_i$ será um número entre *zero* e *um*, implicando num aproveitamento parcial da informação contida no desvio ($\overline{Y}_i - \mu^{0_i}$). Quanto mais esse número aproxima-se de *um* (ϕ_g e/ou n_i elevados) maior será este aproveitamento, haja vista o aumento da herdabilidade associada à estimativa ($\overline{Y}_i - \mu^{0_i}$), ou seja, da confiabilidade de sua informação genotípica. Pode-se concluir, portanto, que uma redução na variância genética em relação à variância ambiental implica num estreitamento da dispersão das respostas genotípicas médias preditas (Figura 2.1), podendo-se chegar ao limite teórico de se igualarem quando $\phi_g = 0$.

Trata-se do chamado efeito *shrinkage*, relatado na literatura de modelos lineares mistos, que nada mais é do que o "encolhimento" da distribuição das médias ajustadas de tratamentos, em torno da média geral, quando se passa de uma análise assumindo-os como de efeitos fixos para outra em que tais efeitos são tidos como aleatórios (Figura 2.1).

Quanto menor a repetibilidade, maior será o efeito *shrinkage*, o qual é tido como uma propriedade desejável do *BLUP*, haja vista que o efeito é notadamente forte sobre as médias $\overline{Y}_i$ extremas (Piepho, 1994; SAS Institute, 1996). Em razão disso, o *BLUP* é também denominado *estimador shrinkage* (Stroup & Mulitze, 1991).

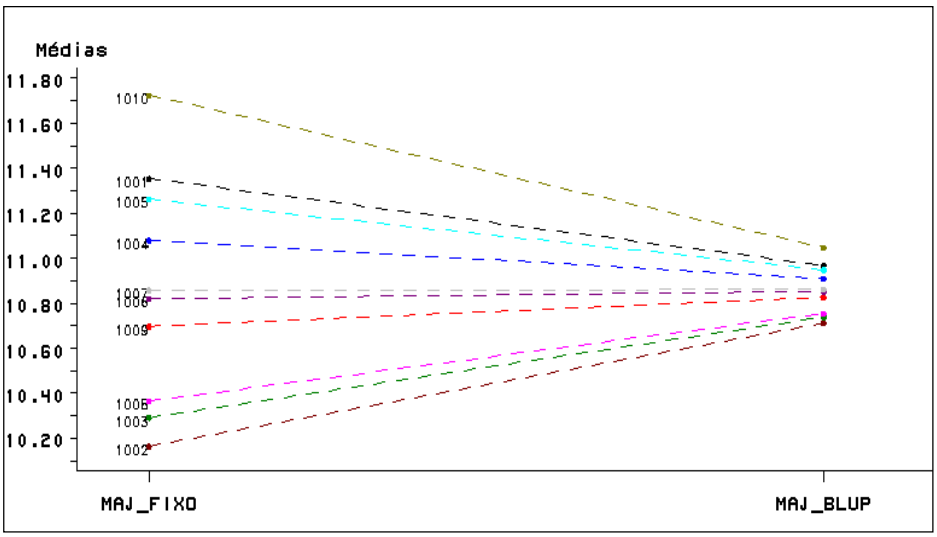

Figura 2.1. Efeito *shrinkage* sobre médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP). Os números (1001 a 1010) identificam os genótipos, num ensaio simulado de blocos completos casualizados com: $\mu=10$; $b_j \sim N(0, S_b^2=0,20)$; $g_i \sim N(0, \sigma_g^2=0,25)$; e $e_{ijr} \sim N(0, \sigma_e^2=2,00)$.

Segundo Duchateau & Janssen (1997), em síntese, o *BLUP* representa uma contração da diferença ($\overline{Y}_i - \mu$). De forma que, se o componente de variância genotípico for bem maior do que o ambiental ($\sigma_g^2 >> \sigma_e^2$), o preditor $\widetilde{g}_i$ será muito próximo de ($\overline{Y}_i - \mu$). Isto significa que a informação dos outros genótipos não é muito útil para se fazer predições acerca do genótipo *i*. Mas, se $\sigma_g^2 << \sigma_e^2$, o preditor encolherá no sentido do valor esperado populacional (zero). Além disso, quanto maior o número de repetições (n_i), mais o valor ($\overline{Y}_i - \mu$) será considerado na predição individual.

Nota-se, portanto, que a abordagem *BLUP* é consistente com a intuição dos melhoristas de se suspeitar de um novo genótipo cujas respostas, em poucas repetições, têm média excepcionalmente alta ou baixa em relação aos demais (Hill & Rosenberger, 1985). Isto pois, a solução *BLUP* leva em conta a informação de que os efeitos g_i têm menor variação do que as respostas dentro de cada genótipo *i* (Robinson, 1991). Resta

ainda avaliar uma possível mudança na ordenação das médias de tratamentos, quando se passa de uma situação de análise para a outra, o que será considerado mais adiante.

6. ORDENAMENTO COMPARATIVO DAS MÉDIAS *BLUP*

De posse dos conhecimentos acerca da predição de médias genotípicas, pode-se agora analisar o seu ordenamento em relação ao das médias ajustadas por um modelo fixo. Como μ_P é comum a todos os tratamentos (amostrados de uma mesma população), a ordenação de suas médias preditas fica determinada, como já referido, apenas pela de $\widetilde{g}_i$, ou seja: *ranking*[*BLUP*(μ_P+g_i)]=*ranking*[($\widetilde{g}_i$)]. Por isso, para fins de seleção de genótipos (tratamentos), em geral, dispensa-se a obtenção das respostas médias preditas de cada tratamento. Em alguns estudos, todavia, o resultado destas médias pode ser de interesse.

Da expressão de $\widetilde{g}_i$ pode-se também escrever: $BLUP(\mu_P+g_i)= \hat{\mu}_P + n_i\lambda_i(\overline{Y}_i - \mu^{0_i})$. Assim, dado que numa situação de blocos completos balanceados μ^{0_i} é comum para todo i ($i=1,2,...,a$) e, inclusive, igual a $\hat{\mu}_P$, tem-se: $ranking(\overline{Y}_i - \mu^{0_i})=ranking(\overline{Y}_i)$. Ademais, sob balanceamento, o peso $n_i\lambda_i$ também é comum para todo i ($n\lambda$), implicando em: $ranking(\widetilde{g}_i)=ranking(\overline{Y}_i)$. E, se a variância genética for bastante elevada ($\phi_g \to \infty$ e $n\lambda \to 1$), tem-se ainda o resultado já obtido: $BLUP(\mu_P+g_i)= \overline{Y}_i$. Por outro lado, se $n\lambda$ afasta-se de *um* (logicamente no sentido de *zero*) esta última igualdade não mais se verifica, embora a dos ordenamentos ainda permaneça, com a peculiaridade do efeito *shrinkage* das médias $BLUP(\mu_P+g_i)$ em relação às médias $\overline{Y}_i$. Pois, à medida que $n\lambda$ tende para *zero*, a amplitude de variação de $n\lambda\,\overline{Y}_i$ (termo determinante do ordenamento) reduz-se sensivelmente. Ademais, a constante $\hat{\mu}_P$ passa a ser multiplicada por $(1-n\lambda)$, de valor também inferior à unidade. Logo: $ranking[BLUP(\mu_P+g_i)]=ranking(\widetilde{g}_i)=ranking(\overline{Y}_i)$. Isto significa que, no caso de blocos completos balanceados, uma seleção baseada em médias marginais ($\overline{Y}_i$) levará à retenção e descarte dos mesmos genótipos que uma seleção baseada em $BLUP(\mu_P+g_i)$, ou simplesmente em $\widetilde{g}_i$ (Figura 2.1).

Resta, portanto, a questão de maior interesse prático relacionada ao processo de seleção de tratamentos: Existem ou não diferenças entre os ordenamentos de médias

produzidas pela abordagem de modelo misto e por uma análise convencional intrablocos (modelo fixo)? A resposta é sim. Em blocos incompletos, balanceados ou não, $\mu^{0}{}_{i}$ não é mais comum para todo *i* e os ajustes para os efeitos de blocos podem fazer com que os ordenamentos dos genótipos por $\widetilde{g}_i$ e $\overline{Y}_i$ não sejam os mesmos. Ressalta-se que a condição de desbalanceamento, por si só (em qualquer delineamento), já é suficiente para não mais garantir a concordância perfeita desses dois tipos de seleção; sobretudo, se os tratamentos diferirem muito em números de repetições. A razão pela qual o nível de desbalanceamento interfere nas médias preditas e no seu ordenamento pode ser buscada diretamente na expressão do preditor $\widetilde{g}_i$, e será discutida mais adiante.

Ratificando o que fora tratado em parágrafos anteriores, pode-se dizer ainda que, independentemente do desenho experimental, do grau de desbalanceamento e dos números de repetições, quando $\phi_g \to \infty$ e $n_i\lambda_i=1$, tem-se: $\widetilde{g}_i=1.(\overline{Y}_i-\mu^{0}{}_{i})=\tau^{0}{}_{i}$; em que $\tau^{0}{}_{i}$ é a solução do sistema de equações normais reduzidas da análise intrablocos, sob $\Sigma n_i\tau^{0}{}_{i}=0$. Em razão disso, as médias preditas da análise de modelo misto, $BLUP(\mu_P+g_i)$, serão iguais às médias ajustadas pela análise intrablocos ($\overline{Y}_{i(\text{ajust./fixo})}$). Mas, à medida que se afasta dessa condição limite ($\phi_g \to \infty$ e $n_i\lambda_i=1$), a igualdade entre $\widetilde{g}_i$ e $\tau^{0}{}_{i}$, bem como entre as médias correspondentes, não mais se verifica.

Embora a relação ϕ_g deva atingir a ordem dos milhares ($\phi_g \to \infty$) para uma igualdade quase absoluta das médias $BLUP(\mu_P+g_i)$ e $\overline{Y}_{i(\text{ajust./fixo})}$, a ordenação das médias pelas duas abordagens pode permanecer idêntica, mesmo sob relações bem menores. Na prática, valores de ϕ_g na casa das centenas, em geral, garantem coincidência absoluta das classificações. E, uma concordância razoável já é conseguida com valores de ϕ_g na casa das dezenas, o que resulta também em seleções muito similares (não obrigatoriamente idênticas) pelos dois procedimentos.

Deve-se ressaltar que a relação ϕ_g não interfere somente no valor de λ_i, mas também no de $(\overline{Y}_i-\mu^{0}{}_{i})$, pois $\mu^{0}{}_{i}$ advém de $\mathbf{X}\beta^{0}$ e $\beta^{0}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Isto é, na abordagem de modelos mistos, a solução para efeitos fixos também leva em conta a estrutura de variâncias e covariâncias das observações. Ademais, esta influência vai além da estimação pontual, interferindo

também nos testes de hipóteses relacionados aos efeitos fixos, o que justifica sempre uma cuidadosa especificação da estrutura de erros (Littell *et al*., 1996; Duchateau & Janssen, 1997).

Diante disso, o costumeiro uso de análises fundamentadas na suposição de tratamentos fixos (ex: análise intrablocos) para fins de seleção de genótipos, quando, na realidade, eles forem aleatórios, desperta especial preocupação se a relação ϕ_g for baixa. E, sem dúvida, esse é o caso de boa parte dos ensaios de avaliação de genótipos em programas de seleção de espécies já bastante melhoradas como a soja, ou seja, com baixa variabilidade genética. É verdade que, nas fases preliminares do processo, a variância genética pode ser consideravelmente alta. Em contrapartida, nestas etapas, os erros experimentais, em geral, são elevados (grande número de tratamentos e pequeno número de repetições), implicando em baixos valores de ϕ_g. Nestes casos, o número de repetições (n_i) e o grau de desbalanceamento voltam a ter influência decisiva na ordenação dos genótipos pela abordagem aqui apresentada (modelo misto com tratamentos aleatórios). Isto porque, sob desbalanceamento, $n_i\lambda_i$ pondera diferentemente o valor do desvio $(\overline{Y}_i - \mu^{0_i})$ de cada genótipo, o que pode resultar em ordenações distintas dos tratamentos pelas médias $BLUP(\mu_P+g_i)$ e $\overline{Y}_{i(\text{ajust./fixo})}$ (ou por $\widetilde{g}_i$ e $\tau^0{}_i$, respectivamente).

É oportuno, então, reconsiderar a influência do peso $n_i\lambda_i$ sobre o desvio $(\overline{Y}_i - \mu^{0_i})$ e, conseqüentemente, sobre $\widetilde{g}_i$, diante de variações em n_i e ϕ_g. Diante do que foi exposto, pode-se estender ainda:

$$n_i\lambda_i = \frac{n_i\sigma_g^2}{\sigma_e^2 + n_i\sigma_g^2} = \frac{n_i\phi_g}{1 + n_i\phi_g} = \frac{\phi_g}{\phi_g + 1/n_i} \quad \Leftrightarrow \quad h^2_{\overline{Y}_i} = \frac{n_i h^2_{Y_{ij}}}{1 + (n_i - 1)h^2_{Y_{ij}}}$$

Atribuindo-se valores seqüenciais à relação ϕ_g é possível construir funções diferentes para cada número de repetição n_i (grupo de tratamentos com n_i comum). A Figura 2.2 ilustra esse tipo de relacionamento.

Observa-se que, para $n_i \geq 4$, os pesos $n_i\lambda_i$ se aproximam rapidamente da *unidade*, mesmo com ϕ_g ainda diminutos ($\sigma_g^2/\sigma_e^2 \leq 1$); isto é, ainda que a herdabilidade seja baixa. Contudo, para $n_i=1$ ou $n_i=2$, esses pesos aproximam-se de *um* de forma bem mais lenta, ou seja, somente para os valores mais altos de ϕ_g (situações de herdabilidade mais elevada). Assim, se o ensaio apresentar uma grande amplitude de desbalanceamento, sob

herdabilidade baixa, o peso atribuído ao desvio $(\overline{Y}_i - \mu^{0_i})$ poderá diferir bastante entre os tratamentos. Para ilustrar tal situação considere-se, por exemplo, um ensaio hipotético com dois grupos de tratamentos, o primeiro com $n_i=2$ e o outro com $n_i=20$ repetições. Sob $\phi_g=0,20$ (baixa herdabilidade em nível de indivíduos), o peso atribuído a $(\overline{Y}_i - \mu^{0_i})$ no primeiro grupo será tão baixo como *0,2857*, enquanto no segundo será igual a *0,80* (Figura 2.2). Isto significa também que, para o primeiro grupo, a herdabilidade de observações individuais ($h^2_{Y_{ij}}=0,1667$) será menos que duplicada para produzir a herdabilidade de médias ($h^2_{\overline{Y}_i}$), já para o segundo, o valor de $h^2_{\overline{Y}_i}$ será quase cinco vezes o de $h^2_{Y_{ij}}$.

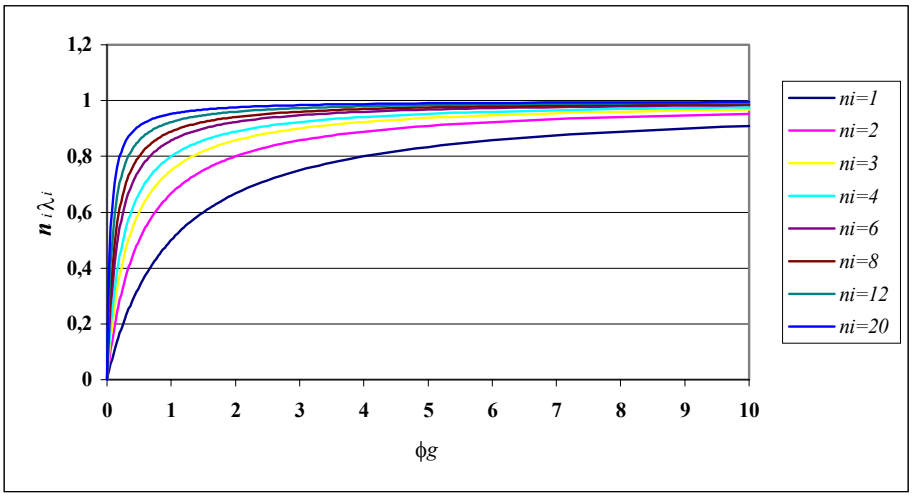

Figura 2.2. Relação entre ϕ_g ($=\sigma_g^2/\sigma_e^2$) e o peso $n_i\lambda_i$ ($=\phi_g/[\phi_g+1/n_i]$) na expressão do preditor $\widetilde{g}_i$, para grupos de tratamentos com diferentes números de repetições (n_i).

O fato mencionado tem implicação direta no valor de $BLUP(\mu_P+g_i)$, pois as médias do primeiro grupo de tratamentos terão maior probabilidade de se concentrarem em torno de μ_P (não diferirem entre si e de μ_P) do que as do segundo. Essa situação é típica dos experimentos em blocos aumentados. Neles as testemunhas são grandemente repetidas, enquanto as progênies (tratamentos novos), via de regra, são aplicadas a uma única parcela. Isso pode determinar uma forte convergência das médias preditas de progênies em torno da média geral da população de origem (μ_P). A menos que ϕ_g assuma, no mínimo, um valor igual a *4* (herdabilidade alta), o que determina um peso $n_i\lambda_i = h^2_{\overline{Y}_i} = h^2_{Y_{ij}}$ superior a *0,80* (linha correspondente a $n_i=1$, na Figura 2.2).

É verdade que, nesses ensaios, as testemunhas normalmente não são tidas como parte da mesma população, mas, como tratamentos de efeitos fixos. Entretanto, em alguns programas de seleção, os cultivares parentais é que constituem essa categoria de tratamentos, as testemunhas (Azevedo Filho *et al.,* 1998). Logo, a preocupação é procedente. Nestes casos, a classificação das progênies em relação às testemunhas pode ser extremamente modificada ao se passar de uma análise que admite o modelo como fixo para outra que o faz como aleatório ou misto (com progênies aleatórias). Isto pode agravar-se quando as testemunhas forem assumidas como fixas, ou seja, representarem outras tantas populações (assunto que será tratado na próxima seção). O posicionamento relativo das progênies entre si (intrapopulacionais) não deve sofrer alterações do mesmo nível, haja vista possuírem número de repetições similares (n_i=*1*, *2* ou *3*) e compartilharem um mesmo valor de ϕ_g.

7. DUAS OUTRAS VARIAÇÕES NO MODELO

Com base no desenvolvimento teórico apresentado procurar-se-á estender algumas constatações anteriores, sem demonstrações, a duas outras variações no modelo estudado. Na primeira, os efeitos de blocos serão admitidos como aleatórios ao lado dos de tratamentos (modelo aleatório). Na outra, os tratamentos (de efeitos aleatórios), são supostamente oriundos de várias populações, cada uma com propriedades específicas em termos de média e variância (ex: genótipos em estrutura de famílias).

7.1. Modelo com blocos aleatórios

A suposição de aleatoriedade para os efeitos de blocos e de tratamentos, num delineamento em blocos, corresponde à adoção de um modelo de análise com recuperação das informações interblocos e intertratamentos (Federer & Wolfinger, 1996; Federer, 1998). Neste caso, apesar da modificação na estrutura da matriz $\mathbf{V}$, o termo μ^{0_i} da expressão do *BLUP*(g_i) torna-se comum a todos os tratamentos ($\mu^{0_i} = \hat{\mu}_p = \hat{m}$), haja vista um único efeito fixo no modelo, a constante *m*. Logo, uma possível modificação na ordem de $\widetilde{g}_i$, em relação à de $\overline{Y}_i$, dependeria somente dos efeitos diferenciados de análogos de $n_i\lambda_i$ (componentes de $\mathbf{CV^{-1}}$), pois: *ranking*($\overline{Y}_i$)=*ranking*$(\overline{Y}_i - \hat{m})$. Sob pequena variação entre blocos, condição para o uso eficiente da informação interblocos (Malheiros, 1982;

Kempton *et al*., 1994), tais efeitos aproximam-se de $n_i\lambda_i$, podendo ser avaliados diretamente na Figura 2.2.

Inspecionando-se os valores que geram a Figura 2.2 (aqui não apresentados), observa-se que a maior diferenciação de pesos $n_i\lambda_i$ entre as classes de tratamentos ocorre, aproximadamente, no intervalo $0,25<\phi_g<0,75$ (ou $0,2< h^2_{\bar{Y}_{ij}} <0,4$); o qual corresponde à região de maior distanciamento entre as linhas, na Figura 2.2. Por exemplo, entre os grupos com $n_i=1$ e $n_i=2$, a maior diferença (distância) ocorre em $\phi_g=0,70$; entre os grupos com $n_i=1$ e $n_i=4$, ocorre em $\phi_g=0,50$; e entre os grupos com $n_i=1$ e $n_i=12$, em $\phi_g=0,28$. Portanto, num experimento com $n_i=1$ e $n_i=4$ (ex: teste de progênies em blocos aumentados), a maior diferenciação possível entre as classificações dos tratamentos, por $BLUP(\mu_P+g_i)$ e por $\overline{Y}_i$, é esperada quando $\phi_g\cong 0,50$. E, quanto maior a amplitude de desbalanceamento, maior será a probabilidade de as duas classificações diferirem. O resultado indica, portanto, que a recuperação da informação intergenotípica, nesse tipo de experimento, exerce maior influência no ordenamento dos genótipos quando a herdabilidade de observações individuais situar-se em torno *0,30*. Logicamente este valor difere de uma situação experimental para a outra, mas, o intervalo anteriormente mencionado serve como uma referência sob baixa variância de blocos (σ^2_b).

É sabido que, nos delineamentos em blocos incompletos (*BIB* e *PBIB*), a alta eficiência da análise com recuperação da informação interblocos requer uma relação $\sigma^2_b/\sigma^2_e<1/k$ (onde k é o tamanho dos blocos). Ou ainda, um valor de $\mathbf{r}<2$, sendo: $\mathbf{r}=1+k\phi_b$ e $\phi_b=\sigma^2_b/\sigma^2_e$ (Malheiros, 1982). Analogamente, os resultados anteriores sugerem que a eficiência do uso da informação intertratamentos exige relações ϕ_g inferiores ao inverso do número de repetições, o que implica, obrigatoriamente, em $\phi_g<1$. Logo, esta informação é especialmente importante quando a discriminação dos tratamentos torna-se dificultada pela baixa variabilidade genética (pequenos valores de ϕ_g). Com efeito, se $\phi_g>4$ até mesmo tratamentos não repetidos ($n_i=1$) apresentam pesos $n_i\lambda_i>0,80$ (Figura 2.2), o que implica numa grande concordância das diferentes estratégias de análise aqui confrontadas. Esse fato aponta, mais uma vez, para os ensaios preliminares dos programas de melhoramento de espécies com uma longa história de seleção artificial, já bastante melhoradas e com pequena variância genotípica.

É necessário esclarecer que a recuperação da informação interblocos, embora possa determinar trocas nas posições relativas das médias dos tratamentos (em relação às abordagens de médias marginais ou de análise intrablocos), não é responsável por *shrinkage* no conjunto das médias. Assim, uma possível maior concentração das médias de tratamentos obtidas a partir do modelo aleatório (médias *BLUP*), decorre do uso da informação intertratamentos.

7.2. Modelo com tratamentos de diferentes populações

Outro questionamento natural que surge ao se discutir o problema da ordenação das médias de tratamentos pelas abordagens de modelo fixo e de modelos mistos é: Como fica o ordenamento comparativo para um conjunto de tratamentos que são oriundos de diversas populações? Nos ensaios de melhoramento genético, os tratamentos podem representar diferentes linhagens ou progênies (genótipos) e as populações suas diferentes procedências, cruzamentos ou famílias. A análise de modelo misto aqui considerada, assume os efeitos de blocos e de populações como fixos e os efeitos de genótipos dentro de populações como aleatórios. Corresponde, portanto, a um modelo de delineamento em blocos com tratamentos hierarquizados em populações.

Dado que somente a abordagem de modelo misto utiliza a informação relativa às variabilidades genotípicas das populações, é possível surgir classificações bastante distintas pelos dois enfoques. Conforme já constatado, é de se esperar que as médias de progênies relacionadas a populações de baixa variabilidade genotípica apresentem valores próximos (*shrinkage*). Isto representa um mecanismo de agrupamento das estimativas de médias do qual a análise intrablocos não pode usufruir, haja vista não levar em conta a informação intergenotípica (Figura 2.3).

Este afunilamento das médias preditas quando se compara os dois enfoques, sobretudo para as populações de baixa variabilidade genotípica (ex: população **P2**, na Figura 2.3), pode determinar a troca de posicionamento relativo entre progênies de populações distintas, mesmo na presença de ortogonalidade e balanceamento. E, nas situações usuais de blocos incompletos, sujeitos a desbalanceamentos planejados ou não, esperam-se, inclusive, mudanças de classificações dentro da mesma população, o que pode, conseqüentemente, ter um forte impacto na seleção.

Finalmente, ainda se poderia perguntar: O uso dessa abordagem de modelos mistos não dificultaria a detecção dos chamados *segregantes transgressivos*, uma vez que há uma

tendência dos $\widetilde{g}_i$'s convergirem para o valor esperado populacional? A resposta é não. Primeiramente, porque, se as exigências da modelagem fixa forem satisfeitas, a de modelos mistos produz resultados equivalentes; mas, se não o forem, esta última reduz a chance de apontar genótipos comuns como transgressivos. Ademais, nesse tipo de abordagem não se pode ignorar a seleção intrapopulacional, concebida pela própria estrutura hierárquica do modelo. Os genótipos segregantes transgressivos caracterizam-se por valores de $\widetilde{g}_i$ discrepantes em relação aos demais genótipos relacionados (da mesma população), podendo ser facilmente identificados. O melhorista deve, portanto, estar atento a este fato, praticando seleção entre e dentro das populações. Caso contrário, não estará explorando adequadamente os recursos da modelagem estatística menos restritiva.

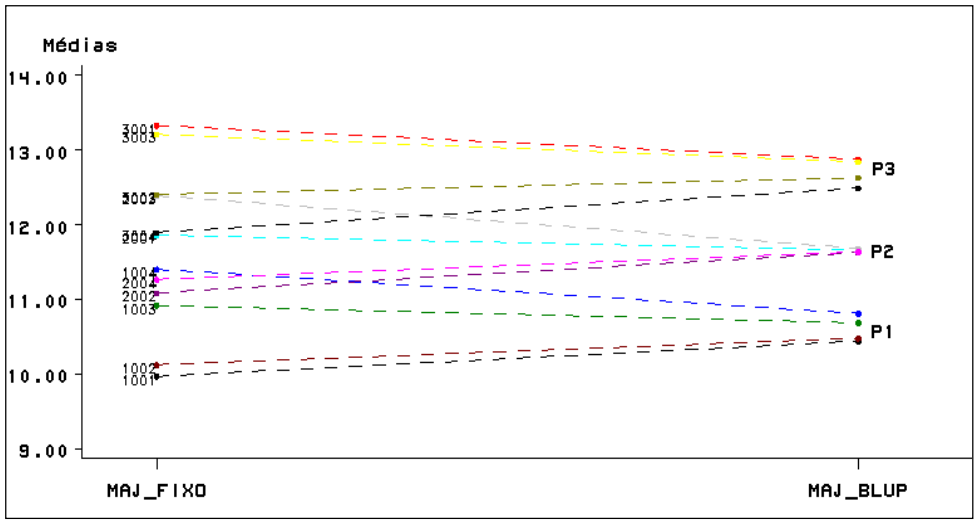

Figura 2.3. Ordenamento de médias ajustadas intrablocos (MAJ_FIXO) em relação às médias ajustadas sob recuperação da informação intergenotípica (MAJ_BLUP), para um conjunto de tratamentos oriundos de três populações (P1, P2 e P3). Dados simulados para um ensaio em blocos ao acaso sob: $\mu=10$; $b_j \sim N(0, S_b^2=0,2)$; $e_{ijr} \sim N(0, \sigma_e^2=2,0)$; e $g_i \sim N(1, \sigma_g^2=0,15)$ se $i \in$ P1; $g_i \sim N(2, \sigma_g^2=0,05)$ se $i \in$ P2; e $g_i \sim N(3, \sigma_g^2=0,2)$ se $i \in$ P3.

8. CONCLUSÕES

A análise teórica desenvolvida neste artigo permite constatar que a abordagem de modelos mistos com tratamentos aleatórios, em geral, produz médias mais uniformes para os tratamentos do que a análise intrablocos e do que o método de médias marginais simples. Além disso, a metodologia de modelos lineares mistos deverá produzir seleções notadamente diferentes em relação às análises tradicionais (médias marginais e análise

intrablocos), quando a variabilidade genética relativa (σ_g^2 / σ_e^2) for baixa e, sobretudo, nos experimentos não ortogonais e desbalanceados. Portanto, é um equívoco admitir que na análise de um modelo com um fator aleatório, ao invés de fixo, apenas os componentes de variância (esperanças de quadrados médios) e os teste "F" podem se alterar.

Pôde-se verificar também que, se os tratamentos forem oriundos de populações diferentes, o fato de a predição *BLUP* levar em conta as variâncias genotípicas específicas de cada população, pode determinar diferentes classificações dos tratamentos em relação a uma análise intrablocos, mesmo sob ortogonalidade e balanceamento. Tais constatações reforçam a preocupação acerca dos problemas de especificação dos modelos de análise estatística, na área do melhoramento vegetal.