

# BIOMETRIA

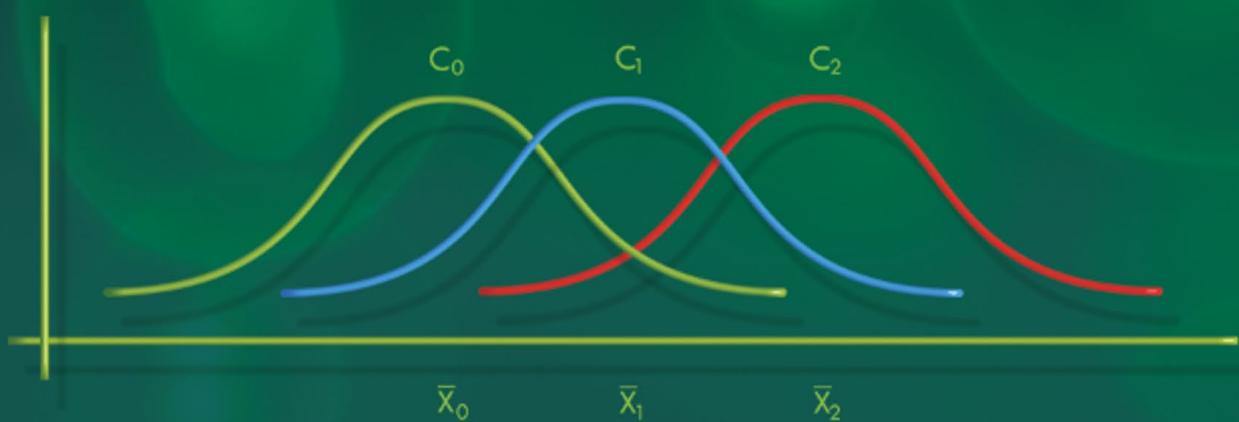
## NO MELHORAMENTO DE PLANTAS



AUTORES  
Manoel Carlos Gonçalves  
Roberto Fritsche-Neto



Com exemplos numéricos e de programação no R



A elevada disponibilidade de textos e de aplicativos computacionais sobre modelos biométricos para a análise de dados experimentais tem estimulado os pesquisadores a aplicar procedimentos de análise mais adequados às suas pesquisas. Entretanto, a utilização consciente desses métodos depende da clara compreensão de seus princípios, aplicações e limitações nos diversos cenários da pesquisa em Melhoramento de Plantas. Deste modo, o objetivo desta publicação é dar uma base aos estudantes de pós-graduação e pesquisadores para a aplicação dos principais procedimentos e modelos biométricos mais “elaborados” no Melhoramento de Plantas. Para todos os tópicos compreendidos neste trabalho, além de uma abordagem teórica e exemplos numéricos, no último capítulo são apresentados exemplos de programação destes procedimentos por meio do software estatístico computacional R.

*Boa leitura!*

Manoel Carlos Gonçalves  
Professor Titular – Faculdade de Ciências Agrárias – UFGD

Roberto Fritsche-Neto  
Assistant Professor – Louisiana State University – LSU

## **AGRADECIMENTO**

Os autores agradecem a valiosa contribuição de Karina Lima Reis Borges que trabalhou na editoração do livro.

# SUMÁRIO

1 Experimentação no Melhoramento de Plantas.....	1
2 Fundamentos de Inferência Estatística.....	7
3 Álgebra de Matrizes.....	60
4 Componentes de Variância.....	96
5 Análise de Covariância.....	119
6 Delineamento Experimental Blocos Completos Casualizados.....	132
7 Delineamentos Experimentais Blocos Incompletos.....	140
8 Delineamentos Experimentais Látice Quadrado.....	147
9 Delineamento Experimental Látice Retangular.....	168
10 Delineamentos Experimentais Alfa.....	184
11 Delineamentos Experimentais Aumentados.....	206
12 Experimentos com Medidas Repetidas.....	220
13 Interação Genótipos x Ambientes.....	257
14 Análise de Experimentos Multiambientes.....	278
15 Metodologia de Modelos Mistos.....	306
16 Seleção de Modelos.....	342
17 Delineamentos Genéticos.....	352
18 Índices de seleção.....	378
19 Estabilidade e adaptabilidade.....	391
20 Análise de Trilha.....	402
21 Componentes Principais.....	412
22 Análise de Agrupamento.....	423
23 Análise AMMI/GGE-Biplot.....	445
24 Análise de Correlações Canônicas.....	486
25 Análise de Fatores.....	498
26 Análise Discriminante.....	525
27 Análise de Variáveis Canônicas.....	545
28 Análise Espacial.....	555
29 Análise de Estabilidade no Melhoramento de Plantas.....	579
30 Exemplos de Análise no R.....	590
Bibliografia.....	591

# CAPITULO 1

## Experimentação no Melhoramento de Plantas

### Importância da experimentação na pesquisa agronômica

A importância da experimentação na pesquisa agronômica está relacionada, principalmente, com o fato de que existe um estreito relacionamento desta com o método científico, o qual consiste das seguintes etapas básicas:

(1)identificar um problema de pesquisa; (2) definir cuidadosamente o problema; (3) estabelecer objetivos claros e precisos; (4) formular uma hipótese científica de trabalho; (5) elaborar um plano para a coleta dos dados adequados ao teste da hipótese formulada; (6) analisar os dados coletados e interpretar os resultados; (6) relatar os resultados obtidos com discussão e conclusões. A estatística tem papel preponderante nas etapas (5) e (6), onde é utilizada de forma bastante extensiva e aprofundada.

A pesquisa agronômica é baseada em experimentos, os quais são desenvolvidos de acordo com os princípios de repetição, casualização e formação de blocos. Com base nestes princípios são formulados delineamentos experimentais, que são planos para a aplicação de tratamentos nas unidades experimentais e que estão associados a modelos estatísticos que são utilizados na análise estatística dos dados coletados nos experimentos.

O principal objetivo de se utilizar os delineamentos experimentais é possibilitar estimar e minimizar o efeito da variância do erro sobre o efeito dos tratamentos avaliados. Quanto menor for a variância do erro maior será a precisão e a acurácia dos experimentos e, consequentemente maior será a eficiência e a eficácia da pesquisa agronômica.

### Princípios Básicos da Experimentação

Os princípios de repetição, casualização e controle local, propostos por R. A. Fisher (1910) são referidos como os três princípios básicos dos delineamentos experimentais e, portanto, da experimentação, os quais são descritos a seguir.

#### 1. Repetição

É a aplicação de um determinado tratamento a duas ou mais unidades experimentais. Um conjunto completo de tratamentos constitui uma repetição do experimento básico. O aumento do número de repetições é um dos procedimentos mais eficientes para se melhorar a exatidão e a precisão de um experimento.

O princípio da repetição permite:

- a) Estimar o erro experimental, o que é necessário para avaliar a significância das diferenças entre as médias de tratamentos;
- b) Aumentar a sensibilidade dos testes pela redução do erro padrão da diferença entre médias de tratamentos;
- c) Aumentar a abrangência do experimento pela incorporação de uma maior diversidade de material experimental, se o controle local apropriado é usado para controlar o erro experimental.

O número de repetições, geralmente, é função de: (a) número de tratamentos; (b) tipo de material experimental (homogeneidade, conhecimento do material); (c) custo do experimento (materiais,

equipamentos, procedimentos e métodos); (d) disponibilidade de espaço físico e de mão-de-obra; (e) magnitude provável da diferença entre médias de tratamento e (f) nível de precisão desejada. Desta forma, é impossível determinar o número exato de repetições necessários para um determinado experimento. Uma regra geral prática, é prover no mínimo 10 graus de liberdade para o erro experimental de qualquer experimento. A redução no tamanho da parcela permite um maior número de parcelas (repetições), para uma determinada área (material) experimental, o que tenderá a aumentar a precisão do experimento.

## 2. Casualização

Consiste na distribuição dos tratamentos ao acaso nas unidades experimentais. Tem a função de assegurar que a estimativa do erro experimental e dos efeitos de tratamentos não tenham tendência ou viés. Apresenta as seguintes características:

- (a) garante que os tratamentos sejam casualmente afetados por fontes desconhecidas de variação, quando eles são distribuídos nas diversas unidades experimentais;
- (b) faz com que os erros (desvios) se tornem aleatórios;
- (c) faz com que os erros associados aos tratamentos sejam independentes entre si, permitindo assim a aplicação dos testes estatísticos.

É necessário ter-se casualizações separadas para cada um dentre vários experimentos. Alguns pesquisadores têm usado o mesmo delineamento, com a mesma casualização para vários experimentos, o que não é recomendável devido a possibilidade de uma leve tendência.

## 3. Controle Local (Formação de Blocos)

Consiste na estratificação (agrupamento) das unidades experimentais (local ou material) em grupos homogêneos denominados de blocos, o que permite a estimação e o controle mais adequado do erro experimental. Em experimentos simples, cada bloco contém o mesmo número de unidades experimentais sobre os quais todos os tratamentos que estão sendo comparados são distribuídos de forma aleatória.

As diferenças entre parcelas de um mesmo tratamento num experimento repetido em blocos são parcialmente devidas ao erro experimental, mas também parcialmente devido à diferença média entre os blocos. Assim, a variação devida a blocos pode ser removida do erro experimental.

Consequentemente, a exatidão/acurácia e a precisão do experimento tornam-se maior quando uma quantidade da variabilidade é removida do erro experimental por meio deste procedimento.

### Relação entre Casualização e Delineamentos Experimentais

Dependendo da maneira de se fazer a casualização dos tratamentos e repetições nas unidades experimentais decorrem os delineamentos experimentais.

Considere três tipos básicos de casualização:

- a) **Sem Restrição** – qualquer uma das unidades experimentais (supostas homogêneas) pode receber (por sorteio) qualquer um dos  $t$  tratamentos em qualquer uma das  $r$  repetições. Esta casualização resulta no delineamento experimental inteiramente casualizado, cujo modelo linear é:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : resposta avaliada no tratamento  $i$  e repetição  $j$ , com  $i=1,2,\dots,t$  e  $j=1,2,\dots,r$ ;

$\mu$ : média geral;

$\tau_i$ : efeito do tratamento  $i$ ;

$\varepsilon_{ij}$ : erro experimental associado ao tratamento  $i$  na repetição  $j$ .

**b) Uma Restrição** – neste caso deve-se organizar  $r$  homogêneos, sendo que cada bloco recebe uma vez todos os  $t$  tratamentos; os  $r$  blocos correspondem a  $r$  repetições; os  $t$  tratamentos são casualizados (sorteados) dentro de cada bloco; os blocos podem estar em condições diferentes ou sofrer manejos diferentes; o importante é que haja homogeneidade dentro dos blocos e heterogeneidade entre os blocos. Esta casualização resulta no delineamento experimental blocos completos casualizados, cujo modelo linear é:

$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$ , em que:

$Y_{ij}$ : resposta avaliada no tratamento  $i$  e bloco  $j$ ;

$\mu$ : média geral;

$\tau_i$ : efeito do tratamento  $i$ , com  $i=1,2,\dots,t$ ;

$\beta_j$ : efeito do bloco  $j$ , com  $j=1,2,\dots,r$ ;

$\varepsilon_{ij}$ : erro experimental associado ao tratamento  $i$  no bloco  $j$ .

**c) Duas Restrições** – neste caso as unidades experimentais são agrupadas segundo dois critérios de heterogeneidade (área ou material); os blocos são formados em duas direções denominadas de linhas (critério L) e colunas (critério C); cada linha e cada coluna recebe uma vez cada tratamento; assim, o número de linhas, de colunas e de tratamentos é igual. Esta casualização resulta no delineamento experimental quadrado latino, cujo modelo linear é:

$Y_{k(ij)} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{k(ij)}$ , em que:

$Y_{k(ij)}$ : resposta observada do tratamento  $k$  na linha  $i$  e coluna  $j$ ;

$\alpha_i$ : efeito da linha  $i$ , com  $i=1,2,\dots,n$ ;

$\beta_j$ : efeito da coluna  $j$ , com  $j=1,2,\dots,n$ ;

$\tau_k$ : efeito do tratamento  $k$ , com  $k=1,2,\dots,n$

$\varepsilon_{k(ij)}$ : erro experimental associado à resposta observada do tratamento  $k$  na linha  $i$  e coluna  $j$ ;

Deve ser observado que o aumento no número de restrições para a casualização diminui o número de graus de liberdade associados ao erro experimental, o que aumenta a variância do mesmo e diminui a precisão experimental.

Então, é esperado que com o uso de restrição na casualização a diminuição no número de graus de liberdade do erro seja compensada e a precisão experimental seja aumentada.

### Aplicações da experimentação no melhoramento de plantas

No melhoramento de plantas o uso de experimentos de campo com repetições tem importância fundamental, principalmente nos processos de avaliação para a seleção de genótipos superiores e para a recomendação de novas cultivares em determinada região de cultivo.

Como a variabilidade genética de determinado conjunto de genótipos sob avaliação é constante, devemos então atuar na redução do efeito ambiental,

para aumentar a proporção  $\frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2 + \sigma_{GxE}^2}$ . Aqui é que entra a experimentação com seus

princípios e procedimentos estatísticos para reduzir a variância do erro e o erro padrão de médias, por meio de:

- Escolha e controle da área experimental;
- Formação de blocos, repetição, casualização;
- Delineamentos experimentais, manejo do experimento;
- Tamanho de parcela adequado: número de plantas que representa determinado genótipo; varia com o caráter; determinado para o caráter mais complexo (produtividade de grãos);
- Delineamentos experimentais especiais (látice, blocos aumentados) para manter a homogeneidade da área dentro dos blocos;
- Metodologias de análise estatística de dados especiais (análise de experimentos multi-ambiente, análise de covariância, metodologia de modelos mistos etc.) para diminuir a variância do erro e aumentar a precisão e a acurácia.

Delineamentos experimentais eficientes são importantes nos principais estágios de um programa de melhoramento de plantas, desde o estágio inicial de seleção com centenas ou milhares de genótipos até o estágio final com poucos genótipos selecionados para recomendação de novas cultivares.

Em cada estágio delineamentos experimentais adequados são críticos para garantir um baixo custo e identificação dos melhores genótipos possíveis. Os delineamentos podem variar desde experimentos sem repetição num único local até experimentos multiambientes que podem envolver dezenas ou centenas de locais em vários anos.

Experimentos de campo são baseados nos conceitos de repetição, casualização e controle da variação entre parcelas. A repetição garante estimativa da variância do erro válida. A casualização garante estimativas de médias e variâncias não viesadas. O controle da variação entre parcelas reduz a variância do erro. O principal foco tem sido como arranjar os genótipos no campo de forma a minimizar o impacto do erro sobre os processos de estimativa e predição.

O principal objetivo de se agrupar parcelas em blocos uniformes é reduzir a variação entre parcelas e melhorar a precisão do experimento. O uso inadequado de blocos num experimento de campo pode resultar em variância do erro inaceitavelmente grande e/ou estimativas viesadas do efeito de genótipos. O controle efetivo da variância do erro geralmente requer blocos relativamente pequenos.

Experimentos com um grande número de genótipos desenvolvidos em blocos completos, onde existe considerável variabilidade entre parcelas dentro de um bloco resultará em informações pobres sobre os genótipos. Para controlar a variação no campo, especialmente com um grande número de genótipos, é essencial fazer uso de delineamentos blocos incompletos.

Experimentos no melhoramento de plantas são desenvolvidos no sentido de fornecer uma avaliação não viesada de todos os genótipos em teste e garantir variâncias iguais para todos os pares de diferenças entre os genótipos. O critério de igual variância do erro resulta leva a delineamentos blocos incompletos balanceados (DBIB). Os DBIBs requerem que todos os pares de genótipos apareçam juntos num bloco com a mesma frequência. O balanceamento é possível em experimentos menores, entretanto, em experimentos maiores não é possível, o que significa que algumas diferenças entre pares de genótipos são estimadas com maior precisão que outras. Isto deixa de ser um problema se os blocos incompletos são eficientes.

Outro conceito importante no delineamento de experimento de campo é resolvabilidade. Um experimento de campo é resolvível se ele está disposto em blocos com repetição completa, sendo cada repetição subdividida em um certo número de blocos incompletos. Experimentos resolvíveis são úteis uma vez que os genótipos num mesmo bloco incompleto em uma repetição são distribuídos ao longo de blocos incompletos em outra repetição. Além disto, delineamentos resolvíveis são fáceis de manejar uma vez que todos os genótipos estão juntos em repetições completas.

Uma categoria de delineamentos blocos incompletos resolvíveis são os delineamentos látice propostos por Yates (1936). Estes delineamentos requerem que o número de tratamentos/genótipos seja um quadrado do tamanho do bloco e o balanceamento é conseguido se for possível ter número de repetições suficiente. Se não for possível, então tem-se os látices que são delineamentos parcialmente balanceados dos tipos simples, triplos, quádruplos, com duas, três ou quatro repetições, respectivamente, os quais têm sido extensivamente utilizados nos experimentos de melhoramento de plantas.

Devido à restrição sobre o número de genótipos que pode ser avaliado nos látices tradicionais, tem sido proposto alguns tipos especiais de delineamentos látice, sendo o mais popular os delineamentos-alfa desenvolvidos por Patterson e Williams (1976). Delineamentos-alfa são delineamentos blocos incompletos resolvíveis, onde o número de tratamentos/genótipos é um múltiplo do tamanho do bloco.

Embora, não seja possível conseguir balanceamento com os delineamentos-alfa, eles têm sido utilizados extensivamente no melhoramento de plantas, principalmente porque eles são muito flexíveis em relação ao número de genótipos a serem avaliados e ao número apropriado de blocos incompletos e também porque permitem um bom controle do erro. A geração destes delineamentos é possível usando o método dado em Patterson e Williams (1976) ou pelo uso de aplicativo computacional (Williams e Talbot, 1993; Multize, 2004).

Para os delineamentos blocos incompletos, tem sido assumido que cada genótipo deve ser alocado em pelo menos duas parcelas. Nos estágios iniciais de um programa de melhoramento de plantas, o ganho genético esperado pode ser aumentado pelo peneiramento de um grande número de genótipos ao invés de comparações mais precisas de um número menor de genótipos. Neste caso, deve ser avaliado um grande número genótipos e pode não ter sementes suficiente para repetir cada um dos genótipos.

Então, Federer propôs os delineamentos aumentados, onde um conjunto de genótipos testemunha/controle são repetidos um igual número de vezes num determinado delineamento de campo e um conjunto adicional de genótipos teste são incluídos no experimento uma única vez (Federer 2002, 2005).

Qualquer tipo de delineamento em blocos pode ser usado para os tratamentos testemunha/controle com os genótipos teste sendo adicionados ou aumentados aos blocos e o erro padrão para a diferença entre genótipos teste ou entre testemunhas é computado de forma simples. Esta abordagem fornece uma forma muito eficiente de peneirar genótipos teste e tem uma quantidade considerável de flexibilidade.

A avaliação de genótipos em multi-ambiente é uma fase crucial no desenvolvimento de genótipos superiores adaptados a uma ampla faixa de condições ambientais. O delineamento de programas eficientes de avaliação para maximizar informações sujeitas a limitações práticas e de recursos disponíveis tem sido considerado por muitos pesquisadores. Por exemplo, Talbot (1984) encontrou que para minimizar a variância de diferenças entre genótipos para vários cultivos diferentes, o uso de 2 anos em 12 locais com duas repetições foi razoável para a maioria dos cultivos, com mais anos aumentando a precisão mais que mais locais.

Geralmente, os resultados são baseados na pressuposição de bons delineamentos de campo em cada local com qualquer tipo de delineamento e delineamentos diferentes em cada local (Federer et. al, 2001).

No melhoramento de plantas é muito importante definir se os fatores incluídos em um modelo de análise estatística de dados são de efeito fixo ou de efeito aleatório. Admite-se que um fator é de efeito fixo quando os níveis deste fator num experimento são os únicos níveis de interesse e os resultados de seu efeito não podem ser extrapolados para uma população destes níveis. Os níveis de um fator de efeito fixo podem ser obtidos repetidas vezes no espaço e/ou tempo. Para fatores de efeito fixo são estimados e comparados quadrados médios (variâncias).

Admite-se que um fator é de efeito aleatório quando os níveis deste fator num experimento representam uma amostra de possíveis níveis e os resultados de seu efeito podem ser extrapolados para uma população destes níveis. Os níveis de um fator de efeito aleatório não podem ser obtidos repetidas vezes no espaço e/ou tempo. Para fatores de efeito aleatório são estimadas e comparadas esperanças de quadrados médios (componentes de variância).

Componentes de variância são as variâncias associadas aos fatores de efeito aleatório e são muito úteis na estimativa de parâmetros genéticos (herdabilidade, ganho com a seleção, acurácia seletiva) no melhoramento de plantas. Os componentes de variância podem ser estimados por meio das metodologias de quadrados mínimos e/ou de modelos mistos.

A metodologia de modelos lineares mistos tem sido bastante utilizada no melhoramento de plantas para a estimativa de componentes de variância em situações de dados desbalanceados, experimentos não ortogonais, variâncias heterogêneas e erros correlacionados. Tem sido utilizada também para a predição de valores genéticos, análise de experimentos multi-ambiente e de delineamentos genéticos.

# CAPITULO 2

## Fundamentos de Inferência Estatística

Nas pesquisas são obtidos dados amostrais ou experimentais que devem ser representativos em relação às populações estudadas. O processo de inferência estatística tem o objetivo de fazer generalizações sobre populações com base em dados de amostras. Assim, todas as inferências estatísticas realizadas são baseadas nas distribuições amostrais e na teoria de probabilidades.

### Probabilidades

#### Conceito

A necessidade de utilização de probabilidades significa que existe um fator de acaso e/ou incerteza quanto à ocorrência ou não de um evento. Em muitos casos, pode ser impossível afirmar por antecipação o que ocorrerá, mas é possível dizer o que poderá ocorrer em relação a um evento futuro.

As probabilidades são úteis porque auxiliam no desenvolvimento de estratégias e tomadas de decisões em importantes estudos.

Por exemplo, um produtor rural sente-se inclinado a comprar a semente de uma nova cultivar se a chance de lucro for boa; uma empresa rural sentir inclinada a investir em novo equipamento agrícola se há boa chance de recuperar o dinheiro investido.

Outro exemplo, os testes estatísticos, que são baseados em probabilidades, são utilizados para a tomada de decisões nas pesquisas científicas experimentais ou observacionais.

O ponto central em todas essas situações é a possibilidade de quantificar o quanto provável é a ocorrência de determinado evento.

As probabilidades são utilizadas para exprimir a chance de ocorrência de determinado evento. A probabilidade de ocorrência de um evento é dada por um número que pode variar de 0 a 1,0.

#### Espaço Amostral e Eventos

Um dos conceitos matemáticos fundamentais utilizado no estudo das probabilidades é o de conjunto. Um conjunto é uma coleção de itens que possuem características comuns. É importante definir cuidadosamente o que constitui o conjunto de interesse, para poder decidir se determinado elemento é ou não membro do conjunto.

A probabilidade só tem sentido no contexto de um espaço amostral, que é o conjunto de todos os resultados possíveis de um experimento aleatório. O termo experimento aleatório sugere incerteza do resultado antes que sejam realizadas as observações. Os resultados de um experimento aleatório podem ser chamados de eventos.

Como exemplos simples de experimento aleatório pode ser considerado o lançamento de uma moeda ou de um dado, a extração de uma carta de um baralho. No caso das cartas há 52 eventos elementares no espaço amostral, mas pode ser considerado outros eventos como combinações dos

eventos elementares. Por exemplo, o evento “sair uma carta de copas” pode obtido por qualquer um dentre os 13 eventos elementares.

Os cálculos de probabilidades devem levar em conta a maneira como os eventos de interesse podem relacionar-se entre si. Algumas dessas relações são descritas pelas expressões “complemento”, “mutuamente excludentes” e “coletivamente exaustivos”.

O complemento de um evento consiste de todos os resultados no espaço amostral que não fazem parte do evento. Por exemplo, o complemento do evento “a carta é de copas” consiste de todas as cartas que não são de copas.

Os eventos são mutuamente excludentes se não têm elementos em comum, ou se não podem ocorrer simultaneamente. Por exemplo, na extração de uma carta, os eventos “a carta é de copas” e “a carta é de ouros” são mutuamente excludentes.

Os eventos são coletivamente exaustivos se nenhum outro resultado é possível (esgotam todas as possibilidades) para o experimento considerado, ou seja, se pelo menos um dos eventos tiver que ocorrer no experimento. Por exemplo, “a carta é preta” e “a carta é vermelha” são coletivamente exaustivos.

Como o espaço amostral consiste de todos os resultados possíveis de um experimento aleatório, a probabilidade do espaço amostral é 1,0 ou 100%. Além disso, porque qualquer evento  $A$  e seu complemento  $A'$  esgotam todas as possibilidades do espaço amostral, tem-se também que  $P(A)+P(A')=1,0$ .

Em resumo, tem-se que:

1-A probabilidade de qualquer evento  $A$  é representada por um número entre 0 e 1,0, ou seja,  $0 \leq P(A) \leq 1,0$ ;

2-A probabilidade representada pelo espaço amostral é de 1,0 ou 100%, ou seja,  $P$  (qualquer evento do espaço amostral) = 1,0;

3-A probabilidade de não ocorrência de um evento é 1,0 menos a probabilidade de sua ocorrência, ou seja,  $P(A')=1,0-P(A)$  ou  $P(A)+P(A')=1,0$ .

## Definição de Probabilidade

### Definição Clássica

A definição clássica aplica-se a situações em que os eventos têm resultados igualmente prováveis.

Por definição, a probabilidade de um evento  $A$  é expressa por:

$P(A)=(\text{Nº de resultados favoráveis ao evento } A)/(\text{Nº total de resultados possíveis})$ , ou

$$P(A)=\frac{F_A}{F_A+C_A}, \text{ em que:}$$

$F_A$ : Número de resultados favoráveis ao evento  $A$

$C_A$ : Número de resultados contrários ao evento  $A$

Exemplo: Para o evento “extração de uma dama”, a probabilidade é

$$P(dama)=\frac{4\text{damas}}{52\text{cartas}}=0,0769 \text{ ou } 7,69\%.$$

## Definição por Frequência Relativa ou Empírica

A definição clássica está limitada às situações em que os resultados/eventos são igualmente prováveis. Entretanto, há casos em que isto não ocorre, como no lançamento de uma moeda viciada um grande número de vezes.

Por exemplo, se os testes de laboratório indicam que, de 25 plantas tratadas com determinado herbicida, 20 passaram a apresentar sintomas de fitotoxicidade, então a percentagem  $\frac{20}{25} = 0,80$

ou 80% pode ser tomada como uma estimativa da probabilidade de ocorrência do evento fitotoxicidade, sob condições idênticas. Então, com base na definição de frequência relativa tem-se:

$$P(A) = (\text{Nº de ocorrências de } A) / (\text{Nº total de testes ou observações})$$

Quando for adotada a definição por frequência relativa é importante reconhecer os seguintes aspectos:

1-A probabilidade assim determinada é apenas uma estimativa do verdadeiro valor;

2-Quanto maior a amostra melhor a estimativa da probabilidade;

3-A probabilidade estimada só é válida para um conjunto de condições idênticas àquelas sob as quais se originaram os dados utilizados para a estimação da probabilidade.

## Teorema do Produto de Probabilidades

Considere dois eventos de interesse,  $A$  e  $B$ , no espaço amostral. Pode ser necessário determinar  $P(A \text{ e } B)$ , ou seja, a probabilidade de ocorrência de ambos os eventos. A probabilidade de ocorrência de dois eventos é chamada de probabilidade conjunta e seu cálculo difere, conforme os eventos sejam ou não independentes.

Dois eventos são considerados independentes se a ocorrência de um não influencia a ocorrência do outro. Se os eventos são dependentes, então o conhecimento da ocorrência de um pode auxiliar a predizer a ocorrência do outro.

Se dois eventos são independentes, então a probabilidade de ocorrência de ambos é igual ao produto de suas probabilidades individuais ou marginais:

$$P(A \text{ e } B) = P(A \cap B) = P(A)P(B)$$

Se dois eventos são dependentes, então o cálculo de  $P(A \text{ e } B)$  deve levar em conta as probabilidades condicionais de cada um dos eventos. Como regra geral, a probabilidade conjunta de dois eventos dependentes é a probabilidade de um deles multiplicada pela probabilidade condicional do outro:

$$P(A \text{ e } B) = P(A \cap B) = P(A)P(B/A) \text{ ou}$$

$P(A \text{ e } B) = P(A \cap B) = P(B)P(A/B)$ , uma vez que não importa qual evento ocorre em primeiro lugar.

Observe que quando dois eventos são independentes, o fato de saber que um deles ocorreu nada informa sobre a ocorrência do outro. Portanto,

$$P(B/A) = P(B) \text{ e } P(A/B) = P(A)$$

## Teorema da Soma de Probabilidades

Considere dois eventos,  $A$  e  $B$ , do espaço amostral. Pode ser necessário determinar a probabilidade de ocorrência de  $A$  ou  $B$ . Neste caso, deve-se calcular a probabilidade de ocorrência de pelo menos um dentre dois eventos, ou seja,  $P(A \cup B)$ . O cálculo desta probabilidade depende de os eventos serem ou não mutuamente excludentes.

Quando dois eventos são mutuamente excludentes (não podem ocorrer conjuntamente), a probabilidade de ocorrência de qualquer um deles é a soma de suas probabilidades individuais:

$$P(A \cup B) = P(A) + P(B)$$

Quando dois eventos não são mutuamente excludentes (podem ocorrer conjuntamente), o cálculo da probabilidade de um ou outro ocorrer deve levar em conta o fato de que um, ou outro ou ambos, podem ocorrer. Neste caso, a probabilidade é expressa por:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

Em resumo, as regras para o cálculo de probabilidades são:

1-  $P(A \cap B)$ , para eventos independentes:

$$P(A \cap B) = P(A)P(B)$$

2-  $P(A \cap B)$ , para eventos dependentes:

$$P(A \cap B) = P(A)P(B/A) \text{ ou}$$

$$P(A \cap B) = P(B)P(A/B)$$

3-  $P(A \cup B)$ , para eventos mutuamente excludentes:

$$P(A \cup B) = P(A) + P(B)$$

4-  $P(A \cup B)$ , para eventos não mutuamente excludentes:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

Exemplo 1: Joga-se duas moedas equilibradas. Qual a probabilidade de ambas darem cara?

Resolução: Uma vez que os resultados/eventos são independentes e

$$P(\text{Cara}) = P(\text{Coroa}) = \frac{1}{2}, \quad \text{tem-se que}$$

$$P(\text{Cara e Cara}) = P(\text{Cara} \cap \text{Cara}) = \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) = \frac{1}{4} \text{ ou } 25\%.$$

Exemplo 2: Sabe-se que um terço das empresas agrícolas de certa região são pequenas e que 40% das empresas compraram um trator novo. Supondo que estes dois eventos sejam independentes, determine a probabilidade de escolher aleatoriamente uma empresa da lista geral, que seja pequena e que tenha comprado um trator novo.

Resolução:  $P(\text{Pequena empresa que comprou um trator novo}) = \left( \frac{1}{3} \right) (0,40) = 0,133$  ou  
13,3%.

Exemplo 3: Considere duas urnas com fichas. A primeira contém 8 fichas vermelhas e 2 brancas. A segunda contém 5 fichas vermelhas e 5 brancas, ou seja:

	Vermelhas	Branca	Total
Urna 1	8	2	10
Urna 2	5	5	10

Considere o experimento “extrair uma ficha de uma das urnas”:

Se for escolhida a Urna1, a probabilidade de a ficha ser vermelha é  $\frac{8}{10}$ . Se for escolhida a Urna2, a

probabilidade de a ficha ser vermelha é  $\frac{5}{10}$ . Então, a  $P(Vermelha)$  depende de qual seja a urna escolhida.

Assim, a probabilidade condicional de escolha de ficha vermelha, admitindo-se que foi escolhida a

Urna1 é  $\frac{8}{10}$  e expressa-se como:

$$P(Vermelha/Urnal) = \frac{8}{10} \text{. Logo,}$$

$$P(Vermelha/Urna2) = \frac{5}{10}, P(Branca/Urnal) = \frac{2}{10} \text{ e } P(Branca/Urna2) = \frac{5}{10}.$$

Considere agora que as duas urnas sejam indistinguíveis e que a probabilidade de escolher qualquer uma delas seja  $P(Urnal) = P(Urna2) = \frac{1}{2}$ . Neste caso, qual a probabilidade de extrair uma

ficha vermelha da Urna2? Agora, para o cálculo desta probabilidade deve-se considerar dois aspectos: 1-a probabilidade de escolher a Urna2 e 2-a probabilidade de extrair uma ficha vermelha supondo-se que a Urna2 tenha sido escolhida. Então, tem-se:

$$P(Urna2) = \frac{1}{2} \text{ e } P(Vermelha/Urna2) = \frac{5}{10}$$

$$P(Urna2 \text{ e FichaVermelha}) = P(Urna2)P(Vermelha/Urna2)$$

$$P(Urna2 \text{ e FichaVermelha}) = \left(\frac{1}{2}\right)\left(\frac{5}{10}\right) = \frac{5}{20} \text{ ou } 25\%.$$

Qual a probabilidade de extrair uma ficha vermelha da Urna1? Neste caso, tem-se:

$$P(UrnaleFichaVermelha) = \left(\frac{1}{2}\right)\left(\frac{8}{10}\right) = \frac{8}{20} \text{ ou } 40\%.$$

Exemplo 4: Considere a probabilidade de extração de uma carta de copas ou um dez de um baralho de 52 cartas.

Neste caso, é possível que uma carta seja simultaneamente/conjuntamente de copas e um dez. Portanto, os dois eventos não são mutuamente excludentes. O cálculo desta probabilidade

considerando apenas a simples adição das probabilidades individuais, aumentará a probabilidade verdadeira porque a carta dez de copas será contada duas vezes, uma vez como dez e outra como copas. Assim, devemos subtrair a probabilidade da interseção. Então, tem-se:

$$P(\text{Copas ou Dez}) = P(\text{Copas}) + P(\text{Dez}) - P(\text{Dez e Copas})$$

$$P(\text{Copas ou Dez}) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

Outra forma de calcular esta probabilidade é verificar se foi incluída a probabilidade de ocorrência de ambos os eventos de duas maneiras, ou seja, como a probabilidade copas e também como a probabilidade de dez. Neste caso, deve-se subtrair a probabilidade conjunta de extração da carta dez de copas. A probabilidade conjunta é o produto das duas probabilidades marginais, ou seja,

$$P(\text{Copas e Dez}) = \left( \frac{13}{52} \right) \left( \frac{4}{52} \right). \text{ Então, tem-se:}$$

$$P(\text{Copas ou Dez}) = \frac{13}{52} + \frac{4}{52} - \left( \frac{13}{52} \right) \left( \frac{4}{52} \right) = \frac{16}{52}$$

### Teorema de Bayes

Se  $n$  eventos  $E_1, E_2, \dots, E_n$  constituem uma partição do espaço amostral  $S$ , o que significa que eles são dois a dois mutuamente excludentes e tais que o evento soma é igual ao próprio espaço amostral e  $P(E_i) > 0 (i=1,2,\dots,n)$ . Além disso, se existe um evento  $B$  que só pode ocorrer como efeito (consequência) de uma das causas mutuamente excludentes  $E_i$ , então:

$$P(E_i/B) = \frac{P(E_i)P(B/E_i)}{\sum_{i=1}^n [P(E_i)P(B/E_i)]}$$

Esta é a equação do teorema de Bayes, que é também denominada de probabilidade das causas ou dos antecedentes. Ela fornece a probabilidade de um particular evento  $E_i$  (uma causa/antecedente)

ocorrer, uma vez que o evento  $B$  (um efeito/consequência) já tenha ocorrido.

O teorema de Bayes procura responder à seguinte questão: Supondo que o evento  $B$  tenha ocorrido, qual a probabilidade de que ele tenha sido proveniente de determinado  $E_i$ ? Ou de outra forma, qual a probabilidade de  $B$  ser consequência/efeito de  $E_i$ , ou que  $E_i$  seja antecedente/causa de  $B$ ?

Exemplo1: Considere que numa fábrica existem três máquinas para a produção de parafusos. A primeira máquina produz diariamente 1000 parafusos, a segunda máquina 4000 e a terceira máquina 5000. Sabendo-se que a primeira máquina produz 4% de parafusos defeituosos, a segunda máquina 3% e a terceira máquina 1%. Ao final do dia, tendo-se encontrado um parafuso defeituoso, qual a probabilidade de ele ter sido produzido em cada uma dessas máquinas?

Os eventos de interesse são:

$A$ : parafuso produzido na primeira máquina

$B$ : parafuso produzido na segunda máquina

$C$ : parafuso produzido na terceira máquina

$X$ : parafuso defeituoso

Então, tem-se:

$$P(A) = \frac{1000}{10000} = 0,10; P(B) = \frac{4000}{10000} = 0,40;$$

$$P(C) = \frac{5000}{10000} = 0,50$$

$$P(X/A) = 0,04; P(X/B) = 0,03; P(X/C) = 0,01$$

Aplicando o teorema de Bayes, tem-se:

$$P(A/X) = \frac{P(A)P(X/A)}{P(A)P(X/A) + P(B)P(X/B) + P(C)P(X/C)}$$

$$P(A/X) = \frac{(0,10)(0,04)}{(0,10)(0,04) + (0,40)(0,03) + (0,50)(0,01)} = 0,19$$

, ou 19%

$$P(B/X) = \frac{P(B)P(X/B)}{P(A)P(X/A) + P(B)P(X/B) + P(C)P(X/C)}$$

$$P(B/X) = \frac{(0,40)(0,03)}{(0,10)(0,04) + (0,40)(0,03) + (0,50)(0,01)} = 0,57$$

, ou 57%

$$P(C/X) = \frac{P(C)P(X/C)}{P(A)P(X/A) + P(B)P(X/B) + P(C)P(X/C)}$$

$$P(C/X) = \frac{(0,50)(0,01)}{(0,10)(0,04) + (0,40)(0,03) + (0,50)(0,01)} = 0,24$$

, ou 24%

Observe que:

$$P(A/X) + P(B/X) + P(C/X) = 19\% + 57\% + 24\% = 100\%$$

Observe que os resultados da aplicação do teorema de Bayes são importantes para testes de significância e estimativa de parâmetros porque relacionam probabilidades "a priori",  $P(A)$ ,  $P(B)$  e  $P(C)$ , com probabilidades "a posteriori",  $P(X/A)$ ,  $P(X/B)$  e  $P(X/C)$ , que são probabilidades de  $A$ ,  $B$  e  $C$ , depois que ocorrer  $X$ .

Exemplo 2: Considere a seguinte situação:

Bolas	Caixas		
	$C_1$	$C_2$	$C_3$
Preta (P)	3	4	2
Branca (B)	1	3	3
Vermelha (V)	5	2	3

Escolheu-se uma caixa ao acaso e dela extraiu-se uma bola ao acaso, verificando-se que a bola é branca. Qual a probabilidade de a bola ter sido extraída da caixa 2? E da caixa 3?  
Neste caso, tem-se:

$$P(C_1) = P(C_2) = P(C_3) = \frac{1}{3}$$

$$P(B/C_1) = \frac{1}{9}; P(B/C_2) = \frac{3}{9} = \frac{1}{3}; P(B/C_3) = \frac{3}{8}$$

Pede-se  $P(C_2/B)$  e  $P(C_3/B)$ , que são probabilidades de antecedentes/causas, ou seja, probabilidades de  $C_2$  e  $C_3$  uma vez que ocorreu bola branca. Aplicando-se o teorema de Bayes tem-se:

$$P(C_2/B) = \frac{P(C_2)P(B/C_2)}{P(C_1)P(B/C_1) + P(C_2)P(B/C_2) + P(C_3)P(B/C_3)}$$

$$P(C_2/B) = \frac{\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)}{\left(\frac{1}{3}\right)\left(\frac{1}{9}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{3}{8}\right)} = 0,407 \text{ ou } 40,7\%$$

$$P(C_3/B) = \frac{P(C_3)P(B/C_3)}{P(C_1)P(B/C_1) + P(C_2)P(B/C_2) + P(C_3)P(B/C_3)}$$

$$P(C_3/B) = \frac{\left(\frac{1}{3}\right)\left(\frac{3}{8}\right)}{\left(\frac{1}{3}\right)\left(\frac{1}{9}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{3}{8}\right)} = 0,458 \text{ ou } 45,8\%$$

Na inferência bayesiana, que é baseada no teorema de Bayes, as probabilidades pré-experimentais são revisadas e dão origem a probabilidades pós-experimentais.

## Distribuições de Probabilidades

Variável aleatória é uma função ou uma regra que atribui um número para cada resultado possível de um experimento. Por exemplo, quando uma moeda é lançada três vezes, o número de caras que ocorre pode ser considerado uma variável aleatória, que pode tomar os valores 0, 1, 2 ou 3. Cada um dos resultados prováveis ocorrerá por chance, ou seja, envolve os fenômenos de acaso e incerteza. Chance pode ser entendida como a interação de grande número de fatores que influem coletivamente no resultado de um experimento ou amostra. Como é virtualmente impossível controlar todos os fatores ou predizer como eles atuarão em conjunto para afetar o resultado, não é possível determinar com precisão qual resultado ocorrerá num experimento. Isto é o que caracteriza a variável aleatória.

A variável aleatória discreta pode assumir apenas valores determinados e inteiros. A variável aleatória contínua pode assumir qualquer valor dentro de um intervalo de valores. As variáveis aleatórias contínuas são representadas por dados provenientes de medições enquanto que as discretas são representadas por dados de contagens. Essas variáveis são de natureza quantitativa.

A distinção clara entre variáveis aleatórias discretas e contínuas é importante porque a utilização dos diferentes tipos de distribuições de probabilidades depende do tipo de variável aleatória considerado.

As variáveis aleatórias nominais e ordinais são chamadas de categóricas e são de natureza qualitativa.

Quando uma medida de probabilidade é atribuída a cada resultado possível de uma variável aleatória  $X$ , produz-se uma distribuição de probabilidade.

A distribuição de probabilidade é simplesmente uma lista de probabilidades que são atribuídas a cada valor possível de uma variável aleatória discreta. A distribuição de probabilidade para uma variável aleatória contínua é dada pela função densidade de probabilidade. As principais distribuições de probabilidades são:

### Distribuição Uniforme

Uma variável aleatória discreta tem distribuição uniforme discreta  $(1, N)$  se  $P(X = x/N) = \frac{1}{N}$ ,

$x = 1, 2, \dots, N$ , onde  $N$  é um número inteiro especificado.

Esta distribuição coloca massa igual em cada um dos resultados  $1, 2, \dots, N$ .

Para calcular a média e a variância da distribuição uniforme, considere as identidades

$$\sum_{i=1}^k i = \frac{k(k+1)}{2} \text{ e } \sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}.$$

Então, tem-se que:

A média desta distribuição é expressa como:

$$E(X) = \sum_{x=1}^N xP(X=x/N) = \sum_{x=1}^N x \frac{1}{N} = \frac{N+1}{2}$$

A variância desta distribuição é expressa como:

$$Var(X) = E(X^2) - [E(X)]^2, \text{ sendo}$$

$$E(X^2) = \sum_{x=1}^N x^2 \frac{1}{N} = \frac{(N+1)(2N+1)}{6}. \text{ Então:}$$

$$Var(X) = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \text{ ou}$$

$$Var(X) = \frac{(N+1)(N-1)}{12}.$$

Esta distribuição pode ser generalizada de modo que o espaço amostral seja qualquer amplitude de números inteiros  $N_0, N_0 + 1, \dots, N_1$ , com  $fp$  (função de probabilidade)

$$P(X=x/N_0, N_1) = \frac{1}{(N_1 - N_0 + 1)}.$$

Uma variável aleatória contínua tem distribuição uniforme contínua definida pela massa uniformemente distribuída sobre um intervalo  $[a, b]$ .

A  $fdp$  (função densidade de probabilidade) desta distribuição é expressa como:

$$f(x/a, b) = \begin{cases} \frac{1}{b-a}, & \text{se } x \in [a, b] \\ 0, & \text{contrário} \end{cases}$$

Pode ser verificado que  $\int_a^b f(x)dx = 1$ . Tem-se também que:

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{b+a}{2} \text{ e}$$

$$Var(X) = \int_a^b \frac{\left(x - \frac{b+a}{2}\right)^2}{b-a} dx = \frac{(b-a)^2}{12}.$$

### Distribuição Binomial

Uma variável aleatória discreta  $X$ , cujos resultados podem ser agrupados em duas classes ou categorias, tem uma distribuição de probabilidade binomial. As duas classes de uma distribuição binomial são referidas como sucesso ou falha, sendo mutuamente excludentes.

Observe que  $P(\text{sucesso}) + P(\text{falha}) = 1,0$ , o que significa que as duas classes são coletivamente exaustivas. As observações de um experimento aleatório binomial são chamadas de testes ou ensaios de Bernoulli.

As pressuposições para a utilização da distribuição binomial são:

1-Existe  $n$  observações ou testes idênticos;

2-Cada teste/ensaio tem apenas dois resultados possíveis, um chamado sucesso e outro falha;  
 3-As probabilidades  $p$  de sucesso e  $q=1-p$  de falha permanecem constantes em todos os testes;

4-Os resultados dos testes são independentes uns dos outros.

A probabilidade de o resultado/evento binomial ocorrer exatamente  $x$  vezes em  $n$  testes, ou seja, ocorrência de  $x$  sucessos e  $n-x$  falhas, é expressa como:

$$P(X=x) = \binom{n}{x} [P(\text{sucesso})]^x [P(\text{falha})]^{n-x} \text{ ou}$$

$$P(X=x) = \binom{n}{x} p^x q^{n-x}, \text{ sendo } \binom{n}{x} = \frac{n!}{x!(n-x)!} \text{ e onde a variável aleatória discreta } X$$

denota o número de sucessos em  $n$  testes, e  $x=0,1,2,\dots,n$ .

Observe que se ocorre  $x$  sucessos em  $n$  testes, então ocorre também  $n-x$  falhas, porque  $x+(n-x)=n$ , que é o número total de testes/observações.

A média de uma distribuição binomial é expressa como:  $\mu_X = np$ .

O desvio padrão de uma distribuição binomial é expresso como:  $\sigma_X = \sqrt{np(1-p)}$ .

### Distribuição Normal

Uma variável aleatória contínua  $X$ , com uma faixa de possíveis valores de  $-\infty$  a  $+\infty$  e com uma média  $\mu_X$  e um desvio padrão  $\sigma_X$ , tem uma distribuição de probabilidade normal, se e somente se, sua função densidade de probabilidade é dada por:

$$f(X) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right]$$

Plotando-se  $f(X)$  contra  $X$  obtemos uma figura em forma de sino. A curva obtida é simétrica em relação ao valor  $\mu_X$  e a área sob a curva é igual a 1. Pode ser verificado que  $E(X) = \mu_X$  e  $E[(X - \mu_X)^2] = \sigma_X^2$ .

A distribuição normal desempenha papel central para os modelos estatísticos usados em análise de variância, regressão e correlação.

## Distribuição Normal Padronizada

Considere a distribuição normal de uma variável aleatória  $X$ , com média  $\mu_X$  e desvio padrão  $\sigma_X$ .

Pode-se definir uma nova variável aleatória  $z$  onde  $z = \frac{X - \mu_X}{\sigma_X}$ . Uma vez que  $\mu_X$  e  $\sigma_X$  são constantes,  $z$  também tem uma distribuição normal e para todo valor de  $X$  há um valor correspondente de  $z$ . Verifica-se que:

$$E(z) = E\left(\frac{X - \mu_X}{\sigma_X}\right) \Rightarrow E(z) = \frac{1}{\sigma_X} [E(X) - \mu_X] = 0$$

$$VAR(z) = E[z - E(z)]^2$$

$$VAR(z) = E\left(\frac{X - \mu_X}{\sigma_X} - 0\right)^2$$

$$VAR(z) = \frac{1}{\sigma_X^2} E[(X - \mu_X)^2]$$

$$VAR(z) = \frac{1}{\sigma_X^2} \sigma_X^2 = 1$$

A variável aleatória  $z$  tem média 0 e desvio padrão 1 e é dita ter uma distribuição normal padronizada. Existem tabelas de probabilidades de  $z$  que podem ser utilizadas para qualquer variável normalmente distribuída.

## Distribuição Amostral da Média

Nas análises de variância, regressão e correlação, faz-se uso frequente de variáveis aleatórias denominadas de estimadores (ou estatísticas). Um estimador é uma variável aleatória cujos valores são calculados a partir de dados amostrais. Ele é uma variável aleatória porque o mesmo cálculo, para diferentes amostras de uma mesma população, pode produzir diferentes valores para o estimador.

A variância deste tipo de variável depende do tamanho da população, da variância da população e

do tamanho da amostra. Por exemplo,  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$  é um estimador chamado de média da amostra.

Um estimador, sendo uma variável aleatória terá uma distribuição de probabilidade, que é chamada de distribuição amostral; seu desvio padrão é chamado de erro padrão do estimador (ou estatística).

A distribuição da média  $(\bar{X})$  é descrita por dois teoremas: um fornece a esperança e a variância e o outro a forma da distribuição.

Teorema: Se uma população infinitamente grande tem média  $\mu_X$  e desvio padrão  $\sigma_X$ , então

$$E(X) = \mu_X \text{ e } Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}.$$

Teorema do Limite Central: Se uma população infinitamente grande tem média  $\mu_X$  e desvio padrão  $\sigma_X$ , a distribuição da média amostral  $\bar{X}$  aproxima-se da distribuição normal com média  $\mu_X$  e desvio padrão  $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$  à medida que o tamanho da amostral ( $n$ ) aumenta.

Quando  $\bar{X}$  tem uma distribuição normal,  $z = \frac{\bar{X} - \mu_X}{\sigma_{\bar{X}}}$  tem uma distribuição normal padronizada e pode ser usada para fornecer probabilidades.

### Distribuição t

Considere a distribuição de  $z = \frac{(\bar{X} - \mu_X)}{\sigma_{\bar{X}}}$ . Supondo-se que  $\sigma_X$  não seja conhecido, mas que

pode ser estimado a partir do desvio padrão amostral, ou seja, usando-se

$$s_X = \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \right]^{1/2} \text{ para estimar } \sigma_X.$$

Devido a  $s_X$  e  $\bar{X}$  serem variáveis aleatórias, usa-se um símbolo diferente, t para a estatística

$$\frac{(\bar{X} - \mu_X)}{(s_X / \sqrt{n})} = \frac{(\bar{X} - \mu_X)}{s_{\bar{X}}}.$$

Tabelas para t não são construídas em termos de  $v = n - 1$ , uma quantidade que é chamada graus de liberdade para a estatística/estimador  $(\bar{X} - \mu_X) / s_{\bar{X}}$

## Distribuição F

Considere a estatística  $S_Y^2 / S_X^2$  que é calculada a partir de dados tomando-se uma amostra de tamanho  $n_Y$  da distribuição normal de  $Y$  e outra de tamanho  $n_X$  da distribuição normal de  $X$ .

Sob a hipótese que  $S_Y^2 = S_X^2$ , a estatística  $S_Y^2 / S_X^2$  tem uma distribuição de probabilidades chamada de distribuição F, cuja forma é função do tamanho das amostras.

A estatística F é uma razão entre duas outras estatísticas, portanto, tem dois valores de graus de liberdade (um para o numerador e outro para o denominador). Tabelas de F são construídas em função dos graus de liberdade do numerador e do denominador.

## Distribuição $\chi^2$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Sabe-se que  $S^2$  é o estimador de  $\sigma^2$

Para fazer inferências (intervalo de confiança, teste de hipóteses) sobre este estimador deve ser considerada a distribuição amostral de  $S^2$  que é representada por uma nova distribuição denominada de  $\chi^2$  (qui-quadrado), cuja forma depende de  $n-1$ .

Definição da Distribuição  $\chi^2$ : Seja  $X_1, X_2, \dots, X_n$  uma variável com distribuição normal, ou seja,  $X \sim N(\mu, \sigma^2)$ .

Então a distribuição de

$$\chi^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1) S^2}{\sigma^2}$$

é chamada de

distribuição  $\chi^2$  com  $n-1$  graus de liberdade.

## Distribuição Gama

Uma variável aleatória  $X$  tem distribuição gama se sua função densidade de probabilidade é expressa como:

$$f(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (\alpha, \beta > 0), \text{ onde } \Gamma(\alpha) \text{ é a função gama. A função gama é}$$

definida como:

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt, n > 0.$$

A média e a variância da distribuição gama são:

$$\mu = \alpha\beta \text{ e } \sigma^2 = \alpha\beta^2.$$

### Distribuição Beta

Uma variável aleatória  $X$  tem distribuição beta se sua função densidade de probabilidade é expressa como:

$$f(x) = \begin{cases} \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}, & 0 < x < 1 \\ 0, \text{contrário } 0 < x < 1 \end{cases} \quad (\alpha, \beta > 0)$$

, sendo  $B(\alpha, \beta)$  a função beta. A função beta é definida como:

$$B(m, n) = \int_0^1 u^{m-1} (1-u)^{n-1} du, m > 0, n > 0.$$

A função beta está relacionada com a função gama por meio da seguinte expressão:

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}. \text{ Desta forma, a distribuição beta pode também ser definida pela seguinte}$$

função densidade de probabilidade:

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & 0 < x < 1 \\ 0, \text{contrário } 0 < x < 1 \end{cases}, \text{ onde}$$

$$(\alpha, \beta > 0).$$

A média e a variância são:

$$\mu = \frac{\alpha}{\alpha+\beta} \text{ e } \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

A distribuição qui-quadrado pode ser definida, para  $x \geq 0$ , por meio da expressão:

$$P(\chi^2 \leq x) = \frac{1}{2^{v/2} \Gamma(v/2)} \int_0^x u^{(v/2)-1} e^{-u/2} du, \text{ onde}$$

$\chi^2 = X_1^2 + X_2^2 + \dots + X_v^2$  é uma variável aleatória obtida de  $X_1, X_2, \dots, X_v$  variáveis aleatórias independentes com distribuição normal de média zero e variância 1 e  $v$  é o número de graus de liberdade.

Então, a função densidade de probabilidade de uma distribuição qui-quadrado é expressa como:

$$f(x) = \begin{cases} \frac{1}{2^{v/2} \Gamma(v/2)} x^{(v/2)-1} e^{-x/2}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Observe que a distribuição qui-quadrado é um caso especial da distribuição gama em que  $\alpha = v/2$  e  $\beta = 2$ , sendo  $\mu = v$  e  $\sigma^2 = 2v$ .

## Inferência por Máxima Verossimilhança

A ideia básica da estimativa por máxima verossimilhança é estimar o valor da população que melhor se ajuste à amostra, ou seja, o valor da população que possa gerar a amostra observada, com maior probabilidade que qualquer outro valor.

### Estimativas de Máxima Verossimilhança para uma População Binomial

Considere que numa amostragem de uma remessa de máquinas agrícolas foram encontradas 3 máquinas defeituosas em um lote de 5. Uma forma de obter uma estimativa da proporção,  $\pi$ , de máquinas defeituosas é  $P=3/5=0,60$ . Um método alternativo é o seguinte: considere uma faixa de valores possíveis de  $\pi$  e procure escolher o que melhor explique a amostra.

O valor  $\pi=0,1$  é um valor possível para a população?

Para  $\pi=0,1$ , a probabilidade de  $S=3$  máquinas defeituosas em uma amostra de  $n=5$  observações é obtida por meio da seguinte equação binomial:

$$\binom{n}{S} \pi^S (1-\pi)^{n-S} = \binom{5}{3} (0,1)^3 (0,9)^3, \text{ cujo valor obtido em tabela de distribuição}$$

binomial é **0,0081** (8 chances em mil de obter a amostra observada). Na mesma tabela, para  $\pi=0,2$ , a probabilidade de  $S=3$  máquinas defeituosas em uma amostra de  $n=5$  observações é **0,0512** (51 chances em mil de obter a amostra observada).

Então, parece natural tomar vários valores de  $\pi$  e verificar na tabela o quanto provável é que cada um dos valores  $\pi$  gere a amostra efetivamente observada. Como o valor amostral  $S=3$  é fixo e a

única variável é o valor hipotético de  $\pi$ , o resultado obtido é designado por função de verossimilhança  $L(\pi)$  e expressa como:

$$L(\pi) = p(3/\pi) = \binom{5}{3} \pi^3 (1-\pi)^2. \text{ Desta forma, tem-se:}$$

Valores de  $\pi$

$$L(\pi) = p(3/\pi) = \binom{5}{3} \pi^3 (1-\pi)^2$$

0	0
0,1	0,008
0,2	0,051
0,3	0,132
0,4	0,230
0,5	0,312
0,6	0,346*
0,7	0,309
0,8	0,205
0,9	0,073
1,0	0

\*máximo=0,346 para  $\pi=0,6$

A estimativa de máxima verossimilhança é o valor de  $\pi$  que maximiza esta função de verossimilhança. Observe que, neste caso, isto ocorre quando  $\pi=0,6$ , o que coincide com o estimador  $P=3/5=0,60$ .

Então, pode-se formular a seguinte definição geral:

O estimador de máxima verossimilhança é o valor populacional hipotético que maximiza a função de verossimilhança da amostra observada.

Generalizando este exemplo tem-se:

Para uma população binomial  $0-1$ , a probabilidade de  $S$  sucessos em  $n$  observações é expressa como:

$$p(S/\pi) = \binom{n}{S} \pi^S (1-\pi)^{n-S}. \text{ Como } S \text{ é um valor observado fixo, esta expressão é descrita}$$

como a função de verossimilhança somente de  $\pi$ :

$$L(\pi) = \binom{n}{S} \pi^S (1-\pi)^{n-S}. \text{ Agora, tente todos os valores possíveis de } \pi, \text{ escolhendo o que}$$

maximize esta função de verossimilhança. O cálculo mostra este valor é  $\pi = \frac{S}{n}$  que é igual a  $P$  (proporção amostral). Logo, o estimador de máxima verossimilhança de  $\pi$  é  $P$ .

Estimativas de Máxima Verossimilhança para uma População Normal

Considere agora uma população normal, com observações  $X_1, X_2, X_3$ , para a qual procura-se o estimador de máxima verossimilhança de  $\mu$ . A probabilidade de a amostra observada resultar de qualquer  $\mu$  dado é expressa como:

$$p(X_1, X_2, X_3 / \mu) = \prod_{i=1}^3 \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2} \right]$$

Como  $(X_1, X_2, X_3)$  são fixados pelos seus valores amostrais, a função de verossimilhança de  $\mu$  é expressa como:

$$L(\mu) = \prod_{i=1}^3 \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2} \right]. \text{ Agora, tente todos os valores possíveis de}$$

$\mu$ , escolhendo o que maximize esta função de verossimilhança. O cálculo mostra que este valor é:  $\mu = \frac{X_1 + X_2 + X_3}{3}$ , que é igual a  $\bar{X}$  (média amostral). Então, o estimador de máxima verossimilhança de  $\mu$  é  $\bar{X}$ .

### Estimativas de Máxima Verossimilhança para um Modelo de Regressão Linear

Considere o seguinte modelo de regressão linear:

$$Y = \alpha + \beta X + \gamma Z + e.$$

Dado um conjunto de observações  $Y_1, Y_2, \dots, Y_n$ , as estimativas de máxima verossimilhança de  $\alpha, \beta, \gamma$  são obtidas da seguinte forma:

A probabilidade de a amostra obtida resultar de qualquer combinação de  $\alpha, \beta, \gamma$  é expressa como:

$$\begin{aligned} p(Y_1, Y_2, \dots, Y_n / \alpha, \beta, \gamma) &= \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum [Y_i - (\alpha + \beta r_i + \gamma z_i)]^2}. \end{aligned}$$

Com  $Y_i$ ,  $X_i$  e  $Z_i$  fixados em seus valores, a equação anterior se torna a função de verossimilhança de  $\alpha, \beta, \gamma$ , designada como  $L(\alpha, \beta, \gamma)$ .

Então, deve-se tentar todas as combinações de  $\alpha$ ,  $\beta$  e  $\gamma$ , escolhendo a combinação que maximize esta função de verossimilhança, minimizando a soma de quadrados no exponente. Desta forma, o estimador de máxima verossimilhança é de quadrados mínimos.

Estimativas de Máxima Verossimilhança para uma População Qualquer

Considere uma população qualquer,  $p(X/\theta)$ .

Dado um conjunto de observações,  $X_1, X_2, \dots, X_n$ , a estimativa de máxima verossimilhança de

$\theta$  da seguinte forma:

A probabilidade de a amostra resultar de qualquer  $\theta$  dado é expressa como:

$$p(X_1, X_2, \dots, X_n/\theta) = p(X_1/\theta)p(X_2/\theta)\cdots p(X_n/\theta)$$

$$p(X_1, X_2, \dots, X_n/\theta) = \prod_{i=1}^n p(X_i/\theta)$$

Com  $(X_1, X_2, \dots, X_n)$  fixos em seus valores, a equação anterior é designada como a função de verossimilhança de  $\theta$  e expressa como:

$$L(\theta) = \prod_{i=1}^n p(X_i/\theta)$$

Então, deve-se escolher o valor de  $\theta$  que maximize esta função de verossimilhança.

Estimativas de Máxima Verossimilhança de Média Normal

Considere uma amostra  $(X_1, X_2, X_3)$ , extraída de uma população que tem distribuição

$$N(\mu, \sigma^2).$$

A estimativa de máxima verossimilhança da média,  $\mu$ , é obtida da seguinte forma:

Uma vez que a população é normal, a densidade de probabilidade de observar qualquer valor  $X$ , dada uma média populacional  $\mu$  é expressa como:

$$p(X/\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X-\mu)^2}$$

As densidades de probabilidade específicas de obtenção dos valores das observações  $X_1$ ,  $X_2$  e

$X_3$  são, respectivamente:

$$p(X_1/\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2},$$

$$p(X_2/\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{1}{2\sigma^2}\right)(X_2-\mu)^2}$$

$$p(X_3/\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{1}{2\sigma^2}\right)(X_3-\mu)^2}$$

Admitindo que  $X_1$ ,  $X_2$  e  $X_3$  sejam independentes, de modo que a densidade de probabilidade conjunta seja o produto das densidades de probabilidade individuais tem-se que:

$$p(X_1, X_2, X_3/\mu) = \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{1}{2\sigma^2}\right)(X_1-\mu)^2} \right] \times \\ \times \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{1}{2\sigma^2}\right)(X_2-\mu)^2} \right] \times \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{1}{2\sigma^2}\right)(X_3-\mu)^2} \right]$$

Mas, os valores amostrais  $X_1$ ,  $X_2$  e  $X_3$  são fixos, enquanto que  $\mu$  é suposto variar em torno de valores hipotéticos. Então, a equação anterior é denominada de função de verossimilhança e designada por  $L(\mu)$ .

Desta forma, a estimativa de máxima verossimilhança de  $\mu$  é definida como o valor hipotético de  $\mu$  que maximiza essa função de verossimilhança, o qual, em uma população normal é  $\bar{X}$ .

## Inferência Bayesiana

Nos métodos de testes de hipótese e estimação de parâmetros a crença *a priori* sobre um parâmetro pode desempenhar um importante papel. A inferência com base no teorema de Bayes ou inferência bayesiana é o meio de considerar uma informação *a priori*, o que pode ser útil para melhorar a compreensão das limitações dos métodos clássicos de inferência estatística.

### Probabilidades *a posteriori*

Suponha que em determinada região ocorra chuva em 40% dos dias e faça sol em 60% dos dias. O serviço de meteorologia, embora digno de confiança, erra em 10% das vezes prevendo sol em dias de chuva e 30% das vezes prevendo chuva em dias de sol. Então tem-se:

Quadro1-Probabilidades *a priori*,  $p(\theta)$

Clima ( $\theta$ )	$p(\theta)$
$\theta_1$ (Chuva)	0,40
$\theta_2$ (Sol)	0,60

Observe que a melhor predição do clima para o dia de amanhã, antes de consultar o serviço de meteorologia, seria a distribuição *a priori* descrita no Quadro1. Mas, após constatar que o serviço de meteorologia prediz chuva, não se poderia esperar chuva mais do que sugere o Quadro1? Intuitivamente, a resposta é afirmativa. Em outras palavras, com esta nova informação, qual seria a distribuição *a posteriori*?

A fidedignidade do serviço de meteorologia é determinada como:

Quadro2-Probabilidades condicionais,  $p(X/\theta)$

Clima ( $\theta$ )	Predição (X)			$\Sigma$
	$X_1(\text{Chuva})$	$X_2(\text{Sol})$		
$\theta_1(\text{Chuva})$	0,90	0,10		1,00
$\theta_2(\text{Sol})$	0,30	0,70		1,00

Observe que as probabilidades condicionais são combinadas com as probabilidades *a priori* para definir o espaço amostral *a posteriori*.

Considere agora a proporção dos dias em que o clima,  $\theta$ , é de chuva e também em que a predição,  $X$ , é de chuva: 40% das vezes o clima é de chuva,  $p(\theta_1)=0,40$ , e quando isto ocorre há 90% de probabilidade de que a predição seja de chuva,  $p(X_1/\theta_1)=0,90$ , consequentemente, esta combinação ocorre em 90% de 40% ou 36% das vezes, ou seja:

$$p(\theta_1, X_1) = p(\theta_1)p(X_1/\theta_1) = (0,4)(0,90) = 0,36$$

Da mesma forma, a probabilidade de sol,  $p(\theta_2)=0,60$ , combinada com a predição de chuva,

$$p(X_1/\theta_2)=0,30, \text{ é obtida como:}$$

$$p(\theta_2, X_1) = p(\theta_2)p(X_1/\theta_2) = (0,60)(0,30) = 0,18$$

Após ter sido previsto chuva, o espaço amostral relevante passa a ser os dias em que há predição de chuva. Nesse espaço amostral, menor que o espaço amostral total, observa-se que chuva é duas vezes mais provável do que sol (0,36 contra 0,18), o que origina a distribuição *a posteriori*, que pode ser obtida como a seguir.

Neste contexto, a predição de chuva é expressa como:

$$p(X_1) = 0,36 + 0,18 = 0,54$$

Então, tem-se que:

$$p(\theta_1/X_1) = \frac{p(\theta_1, X_1)}{p(X_1)} = \frac{p(\theta_1)p(X_1/\theta_1)}{p(X_1)} = \frac{0,36}{0,54} = 0,67$$

$$p(\theta_2/X_1) = \frac{p(\theta_2, X_1)}{p(X_1)} = \frac{p(\theta_2)p(X_1|\theta_2)}{p(X_1)} = \frac{0,18}{0,54} = 0,33$$

Quadro3-Distribuição de probabilidades *a posteriori*,  $p(\theta/X_1)$

Clima( $\theta$ )	$p(\theta/X_1)$
$\theta_1(\text{Chuva})$	0,67
$\theta_2(\text{Sol})$	0,33

Observe que, antes de conhecer a evidência/predição (probabilidade de chuva do serviço de meteorologia), as probabilidades *a priori*,  $p(\theta)$ , fornecem as chances de aposta no clima. Mas, depois de conhecer a evidência/predição,  $X_1(\text{Chuva})$ , as chances de aposta no clima podem ser atualizadas/melhoradas por meio das probabilidades *a posteriori*,  $p(\theta/X_1)$ , que fornecem as verdadeiras chances de aposta no clima.

Este exemplo pode ser generalizado por meio das seguintes expressões:

$$p(\theta/X_1) = \frac{p(\theta, X_1)}{p(X_1)} \text{ ou } p(\theta/X_1) = \frac{p(\theta)p(X_1|\theta)}{p(X_1)}$$

Observe no Quadro3 e nesta equação que  $\theta$  varia sobre todas as possíveis condições do clima, enquanto que  $X_1$  é a observação fixa/constante. Então tem-se que:

$$p(\theta/X_1) = c \left[ p(\theta)p(X_1|\theta) \right], \text{ onde } c = \frac{1}{p(X_1)}. \text{ A constante } c \text{ faz com que a}$$

probabilidade total seja igual a 1,0 sobre todos os  $\theta$ .

Observe ainda que,  $p(X_1|\theta)$  é denominada de função de verossimilhança. Então, a equação anterior pode ser descrita como a seguir:

(distribuição *a posteriori*)  $\propto$  (distribuição *a priori*) (função de verossimilhança)

Este é o princípio geral da inferência bayesiana, que indica como a distribuição final/atualizada (*a posteriori*) é calculada combinando-se a distribuição *a priori* com a informação amostral/função de verossimilhança.

### Distribuição *a posteriori* para uma proporção $\pi$

Considere o caso de mais de duas condições naturais ( $\theta$ ) como sendo uma proporção populacional  $\pi$ . Por exemplo, seja  $\pi$  a proporção de rádios defeituosos em cada remessa de uma fábrica. Os

valores máximos de  $\pi$  em todas as encomendas enviadas no passado, ou seja, a distribuição *a priori* de  $\pi$  estão apresentados no quadro a seguir:

Quadro4-Probabilidades *a priori* de  $\pi$

(1) Proporção (%) de rádios defeituosos ( $\pi$ )	(2) Número de encomendas	(3) Número relativo de encomendas
0	2	0,01
10	30	0,15
20	40	0,20
30	42	0,21
40	34	0,17
50	26	0,13
60	16	0,08
70	8	0,04
80	2	0,01
90		0
100		0
	200	1,00

Suponha que um comprador precisa tomar uma decisão sobre se devolve ou não uma remessa de rádios da fábrica que se encontra em sua loja a três semanas. Para tomar a decisão ele precisa obter uma estimativa da proporção de rádios defeituosos ( $\pi$ ).

Considere que, para obter uma evidência amostral sobre  $\pi$ , ele examina 5 rádios ao acaso e que 3 desses são defeituosos. Qual é agora a distribuição *a posteriori* de  $\pi$ ?

Para calcular a distribuição *a posteriori* deve-se utilizar a seguinte identidade:

(distribuição *a posteriori*)  $\propto$  (distribuição *a priori*) (função de verossimilhança)

Então, primeiro é necessário obter a função de verossimilhança, ou seja, a verossimilhança de 3 rádios defeituosos observados na amostra de 5 rádios. Essa função é obtida por meio da equação binomial, para  $S$  fixo, igual a 3, e vários valores de  $\pi$ , em tabela de Probabilidades Binomiais Individuais,  $p(s)$

Em seguida, multiplica-se a função de verossimilhança pela distribuição *a priori* para obter a distribuição *a posteriori*, apresentada no quadro a seguir:

Quadro5-Cálculos para obter a distribuição *a posteriori*

(4) Verossimilhança de  $\pi$ , (5) *Priori* vezes (6) Divisão da *posteriori*  
tabelado para  $n=5$  e verossimilhança  $(3)x(4)$  por 0,160

$S=3$

0	0	0
0,01	0,002	0,01
0,05	0,010	0,06
0,13	0,027	0,17
0,23	0,039	0,24
0,31	0,040	0,25
0,35	0,028	0,18
0,31	0,012	0,08
0,20	0,002	0,01

0,07	0	0
0	0	0
	0,160	1,00

Suponha que a remessa é considerada satisfatória apenas no caso de  $\pi$  ser inferior a 25%.

Qual a probabilidade de isto acontecer antes de obter a amostra?

Com base no Quadro4 tem-se:

$$P(\pi < 25\%) = 0,01 + 0,15 + 0,20 = 0,36$$

Qual a probabilidade de isto acontecer depois de obter a amostra?

Com base no Quadro5 tem-se:

$$P(\pi < 25\% / S=3) = 0 + 0,01 + 0,06 = 0,07$$

Observe que, o uso da informação obtida com a amostra reduziu de 0,36 para 0,07 a probabilidade de uma remessa ser considerada satisfatória.

Os aspectos gerais da inferência bayesiana mostrados por meio deste exemplo são:

1-A distribuição *a posteriori* situa-se entre a distribuição *a priori* e a função de verossimilhança;

2-A multiplicação da distribuição *a priori* ou da função de verossimilhança por uma constante conveniente não afeta a distribuição *a posteriori*.

### Distribuição *a posteriori* para a média $\mu$

Considere que uma empresa vende vigas de aço. Em cada lote remetido, as resistências à tensão das vigas tem distribuição normal em torno da média  $\mu$ , com variância  $\sigma^2 = 300$ . Mas,  $\mu$  varia de um lote para outro devido a um controle de qualidade deficiente. Suponha que as médias  $\mu$  de todos os lotes tenha distribuição aproximadamente normal, com média  $\mu_0 = 60$  e variância

$$\sigma_0^2 = 100.$$

Suponha que um cliente precisa tomar uma decisão quanto à devolução de um lote específico de vigas adquirido na referida empresa. A decisão do cliente dependerá de sua suposição sobre a resistência média,  $\mu$ . Considere que o cliente dispõe de uma amostra de 12 vigas desse lote, com  $\bar{X} = 70$ . Pergunta-se:

1-Qual é a função de verossimilhança? Ou, qual a verossimilhança de obter  $\bar{X} = 70$ ?

A distribuição de  $\bar{X}$  é normal, com variância  $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{300}{12} = 25$ .

A função da distribuição normal é expressa como

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{1}{2}\right)\frac{(x-\mu)^2}{\sigma^2}}$$

Para simplificar, ignore o multiplicador constante  $\frac{1}{\sqrt{2\pi}\sigma}$ .

Desta forma, a função de distribuição de  $\bar{X}$  é expressa como:

$$p(\bar{X}/\mu) \propto e^{-\left(\frac{1}{2}\right)\frac{(\bar{X}-\mu)^2}{25}}, \text{ onde } \bar{X}=70 \text{ e } \mu \text{ varia.}$$

Designando esta equação por  $L(\mu)$  tem-se:

$$L(\mu) \propto e^{-\left(\frac{1}{2}\right)\frac{(70-\mu)^2}{25}}. \text{ Pode-se reconhecer nesta função de verossimilhança uma curva normal centrada em 70 com desvio padrão } \sqrt{25}=5.$$

2-Qual é a distribuição *a posteriori* obtida multiplicando a distribuição *a priori* pela função de verossimilhança?

Deve-se fazer esta multiplicação para vários valores de  $\mu$  (por exemplo 55, 60, 65, 70, 75, 80 e 85) para obter os diversos pontos e esboçar o gráfico da distribuição *a posteriori*, que pode ser utilizado para estimar a probabilidade de  $\mu$  estar abaixo ou acima de determinado valor.

A distribuição *a posteriori* é calculada a partir de dois tipos de informação:

1-A amostra efetiva centrada em  $\bar{X}$ , que é a base da função de verossimilhança;

2-A distribuição *a priori*, que é equivalente à informação em outra amostra, chamada de hipotética. Esta amostra hipotética é centrada na média *a priori*  $\mu_0$  e constituída de  $n_0$  observações hipotéticas. Para determinar o tamanho  $n_0$ , deve-se observar que quanto menor a variância,  $\sigma_0^2$ ,

da distribuição *a priori* em relação a  $\sigma^2$ , maior será a influência dessa distribuição e, consequentemente, maior será  $n_0$ . Desta forma, o tamanho  $(n_0)$  da amostra hipotética é

$$\text{expresso como } n_0 = \frac{\sigma^2}{\sigma_0^2}, \text{ ou seja: } \sigma_0^2 = \frac{\sigma^2}{n_0}$$

Quando as  $n_0$  observações hipotéticas centradas na média  $(\mu_0)$  *a priori* são combinadas com as  $n$  observações efetivas centradas em  $\bar{X}$ , a média global é expressa como:

$$\text{média } a \text{ posteriori} = \frac{n_0\mu_0 + n\bar{X}}{n_0 + n}$$

A variância da distribuição *a posteriori* é expressa como:

$$\text{variância } a \text{ posteriori} = \frac{\sigma^2}{n_0 + n}$$

Portanto, tem-se que:

$$\mu \square N\left( \frac{n_0\mu_0 + n\bar{X}}{n_0 + n}, \frac{\sigma^2}{n_0 + n} \right)$$

Reconsiderando o exemplo da empresa que vende vigas de aço, pede-se:

1-Exprimir o valor da distribuição a priori em termos de “observações hipotéticas”.

Tem-se que  $n_0 = \frac{\sigma^2}{\sigma_0^2} = \frac{300}{100} = 3$ . Portanto, a distribuição *a priori* é equivalente a uma amostra hipotética de 3 observações.

2-Calcular a distribuição *a posteriori*.

A distribuição *a posteriori* é normal com:

$$\text{média } a \text{ posteriori} = \frac{n_0\mu_0 + n\bar{X}}{n_0 + n} = \frac{3(60) + 12(70)}{3+12} = 68$$

$$\text{variância } a \text{ posteriori} = \frac{\sigma^2}{n_0 + n} = \frac{300}{3+12} = 20$$

3-Calcular a probabilidade de  $\mu$  estar abaixo de 62,5, antes de extrair a amostra.

Para antes de extrair a amostra utiliza-se a distribuição *a priori*. Então tem-se:

$$\begin{aligned} Pr(\mu < 62,5) &= Pr\left(\frac{\mu - \mu_0}{\sigma_0} < \frac{62,5 - 60}{\sqrt{100}}\right) \\ &= Pr(z < 0,25) = 1 - 0,40 = 0,60 \end{aligned}$$

4- Calcular a probabilidade de  $\mu$  estar abaixo de 62,5, após ser extraída a amostra.

Para após ser extraída a amostra utiliza-se a distribuição *a posteriori*. Então tem-se:

$$\begin{aligned} Pr(\mu < 62,5) &= Pr\left(\frac{\mu - 68}{\sqrt{20}} < \frac{62,5 - 68}{\sqrt{20}}\right) \\ &= Pr(z < -1,23) = 0,11 \end{aligned}$$

A distribuição *a posteriori* pode ser utilizada para calcular um intervalo de confiança para  $\mu$ . O intervalo de confiança bayesiano que tem 95% de chance de conter  $\mu$  é expresso como:

$$\mu = \frac{n_0\mu_0 + n\bar{X}}{n_0 + n} \pm 1,96 \sqrt{\frac{\sigma^2}{n_0 + n}}$$

Em comparação com o intervalo de confiança clássico,  $\mu = \bar{X} \pm 1,96 \sqrt{\frac{\sigma^2}{n}}$ , o intervalo de confiança bayesiano está centrado em um valor intermediário entre  $\mu_0$  e  $\bar{X}$ , e, é mais estreito por causa das  $n_0$  observações hipotéticas extras, proporcionadas pela informação *a priori*.

Para o exemplo da empresa que vende vigas de aço, os intervalos de confiança são:

1-Intervalo de confiança bayesiano de 95%

$$\mu = \frac{n_0\mu_0 + n\bar{X}}{n_0 + n} \pm 1,96 \sqrt{\frac{\sigma^2}{n_0 + n}}, \text{ ou seja,}$$

$$\mu = \text{média } a \text{ posteriori} \pm 1,96 \text{ desvios padrão } a \text{ posteriori}$$

$$\mu = 68 \pm 1,96 \sqrt{20} = 68 \pm 8,8$$

2-Intervalo de confiança clássico de 95%

$$\mu = \bar{X} \pm 1,96 \sqrt{\frac{\sigma^2}{n}} = 70 \pm 1,96 \sqrt{\frac{300}{12}} = 70 \pm 9,8$$

Observe que o intervalo de confiança clássico é mais amplo porque não considera a informação *a priori*.

### Análise de Regressão Bayesiana

Suponha que no modelo de regressão linear simples,  $Y = \alpha + \beta X$ , o coeficiente de regressão  $\beta$ , tenha a distribuição *a priori*,  $\beta \square N(\beta_0, \sigma_0^2)$ .

Para uma amostra, tem-se um estimador,  $\hat{\beta}$ , que também tem distribuição normal expressa como:

$$\hat{\beta} \square N\left(\beta, \frac{\sigma^2 / \sigma_X^2}{n}\right), \text{ em que:}$$

$\sigma^2$ : variância do desvio em relação à reta de regressão;

$\sigma_X^2 = \frac{1}{n} \sum x^2$ : variância dos valores de  $X$ .

Quando são combinadas informação *a priori* com informação amostral, obtém-se a distribuição *a posteriori*:

$$\beta \square N\left(\frac{n_0\beta_0 + n\hat{\beta}}{n_0 + n}, \frac{\sigma^2 / \sigma_X^2}{n_0 + n}\right), \text{ onde } n_0 = \frac{\sigma^2 / \sigma_X^2}{\sigma_0^2}.$$

Então, pode ser construído um intervalo de confiança bayesiano de 95% para  $\beta$ , expresso como:

$$\beta = \frac{n_0\beta_0 + n\hat{\beta}}{n_0 + n} \pm 1,96 \sqrt{\frac{\sigma^2/\sigma_X^2}{n_0 + n}}$$

Considere, como exemplo, que a distribuição *a priori* de  $\beta$  seja normal, em torno de uma média 5, com variância 0,25. Para estimar  $\beta$ , uma amostra de 8 observações fornece as seguintes estatísticas:

$$\sum xy = 2400, \sum x^2 = 400 \text{ e variância do desvio de regressão, } S^2 = 25.$$

Usando  $S^2$  como aproximação de  $\sigma^2$ , calcule:

1-O tamanho da amostra hipotética,  $n_0$ , que é equivalente ao da distribuição *a priori*.

A informação *a priori* é fornecida por  $\beta_0 = 5$  e  $\sigma_0^2 = 0,25$ .

$$\text{Com base na amostra pode-se calcular } \hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{2400}{400} = 6 \quad \text{e}$$

$$\sigma_X^2 = \frac{1}{n} \sum x^2 = \frac{400}{8} = 50. \text{ Então, tem-se que:}$$

$$n_0 = \frac{\sigma^2/\sigma_X^2}{\sigma_0^2} = \frac{25/50}{0,25} = 2$$

2-O intervalo de confiança bayesiano de 95%.

Tem-se que:

$$\begin{aligned} \beta &= \frac{n_0\beta_0 + n\hat{\beta}}{n_0 + n} \pm 1,96 \sqrt{\frac{\sigma^2/\sigma_X^2}{n_0 + n}} \\ &= \frac{2(5) + 8(6)}{2+8} \pm 1,96 \sqrt{\frac{25/50}{2+8}} \\ &= 5,80 \pm 0,44 \end{aligned}$$

3-O intervalo de confiança bayesiano de 95%, substituindo 1,96 por  $t_{0,05}$  com  $gl = n_0 + n - 2$ .

Com  $gl = n_0 + n - 2 = 2 + 8 - 2 = 8$ , numa tabela de t bilateral, obtém-se  $t_{0,05} = 2,306$ .

Então, tem-se:

$$\begin{aligned} \beta &= \frac{2(5) + 8(6)}{2+8} \pm 2,306 \sqrt{\frac{25/50}{2+8}} \\ &= 5,80 \pm 2,306(0,224) \end{aligned}$$

$$= 5,80 \pm 0,52$$

4-O intervalo de confiança clássico de 95%.

Neste caso,  $gl = n - 2 = 8 - 2 = 6$ , e, numa tabela de t bilateral obtém-se  $t_{0,05} = 2,447$ .

Então, tem-se:

$$\beta = \hat{\beta} \pm t_{0,05} \sqrt{\frac{s^2}{\sum x^2}}$$

$$= 6 \pm 2,447 \sqrt{\frac{25}{400}}$$

$$= 6 \pm 0,61$$

Observe que o intervalo de confiança bayesiano em (3) é mais estreito que o intervalo de confiança clássico em (4), o que reflete o valor da informação *a priori*.

O intervalo de confiança bayesiano está centrado entre  $\beta_0 = 5$  e  $\hat{\beta} = 6$ . Entretanto, como a amostra contém quatro vezes mais informação,  $n = 8$  comparado com  $n_0 = 2$ , o intervalo de confiança bayesiano está quatro vezes mais próximo de  $\hat{\beta}$ .

Quando  $S^2$  for utilizado como estimador de  $\sigma^2$  deve-se substituir  $z_{0,05}$  para  $t_{0,05}$  no cálculo do intervalo de confiança bayesiano.

### Teoria da Decisão Bayesiana

Considere que um vendedor tenha a concessão de vender refrigerantes num estádio de futebol, mas ele está vendendo refrigerantes e guarda-chuvas. Ele recebe uma taxa líquida de 100 reais por jogo, da qual deve deduzir seu prejuízo, que por sua vez depende da escolha que fizer das mercadorias, em função do clima. Suponha que ele tenha apenas três opções/ações:

$a_1$ : vender apenas refrigerantes;

$a_2$ : vender refrigerantes e guarda-chuvas;

$a_3$ : vender apenas guarda-chuvas.

Se ele vender apenas refrigerantes ( $a_1$ ) e chover, sua perda será de 70 reais. Entretanto, se fizer sol ele perderá somente 10 reais. Se ele escolher as ações  $a_2$  ou  $a_3$  também incorrerá em certas perdas. Todas as informações estão resumidas no quadro a seguir:

Quadro6-Função de perda,  $l(\theta, a)$

Clima ( $\theta$ )	Ação		
	$a_1$	$a_2$	$a_3$

$\theta_1$ (Chuva)	70	40	20
$\theta_2$ (Sol)	10	40	50

Suponha ainda que a distribuição de probabilidade do clima (a longo prazo) seja como no quadro a seguir:

Quadro7-Distribuição de probabilidade de  $\theta$

Clima ( $\theta$ )	$p(\theta)$
$\theta_1$ (Chuva)	0,40
$\theta_2$ (Sol)	0,60

No caso de o vendedor desejar minimizar sua perda (a longo prazo), qual a melhor ação a escolher?

A perda, em média,  $L(a)$ , para cada uma das ações escolhidas é calculada ponderando cada perda com sua frequência relativa, obtendo-se:

$$L(a_1) = 70(0,40) + 10(0,60) = 34$$

$$L(a_2) = 40(0,40) + 40(0,60) = 40$$

$$L(a_3) = 20(0,40) + 50(0,60) = 38$$

No quadro a seguir estão descritos todos os cálculos:

Quadro8-Cálculo da ação ótima usando a distribuição *a priori*  $p(\theta)$

$\theta$	$p(\theta)$	$a$		
		$a_1$	$a_2$	$a_3$
$\theta_1$	0,40	70	40	20
$\theta_2$	0,60	10	40	50
Média de Perdas, $L(a)$		34	40	38

Observe que a média de perdas,  $L(a)$ , é mínima em  $a_1$ . Então, a ação ótima consiste em vender somente refrigerantes.

Este exemplo pode ser generalizado para um número qualquer de estados,  $\theta$ , e/ou de ações,  $a$ . Para cada ação( $a$ ), sempre que ocorrer um determinado estado( $\theta$ ), haverá uma perda correspondente,  $l(\theta, a)$ . A decisão é chamada bayesiana, se for escolhida a ação  $a$  que minimiza a média de perdas esperada, ou seja:

Escolher  $a$  que minimiza  $L(a) \sum_{\theta} l(\theta, a) p(\theta)$ .

As probabilidades,  $p(\theta)$ , devem representar o melhor conhecimento possível da conjuntura. No exemplo anterior, se o vendedor não puder predizer o clima, terá que usar as probabilidades *a priori*, mas, se ele puder prevê-lo de forma confiável, então deverá usar as probabilidades *a posteriori*, que resultam da utilização dessa informação amostral.

Suponha que, no exemplo anterior, o vendedor tenha acesso ao serviço de meteorologia parcialmente confiável, que prediz chuva. Então, a probabilidade de chuva e de sol *a posteriori* é conforme o quadro a seguir:

Quadro9-Distribuição de probabilidades *a posteriori*,  $p(\theta/X_1)$

Clima( $\theta$ )	$p(\theta/X_1)$
$\theta_1$ (Chuva)	0,67
$\theta_2$ (Sol)	0,33

Se o serviço de meteorologia prediz chuva, o que o vendedor deverá fazer?

Agora, a perda, em média,  $L(a)$ , para cada uma das ações escolhidas é calculada ponderando cada perda com as probabilidades *a posteriori*, obtendo-se os resultados descritos no quadro a seguir:

Quadro10-Cálculo da ação ótima,  $a$ , usando a distribuição *a posteriori*,  $p(\theta/X_1)$

$\theta$	$p(\theta/X_1)$	$a$		
		$a_1$	$a_2$	$a_3$
$\theta_1$	0,67	70	40	20
$\theta_2$	0,33	10	40	50
Média de Perdas, $L(a)$		50	40	30

Observe que a média de perdas,  $L(a)$ , é mínima em  $a_3$ . Então, a ação ótima consiste em vender somente guarda-chuvas.

### Teste de hipóteses bayesiano

Assim como é possível deduzir estimadores ótimos, tratando as estimativas como forma de decisão, para muitas situações, também é possível deduzir testes de hipóteses ótimos.

Considere que duas espécies de insetos estejam infestando uma plantação. A espécie  $S_0$ , menos daninha (inofensivo), causa um prejuízo de 10 mil reais e a  $S_1$ , mais daninha, causa um prejuízo de 100 mil reais. O agricultor dispõe de um inseticida eficiente, cujo custo de aplicação é de 30 mil reais. Suponha que está prevista uma invasão de insetos, mas o agricultor não sabe de que espécie.

1-Qual deve ser a decisão do agricultor (plicar o inseticida ou não), se a única informação de que dispõe indica que a espécie  $S_0$  corre com frequência igual a três vezes a frequência da  $S_1$ ?

As informações disponíveis e a solução estão resumidas no quadro a seguir:

Quadro11-Probabilidades *a priori*, Função de Perda e Perda Média,  $L(a)$

$p(\theta)$	Estado $\theta$	Ação $a$	
		$a_0$ (Não aplicar)	$a_1$ (Aplicar)
0,75	$S_0$ (Inofensivo)	10	30
0,25	$S_1$ (Daninho)	100	30
	Perda Média $L(a)$	32,5	30

Observe que a perda média mínima de 30 mil reais ocorre na ação  $a_1$ . Desta forma, a ação ótima é  $a_1$  (aplicar o inseticida).

2-Qual deve ser a decisão do agricultor (plicar o inseticida ou não), se a única informação de que dispõe indica que a espécie  $S_0$  é nove vezes mais frequente que a  $S_1$ ?

As informações disponíveis e a solução estão resumidas no quadro a seguir:

Quadro12-Probabilidades *a priori*, Função de Perda e Perda Média,  $L(a)$

$p(\theta)$	Estado $\theta$	Ação $a$	
		$a_0$ (Não aplicar)	$a_1$ (Aplicar)
0,90	$S_0$ (Inofensivo)	10	30
0,10	$S_1$ (Daninho)	100	30
	Perda Média $L(a)$	19,0	30,0

Neste caso, a ação ótima seria  $a_0$  (não aplicar o inseticida) com perda mínima de 19 mil reais.

Observe que é tão pouco provável que os insetos sejam daninhos,  $p(\theta)=0,10$ , que vale a pena correr o risco, ou seja, não aplicar o inseticida.

Observe também que a ação ótima seria  $a_0$ , mesmo se não houver certeza sobre a espécie  $S_0$ , mas apenas suficiente probabilidade, por exemplo, 9 vezes mais provável que  $S_1$ . Ou seja, se  $S_0$

for suficientemente provável, a ação seria como se houvesse certeza. Então,  $S_0$  é aceita como hipótese de trabalho, pois mesmo sabendo que ela pode não ser verdadeira é melhor que a hipótese alternativa,  $S_1$ .

Este exemplo, demonstra em que consiste o teste de hipótese: uma busca da melhor hipótese de trabalho, e não da plena verdade. Para generalizar este exemplo, considere as seguintes notações:  
 1-Suponha um conjunto de dados observados  $X$ , de forma que seja possível dispor das probabilidades *a priori*;

2-Semelhante a  $\theta_0$  e  $\theta_1$ , os estados hipotéticos são chamados de  $H_0$  (hipótese nula) e  $H_1$  (hipótese alternativa), e, então as possíveis ações são “aceitar  $H_0$ ” e “aceitar  $H_1$ ”;

3-Abreviar a função perda  $l(\theta_i, a_j)$  para  $l_{ij}$ .

Então, obtém um quadro com a seguinte forma:

Quadro13-Formato para o teste de hipóteses bayesiano

$p(\theta/X)$	Estado $\theta$	Ação $a$	
		$a_0$ (Aceitar $H_0$ )	$a_1$ (Aceitar $H_1$ )
$p(\theta_0/X)$	$H_0$	$l_{00}$	$l_{01}$
$p(\theta_1/X)$	$H_1$	$l_{10}$	$l_{11}$

4-Calcular as perdas médias, da seguinte forma:

$$L(a_0) = p(\theta_0/X)l_{00} + p(\theta_1/X)l_{10}$$

$$L(a_1) = p(\theta_0/X)l_{01} + p(\theta_1/X)l_{11}$$

5-O critério de decisão consiste em escolher  $a_0$  se, e somente se:

$$L(a_0) < L(a_1), \text{ ou seja,}$$

$$p(\theta_1/X)[l_{10} - l_{11}] < p(\theta_0/X)[l_{01} - l_{00}].$$

Observe que, os valores entre colchetes são chamados *regrets* a  $r_0$  e  $r_1$ . O  $r_0 \square l_{01} - l_{00}$ , é a medida do quanto  $l_{01}$  excede  $l_{00}$ , ou seja, é a perda extra ocorrida pela decisão errada, quando  $H_0$  é verdadeira. De forma semelhante, o  $r_1 \square l_{10} - l_{11}$ , é a perda extra (*regret*) acarretada pela decisão errada, quando  $H_1$  é verdadeira.

Observe ainda que, tem-se:

$p(\theta_1/X)r_1 < p(\theta_0/X)r_0$ , ou seja,  $\frac{p(\theta_1/X)}{p(\theta_0/X)} < \frac{r_0}{r_1}$ . Utilizando as probabilidades *a posteriori* nesta equação obtém-se:

$\frac{p(\theta_1)p(X/\theta_1)}{p(\theta_0)p(X/\theta_0)} < \frac{r_0}{r_1}$ . Resolvendo em relação à razão de verossimilhanças tem-se o critério da Razão de Verossimilhança bayesiano que é: aceitar  $H_0$  se, e somente se:

$$\frac{p(X/\theta_1)}{p(X/\theta_0)} < \frac{r_0 p(\theta_0)}{r_1 p(\theta_1)}.$$

Observe que um teste bayesiano usa a mesma informação de um teste clássico e ainda utiliza a distribuição *a priori*,  $p(\theta)$ , e os *regrets* (função perda). O teste clássico fixa a probabilidade de erro Tipo I arbitrariamente ou às vezes com referência implícita a considerações vagas sobre perda e crença *a priori*. Na inferência bayesiana, argumenta-se que tais considerações deveriam ser introduzidas explicitamente tornando todas as hipóteses abertas e expostas a crítica e aperfeiçoamento.

## Estimação de Parâmetros Populacionais

### Conceitos básicos

**Parâmetro:** é uma constante inerente a uma população, cujo valor real desconhecido descreve uma característica da população. Por exemplo, a média ( $\mu$ ) e o desvio padrão ( $\sigma$ ).

**Estatística:** é uma medida descritiva da amostra, ou seja, uma estatística  $M$  é uma função de  $X_1, X_2, \dots, X_n$ , ou seja,  $M = f(X_1, X_2, \dots, X_n)$ .

**Estimador:** é qualquer estatística  $M = f(X_1, X_2, \dots, X_n)$  que é usada para estimar um valor desconhecido  $g(\theta)$ .

**Estimativa:** é o valor numérico assumido pelo estimador, quando os valores observados na amostra  $x_1, x_2, \dots, x_n$  são considerados.

Retirando-se uma amostra de uma população, podem ser obtidas estimativas dos parâmetros populacionais com base nessa amostra. Por exemplo, se o parâmetro sob estudo é a média

populacional ( $\mu$ ), então o estimador (ou estatística) apropriado é fornecido por  $m = \frac{\sum_i X_i}{n}$ , sendo  $n$  o tamanho da amostra.

Como podem ser retiradas várias amostras de uma população, podem ser obtidas inúmeras estimativas de  $m$ . Demonstra-se que a distribuição de  $m$ , independentemente da distribuição populacional, será aproximadamente normal com média  $\mu_m = \mu$  e variância  $\sigma_m^2 = \sigma^2/n$ .

### Estimador e estimativa

Na Estatística Descritiva o objetivo fundamental é o de caracterizar conjuntos restritos de dados. Por exemplo, ao calcular uma média amostral pretende-se apenas definir uma medida de tendência central (ou de localização) para o conjunto de dados. No âmbito da Inferência Estatística, ao calcular estatísticas, existe o objetivo adicional de caracterizar a população a partir da qual a amostra foi retirada, procurando designadamente estimar parâmetros desta população. Para fazer isso, são definidas estatísticas (estimadores) cujos valores particulares constituem estimativas dos parâmetros populacionais que estão sendo estudados. Por exemplo, seria razoável estimar  $\mu$  a partir da média

amostral  $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ . A estatística  $\bar{Y}$  é um estimador pontual do parâmetro em estudo (valor esperado), e uma realização particular  $\bar{y}$  desta estatística, obtida a partir de um conjunto particular de valores observados  $(y_1, y_2, \dots, y_n)$ , constitui uma estimativa do parâmetro  $\mu$ .

Designa-se por estimador pontual de um parâmetro  $\theta$  uma estatística (variável aleatória)  $\hat{\Theta}$  cujos valores particulares  $\hat{\theta}$  constituem estimativas do parâmetro considerado.

### Propriedades desejáveis dos estimadores pontuais

#### a) Não viesamento

O viesamento (ou viés) do estimador  $\hat{\Theta}$  é definido como a diferença entre o valor esperado do estimador,  $E(\hat{\Theta}) = \mu_{\hat{\Theta}}$ , e o valor do parâmetro,  $\theta$ , ou seja:

$Viesamento_{\hat{\Theta}} = E(\hat{\Theta}) - \theta = \mu_{\hat{\Theta}} - \theta$ . Um estimador é dito ser não-viesado quando o viesamento for nulo, como mostrado da figura a seguir.

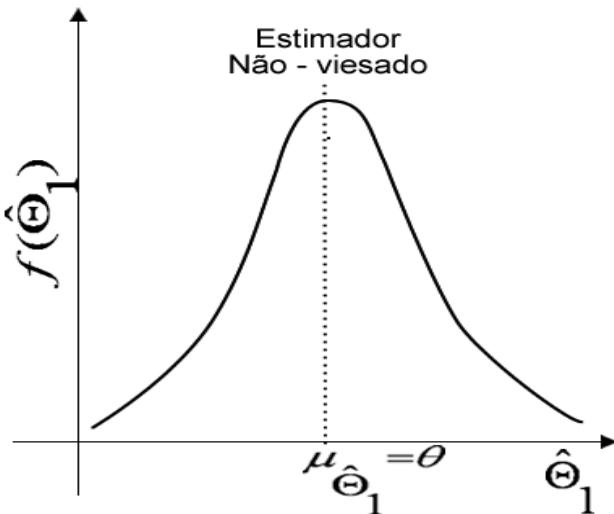


Figura 2.1. Estimadores não-viesado ( $\hat{\Theta}_1$ ) do parâmetro  $\theta$

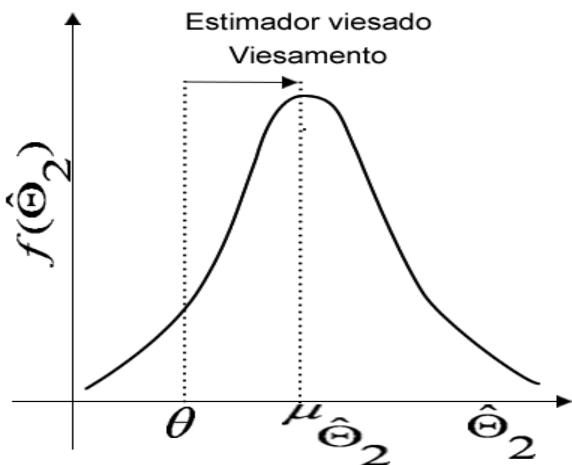


Figura 2.2. Estimadores não-viesado ( $\hat{\Theta}_1$ ) e viesado ( $\hat{\Theta}_2$ ) do parâmetro  $\theta$

Pode ser demonstrado que a média ( $\bar{Y}$ ) e a variância ( $S_Y^2$ ) amostrais são estimadores não-viesados da média e da variância populacional, qualquer que seja a distribuição da população.

### b) Eficiência

Considere dois estimadores não-viesados  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$ , em relação aos quais se admite que, para amostras da mesma dimensão,  $N$ , as suas variâncias são, respectivamente,  $\sigma_1^2$  e  $\sigma_2^2$ , com  $\sigma_1^2 < \sigma_2^2$ , como mostrado da figura a seguir.

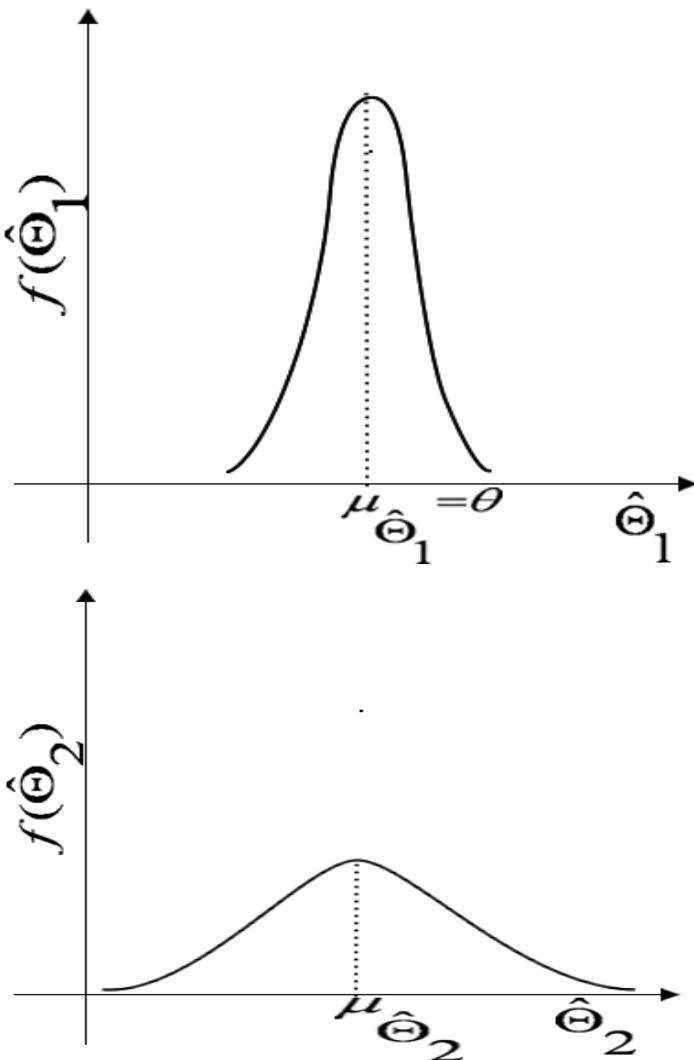


Figura 2.3. Comparação entre estimadores não-viesados de um parâmetro  $\theta$

Neste caso, o estimador  $\hat{\Theta}_1$  é melhor, mais preciso ou mais eficiente do que  $\hat{\Theta}_2$ , pois a dispersão dos erros de estimação que podem ser cometidos é menor quando se recorre ao primeiro estimador.

Compare agora a eficiência de estimadores admitindo que estes podem ser viesados, como ocorre com dois estimadores cujas funções densidades de probabilidade são representadas na figura a seguir.

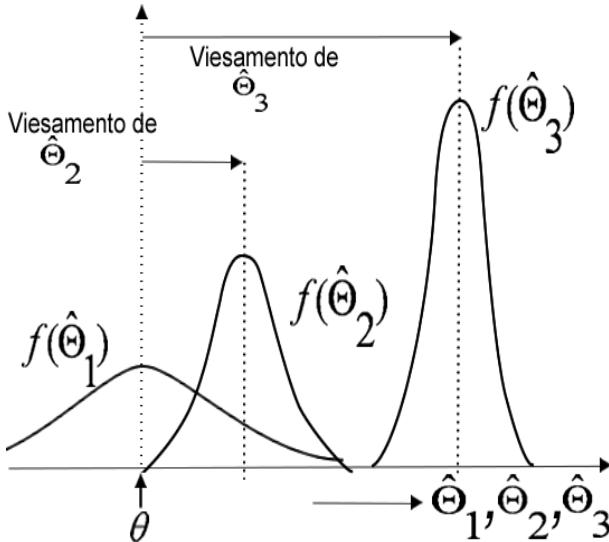


Figura 2.4. Comparação entre estimadores viesados e não-viesados de um parâmetro  $\theta$ .

Dentre os três estimadores representados, aquele que tem menor variância ( $\hat{\Theta}_3$ ) parece estar longe de ser o melhor, dado o seu significativo viesamento. O melhor também não parece ser o que apresenta viesamento nulo ( $\hat{\Theta}_1$ ), pois tem uma variância muito elevada. Restará possivelmente  $\hat{\Theta}_2$  como o melhor compromisso para a estimativa desejada.

Este caso sugere a adoção de um critério de eficiência que tome em conta não apenas a variância de cada estimador (que mede a sua dispersão em torno do seu valor esperado), mas também uma medida de dispersão do estimador em torno do parâmetro estimado. A eficiência de um estimador  $\hat{\Theta}$  (que reflete a sua precisão potencial) pode ser medida por meio do erro quadrático médio, definido pela expressão seguinte:

$$\begin{aligned} E\left[\left(\hat{\Theta}-\theta\right)^2\right] &= E\left\{\left[\left(\hat{\Theta}-\mu_{\hat{\Theta}}\right)-\left(\theta-\mu_{\hat{\Theta}}\right)\right]^2\right\} \\ &= E\left[\left(\hat{\Theta}-\mu_{\hat{\Theta}}\right)^2\right]+\left(\theta-\mu_{\hat{\Theta}}\right)^2-2\left(\theta-\mu_{\hat{\Theta}}\right)E\left(\hat{\Theta}-\mu_{\hat{\Theta}}\right). \end{aligned}$$

Dado que  $E\left(\hat{\Theta}-\mu_{\hat{\Theta}}\right)=0$ , que o primeiro termo do segundo membro representa a variância do estimador, e que o segundo termo é igual ao quadrado do viesamento, então, a eficiência do estimador  $\hat{\Theta}$  é dada por:

$$Eficiencia_{\hat{\Theta}} = E\left[\left(\hat{\Theta}-\theta\right)^2\right] = \sigma_{\hat{\Theta}}^2 + \left(Viesamento_{\hat{\Theta}}\right)^2.$$

A comparação entre a eficiência de diferentes estimadores é feita com base no conceito de eficiência relativa. Para dois estimadores quaisquer  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$  de um mesmo parâmetro  $\theta$ , a eficiência do primeiro relativamente ao segundo é dada pela razão:

$$Eficácia Relativa \hat{\Theta}_1/\hat{\Theta}_2 = \frac{E[(\hat{\Theta}_1 - \theta)^2]}{E[(\hat{\Theta}_2 - \theta)^2]}.$$

Para um determinado parâmetro, se existir algum estimador que seja mais eficiente do que qualquer outro, então, este estimador é dito ser absolutamente eficiente.

c) Consistência

Seja  $\hat{\Theta}$  um estimador do parâmetro  $\theta$  e  $N$  a dimensão da amostra com base na qual  $\hat{\Theta}$  é calculado. Este estimador é dito ser consistente quando, para qualquer valor positivo  $\delta$ , se verifica a condição seguinte:

$$\lim_{N \rightarrow \infty} Probabilidade(|\hat{\Theta} - \theta| < \delta) = 1.$$

Esta condição – dita de convergência em probabilidade do estimador para o parâmetro – significa que, quando a dimensão da amostra tende para infinito, o estimador consistente se concentra sobre o seu alvo, isto é, toma o valor do parâmetro estimado. A figura a seguir representa esta condição.

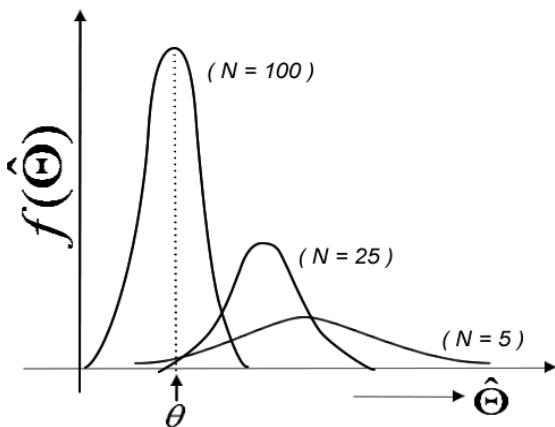


Figura 2.5. Funções densidade de probabilidade de um estimador consistente, para diferentes dimensões da amostra.

Pode ser demonstrado que, se o viesamento e a variância de um estimador tender para zero – e, portanto, tender para zero o valor esperado de  $(\hat{\Theta} - \theta)^2$ , quando a dimensão da amostra tender para infinito, então o estimador será consistente. Em notação simbólica, estas condições suficientes para a consistência de um estimador podem ser expressas nos termos seguintes:

$$\lim_{N \rightarrow \infty} (\mu_{\hat{\Theta}} - \theta) = 0$$

$$\lim_{N \rightarrow \infty} \sigma_{\hat{\Theta}}^2 = 0 \text{ ou}$$

$$\lim_{N \rightarrow \infty} E[(\hat{\Theta} - \theta)^2] = 0.$$

No âmbito da estimação pontual, a suficiência de uma estatística (estimador) traduz a capacidade que ela tem de condensar toda a informação que, relativamente ao parâmetro estimado, esteja contida no conjunto das observações que integram a amostra. Expressando de outra forma, uma amostra (constituída por N observações) não contém mais informação relativamente ao parâmetro estimado do que um estimador suficiente calculado a partir dela.

O não-viesamento e a eficiência de certos estimadores dependem criticamente da validade de certas hipóteses formuladas sobre a forma das distribuições populacionais. Existem, no entanto, estimadores caracterizados por uma considerável **robustez**, propriedade que se traduz pelo fato de tais estimadores se manterem aproximadamente não-viesados e eficientes para uma vasta gama potencial de distribuições populacionais.

## Métodos de estimação

Uma vez descritas algumas propriedades desejáveis dos estimadores, a questão que se coloca em seguida é a de saber como os definir, ou seja, como os estimar. Não existe um método geral e único que permita especificar estimadores ideais em todas as circunstâncias. Em seguida serão descritos alguns métodos alternativos de estimação pontual.

### Método da máxima verossimilhança

Com base na ilustração apresentada a seguir será formulada a idéia-base do método. Considere que, a partir de uma amostra obtida aleatoriamente com base numa população normal, se pretendem estimar os parâmetros desta população. Na figura a seguir estão representadas as observações que constituem a amostra (no eixo horizontal inferior) e três funções densidade de probabilidade normais,

com

valores

esperados

variâncias

diferentes.

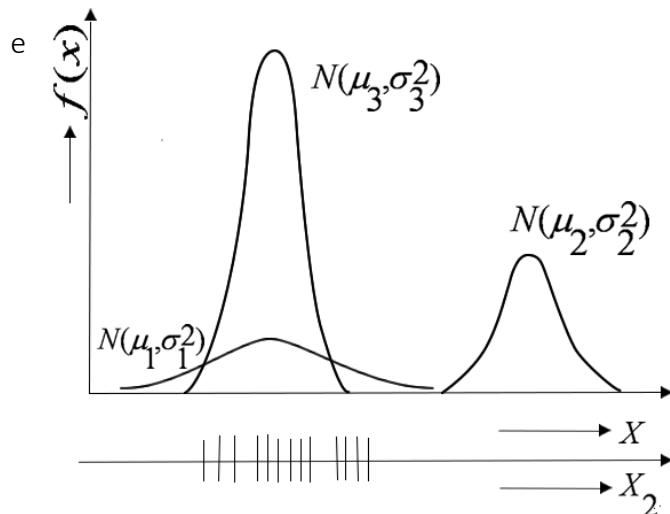


Figura 2.6. Idéia-base do método de máxima verossimilhança.

Dado que as caudas das distribuições normais se estendem indefinidamente, a amostra considerada poderia provir de qualquer uma das distribuições representadas. No entanto, será muito mais verossímil (provável, semelhante à verdade) que a amostra tenha sido obtida a partir da distribuição  $N(\mu_1, \sigma_1^2)$  do que das duas restantes. De fato, a amostra está claramente descentrada em relação à distribuição  $N(\mu_2, \sigma_2^2)$  e tem uma dispersão muito superior à da distribuição  $N(\mu_3, \sigma_3^2)$ .

Dos três pares de parâmetros, o primeiro parece ser o mais adequado para representar a população da qual a amostra foi obtida. A idéia-base do método de máxima verossimilhança consiste em selecionar, dentre todos os valores possíveis dos parâmetros populacionais, aqueles que tornem mais verossímil (provável) a ocorrência de uma amostra idêntica àquela que efetivamente se obteve.

Considerando a idéia-base do método de máxima verossimilhança, este pode ser formalizado considerando inicialmente o caso de distribuições populacionais discretas. Admita uma população discreta representada pela variável aleatória  $Y$  cuja função de probabilidade  $p(y)$  é conhecida para  $R$  parâmetros,  $\theta_1, \theta_2, \dots, \theta_R$  e denote por  $(y_1, y_2, \dots, y_N)$  uma amostra aleatória simples obtida a partir daquela população. A probabilidade de ocorrência de uma amostra aleatória simples idêntica àquela que efetivamente se obteve a partir da população é dada por:

$$L = \text{Probabilidade}(Y_1 = y_1, y_2, \dots, y_N | \theta_1, \theta_2, \dots, \theta_R)$$

$$L = \prod_{n=1}^N \text{Probabilidade}(Y_n = y_n | \theta_1, \theta_2, \dots, \theta_R).$$

Esta probabilidade, que é função dos parâmetros  $\theta_1, \theta_2, \dots, \theta_R$ , é designada por função de verossimilhança.

Note que, quanto maior for o valor da função, mais verossímil (provável) se tornará a ocorrência de uma amostra idêntica à  $(y_1, y_2, \dots, y_N)$ .

As estimativas de máxima verossimilhança (estimativas MV ou ML) dos parâmetros  $\theta_1, \theta_2, \dots, \theta_R$  são os valores destes parâmetros que maximizam a função de verossimilhança. Dito de outra forma, as estimativas MV são os valores dos parâmetros que tornam máxima a probabilidade de ocorrência de uma amostra idêntica àquela que efetivamente ocorreu. Em notação simbólica, as estimativas MV são obtidas por meio de:

$$\underset{\theta_1, \theta_2, \dots, \theta_R}{\text{Max}} : L(\theta_1, \theta_2, \dots, \theta_R) = \prod_{n=1}^N \text{Prob}(Y_n = y_n | \theta_1, \theta_2, \dots, \theta_R)$$

Admitindo que a função de verossimilhança possa ser diferenciada, as estimativas MV podem ser obtidas resolvendo o sistema de equações definido por:

$$\frac{\partial L}{\partial \theta_i} = 0 \quad (i = 1, 2, \dots, R).$$

Geralmente, as soluções assim obtidas (que correspondem a pontos estacionários das funções de verossimilhança) são máximos destas funções.

Na estimativa de parâmetros de distribuições contínuas, não faz sentido tentar maximizar a probabilidade de ocorrência de uma amostra idêntica àquela que efetivamente ocorreu, pois tal probabilidade é sempre nula. Nesta situação, se procura maximizar a densidade de probabilidade definida para os valores observados da variável, atuando sobre os valores dos parâmetros. Em outras palavras, se procura maximizar a função de verossimilhança que, para uma variável contínua X, com uma distribuição conhecida a menos de R parâmetros,  $\theta_1, \theta_2, \dots, \theta_R$ , e para uma amostra  $(x_1, x_2, \dots, x_n)$  é definida por:

$$L(\theta_1, \theta_2, \dots, \theta_R) = f_{x_1, x_2, \dots, x_N | \theta_1, \theta_2, \dots, \theta_R}(x_1, x_2, \dots, x_N)$$

$$= \prod_{n=1}^N f(x_n | \theta_1, \theta_2, \dots, \theta_R)$$

### Método dos quadrados mínimos

Considere uma situação, na qual diferentes repetições de uma variável X podem ser expressas sob a forma seguinte:

$X_n = f_n(\theta_1, \theta_2, \dots, \theta_R) + E_n$ , em que  $f_n(\theta_1, \theta_2, \dots, \theta_R)$  ( $n=1, 2, \dots, N$ ) representam funções conhecidas de R parâmetros  $\theta_1, \theta_2, \dots, \theta_R$  e  $E_n$  ( $n=1, 2, \dots, N$ ) representam repetições independentes de uma variável aleatória E (geralmente designada por erro) com valor esperado nulo ( $\mu_E = 0$ ) e variância  $\sigma_E^2$ .

Admita que se disponha de um conjunto de  $N$  ( $N > R$ ) observações de X,  $(X_1, X_2, \dots, X_N)$ . Os estimadores dos parâmetros  $\theta_1, \theta_2, \dots, \theta_R$  obtidos segundo o método dos quadrados mínimos (estimadores QM ou OLS) são os valores destes parâmetros que minimizam o somatório dos quadrados dos erros (ou seja, das diferenças entre os valores  $X_n$  e os valores das funções  $f_n(\theta_1, \theta_2, \dots, \theta_R)$ ). Em notação simbólica, os estimadores  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_R$  são obtidos recorrendo

a:  $\text{Min}_{\theta_1, \theta_2, \dots, \theta_R} : \text{SQE}(\theta_1, \theta_2, \dots, \theta_R) = \sum_{n=1}^N E_n^2 = \sum_{n=1}^N [X_n - f_n(\theta_1, \theta_2, \dots, \theta_R)]^2$

Admitindo que a função  $\text{SQE}(\theta_1, \theta_2, \dots, \theta_R)$  possa ser diferenciada, os estimadores ou as estimativas de QM podem ser obtidos resolvendo o sistema de equações definido por:

$$\frac{\partial \text{SQE}}{\partial \theta_i} = 0 \quad (i=1, 2, \dots, R).$$

Geralmente, as soluções assim obtidas (que correspondem a pontos estacionários dos somatórios dos erros quadráticos) são pontos de mínimos destas funções.

### Estimação de parâmetros para modelos lineares mistos

Considere o modelo misto  $Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$  para o DBC, com tratamento ( $i$ ) como fator aleatório e bloco ( $j$ ) como fator fixo. Nesse modelo é assumido que  $\tau_i \sim NID(0, \sigma_t^2)$ ,  $\varepsilon_{ij} \sim NID(0, \sigma_e^2)$  e que  $\tau_i$  e  $\beta_j$  são independentes.

O vetor observação (resposta) para o modelo misto anterior pode ser escrito como:

$$Y_i = X_i B + Z_i u_i + e_i, \text{ em que:}$$

$X_i$ : matriz delineamento, que pode variar ao longo dos tratamentos;

$B$ : vetor de parâmetros comum para os efeitos fixos;

$Z_i$ : matriz delineamento para os efeitos aleatórios;

$u_i$ : efeitos aleatórios para o  $i$ -ésimo tratamento;

$e_i$ : vetor de resíduos para o  $i$ -ésimo tratamento.

Para três blocos ( $j = 1, 2, 3$ ), os dados de um determinado tratamento ( $i$ ) podem ser coletados no

$$\text{vetor } Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix}.$$

As diferentes medições dentro de um tratamento podem ser correlacionadas, mas é assumido que os diferentes tratamentos são independentes. Denotando a matriz covariância de  $u_i$  por  $B_i$  e a de  $e_i$  por  $E_i$ , segue que a matriz covariância de  $y_i$  é:

$$\begin{aligned} Var(y_i) &= \Sigma_i = Var(Z_i u_i) + Var(e_i) \\ &= Z_i B_i Z_i' + E_i, \text{ sendo assumido que } E_i = \sigma^2 I. \end{aligned}$$

### Estimação de máxima verossimilhança (ML)

Os estimadores de máxima verossimilhança dos parâmetros da equação de modelos mistos anterior podem ser obtidos pela maximização da função verossimilhança:

$$l = -\frac{1}{2} \left[ \sum_i \log |\Sigma_i| + \sum_i (\mathbf{y}_i - \mathbf{X}_i \mathbf{B})' \Sigma_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{B}) \right], \text{ o que fornece:}$$

$\hat{\mathbf{B}} = \left( \sum_i \mathbf{X}_i' \hat{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \left( \sum_i \mathbf{X}_i' \Sigma_i^{-1} \mathbf{y}_i \right)$ , onde  $\hat{\Sigma}_i$  é o estimador ML de  $\Sigma_i$ . Essa solução é obtida usando métodos iterativos, tais como o algoritmo Newton-Raphson.

Em muitos casos é interessante modelar  $\Sigma_i$  como uma função de um menor conjunto de parâmetros:  $\Sigma_i = f(\phi)$ . Isto pode ser usado para estabelecer uma estrutura mais simples para as covariâncias, o que frequentemente pode fornecer convergência mais rápida e interpretação mais simples.

### Estimação de máxima verossimilhança restrita (REML)

É sabido que os estimadores de variâncias ML são viesados. Por exemplo, o estimador ML de variância baseado numa amostra de tamanho  $n$  de uma distribuição normal é

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_i (\mathbf{y}_i - \bar{\mathbf{y}})^2. \text{ Entretanto, o estimador não-viesado de } \sigma^2 \text{ é } \hat{\sigma}^2 = \frac{1}{n-1} \sum_i (\mathbf{y}_i - \bar{\mathbf{y}})^2.$$

Assim,  $E(\tilde{\sigma}^2) = \frac{n-1}{n} \sigma^2$ , então o estimador ML é viesado em amostras finitas. Este fato levou ao

desenvolvimento do método ML Restrito (ou Residual), onde são obtidos estimadores de variâncias não-viesados. A ideia básica do método é maximizar a verossimilhança de  $L_{\mathbf{y}_i}$ , onde  $L$  é escolhido de tal forma que apenas a parte da verossimilhança que é invariante em relação à  $\mathbf{B}$  é maximizada.

As propriedades dos estimadores REML são similares às dos estimadores ML: normalidade assintótica e consistência. As estimativas de erros padrões dos parâmetros estimados são obtidas por meio da matriz de derivadas de segunda ordem da função verossimilhança, a chamada informação de Fisher. O REML é o método de estimação padrão no procedimento Mixed do SAS.

### Estimação de quadrados mínimos generalizados (GLS)

Uma alternativa ao ML e REML é basear a estimação nos Quadrados Mínimos Generalizados. Isto significa que são usados como estimadores aqueles valores dos parâmetros que minimizam

$$(\mathbf{y} - \mathbf{XB})' V^{-1} (\mathbf{y} - \mathbf{XB}).$$

A matriz  $V$  é assumida ser conhecida. Se isto não ocorre, pode ser inserida alguma estimativa da matriz covariância para obter os estimadores dos parâmetros fixos.

## Padrões de correlações

Uma parte importante da modelagem de dados usando modelos mistos é a escolha de uma estrutura realística para as matrizes covariância  $\Sigma_i$ . Embora, em princípio possa ser permitido que essas matrizes sejam livres (não-estruturadas), frequentemente é vantajoso impor uma estrutura mais limitada.

Existem várias razões para esta imposição:

- (a) Se as matrizes covariância são não-estruturadas o modelo incluirá um grande número de parâmetros. Especialmente se o tamanho da amostra é pequeno, isto pode causar problemas na convergência do procedimento iterativo.
- (b) Mesmo quando é tecnicamente possível deixar que as matrizes covariância sejam livres, um modelo mais simples com poucos parâmetros deve ser preferido, se a perda de informação não é muito grande.
- (c) Critérios tais como o Critério de Informação de Akaike (AIC) podem ser usados para avaliar se a simplificação do modelo deve ser adotada.

Alguns exemplos de estruturas de covariâncias, úteis para modelos mistos, que estão disponíveis, por exemplo, no procedimento Mixed do SAS são fornecidas a seguir:

$$\text{Simetria Composta (CS): } \Sigma = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix}$$

$$\text{Não-Estruturada (UN): } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

$$\text{Bandeada [UN(1)]: } \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

Autocorrelação de Primeira Ordem [AR(1)]:  $\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$

Potência Espacial [SP(POW)(c)]:  $\Sigma = \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}$

## Algumas aplicações do método de modelos mistos

### Experimentos multiambientes

Em experimentos multiambientes, os ambientes estudados podem às vezes ser vistos como selecionados aleatoriamente de uma população de ambientes. Então, nesse caso, um modelo para os dados deve incluir todos os efeitos fixos dentro de ambiente e um efeito aleatório de ambiente. Uma vez que a interação entre ambiente e tratamentos pode ser de interesse, um termo interação deve ser incluído no modelo.

### Experimentos transversais (Cross Designs)

Em experimentos transversais (ou cruzados), o mesmo indivíduo (parcela experimental) é exposto a vários tratamentos. Assim, existem pelo menos duas fontes de variação aleatória: entre indivíduos e dentro de indivíduos. Isto pode ser levado em conta considerando o indivíduo como um fator aleatório. As matrizes covariância,  $\Sigma_i$ , são frequentemente deixadas livres nesse tipo de aplicação, embora seja possível considerar a modelagem das covariâncias entre medições ao longo do tempo, numa forma similar a dados de medidas repetidas.

### Medidas repetidas

Experimentos em que medições tenham sido tomadas em várias ocasiões sobre os mesmos indivíduos (ou parcelas) são chamados de experimentos com medidas repetidas (ou com dados longitudinais). Este tipo de experimento também tem no mínimo duas fontes de variação aleatória: entre indivíduos e dentro de indivíduos. Assim, a análise de medidas repetidas pode ser realizada com base na metodologia de modelos mistos.

### Outras considerações sobre modelos mistos

A utilização de modelos mistos para estimação de parâmetros tem várias justificativas:

- (a) Melhora o entendimento de mecanismos genéticos;

- (b) Necessária para predição de valores genéticos (índice de seleção/BLUP);
- (c) Necessária para otimização de programas de melhoramento e predição de respostas;
- (d) Necessária para estimação de componentes de variância com dados desbalanceados ou com parcelas perdidas;
- (e) Estimação de componentes de variância para caracteres novos;
- (f) Reestimarão de componentes quando as variâncias (medida do grau de diferenças) ou covariâncias (medida de diferenças em comum) mudam ao longo do tempo devido a mudanças genéticas e/ou ambientais (seleção, melhoria de condições, redefinição de caracteres).

Os métodos para estimação de componentes de variância podem ser classificados da seguinte forma:

- (1) ANOVA – dados balanceados
- (2) ANOVA – dados desbalanceados (Métodos de Henderson, implementados no SAS)
- (3) Métodos de Verossimilhança (Máxima Verossimilhança – ML e Máxima Verossimilhança Restrita – REML)
- (4) Métodos Bayesianos (Amostragem Gibbs)

### **ANOVA para dados desbalanceados**

Tem como base a mesma ideia dos dados balanceados, mas usa um número ponderado para “ $n$ ” em:  $E(QM_A) = \sigma_e^2 + n\sigma_a^2$ .

Necessita de notação matricial para obter SQ e E (QM). Existem métodos padrão em programas computacionais tais como SAS, STATISTICA e outros. O mais geral desses métodos é chamado de “Método III de Henderson”.

### **Métodos de verossimilhança**

Nesses métodos cada observação tem uma densidade de probabilidade, determinada por sua distribuição, valor esperado (parâmetros locação) e variância (parâmetros dispersão). Por exemplo, para uma observação com distribuição normal, média  $\mu$  e variância  $\sigma^2$  tem-se:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}}$$

Esta é uma função densidade de probabilidade para a observação  $y$ . Ela fornece a probabilidade de ocorrência da observação, dados os parâmetros  $\mu$  e  $\sigma^2$ . Mas, pode-se reverter este raciocínio e

fornecer a verossimilhança de parâmetros, dado a observação  $y$ , o que constitui a ideia base dos métodos de verossimilhança.

### Método da máxima verossimilhança (ML)

Os valores de probabilidade, obtidos segundo o raciocínio anterior, podem ser multiplicados sobre todas as observações (dados), e incluir ainda o fato que algumas das observações podem estar relacionadas, ou seja, tem-se uma distribuição conjunta. Dessa forma, considerando o vetor de observações  $y$  com valores esperados  $E(y) = Xb$  e variância  $Var(y) = V$ , tem-se o logaritmo da verossimilhança dado por:

$$L(b, V | X, y) = -\frac{1}{2}N \log(2\pi) - \frac{1}{2} \log(|V|) - \frac{1}{2}(y - Xb)'V^{-1}(y - Xb)$$

Essa expressão fornece a máxima verossimilhança dos parâmetros  $(b, V)$ , uma vez fornecidos os dados  $(X, y)$  no lado direito: os dois primeiros termos são esperanças e o último termo é uma soma de quadrados.

### Método da máxima verossimilhança restrita (REML)

Primeiro todos os dados são corrigidos para os efeitos fixos.

Após essas correções obtém-se a máxima verossimilhança (solução para componentes de variância).

Usualmente é utilizado um procedimento iterativo para solucionar o problema.

São necessários valores iniciais (para os parâmetros) para realizar o procedimento iterativo.

Exemplo de um algoritmo utilizado no REML (algoritmo EM – maximização da esperança):

(1) Resolve as equações de modelos mistos (MME) usando um valor inicial para os componentes de variância. O sistema de equações MME é dado por:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

(2) Resolve para os componentes de variância à partir das soluções de MME obtendo-se:

$$\sigma_a^2 = [\hat{a}'A^{-1}\hat{a} + tr(A^{-1}C)\sigma_e^2] / q \quad \sigma_e^2 = [y'y - \hat{b}'X'y - \hat{a}'Z'y] / (N - r(X))$$

(3) Usa um novo  $\lambda = \sigma_e^2 / \sigma_a^2$  e iteração novamente seguindo os passos (1) e (2). O processo é concluído quando ocorre convergência.

## Vantagens da estimação por REML

- (1) É por definição um método mais acurado (exato);
- (2) Usa equações de modelos mistos completas; então pode usar todas as relações entre plantas num modelo por planta;
- (3) Tem muitas propriedades similares ao BLUP, por exemplo, ele considera a seleção;
- (4) Admite modelos mistos mais complexos (efeitos maternos, múltiplos caracteres) como ocorre no BLUP;
- (5) Se for utilizado um modelo por planta, a herdabilidade é estimada a partir da combinação de informação entre famílias (MI/IC), de informação de regressão progenitor – progénie;
- (6) O método e o modelo são bastante flexíveis, mas pode ser difícil avaliar as estimativas baseadas nos dados e a estrutura dos dados.

## Avaliação da qualidade das estimativas de parâmetros

- (1) Acurácia

Observar os erros padrões (EP) das estimativas, embora estes sejam aproximados.

Avaliar o efeito do número de observações e da estrutura de dados (número de grupos versus número de observações por grupo).

- (2) Não-viesamento

Verificar o viesamento dos dados e os possíveis efeitos; avaliar se existiu viés da seleção ou o confundimento de efeitos.

- (3) Comparação de diferentes modelos

Pode ser utilizado o Teste de Razão de Verossimilhança, que é dado por:

$$LR = -2 \ln \frac{\text{Max\_Verossim}(\text{mod.reduzido})}{\text{Max\_Verossim}(\text{mod.completo})}, \text{ que tem distribuição qui-quadrado com } gl = \text{diferença no número de parâmetros dos modelos.}$$

## Valores Esperados, Variâncias e Covariâncias

### Valores Esperados

O valor esperado ou esperança de uma variável aleatória,  $E(Y)$  ou  $\mu_Y$  é a verdadeira média ponderada desta variável aleatória. Por exemplo, quantas caras seriam esperadas se uma moeda é lançada duas vezes?

$$Y = \text{evento Cara} = \{0, 1, 2\} \quad \begin{cases} 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{cases} \quad \begin{array}{l} P(0) = 1/4 \\ P(1) = 1/2 \\ P(2) = 1/2 \end{array}$$

$$\text{Média Ponderada} = 0(1/4) + 1(1/2) + 2(1/4) = 1$$

Seja  $Y = Y_1, Y_2, Y_3, \dots, Y_n$  e

$$P(Y_1), P(Y_2), P(Y_3), \dots, P(Y_n), \text{ então}$$

$$E(Y) = \sum_i Y_i P(Y_i) \quad \text{e} \quad E(Y) \quad \text{é} \quad \mu_Y = \left( \sum_i Y_i \right) / N \quad \text{quando}$$

$$P(Y_1) = P(Y_2) = P(Y_3) = \dots = P(Y_N) = \frac{1}{N}, \text{ sendo } N : \text{número total de indivíduos na população.}$$

Esperança como uma função de uma variável aleatória:

Seja  $Y = Y_1, Y_2, Y_3, \dots, Y_n$  e

$$P(Y_1), P(Y_2), P(Y_3), \dots, P(Y_n), \quad \text{então} \quad \text{a função} \quad g(Y) \quad \text{tem}$$

$$E[g(Y)] = \sum_i g(Y_i) P(Y_i)$$

Exemplo 1

$$Z = aY + b$$

$$E(Z) = E(aY + b)$$

$$= \sum_i (aY_i + b) P(Y_i)$$

$$= a \sum_i Y_i P(Y_i) + b \sum_i P(Y_i)$$

$$= aE(Y) + b$$

Exemplo 2

Considere duas variáveis aleatórias  $bX$  e  $cY$ , então  $E(bX + cY) = bE(X) + cE(Y)$

Variâncias

$$V(Y) = E[Y - E(Y)]^2$$

$$= E\{Y^2 - 2YE(Y) + [E(Y)]^2\}$$

$$= E(Y^2) - 2E(Y)E(Y) + [E(Y)]^2$$

$$\begin{aligned}
&= E(Y^2) - 2[E(Y)]^2 + [E(Y)]^2 \\
V(Y) &= E(Y^2) - [E(Y)]^2 \\
\text{Exemplo 1} \\
Z &= aY + b \\
V(Z) &= V(aY + b) \\
&= E\left\{(aY + b)^2 - [E(aY + b)]^2\right\} \\
&= E\left\{a^2Y^2 + 2abY + b^2 - [aE(Y) + b]^2\right\} \\
&= a^2E(Y^2) + 2abE(Y) + b^2 - a^2[E(Y)]^2 - 2abE(Y) - b^2 \\
&= a^2\left\{E(Y^2) - [E(Y)]^2\right\} \\
V(Z) &= a^2V(Y)
\end{aligned}$$

Transformação Linear de Variâncias:

$$\begin{aligned}
V(Y+b) &= V(Y) \\
V(bY) &= b^2V(Y) \\
V(bY+c) &= b^2V(Y) \\
V(X+Y) &= E(X+Y)^2 - [E(X+Y)]^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - [E(X)]^2 \\
&\quad - 2E(X)E(Y) - [E(Y)]^2 \\
&= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 \\
&\quad + 2[E(XY) - E(X)E(Y)] \\
V(X+Y) &= V(X) + V(Y) + 2\text{Cov}(X, Y) \\
V(bX+cY) &= b^2V(X) + c^2V(Y) + 2bc\text{Cov}(X, Y)
\end{aligned}$$

### Covariância e Correlação

A covariância entre duas variáveis aleatórias  $X$  e  $Y$  é definida como

$$\text{Cov}(X, Y) = E\left\{(X - E(X))(Y - E(Y))\right\}$$

Da mesma forma que foi feito para variância, pode-se reescrever esta equação,

$$\begin{aligned}
\text{Cov}(X, Y) &= E\{(X - E(X))(Y - E(Y))\} \\
&= E\{[XY - E(X)Y - XE(Y) + E(X)E(Y)]\} \\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y) = \text{Cov}(X, Y)
\end{aligned}$$

Observe que, se  $X$  e  $Y$  são independentes, então

$$\begin{aligned}
\text{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\
&= E(X)E(Y) - E(X)E(Y) = 0
\end{aligned}$$

Entretanto, o contrário não é verdadeiro. Considere o seguinte exemplo contrário: Define-se  $X$  e  $Y$  de forma que

$$P(X=0)=P(X=1)=P(X=-1)=1/3$$

$$Y = \begin{cases} 0, & \text{se } X \neq 0 \\ 1, & \text{se } X = 0 \end{cases}, \text{ então } X \text{ e } Y \text{ são dependentes.}$$

Entretanto, tem-se que  $XY=0$  e

$$E(XY)=E(X)=0, \text{ então}$$

$$\text{Cov}(X, Y)=E(XY)-E(X)E(Y)=0$$

Propriedades de Covariância:

$$1) \text{Cov}(X, Y)=\text{Cov}(Y, X)$$

$$2) \text{Cov}(X, X)=\text{Var}(X)$$

$$3) \text{Cov}(aX, Y)=a\text{Cov}(X, Y)$$

$$4) \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right)=\sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$$

$$\text{A correlação entre duas variáveis aleatórias } X \text{ e } Y \text{ como } \rho(X, Y)=\frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Pode ser mostrado que  $-1 \leq \rho(X, Y) \leq 1$ , somente se  $Y=a+bX$  e assumindo que  $E(X^2)$  e  $E(Y^2)$  são ambas finitas. Esta é a chamada desigualdade de Cauchy-Schwarz.

Portanto, o coeficiente de correlação é uma medida do grau de linearidade entre  $X$  e  $Y$ . Se  $\rho(X, Y)=0$ , então isto indica não linearidade e diz-se que  $X$  e  $Y$  são não correlacionados.

Transformação Linear de Covariâncias e Correlações

$$\text{Cov}(Y, Y)=\text{Var}(Y)$$

$$\text{Corr}(Y, Y)=1$$

$$\text{Cov}(bX, cY)=bc\text{Cov}(X, Y)$$

$$\text{Corr}(bX, cY) = \text{Corr}(X, Y)$$

$$\text{Cov}(bX + d, cY + e) = bc\text{Cov}(X, Y)$$

$$\text{Corr}(bX + d, cY + e) = \text{Corr}(X, Y)$$

$$\begin{aligned} \text{Cov}(X + Y, W + Z) &= \text{Cov}(X, W) + \text{Cov}(X, Z) + \\ &+ \text{Cov}(Y, W) + \text{Cov}(Y, Z) \end{aligned}$$

### Esperança Condisional

Tem-se que, se  $X$  e  $Y$  são variáveis aleatórias discretas, a magnitude a função de probabilidade condicional de  $X$ , dado  $Y = y$ , para todo  $y$  tal que  $P(Y = y) > 0$ , é definido por:

$$p_{X/Y}(x/y) = P(X = x/Y = y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

Definição 1: Se  $X$  e  $Y$  são variáveis aleatórias discretas, a esperança condicional de  $X$ , dado  $Y = y$ , para todo  $y$  tal que  $P(Y = y) > 0$ , é definida por:

$$E(X/Y = y) = \sum_x x P(X = x/Y = y) = \sum_x x p_{X/Y}(x/y)$$

Sabe-se que, se  $X$  e  $Y$  são variáveis aleatórias contínuas, a função densidade de probabilidade (fdp) condicional de  $X$ , dado  $Y = y$ , para todo  $y$  tal que  $f_Y(y) > 0$ , é definido por:

$$f_{X/Y}(x/y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Definição 2: Se  $X$  e  $Y$  são variáveis aleatórias contínuas, a esperança condicional de  $X$ , dado  $Y = y$ , para todo  $y$  tal que  $f_Y(y) > 0$ , é definido por:

$$E(X/Y = y) = \int_{-\infty}^{\infty} x P(X = x/Y = y) = \int_{-\infty}^{\infty} x f_{X/Y}(x/y) dx$$

As esperanças condicionais são em si variáveis aleatórias. A esperança condicional de  $X$  dado  $Y = y$ , é exatamente o valor esperado em um espaço amostral reduzido que consiste apenas de respostas onde  $Y = y$ .

É importante observar que as esperanças condicionais satisfazem todas as propriedades das esperanças regulares, ou seja:

$$1. E[g(X)/Y = y] = \sum_x g(x) p_{X/Y}(x/y)$$

$$2. E[g(X)/Y = y] = \int_{-\infty}^{\infty} g(x) f_{X/Y}(x/y) dx$$

# CAPITULO 3

## Álgebra de Matrizes

### Conceitos básicos de álgebra de matrizes

Matriz é uma ordenação retangular de elementos arranjados em linhas e colunas. Um exemplo de matriz é dado por

$$\begin{bmatrix} 1285 & 42 \\ 2930 & 65 \\ 1846 & 84 \end{bmatrix}$$

Outros exemplos de matrizes são

$$\begin{bmatrix} 2 & 0 \\ 6 & 12 \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} 5 & 8 & 13 & 20 \\ 3 & 7 & 15 & 8 \end{bmatrix}$$

Podem-se utilizar símbolos para identificar os elementos de uma matriz:  $j=1 \quad j=2 \quad j=3$ .

$$i=1 \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

A matriz pode ser representada por um símbolo tal como  $H$ ,  $X$ ,  $\beta$ . Por exemplo,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Outra notação da matriz A é  $A = [a_{ij}] \quad i=1, 2; \quad j=1, 2, 3$

Resumindo, uma matriz com r linhas e c colunas será amplamente representada por

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2c} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{ic} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{r1} & a_{r2} & \dots & a_{rj} & \dots & a_{rc} \end{bmatrix}$$

ou na forma abreviada como

$$A = \begin{bmatrix} a_{ij} \end{bmatrix} \quad i=1, \dots, r; \quad j=1, \dots, c$$

### Matriz Quadrada

Dois exemplos são:

$$\begin{bmatrix} 5 & 3 \\ 7 & 9 \end{bmatrix} \text{ e } \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

### Vetor

Dois exemplos são:

$$A = \begin{bmatrix} 5 \\ 7 \\ 11 \end{bmatrix} \quad \text{e} \quad C = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{bmatrix}$$

### Transposta

Por exemplo, se

$$A(3 \times 2) = \begin{bmatrix} 2 & 5 \\ 7 & 12 \\ 3 & 4 \end{bmatrix}, \text{ então a transposta } A' \text{ é dada por}$$

$$A'(2 \times 3) = \begin{bmatrix} 2 & 7 & 3 \\ 5 & 12 & 4 \end{bmatrix}$$

Como outro exemplo, considere:

$$C = \begin{bmatrix} 4 \\ 3 \\ 7 \end{bmatrix} \quad C' = [4 \ 3 \ 7]$$

Em geral, têm-se:

$$A(r \times c) = \begin{bmatrix} a_{11} & \dots & a_{1c} \\ a_{r1} & \dots & a_{rc} \end{bmatrix} = \begin{bmatrix} a_{ij} \end{bmatrix} i=1, \dots, r \quad j=1, \dots, c$$

$$A'(c \times r) = \begin{bmatrix} a_{11} & \dots & a_{r1} \\ a_{1c} & \dots & a_{rc} \end{bmatrix} = \begin{bmatrix} a_{ij} \end{bmatrix} j=1, \dots, c \quad i=1, \dots, r$$

A operação transposta é reflexiva, ou seja, a transposta de uma matriz transposta é a própria matriz original. Tem-se então que  $(A')'=A$ . Uma matriz quadrada é simétrica quando ela é igual à sua transposta, isto é,  $A$  é simétrica quando  $A=A'$ .

### Igualdade de Matrizes

Por exemplo, se

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad B = \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}, \text{ então } A=B \text{ implica que } a_1=4, a_2=7 \text{ e } a_3=3$$

### Adição e Subtração de Matrizes

Suponhamos,

$$A(3 \times 2) = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \text{ e } B(3 \times 2) = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}, \text{ então}$$

$$A+B(3 \times 2) = \begin{bmatrix} 1+1 & 4+2 \\ 2+2 & 5+3 \\ 3+3 & 6+4 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 4 & 8 \\ 6 & 10 \end{bmatrix}$$

Similarmente, tem-se que

$$A-B(3 \times 2) = \begin{bmatrix} 1-1 & 4-2 \\ 2-2 & 5-3 \\ 3-3 & 6-4 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 2 \\ 0 & 2 \end{bmatrix}$$

Em geral, se

$$A(r \times c) = \begin{bmatrix} a_{ij} \end{bmatrix} \text{ e } B(r \times c) = \begin{bmatrix} b_{ij} \end{bmatrix}, i=1, \dots, r; j=1, \dots, c$$

### Multiplicação de Matrizes

a) Multiplicação de uma matriz por um escalar

Por exemplo, suponha a matriz  $A$ , dada por  $A = \begin{bmatrix} 2 & 7 \\ 5 & 3 \end{bmatrix}$ , então  $5A$

$$\text{é igual a } 5A = 5 \begin{bmatrix} 2 & 7 \\ 5 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 35 \\ 25 & 15 \end{bmatrix}.$$

Similarmente, tem-se que  $\beta A = \beta \begin{bmatrix} 2 & 7 \\ 5 & 3 \end{bmatrix} = \begin{bmatrix} 2\beta & 7\beta \\ 5\beta & 3\beta \end{bmatrix}$ , em que representa um escalar.

Se todos os elementos de uma matriz têm um fator comum, este fator pode ser retirado para fora da matriz e tratado como um escalar. Por exemplo,

$$\begin{bmatrix} 6 & 36 \\ 12 & 18 \end{bmatrix} = 3 \begin{bmatrix} 2 & 12 \\ 4 & 6 \end{bmatrix}.$$

Em geral, se  $A = \begin{bmatrix} a_{ij} \end{bmatrix}$  e  $\lambda$  é um escalar, então  $\lambda A = A \lambda = \begin{bmatrix} \lambda a_{ij} \end{bmatrix}$ .

b) Multiplicação de uma matriz por outra matriz

Consideremos as matrizes  $A (2 \times 2) = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix}$  e  $B (2 \times 2) = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix}$ ,

o produto  $AB$  será uma matriz  $2 \times 2$  cujos elementos são obtidos como a seguir:

$$A = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix} \quad AB = \begin{bmatrix} 33 & ? \\ ? & ? \end{bmatrix}$$

Toma-se o produto cruzado e soma-se  $(2)(4) + (5)(5) = 33$

$$A = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix} \quad AB = \begin{bmatrix} 33 & 52 \\ ? & ? \end{bmatrix}$$

Os produtos cruzados somados são:  $(2)(6) + (5)(8) = 52$

Continuando o processo, encontramos o produto  $AB$  como sendo:

$$AXB (2 \times 2) = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix} = \begin{bmatrix} 33 & 52 \\ 21 & 32 \end{bmatrix}$$

Consideremos outro exemplo:

$$A (2 \times 3) = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \quad e \quad B (3 \times 1) = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$$

$$AXB (2 \times 1) = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 26 \\ 41 \end{bmatrix}$$

Quando se obtém o produto  $AB$ , diz-se que  $A$  está pós-multiplicada por  $B$ , ou que  $B$  está pré-multiplicada por  $A$ .

Em geral, o produto  $AB$  está definido apenas quando o número de colunas em  $A$  é igual ao número de linhas em  $B$ , de forma que haverá termos correspondentes nos produtos cruzados. Além disso, se  $A$  tem dimensão  $r \times c$  e  $B$  tem dimensão  $c \times s$ , o produto  $AB$  é uma matriz de dimensão  $r \times s$ , cujo elemento na  $i$ -ésima linha e  $j$ -ésima coluna é dado por

$$\sum_{k=1}^c a_{ik} b_{kj}, \text{ de forma que}$$

$$AB = \left[ \sum_{k=1}^c a_{ik} b_{kj} \right] \quad i=1, \dots, r; \quad j=1, \dots, s.$$

### Tipos especiais de Matrizes

#### a) Matriz Simétrica

Se  $A = A'$ , A é dita ser simétrica. A matriz A, dada a seguir, é simétrica:

$$A = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix} \quad e \quad A' = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix}$$

Evidentemente que uma matriz simétrica necessariamente é quadrada.

#### b) Matriz Diagonal

É uma matriz quadrada cujos elementos  $a_{ij}$  fora da diagonal são todos iguais a zero, tal como:

$$A = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \quad B = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

Dois tipos importantes de matriz diagonal são a matriz identidade e a matriz escalar.

#### c) Matriz Identidade

A matriz identidade ou matriz unidade é denotada por I. É uma matriz quadrada cujos elementos sobre a diagonal principal são todos iguais a 1 e cujos elementos fora da diagonal são todos zeros. Pré-multiplicando ou pós-multiplicando qualquer matriz  $A_{rxr}$  pela matriz identidade  $I_{rxr}$  torna A não modificada. Por exemplo,

$$IA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Similarmente, tem-se:

$$AI = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Assim, a matriz identidade pode ser introduzida ou retirada de uma expressão matricial se isto for conveniente.

#### d) Matriz Escalar

É uma matriz quadrada cujos elementos da diagonal principal são uma quantidade escalar e cujos elementos fora da diagonal são zeros.

Exemplos:  $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$  e  $\begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$

Uma matriz escalar pode ser representada por  $\lambda I$ , em que  $\lambda$  é o escalar. Por exemplo,

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ e } \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Multiplicar uma matriz  $A_{rXr}$  pela matriz escalar  $\lambda I_{rXr}$  é equivalente a multiplicar a matriz A pelo escalar  $\lambda$ .

d) Vetor zero

É um vetor coluna que contém apenas zeros como elementos. Por exemplo:

$$0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

### Dependência linear e posto (ou grau) de uma matriz

a) Dependência Linear

Considere a seguinte matriz:

$$A = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix}$$

Note que o terceiro vetor coluna é um múltiplo do primeiro vetor coluna, ou seja:

$$\begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Neste caso, diz-se que as colunas de A são linearmente dependentes. Elas contêm informação redundante, uma vez que uma coluna pode ser obtida como uma combinação linear de outra.

b) Posto (ou grau) de uma Matriz

O posto de uma matriz é definido como o número máximo de colunas linearmente independentes na matriz. O posto de uma matriz é único e pode ser definido equivalentemente como o número máximo de linhas linearmente independentes. Isto implica que o posto de uma matriz  $rXc$  não pode exceder  $\min(r, c)$ , ou seja, o mínimo dos dois valores  $r$  e  $c$ .

O posto da matriz A anterior não pode ser 4, uma vez que as quatro colunas são linearmente dependentes. Podem-se, entretanto, encontrar 3 colunas (1, 2 e 4) que são linearmente independentes. Assim, o posto de A é 3.

### Inversa de uma matriz

Em álgebra comum, o inverso de um número é seu recíproco. Um número multiplicado pelo seu inverso é sempre igual a 1, ou seja:

$$(6) \frac{1}{6} = 1 \quad e \quad (x) \frac{1}{x} = x \cdot x^{-1} = x^{-1} \cdot x = 1.$$

Em álgebra matricial, o inverso de uma matriz A é outra matriz, representada por

$$A^{-1}, \text{ tal que } A^{-1} \cdot A = A \cdot A^{-1} = I, \text{ em que } I \text{ é a matriz identidade.}$$

A inversa de uma matriz é definida apenas para matrizes quadradas. Mesmo assim, muitas matrizes quadradas não têm uma inversa. Se uma matriz quadrada tem uma inversa, esta inversa é única.

Alguns exemplos de inversa são:

a) Inversa da Matriz

$$A = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \quad \text{é} \quad A^{-1} = \begin{bmatrix} -0,1 & 0,4 \\ 0,3 & -0,2 \end{bmatrix},$$

$$\text{uma vez que } A^{-1} \cdot A = \begin{bmatrix} -0,1 & 0,4 \\ 0,3 & -0,2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\text{ou ainda } A \cdot A^{-1} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} -0,1 & 0,4 \\ 0,3 & -0,2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

b) A Inversa da Matriz

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{é} \quad A^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}, \text{ uma vez que}$$

$$A^{-1} \cdot A = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

### Como encontrar a inversa e quando ela existe

A inversa de uma matriz quadrada  $r \times r$  existe se e somente se o posto da matriz for  $r$ . Tal matriz é dita ser não singular. Uma matriz  $r \times r$  com posto menor que  $r$  é dita ser singular e não tem inversa.

Os cálculos para matrizes  $2 \times 2$  e  $3 \times 3$  são como a seguir:

$$1) \text{ Se } A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \text{ então } A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} \frac{d}{D} & \frac{-b}{D} \\ \frac{-c}{D} & \frac{a}{D} \end{bmatrix}, \text{ em que } D = ad - bc \text{ é chamado de}$$

determinante da matriz A. Se A for singular, seu determinante seria igual a zero e não existiria nenhuma inversa de A.

2) Se  $B = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}$ , então

$$B^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & K \end{bmatrix}, \text{ em que:}$$

$$A = (ek - fh)/Z$$

$$B = -(bk - ch)/Z$$

$$C = (bf - ce)/Z$$

$$D = -(dk - fg)/Z$$

$$E = (ak - cg)/Z$$

$$F = -(af - cd)/Z$$

$$G = (dh - eg)/Z$$

$$H = -(ah - bg)/Z$$

$$K = (ae - bd)/Z$$

$$Z = a(ek - fh) - b(dk - fg) + c(dh - eg)$$

Z é chamado de determinante da matriz B.

Por exemplo, seja calcular a inversa da matriz  $A = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$ .

Temos que a=2; b=4; c=3; d=1;

$$D = ad - bc = (2)(1) - (4)(3) = -10$$

$$A^{-1} = \begin{bmatrix} \frac{1}{-10} & \frac{-4}{-10} \\ \frac{-3}{-10} & \frac{2}{-10} \end{bmatrix} = \begin{bmatrix} -0,1 & 0,4 \\ 0,3 & -0,2 \end{bmatrix}$$

Quando for obtida uma inversa  $A^{-1}$ , seja por cálculos manuais ou por computador, recomenda-se calcular  $A^{-1} \cdot A$  para checar se o produto é igual à matriz identidade.

### Usos da matriz inversa

Em álgebra comum, resolve-se uma equação do tipo  $5x=20$  multiplicando ambos os lados da equação pelo inverso de 5:

$$\frac{1}{5}(5x) = \frac{1}{5}(20), \text{ donde } x = \frac{1}{5}(20) = 4$$

Em álgebra matricial, quando se tem uma equação  $AY = C$ , correspondentemente, se ela for pré-multiplicada em ambos os lados por  $A^{-1}$ , assumindo que A tenha uma inversa, obtém-se  $A^{-1}AY = A^{-1}C$ , e uma vez que  $A^{-1}AY = IY$ , tem-se que  $Y = A^{-1}C$ .

Por exemplo, suponha que temos duas equações simultâneas,

$$\begin{cases} 2x + 4y = 20 \\ 3x + y = 10 \end{cases}, \text{ que podem ser escritas em notação matricial da seguinte forma:}$$

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}.$$

A solução destas equações é:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 20 \\ 10 \end{bmatrix}, \text{ donde}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0,1 & 0,4 \\ 0,3 & -0,2 \end{bmatrix} \begin{bmatrix} 20 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \text{ logo } x=2 \text{ e } y=4 \text{ satisfazem estas duas equações.}$$

### Propriedades da matriz inversa

Se A é uma matriz quadrada não singular, a sua inversa  $A^{-1}$  tem as seguintes propriedades:

a) A inversa comuta com A, sendo ambos os produtos a matriz identidade, ou seja,  $A^{-1}A = AA^{-1} = I$ .

b) A inversa de A é única.

c) O determinante da inversa de A é o recíproco do determinante de A, ou seja,

$$|A^{-1}| = \frac{1}{|A|}.$$

d) A matriz inversa é não singular.

e) A inversa de uma transposta é a transposta da inversa, isto é,  $(A')^{-1} = (A^{-1})'$ .

f) A inversa de  $A^{-1}$  é A, ou seja,  $(A^{-1})^{-1} = A$ .

g) Se A é simétrica, sua inversa também será, ou seja, se  $A' = A$  então  $(A^{-1})' = A^{-1}$ .

h) A inversa de um produto é o produto das inversas tomado em ordem inversa quando as inversas existem, ou seja, se  $A^{-1}$  e  $B^{-1}$  existem, então  $(AB)^{-1} = B^{-1}A^{-1}$ .

i) Se A é tal que sua inversa é igual sua transposta, ou seja,  $A^{-1} = A'$ , A é dita ser uma matriz ortogonal e  $AA' = I$ .

### Matrizes e vetores aleatórios

Uma matriz aleatória ou um vetor aleatório contém elementos que são variáveis aleatórias. Por exemplo, o vetor observação (ou resposta)  $Y$  é um vetor aleatório, uma vez que os  $Y_i$  elementos são variáveis aleatórias.

Suponha que se têm  $n=3$  observações que estão relacionados no vetor resposta

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

Então, define-se a esperança matemática de um vetor ou matriz aleatória como um vetor, ou matriz, cujos elementos são os valores esperados dos elementos correspondentes no vetor ou matriz original.

$$\text{Assim, } E(Y) = \begin{bmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \end{bmatrix}.$$

Em geral, para um vetor aleatório  $Y$ , a esperança é  $E(Y) = [E(Y_i)]_{i=1,2,\dots,n}$  e para uma matriz aleatória  $Y$  a esperança é  $E(Y) = [E(Y_{ij})]_{i=1,2,\dots,n; j=1,2,\dots,p}$ .

No vetor aleatório  $Y$  dado anteriormente, cada variável tem uma variância  $\sigma^2(Y_i)$  e duas variáveis aleatórias quaisquer têm uma covariância  $\sigma(Y_i, Y_{i'})$ . Pode-se colocá-las numa matriz chamada de matriz variância-covariância de  $Y$ , que é denotada por  $\sigma^2(Y)$  e tem a seguinte forma:

$$\sigma^2(Y) = \begin{bmatrix} \sigma^2(Y_1) & \sigma(Y_1, Y_2) & \sigma(Y_1, Y_3) \\ \sigma(Y_2, Y_1) & \sigma^2(Y_2) & \sigma(Y_2, Y_3) \\ \sigma(Y_3, Y_1) & \sigma(Y_3, Y_2) & \sigma^2(Y_3) \end{bmatrix}$$

Observe que nesta matriz as variâncias estão sobre a diagonal principal e as covariâncias  $\sigma(Y_i, Y_{i'})$  são encontradas nas  $i$ -ésimas linhas e  $j$ -ésimas colunas da matriz. Note que  $\sigma(Y_2, Y_1) = \sigma(Y_1, Y_2)$ , o que significa que  $\sigma^2(Y)$  é uma matriz simétrica, uma vez que as covariâncias abaixo da diagonal principal são as mesmas das existentes acima da diagonal principal.

Neste exemplo, suponha que os três erros tenham variância constante  $\sigma^2(\varepsilon_i) = \sigma^2$  e que são não correlacionados  $\sigma(\varepsilon_i, \varepsilon_{i'}) = 0$ , para  $i \neq i'$ . Pode-se então escrever que a matriz variância-covariância para o vetor aleatório  $\varepsilon$  é como se segue,  $\sigma^2(\varepsilon) = \sigma^2 I$ , uma vez que

$$\sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

## Análise de regressão linear em termos de matrizes

Para ilustrar a aplicação de álgebra matricial na análise de regressão linear simples, serão utilizados os dados resultantes de um estudo da relação entre o tamanho do lote produzido e a quantidade despendida de homens/hora numa determinada fábrica (extraído de NETER & WASSERMAN, 1977). Os dados são os seguintes:

Seção de Produção ( $i$ )	Tamanho do Lote ( $X_i$ )	Homens-hora ( $Y_i$ )
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Em análise de regressão uma matriz básica é o vetor Y que consiste de n observações da variável dependente

$$Y (n \times 1) = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

A transposta Y' é o vetor linha  $Y' (1 \times n) = [Y_1 \quad Y_2 \quad \dots \quad Y_n]$ .

Outra matriz básica na análise de regressão é a matriz X, definida da seguinte forma para regressão simples:

$$X(n \times 2) = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

A matriz  $X$  consiste de uma coluna de números 1 e uma coluna contendo  $n$  valores da variável independente  $X_i$ . A transposta de  $X$  é

$$X'(2 \times n) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix}$$

Para o exemplo considerado, as matrizes  $Y$  e  $X$  são:

$$Y = \begin{bmatrix} 73 \\ 50 \\ 128 \\ \vdots \\ 132 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 30 \\ 1 & 20 \\ 1 & 60 \\ \vdots & \vdots \\ 1 & 60 \end{bmatrix}$$

O modelo de regressão  $Y_i = E(Y_i) + \varepsilon_i$ ,  $i=1, \dots, n$  pode ser escrito em notação matricial como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Assim, a matriz  $Y$  é igual à soma de duas matrizes, uma matriz que contém os valores esperados e outra que contém os erros.

Define-se o vetor  $\beta$ , que representa os coeficientes de regressão, da seguinte forma:

$$\beta(2 \times 1) = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

O produto  $X\beta$  é uma matriz  $n \times 1$

$$X\beta(n \times 1) = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix}$$

Uma vez que  $\beta_0 + \beta_1 X_i = E(Y_i)$ , observa-se que  $X\beta$  é o vetor dos valores esperados  $E(Y_i)$  para o modelo de regressão linear simples.

Outro produto necessário é  $Y'Y$ :

$$Y'Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = [Y_1^2 + Y_2^2 + \dots + Y_n^2]$$

$$Y'Y = [\sum Y_i^2]$$

Observe que  $Y'Y$  é uma matriz  $1 \times 1$ , ou seja, um escalar, sendo então uma forma compacta de escrever uma soma de quadrados.

Em regressão, necessita-se também de  $X'X$ , que é uma matriz  $2 \times 2$ , em regressão simples:

$$X'X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

e ainda de  $X'Y$ , que é uma matriz  $2 \times 1$ :

$$X'Y (2 \times 1) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

A principal matriz inversa utilizada na análise de regressão é a inversa da matriz  $X'X$ :

$$X'X = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

Usando a regra do determinante, tem-se:

$$a = n \quad b = \sum X_i \quad c = \sum X_i^2 \quad d = \sum X_i^2, \text{ logo}$$

$$D = n \sum X_i^2 - (\sum X_i)(\sum X_i)$$

$$D = n \left[ \sum X_i^2 - \frac{(\sum X_i)^2}{n} \right] = n \sum (X_i - \bar{X})^2$$

$$\text{Portanto, } (X'X)^{-1} = \begin{bmatrix} \frac{\Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2} & \frac{-\Sigma X_i}{n\Sigma(X_i - \bar{X})^2} \\ \frac{-\Sigma X_i}{n\Sigma(X_i - \bar{X})^2} & \frac{n}{n\Sigma(X_i - \bar{X})^2} \end{bmatrix}$$

Uma vez que  $\Sigma X_i = n\bar{X}$ , pode-se simplificar a matriz interior:

$$(X'X)^{-1} = \begin{bmatrix} \frac{\Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2} & \frac{-\bar{X}}{\Sigma(X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\Sigma(X_i - \bar{X})^2} & \frac{1}{\Sigma(X_i - \bar{X})^2} \end{bmatrix}$$

Numa aplicação de regressão, para  $n=3$  observações, os três erros  $\varepsilon_1, \varepsilon_2, \varepsilon_3$ , cada um tem esperança zero. Define-se o vetor erro como:

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}, \text{ logo... } E(\varepsilon) = 0, \text{ uma vez que } E(\varepsilon) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ E(\varepsilon_3) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Admite-se que os erros tenham variância constante,  $\sigma^2(\varepsilon_i) = \sigma^2$  e que sejam não correlacionados,  $\sigma(\varepsilon_i, \varepsilon_j) = 0$ . Portanto, a matriz variância-covariância para o vetor  $\varepsilon$  pode ser escrita como  $\sigma^2(\varepsilon) = \sigma^2 I$ , uma vez que

$$\sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

## O modelo de regressão linear simples

Este modelo de regressão é definido como

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i=1, 2, \dots, n. \text{ Portanto, tem-se que}$$

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$

$$Y_3 = \beta_0 + \beta_1 X_3 + \varepsilon_3$$

$$\vdots \quad \vdots$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

Define-se o vetor resposta ou observação  $Y$ , a matriz  $X$  ou matriz-modelo, o vetor  $\beta$  ou vetor coeficientes de regressão e o vetor erro  $\varepsilon$  como:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Assim, o modelo é escrito, compactamente, em termos matriciais, como:

$$Y(nX1) = X\beta(nX1) + \varepsilon(nX1), \text{ logo vem que}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \beta_0 + \beta_1 X_3 + \varepsilon_3 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix}$$

A coluna de números 1 na matriz  $X$  pode ser vista como consistindo da variável “dummy”  $X_0 = 1$  no modelo de regressão.

$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i$ . Desta forma, a matriz  $X$  contém um vetor coluna da variável “dummy”  $X_0$  e outro vetor coluna que consiste dos valores da variável independente  $X_i$ .

No modelo de regressão,  $\varepsilon$  é um vetor de variáveis aleatórias normais, independentes, com  $E(\varepsilon) = 0$  e  $\sigma^2(\varepsilon) = \sigma^2 I$ .

Estimação de coeficientes de regressão pelo método dos quadrados mínimos ordinários

As equações normais de quadrados mínimos:

$$nb_0 + b_1 \sum X_i = \sum Y_i$$

$$b_0 \sum X_i + b_1 \sum X_i^2 = \sum X_i Y_i$$

Em termos matriciais, são dadas por  $X'Xb = X'Y$  em que

$$b = \begin{bmatrix} b \\ 0 \\ b \\ 1 \end{bmatrix}$$

Esta equação matricial estabelece que:

$$\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \text{ ou}$$

$$\begin{bmatrix} nb_0 + b_1 \sum X_i \\ b_0 \sum X_i + b_1 \sum X_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}, \text{ o que representa precisamente as equações normais.}$$

Para obter os coeficientes de regressão estimados a partir das equações normais  $X'Xb = X'Y$  por métodos matriciais, deve-se pré-multiplicar ambos os lados da equação pela inversa de  $X'X$  (assume-se que a inversa exista):

$$(X'X)^{-1} X'Xb = (X'X)^{-1} X'Y, \text{ e uma vez que}$$

$$(X'X)^{-1} X'X = I, \text{ obtém-se } b = (X'X)^{-1} X'Y.$$

Por exemplo, seja encontrar os coeficientes de regressão do exemplo anterior:

$$n = 10 \quad \sum Y_i = 1100 \quad \sum X_i = 500 \quad \sum X_i^2 = 28400 \quad \sum X_i Y_i = 61800$$

$$n \sum (X_i - \bar{X})^2 = n \sum X_i^2 - (\sum X_i)^2 = 34000$$

$$\text{Logo: } (X'X)^{-1} = \begin{bmatrix} \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} & \frac{-(\sum X_i)}{n \sum (X_i - \bar{X})^2} \\ \frac{-\sum X_i}{n \sum (X_i - \bar{X})^2} & \frac{n}{n \sum (X_i - \bar{X})^2} \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{28.400}{34.000} & \frac{-500}{34.000} \\ \frac{-500}{34.000} & \frac{10}{34.000} \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0,8352941 & -0,0147059 \\ -0,0147059 & 0,0002941 \end{bmatrix}$$

$$\text{e } X'Y = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix} = \begin{bmatrix} 1.100 \\ 61.800 \end{bmatrix}$$

$$\text{Então: } b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X'X)^{-1} X'Y = \begin{bmatrix} 0,8352941 & -0,0147059 \\ -0,0147059 & 0,0002941 \end{bmatrix} \begin{bmatrix} 1.100 \\ 61.800 \end{bmatrix} = \begin{bmatrix} 10 \\ 2,0 \end{bmatrix}$$

$$\text{ou } b_0 = 10,0 \text{ e } b_1 = 2,0$$

Observação: Para evitar erros de arredondamento, recomenda-se retirar a constante do denominador dos elementos de  $(X'X)^{-1}$  para fora da matriz e fazer a divisão como o último passo.

## Análise de variância da regressão

As somas de quadrados para a análise de variância, em notação matricial, são como a seguir:

$$SQ\ Total = Y'Y - n\bar{Y}^2$$

$$SQ\ Reg.= b'X'Y - n\bar{Y}^2$$

$$SQ\ Erro = e'e = Y'Y - b'X'Y$$

Seja o vetor dos valores estimados  $\hat{Y}_i$ , denotado por  $\hat{Y}$ :

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \text{ e o vetor dos erros } e_i = Y_i - \hat{Y}_i, \text{ denotado por}$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \text{ em que } e = Y - \hat{Y}.$$

Em notação matricial, tem-se que  $\hat{Y} = Xb$  porque

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix}$$

No exemplo em consideração, temos que

$$Y'Y = \sum Y_i^2 = 134660$$

$$b = \begin{bmatrix} 10,0 \\ 2,0 \end{bmatrix} \quad e \quad X'Y = \begin{bmatrix} 1100 \\ 61800 \end{bmatrix}$$

$$\text{Logo, } b'X'Y = [10,0 \ 2,0] \begin{bmatrix} 1.100 \\ 61.800 \end{bmatrix} = 134.600$$

$$S.Q. Erro = Y'Y - b'X'Y = 134.660 - 134.600 = 60$$

$$S.Q. Total = Y'Y - n\bar{Y}^2 = 134.660 - 121.000 = 13.660$$

$$S.Q. Re g. = b'X'Y - n\bar{Y}^2 = 134.600 - 121.000 = 13.600$$

## Inferências em análise de regressão

a) Coeficientes de Regressão: a matriz variância-covariância de b:

$$\sigma^2(b) = \begin{bmatrix} \sigma^2(b_0) & \sigma(b_0, b_1) \\ \sigma(b_1, b_0) & \sigma^2(b_1) \end{bmatrix} \text{ é dada por}$$

$$\sigma^2(b) = \sigma^2(X'X)^{-1} \text{ ou}$$

$$\sigma^2(b) = \begin{bmatrix} \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2} & \frac{-\bar{X} \sigma^2}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X} \sigma^2}{\sum (X_i - \bar{X})^2} & \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \end{bmatrix}$$

Quando então  $\sigma^2$  é substituído por  $Q.M. Erro$ , tem-se:

$$S^2(b) = Q.M. Erro (X'X)^{-1} = \begin{bmatrix} \frac{Q.M.E \sum X_i^2}{n \sum (X_i - \bar{X})^2} & \frac{-\bar{X} Q.M.E}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X} Q.M.E}{\sum (X_i - \bar{X})^2} & \frac{Q.M.E}{\sum (X_i - \bar{X})^2} \end{bmatrix}$$

em que  $S^2(b)$  é a matriz variância – covariância estimada de b. Nesta matriz, são reconhecidas as variâncias de  $b_0$  e  $b_1$  e a covariância ( $b_0, b_1$ ).

b) Faixa de confiança conjunta para  $\beta_0$  e  $\beta_1$ :

O limite para a faixa de confiança conjunta para  $\beta_0$  e  $\beta_1$ , dado como

$$\frac{n(b_0 - \beta_0)^2 + 2(\sum X_i)(b_0 - \beta_0)(b_1 - \beta_1) + (\sum X_i)^2(b_1 - \beta_1)^2}{2QME} = F(1-\alpha; n-2),$$

é expresso em termos matriciais como:

$$\frac{(b-\beta)'X'X(b-\beta)}{2QME} = F(1-\alpha; n-2).$$

c) Resposta Média: para estimar a resposta média em  $X_h$  (um valor qualquer de X, dentro do intervalo do modelo), define-se o vetor

$$X_h = \begin{bmatrix} 1 \\ X_h \end{bmatrix} \text{ ou } X'_h = \begin{bmatrix} 1 & X_h \end{bmatrix}.$$

A resposta média estimada, em notação matricial, é  $\hat{Y}_h = X'_h b$ , uma vez que

$$X'_h b = \begin{bmatrix} 1 & X_h \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_h \end{bmatrix} = \hat{Y}_h$$

A variância de  $\hat{Y}_h$ , dada anteriormente como

$$\sigma^2(\hat{Y}_h) = \sigma^2 \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right], \text{ em notação matricial, fica } \sigma^2(\hat{Y}_h) = \sigma^2 X'_h (X'X)^{-1} X_h.$$

A variância estimada de  $\hat{Y}_h$ , dada anteriormente como

$$S^2(\hat{Y}_h) = QME \left[ \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right], \text{ em notação matricial, fica:}$$

$$S^2(\hat{Y}_h) = QME \left[ X'_h (X'X)^{-1} X_h \right].$$

d) Predição de novas observações individuais: a variância estimada

$$S^2[\hat{Y}_h(novo)] \text{ dada anteriormente como}$$

$$S^2[\hat{Y}_h(novo)] = QME \left[ 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right], \text{ em notação matricial, fica:}$$

$$S^2[\hat{Y}_h(novo)] = QME \left[ 1 + X'_h (X'X)^{-1} X_h \right].$$

Por exemplo, suponha que se deseja encontrar

$S^2(b_0)$  e  $S^2(b_1)$ , do exemplo anterior, por métodos matriciais, sendo que QME=7,5 e

$$(X'X)^{-1} = \begin{bmatrix} 0,8352941 & -0,0147059 \\ -0,0147059 & 0,0002941 \end{bmatrix}$$

$$S^2(b) = QME(X'X)^{-1}$$

Logo,

$$= 7,5 \begin{bmatrix} 0,8352941 & -0,0147059 \\ 0,0147059 & 0,0002941 \end{bmatrix} = \begin{bmatrix} 6,26471 & -0,110294 \\ -0,110294 & 0,002206 \end{bmatrix}$$

Assim,  $S^2(b_0) = 6,26471$  e  $S^2(b_1) = 0,002206$

Observe como é simples encontrar as variâncias estimadas dos coeficientes de regressão, uma vez obtida a matriz  $(X'X)^{-1}$ .

Seja encontrar  $S^2(Y_h)$ , no exemplo anterior, para  $X_h = 55$ . Neste caso, tem-se  $X'_h = [1 \ 55]$  e obtém-se:

$$\begin{aligned} S^2(Y_{55}) &= QME \left[ X'_h (X'X)^{-1} X_h \right] \\ &= 7,5 [1 \ 55] \begin{bmatrix} 0,8352941 & -0,0147059 \\ -0,0147059 & 0,0002941 \end{bmatrix} \begin{bmatrix} 1 \\ 55 \end{bmatrix} \end{aligned}$$

$$S^2(Y_{55}) = QME \left[ X'_h (X'X)^{-1} X_h \right] = 0,80520$$

## Decomposição espectral (DE)

Toda matriz simétrica  $A_{(pxp)}$  (ou seja,  $A = A'$ ) pode ser decomposta como

$$A = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \dots + \lambda_p e_p e_p' = \sum_i \lambda_i e_i e_i'$$

Isto pode ser escrito como  $A = P \Lambda P'$ , onde  $P = P^{-1}$  (ou seja,  $P$  é uma matriz ortogonal com colunas  $e_1, e_2, \dots, e_p$ , que são os autovetores correspondentes aos autovalores  $\lambda_1, \lambda_2, \dots, \lambda_p$ ) e  $\Lambda = \text{diag}(\lambda_i)$ .

A decomposição espectral ou decomposição de Jordan vincula a estrutura de uma matriz aos autovalores e autovetores e constitui um ponto importante na análise de matrizes. Relembre que o teorema da decomposição de Jordan estabelece que cada matriz simétrica  $A_{(pxp)}$  pode ser escrita

como  $A = P \Lambda P' = \sum_{i=1}^p \lambda_i e_i e_i'$  em que  $\Lambda = \text{diag}(\lambda_i) = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$  e

$P = (e_1, e_2, \dots, e_p)$  é uma matriz ortogonal consistindo dos autovetores  $e_i$  de  $A$ .

### Exemplo numérico

Considere uma matriz  $A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$ . Os autovalores de  $A$  são obtidos pela resolução de

$$|A - \lambda I| = 0, \text{ o que é equivalente a } \begin{vmatrix} 1-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = (1-\lambda)(3-\lambda) - 4 = 0.$$

Portanto, os autovalores são  $\lambda_1 = 2 + \sqrt{5}$  e  $\lambda_2 = 2 - \sqrt{5}$ . Substituindo os autovalores  $\lambda_1$  e  $\lambda_2$  em  $|A - \lambda_i I| e_i = 0$ , são obtidos os autovetores  $e_1 = (0,5257; 0,8506)'$  e  $e_2 = (0,8506; -0,5257)'$ , que são ortogonais, uma vez que  $e_1' e_2 = 0$ .

Usando a decomposição espectral, podem-se definir potências de uma matriz  $A(p \times p)$ : suponha que  $A$  seja uma matriz simétrica, então, pelo teorema da decomposição espectral,  $A = P \Lambda P'$  pode ser definido  $A^\alpha = P \Lambda^\alpha P'$ , em que  $\Lambda^\alpha = \text{diag}(\lambda_1^\alpha, \lambda_2^\alpha, \dots, \lambda_p^\alpha)$ .

### Autovalores

Para extrair os autovalores da equação característica, faz-se  $|R - \lambda I| = 0$ . Por exemplo, dada a

matriz  $R = \begin{bmatrix} 1,00 & 0,30 \\ 0,30 & 1,00 \end{bmatrix}$ , tem-se

$$\begin{bmatrix} 1,00 & 0,30 \\ 0,30 & 1,00 \end{bmatrix} - \lambda \begin{bmatrix} 1,00 & 0 \\ 0 & 1,00 \end{bmatrix} = 0 \text{ ou}$$

$$\begin{bmatrix} 1,00 & 0,30 \\ 0,30 & 1,00 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 0,30 \\ 0,30 & 1-\lambda \end{bmatrix} = 0. \text{ Então segue que}$$

$$\begin{vmatrix} 1-\lambda & 0,30 \\ 0,30 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 0,09 = 1^2 - 2\lambda + \lambda^2 - 0,09 = 0. \text{ Logo se tem}$$

$$\begin{vmatrix} 1-\lambda & 0,30 \\ 0,30 & 1-\lambda \end{vmatrix} = \lambda^2 - 2\lambda + 0,91 = 0. \text{ Então,}$$

$$\lambda_{1,2} = \frac{2 \pm \sqrt{2^2 - 4(0,91)}}{2} = \frac{2 \pm 0,6}{2} \text{ e } \begin{cases} \lambda_1 = 1,30 \\ \lambda_2 = 0,70 \end{cases}, \text{ que são os autovalores da matriz R.}$$

### Propriedades dos autovalores

$$1) \sum_i \lambda_i = \text{traço} R = \text{tr} R$$

$$2) \prod_i \lambda_i = \det R = |R|$$

3)  $\varphi_i = \frac{\lambda_i}{trR}$ , em que  $\varphi_i$  representa a proporção da variância devida a cada autovalor e  $\sum_i \varphi_i = 1,0$ .

### Autovetores

A substituição dos autovalores extraídos da matriz de correlação R na equação característica  $[R - \lambda I]v = 0$  permite resolver o sistema de equações resultantes em relação a  $v$ , que representa a matriz dos autovetores associados a cada um dos autovalores obtidos.

Desta forma, considerando o exemplo anterior, para  $\lambda_1 = 1,30$ , tem-se

$$\begin{bmatrix} 1,00 & 0,30 \\ 0,30 & 1,00 \end{bmatrix} - 1,30 \begin{bmatrix} 1,00 & 0 \\ 0 & 1,00 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ ou } \begin{bmatrix} 1,00 & 0,30 \\ 0,30 & 1,00 \end{bmatrix} - \begin{bmatrix} 1,30 & 0 \\ 0 & 1,30 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

ou ainda  $\begin{bmatrix} -0,30 & 0,30 \\ 0,30 & -0,30 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ , o que resulta no sistema de equações lineares

$$\begin{cases} -0,30v_1 + 0,30v_2 = 0 \\ 0,30v_1 - 0,30v_2 = 0 \end{cases}, \text{ que pode ser simplificado para } v_1 = v_2 \text{ (número infinito de soluções).}$$

Para o autovalor  $\lambda_2 = 0,70$ , tem-se  $\begin{bmatrix} 1-0,70 & 0,30 \\ 0,30 & 1-0,70 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$  e  $\begin{cases} 0,30v_1 + 0,30v_2 = 0 \\ 0,30v_1 + 0,30v_2 = 0 \end{cases}$ ,

sendo  $v_1 = -v_2$ .

Como ambos os sistemas de equações são homogêneos, diz-se que esta solução é apenas estrutural

com estrutura de autovetor S dada por  $S = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ . Para completar a extração de autovetores,

deve-se impor alguma restrição sobre a solução estrutural (por exemplo, os componentes principais são extraídos de forma ortogonal, maximizando a variância de cada componente).

No caso dos componentes principais (CP), têm-se, por exemplo, os critérios seguintes:

---

CP1 CP2  $h^2$  (Comunalidade)

---

X<sub>1</sub> 1,00 Soma de linhas de elementos

---

X<sub>2</sub> 1,00 quadrados igual a 1,0

---

$\lambda_1 \lambda_2$  Traço da matriz

---

Soma de colunas de  
elementos quadrados  
igual aos autovalores

---

No caso do exemplo considerado tem-se:

---

CP1 CP2  $h^2$

---

X<sub>1</sub> 0,65 0,35 1,00

---

---

0,65 e 0,35 são os valores  
X<sub>2</sub> 0,65 0,35 1,00 dos elementos ao quadrado  
na matriz com restrições

---

$\lambda_1=1,30 \quad \lambda_2=0,70$

---

## Decomposição em valor singular (DVS)

Eckart e Young (1936) mostraram que qualquer matriz retangular  $Y$  pode ser decomposta da seguinte forma:

$Y_{(n \times p)} = U_{(n \times p)} W_{(diagonal, p \times p)} V'_{(p \times p)}$ , em que, tanto  $U$  quanto  $V$  são matrizes coluna-ortogonais, ou seja, matrizes que contêm vetores coluna que são normalizados e ortogonais uns em relação aos outros, e  $W$  é uma matriz diagonal  $D(w_i)$ . Esse método é conhecido como decomposição em valor singular (DVS). A ilustração a seguir mostra mais claramente os tipos dessas matrizes.

$$\left[ Y_{(n \times p)} \right] = \left[ U_{(n \times p)} \right] \begin{bmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & w_p \end{bmatrix} \left[ V'_{(p \times p)} \right].$$

Os valores diagonal  $w_i$  em  $W$  são não negativos e chamados de valores singulares de  $Y$ . A DVS fornece uma forma de manejar matrizes que são singulares ou numericamente muito próximas de singular. A DVS pode ainda fornecer um diagnóstico claro do problema, ou resolvê-lo. A singularidade pode ocorrer quando da resolução de um conjunto de equações lineares simultâneas representadas pela equação matricial  $Ab=c$ , sendo a matriz  $A$  quadrada,  $A$  e  $c$  são conhecidos e  $b$  desconhecido. A matriz  $A$  deve ser invertida para encontrar  $b$ . A matriz  $A$  sempre pode ser decomposta (com base em DVS) como:

$$A = UD(w_i)V'$$

Neste caso,  $U$ ,  $D(w_i)=W$  e  $V$  são todas matrizes quadradas com a mesma dimensão de  $A$ .

Usando as propriedades de inversas, pode-se obter a inversa de  $A$  computando:

$$A^{-1} = [UD(w_i)V']^{-1} = [V']^{-1} [D(w_i)]^{-1} [U]^{-1}$$

Uma vez que  $U$  e  $V$  são ortogonais, suas inversas são iguais a suas transpostas, enquanto a inversa de uma matriz diagonal é uma matriz diagonal cujos elementos são os recíprocos dos elementos originais. Então, pode-se escrever:

$$A^{-1} = VD(1/w_i)U'$$

Pode-se esperar que um ou mais valores  $w_i$  sejam zero, então seus recíprocos são infinito; então  $A$  é uma matriz singular. Pode-se esperar também que um ou mais valores  $w_i$  sejam numericamente tão pequenos que não possa ser computado devido à precisão da máquina; neste caso, a matriz  $A$  é dita estar mal-condicionada. Quando  $A$  é singular, as colunas de  $V$ , que correspondem aos elementos zero em  $D(w_i)$ , formam uma base ortogonal para o espaço em que o sistema de equações não tem nenhuma solução, enquanto as colunas de  $U$ , que correspondem aos elementos zero em  $D(w_i)$ , constituem uma base ortogonal para o espaço em que o sistema tem uma solução.

### Exemplos numéricos

a) Admitindo que a inversa generalizada de  $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$  é  $A^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ , segue-se que  $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = A$ , então a matriz  $\begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix}$  também é uma inversa generalizada de  $A$ .

b) Sejam as matrizes  $A$  e  $A'A$ , dadas por  $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 1 \end{bmatrix}$  e  $A'A = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$ . Com base na expressão  $A = U_{(n \times p)} W_{(p \times p)} V'_{(p \times p)}$ , obtém-se  $U = \begin{bmatrix} 1 & 0 \\ 0 & \frac{2}{\sqrt{5}} \\ 0 & \frac{1}{\sqrt{5}} \end{bmatrix}$ ,  $W = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{5} \end{bmatrix}$  e  $V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ .

Agora, é comum verificar se  $AV = UW$ . Na prática, a matriz  $W$  é arranjada de tal forma que os  $w_i = \sqrt{\lambda_i}$  são listados em ordem decrescente na sua diagonal, o que necessitará de permutações nos elementos de  $U$ .

c) Considere a matriz  $A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix}$ , então  $AA' = A'A = \begin{bmatrix} 6 & 10 & 6 \\ 10 & 17 & 10 \\ 6 & 10 & 6 \end{bmatrix}$ . Os autovalores de

$AA'$ ,  $A'A$  são:  $\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 28,86 \\ 0,14 \\ 0 \end{bmatrix}$ . Os autovetores de  $AA'$ ,  $A'A$  são:

$$u_1 = v_1 = \begin{bmatrix} 0,454 \\ 0,766 \\ 0,454 \end{bmatrix}, u_2 = v_2 = \begin{bmatrix} 0,542 \\ -0,643 \\ 0,542 \end{bmatrix}, u_3 = v_3 = \begin{bmatrix} -0,707 \\ 0 \\ -0,707 \end{bmatrix}$$

Fazendo a decomposição de  $A$ , dada por  $A = \sum_{i=1}^2 w_i u_i v_i'$ , obtém-se:

$$A = \begin{bmatrix} 1,11 & 1,87 & 1,11 \\ 1,87 & 3,15 & 1,87 \\ 1,11 & 1,87 & 1,11 \end{bmatrix}.$$

O posto de uma matriz é igual ao número de valores singulares não zero. Uma matriz quadrada  $A$  é não singular se  $w_i \neq 0$  para todo  $i$ . Se  $A$  é uma matriz  $n \times n$  não singular, então sua inversa é dada por  $A = UWV'$  ou  $A^{-1} = VW^{-1}U'$ , em que

$W^{-1} = \text{diag}\left(\frac{1}{w_1}, \frac{1}{w_2}, \dots, \frac{1}{w_n}\right)$ . Se  $A$  é singular ou mal-condicionada, então se pode

usar a DVS para aproximar sua inversa com base nas matrizes seguintes:

$$A^{-1} = (UWV')^{-1} = VW_o^{-1}U'$$

$$W_o^{-1} = \begin{cases} 1/w_i, & w_i > t \\ 0, & \text{contrario} \end{cases} \quad (\text{em que } t \text{ é um pequeno valor inicial})$$

### Decomposição em valor singular (DVS) e componentes principais (CP)

Considere as  $p$  variáveis  $X_1, X_2, \dots, X_p$ , observadas em  $n$  indivíduos ( $n \geq p$ ), cujos valores resultantes são dispostos em uma matriz  $X_{n \times p}$ . Seja  $S_{p \times p}$ , a matriz de variâncias e covariâncias amostrais associada à matriz de dados  $X$  e  $R_{p \times p}$  à correspondente matriz de correlações

amostrais. A decomposição por valor singular da matriz  $X$  é definida por  $X = UDV'$ , em que  $U'U = VV' = VV = I_p$ ,

$D = diag(r_1, r_2, \dots, r_p)$ , matriz diagonal formada pela raiz quadrada dos autovalores de  $X'X$ , sendo  $r_1 \geq r_2 \geq \dots \geq r_p$ , as colunas de  $U$  são os  $p$  autovetores ortonormalizados de  $X'X$  e as colunas de  $V$ , os autovetores normalizados de  $X'X$ .

Tomando como base os valores observados centrados na média ( $y_{ij} = X_{ij} - \bar{X}_j$ ),  $i = 1, 2, \dots, n$  e  $j = 1, 2, \dots, p$ , tem-se então a matriz de dados  $Y_{n \times p}$  com decomposição por valor singular  $Y = U^* D_y V^{*'} \cdot$ , sendo  $V^*$  e  $U^*$  as matrizes cujas colunas são os autovetores ortonormalizados de  $YY'$  e de  $Y'Y$ , respectivamente, e  $D_y = diag(d_1, d_2, \dots, d_p)$ , matriz diagonal formada pela raiz quadrada dos autovalores de  $YY'$ , sendo  $d_1 \geq d_2 \geq \dots \geq d_p$ .

Considere  $Y_{n \times p}$  uma matriz de observações amostrais com valores centrados na média. Sejam  $u_{ij}$  os elementos  $(i, j)$  de  $U^*$  e  $v_{ij}$  os elementos  $(i, j)$  de  $V^{*'} \cdot$ . A decomposição por valor singular de  $Y$  é dada por:

$$Y = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix}$$

Pode-se observar que o elemento  $y_{ij}$  de  $Y$  pode ser escrito como  $y_{ij} = \sum_{t=1}^p u_{it} d_t v_{tj}$ .

Na Análise por Componentes Principais para estes dados, tem-se que  $S_y = \frac{1}{n-1} Y'Y$ , portanto,

$C_Y = Y_{n \times p} V^{*'} \cdot p \times p$ , uma vez que os autovetores ortonormalizados de  $Y'Y$  e de  $\frac{1}{n-1} Y'Y$  são os mesmos.

Como, pela decomposição por valor singular  $Y = U^* D_y V^{*'} \cdot$ , então  $C_Y = U^* D_y V^{*'} \cdot V^{*'} V^{*'} \cdot p \times p = U^* D_y V^{*'} \cdot$  dado que  $V^{*'} V^{*'} = I_p$ . Também pode ser observado desta expressão que  $Y = C_Y V^{*'} \cdot$ .

Se a Análise por Componentes Principais é utilizada com o objetivo de redução de dimensionalidade, a questão que se impõe é quantos componentes devem ser retidos para garantir que uma grande proporção da variabilidade dos dados seja explicada por eles. Retirar os  $m$  primeiros componentes é

equivalente a modelar os elementos de  $Y$  por  $y_{ij} = \sum_{t=1}^M u_{it} d_t v_{tj} + \varepsilon_{ij}$ , em que  $\varepsilon_{ij}$  é o termo

residual, o que reduz a questão à escolha do modelo a ser utilizado. A metodologia de análise AMMI – Biplot adota essa abordagem no estudo da interação genótipo x ambiente.

No método AMMI (Efeito Principal Aditivo e Interação Multiplicativa) de análise da interação genótipo x ambiente, a resposta média do genótipo  $i$  em um ambiente  $j$  é modelada por:

$$\bar{Y}_{ij} = \mu + g_i + e_j + \sum_{k=1}^m \alpha_{ik} \lambda_k \gamma_{jk} + \rho_{ij} + \varepsilon_{ij}, \text{ em que } [(ge)_{ij}] \text{ é representado por} \\ \sum_{k=1}^m \alpha_{ik} \lambda_k \gamma_{jk} + \rho_{ij}, \text{ sob as restrições } \sum_i g_i = \sum_j e_j = \sum_i [(ge)_{ij}] = \sum_j [(ge)_{ij}] = 0.$$

As estimativas da média geral ( $\mu$ ) e dos efeitos principais ( $g_i$  e  $e_j$ ) são obtidas por meio de ANOVA simples com dois fatores a partir da matriz de médias  $Y_{(ge)} = [\bar{Y}_{ij}]$ . Os resíduos obtidos dessa matriz constituem a matriz de interações

$GE_{(ge)} = [(\hat{g}e)_{ij}]$ , em que  $GE = \bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{..}$ . Os termos da interação multiplicativa são estimados com base na decomposição em valor singular (DVS) da matriz  $GE_{(ge)} = [(\hat{g}e)_{ij}]$  junto com a análise de componentes principais (ACP). Assim,  $\lambda_k$  é estimado pelo k-ésimo valor singular de  $GE$ ,  $\alpha_{ik}$  é estimado pelo i-ésimo elemento do vetor singular esquerdo do  $\alpha'_{k(g \times l)}$  e  $\gamma_{jk}$  é estimado pelo j-ésimo elemento do vetor singular direito  $\gamma'_{k(l \times e)}$  associado com  $\lambda_i$ .

A análise de componentes principais (ACP) auxilia no processo de redução da dimensionalidade, ou seja, na determinação do número de termos de interação multiplicativa ( $m$ ) a serem incluídos no modelo AMMI. Como o primeiro componente (CP1) explica maior proporção da variação dos dados que o segundo componente (CP2), e o segundo explica maior proporção que o terceiro (CP3), e assim por diante, determinam-se como mais explicativos os modelos AMMI1 > AMMI2 > AMMI3 e assim sucessivamente. O biplot é um gráfico formado pelos eixos dos componentes CP1 e CP2, contendo dois tipos de pontos, um para genótipos e outro para ambientes.

## Propriedades e operações especiais de matrizes

As matrizes possuem algumas propriedades que são mais especializadas, tais como determinantes, inversas, autovalores e posto. Estas propriedades são aplicadas apenas para matrizes quadradas. A álgebra de matrizes possui algumas operações especializadas, tais como ortonormalização, decomposição espectral e decomposição em valor singular.

## Determinantes

O determinante de uma matriz quadrada  $n \times n A = [a_{ij}]$  é um número (escalar) denotado por  $|A|$  ou  $\det(A)$ , por meio do qual, propriedades importantes, como a singularidade, pode ser caracterizada. Este número é definido como uma função dos elementos da matriz

$|A| = \sum_{j=1}^n a_{ij} |A_{1j}| (-1)^{1+j}$ , em que  $A_{1j}$  é a matriz  $(n-1)(n-1)$  obtida pela exclusão da primeira linha e j-ésima coluna de  $A$ . Igualmente,  $|A| = \sum_{j=1}^n a_{ij} |A_{ij}| (-1)^{i+j}$ , com a i-ésima linha em lugar da primeira linha.

$$\text{Exemplo 1} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22}(-1)^2 + a_{12}a_{21}(-1)^3 = a_{11}a_{22} - a_{12}a_{21}$$

$$\text{Exemplo 2} \begin{vmatrix} 1 & 3 \\ 6 & 4 \end{vmatrix} = 1|4|(-1)^2 + 3|6|(-1)^3 = 1(4) - 3(6)$$

Observação: O cálculo de determinantes, usando este método, é impraticável para matrizes cuja ordem seja maior que 4 devido ao grande aumento no número de termos e de multiplicações necessárias. Existem métodos mais práticos, baseados na decomposição (fatoração) de matrizes.

## Algumas propriedades de determinantes

1. As posições de linhas e colunas podem ser modificadas sem afetar o valor de um determinante. Ou seja,  $|A| = |A'|$ .

2. Se duas linhas (ou colunas) são mudadas de posição, o sinal do determinante é modificado. Por exemplo,

$$\begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} = - \begin{vmatrix} 1 & -2 \\ 3 & 4 \end{vmatrix}$$

3. Se uma linha (ou coluna) é modificada adicionando ou subtraindo a seus elementos os elementos correspondentes de qualquer outra linha (ou coluna), o determinante permanece inalterado. Por exemplo,

$$\begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} = \begin{vmatrix} 3+1 & 4-2 \\ 1 & -2 \end{vmatrix} = \begin{vmatrix} 4 & 2 \\ 1 & -2 \end{vmatrix} = -10$$

4. Se os elementos em qualquer linha (ou coluna) têm um fator comum  $\alpha$ , então o determinante é igual ao determinante da matriz correspondente, na qual  $\alpha = 1$  multiplicado por  $\alpha$ . Por exemplo,

$$\begin{vmatrix} 6 & 8 \\ 1 & -2 \end{vmatrix} = 2 \begin{vmatrix} 3 & 4 \\ 1 & -2 \end{vmatrix} = 2(-10) = -20$$

5. Quando pelo menos uma linha (ou coluna) de uma matriz é uma combinação linear de outras linhas (ou colunas), o determinante é zero. Inversamente, se o determinante é zero, então

pelo menos uma linha e uma coluna são linearmente dependentes de outras linhas e colunas, respectivamente. Por exemplo,

$$\begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & -1 \\ 2 & -1 & 3 \end{bmatrix}$$

Este determinante é zero porque a primeira coluna é uma combinação linear da segunda e terceira coluna, ou seja,

coluna 1 = coluna 2 + coluna 3. Similarmente, existe uma dependência linear entre as linhas, que é dada pela relação linha 1= 7/8linha 2 + 4/5 linha 3.

6. O determinante de uma matriz triangular superior ou inferior é o produto dos elementos da diagonal principal. Por exemplo,

$$\begin{bmatrix} 3 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 4 \end{bmatrix} = (3)(2)(4) = 24$$

7. O determinante do produto de duas matrizes quadradas é o produto dos determinantes individuais, ou seja,  $|AB| = |A||B|$

Esta regra pode ser generalizada para qualquer número de fatores. Uma aplicação imediata é para potências de matrizes  $|A^2| = |A||A| = |A|^2$  e de forma mais generalizada,  $|A^n| = |A|^n$ , para qualquer  $n$  inteiro.

8. O determinante da transposta de uma matriz é o mesmo da matriz original, ou seja,  $|A'| = |A|$ .

Observação 2: As propriedades 6 e 7 são as chaves para a avaliação prática de determinantes. Qualquer matriz quadrada não singular  $A$  pode ser decomposta como o produto de dois fatores triangulares  $|A| = |L||U|$ , em que  $L$  é triangular inferior unitária e  $U$  é triangular superior. Isto é chamado de triangulação  $LU$ , fatoração ou decomposição  $LU$ . Este processo pode ser conduzido em  $O(n^3)$  operações pontos-flutuantes. De acordo com a regra 7, tem-se  $|A| = |L||U|$ , e de acordo com a regra 6,  $|L| = 1$  e  $|U| = u_{11}u_{22}\dots u_{nn}$ . Esta última operação requer apenas  $O(n^3)$  operações pontos-flutuantes. Assim, a avaliação de  $|A|$  é dominada pelo esforço envolvido na computação da fatoração.

### Matrizes singulares e posto

Se o determinante  $|A|$  de uma matriz quadrada  $n \times n$ ,  $A \equiv A_n$ , é zero, então a matriz é dita **singular**. Isto significa que pelo menos uma linha e uma coluna são linearmente dependentes uma da outra. Se esta linha e coluna são removidas, ficamos com outra matriz,  $A_{n-1}$ , para a qual podemos aplicar o mesmo critério. Se o determinante  $|A_{n-1}|$  é zero, podemos remover outra linha e coluna para

obter a matriz  $A_{n-2}$ , e assim sucessivamente. Suponha que eventualmente chegemos a uma matriz  $A_r$   $r \times r$  cujo determinante é diferente de zero. Então a matriz  $A$  é dita ter **posto**  $r$ , e escrevemos  $\text{posto}(A)=r$ .

Se o determinante de  $A$  é diferente de zero, então  $A$  é dita ser **não singular**. O posto de uma matriz não singular  $n \times n$  é igual a  $n$ . Obviamente que o posto de  $A'$  é o mesmo de  $A$ , uma vez que se tem apenas uma transposição de linhas e colunas.

**Exemplo 1**  $A = \begin{bmatrix} 3 & 2 & 2 \\ 1 & 2 & -1 \\ 2 & -1 & 3 \end{bmatrix}$ , tem posto  $r = 3$  porque  $|A| = -3 \neq 0$

**Exemplo 2**  $A = \begin{bmatrix} 3 & 2 & 2 \\ 1 & 2 & -1 \\ 2 & -1 & 3 \end{bmatrix}$ , é singular porque a primeira linha e coluna podem ser expressas

como combinações lineares das outras. Removendo a primeira linha e coluna, ficamos então com uma matriz  $2 \times 2$  cujo determinante é  $2 \times 3 - (-1) \times (-1) = 5 \neq 0$ . Consequentemente,  $A$  tem posto  $r = 2$ .

### Deficiência de posto

Se a matriz quadrada  $A$  é suposta ser de posto  $r$ , mas de fato tem um posto menor  $\bar{r} < r$ , a matriz é dita ser deficiente de posto. O número  $r - \bar{r}$  é chamado de deficiência de posto.

### Posto de matrizes somas e produtos

Em análises, as matrizes são frequentemente construídas por meio de combinações de somas e de produtos de matrizes mais simples. Duas regras importantes são aplicadas para a propagação do posto através destas combinações:

a) O posto do produto de duas matrizes  $A$  e  $B$  não pode exceder o menor posto das matrizes multiplicando. Ou seja, se o posto de  $A$  é  $r_a$  e o posto de  $B$  é  $r_b$ , então  $\text{posto}(AB) \leq \min(r_a, r_b)$ .

b) O posto de uma matriz soma não pode exceder a soma dos postos das matrizes somando. Ou seja, se o posto de  $A$  é  $r_a$  e o posto de  $B$  é  $r_b$ , então  $\text{posto}(A+B) \leq r_a + r_b$ .

### Sistemas singulares: soluções particulares e homogêneas

Seja o sistema usual de equações simultâneas denotado por  $Ax = y$ , em que  $A$  é uma dada matriz  $n \times n$ ,  $y$  é um dado vetor  $n \times 1$  e  $x$  é o vetor de incógnitas  $n \times 1$ . Uma receita para solucionar equações algébricas em termos de determinantes é dada pela regra de Cramer, para  $n < 3$ . A solução explícita para os componentes  $x_1, x_2, \dots, x_n$  de  $x$  em termos de determinantes é

$$x_1 = \frac{\begin{vmatrix} y_1 & a_{12} & a_{13} & \dots & a_{1n} \\ y_2 & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ y_n & a_{n2} & a_{n3} & \dots & a_{nn} \end{vmatrix}}{|A|}, x_2 = \frac{\begin{vmatrix} a_{11} & y_1 & a_{13} & \dots & a_{1n} \\ a_{21} & y_2 & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & y_n & a_{n3} & \dots & a_{nn} \end{vmatrix}}{|A|}, \dots$$

### Exemplo 1

A solução do sistema linear  $3 \times 3$ ,  $\begin{bmatrix} 5 & 2 & 1 \\ 3 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ 5 \\ 3 \end{bmatrix}$ , pela regra de Cramer, é dada por

$$x_1 = \frac{\begin{bmatrix} 8 & 2 & 1 \\ 5 & 2 & 0 \\ 3 & 0 & 2 \end{bmatrix}}{\begin{bmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}} = \frac{6}{6} = 1, \quad x_1 = \frac{\begin{bmatrix} 5 & 8 & 1 \\ 3 & 5 & 0 \\ 1 & 3 & 2 \end{bmatrix}}{\begin{bmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}} = \frac{6}{6} = 1, \quad x_1 = \frac{\begin{bmatrix} 5 & 2 & 8 \\ 3 & 2 & 5 \\ 1 & 0 & 3 \end{bmatrix}}{\begin{bmatrix} 5 & 2 & 1 \\ 3 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}} = \frac{6}{6} = 1$$

Uma consequência imediata da regra de Cramer é o que ocorre se  $y_1 = y_2 = \dots = y_n = 0$ . Os sistemas de equações lineares com o lado direito nulo,  $Ax = 0$ , são chamados de sistemas homogêneos. Pela regra de Cramer, vemos que se  $|A|$  é diferente de zero, todos os componentes da solução do sistema são zero e, consequentemente, a única solução possível é a trivial  $x = 0$ .

Com base no conceito de posto, podemos discutir o que acontece com o sistema linear quando o determinante de  $A$  desaparece, significando que seu posto é menor que  $n$ . Neste caso, o sistema linear tem nenhuma solução ou um número infinito de soluções. A regra de Cramer tem ajuda limitada ou nenhuma ajuda nesta situação.

Discutindo mais este caso, notamos que se  $|A| = 0$  e o posto de  $A$  é  $r = n = d$ , em que  $d \geq 1$  é a deficiência de posto, então existe  $d$  vetores independentes diferentes de zero  $Z_i$ ,  $i = 1, \dots, d$ , tais que  $AZ_i = 0$ . Estes  $d$  vetores, apropriadamente ortonormalizados, são chamados autovetores nulos de  $A$ , e formam uma base para seu espaço nulo. Seja  $Z$  a matriz  $n \times d$  obtida colecionando os  $Z_i$  como colunas. Se  $y$  está na faixa de  $A$ , isto é, existe uma matriz não nula  $x_p$  tal que  $y = Ax_p$ , sua solução geral é  $x = x_p + x_h + Zw$ , em que  $w$  é um vetor de ponderação arbitrário  $d \times 1$ . Os componentes  $x_p$  e  $x_h$  são chamados de parte particular e homogênea, respectivamente, da solução  $x$ . Se  $y = 0$ , apenas a parte homogênea permanece.

### Inversão de matrizes

A inversa de uma matriz quadrada não singular  $A$  é representada pelo símbolo  $A^{-1}$  e definida pela relação  $A^{-1}A = AA^{-1} = I$ . A aplicação mais importante do conceito de inversa é a solução de sistemas lineares. Suponha o sistema seguinte  $AX = y$ . Pré-multiplicando ambos os lados por  $A^{-1}$ , obtemos a relação  $x = A^{-1}y$ . De forma geral, considere a equação matricial para o lado direito múltiplo ( $m$ ),  $A_{n \times n}X_{n \times m} = Y_{n \times m}$ , que se reduz ao sistema anterior para  $m=1$ . A relação inversa que fornece  $X$  como função de  $Y$  é  $X = A^{-1}y$ . Em particular, a solução de  $AX = I$  é  $X = A^{-1}$ .

### Computação de inversas

Uma fórmula geral para os elementos da inversa pode ser obtida especializando a regra de Cramer para  $AX = I$ . Seja, por exemplo,  $B = [b_{ij}] = A^{-1}$ , então,  $b_{ij} = \frac{A_{ji}}{|A|}$ , em que  $A_{ji}$  denota a chamada adjunta do elemento  $a_{ij}$  de  $A$ . A adjunta  $A_{ji}$  é definida como o determinante da submatriz de ordem  $(n-1)(n-1)$  obtida pela eliminação da  $j$ -ésima linha e  $i$ -ésima coluna de  $A$ , multiplicado por  $(-1)^{i+j}$ .

### Algumas propriedades da inversa

1. Inversa da transposta é igual à transposta da inversa  $(A')^{-1} = (A^{-1})'$ , porque  $(AA^{-1}) = (AA^{-1})' = (A^{-1})'A = I$ .
2. A inversa de uma matriz simétrica é também simétrica. Em função dessa regra, tem-se  $(A')^{-1} = A^{-1}(A^{-1})'$ , portanto,  $A^{-1}$  é também simétrica.
3. A inversa de uma matriz produto é o produto reverso das inversas dos fatores:  $(AB)^{-1} = B^{-1}A^{-1}$ .
4. Para uma matriz diagonal  $D$ , na qual todos os elementos são diferentes de zero,  $D^{-1}$  é também uma matriz diagonal com elementos  $1/d_{ij}$ .
5. A inversa de uma matriz triangular superior é também uma matriz triangular superior. A regra é válida também para uma matriz triangular inferior.

### Matrizes reais simétricas

As matrizes simétricas reais são de especial importância em álgebra linear que lida com problemas característicos, podendo ser demonstrado que:

- a) Os  $n$  autovalores de uma matriz simétrica real de ordem  $n$  são reais;
- b) Os autovetores que correspondem a distintos autovalores são ortogonais. Os autovetores correspondentes a múltiplas raízes podem ser ortogonalizados uns em relação aos outros;

c) Os  $n$  autovetores formam uma base ortogonal completa para o espaço Euclidiano  $E_n$ .

### Positividade

Seja  $A$  uma matriz quadrada  $n \times n$ .  $A$  é dita ser definida positiva se  $x'Ax > 0$ ,  $x \neq 0$ . Essa propriedade pode ser checada computando os  $n$  autovalores  $\lambda_i$  de  $Az = \lambda z$ . Se todos os  $\lambda_i > 0$ ,  $A$  é dita ser definida positiva.  $A$  é dita ser não negativa (semidefinida positiva) se  $x'Ax \geq 0$ ,  $x \neq 0$ . Uma matriz definida positiva é também não negativa, mas o inverso não é necessariamente verdadeiro. Essa propriedade pode ser checada computando os  $n$  autovalores  $\lambda_i$  de  $Az = \lambda z$ . Se  $r$  autovalores  $\lambda_i > 0$  e  $n-r$  autovalores são zero,  $A$  é não negativa com posto  $r$ . Uma matriz quadrada  $A$  é dita ser semidefinida negativa se  $x'Ax < 0$ ,  $x \neq 0$ . Uma matriz quadrada  $A$  que pode ter autovalores negativos, positivos ou nulos é dita ser indefinida.

### Matrizes normais e ortogonais

Seja  $A$  uma matriz quadrada  $n \times n$ . Essa matriz é chamada normal se  $A'A = AA'$ . Uma matriz normal é chamada ortogonal se  $A'A = AA' = I$  ou  $A' = A^{-1}$ .

Todos os autovalores de uma matriz ortogonal têm módulos um, e a matriz tem posto  $n$ .

### Algumas definições e resultados

Definição 1: Uma matriz quadrada  $Q$   $p$ -dimensional é ortogonal se  $QQ' = Q'Q = I_p$  ou equivalentemente  $Q' = Q^{-1}$ . Isso implica que as linhas e colunas de  $Q$  têm norma unitária e são ortogonais.

Definição 2: Uma matriz quadrada  $A$   $p$ -dimensional tem um autovalor  $\lambda$  com autovetor correspondente  $x \neq 0$  se  $Ax = \lambda x$ . Se o autovetor  $x$  é normalizado, o que significa que  $\|x\| = 1$ , então o autovetor normalizado será denotado por  $\varepsilon$ .

Resultado 1: Uma matriz quadrada simétrica  $A$   $p$ -dimensional tem  $p$  pares de autovalores e autovetores

$(\lambda_1 \varepsilon_1), (\lambda_2 \varepsilon_2), \dots, (\lambda_p \varepsilon_p)$ . Os autovetores podem ser escolhidos de forma a serem normalizados  $(\varepsilon'_1 \varepsilon_1 = \varepsilon'_2 \varepsilon_2 = \dots = \varepsilon'_p \varepsilon_p = 1)$  e ortogonais ( $\varepsilon'_i \varepsilon_j = 0$  se  $i \neq j$ ). Se todos os autovalores são diferentes, então os autovetores são únicos.

Resultado 2: A decomposição espectral de uma matriz quadrada simétrica  $A$   $p$ -dimensional é dada por  $A = \lambda_1 \varepsilon_1 \varepsilon'_1 + \lambda_2 \varepsilon_2 \varepsilon'_2 + \dots + \lambda_p \varepsilon_p \varepsilon'_p$ , em que  $(\lambda_1 \varepsilon_1), (\lambda_2 \varepsilon_2), \dots, (\lambda_p \varepsilon_p)$  são os pares autovalor / autovetor normalizado de  $A$ .

Definição 3: Uma matriz simétrica  $A$   $p \times p$  é chamada definida não negativa se  $0 \leq x'Ax$  para todo  $x' \in \mathbb{R}^p$ ; e  $A$  é chamada definida positiva se  $0 < x'Ax$  para todo  $x \neq 0$ . Segue (da decomposição espectral) que  $A$  é definida positiva se e somente se todos os autovalores de  $A$  são positivos e  $A$  é definida não negativa se e somente se todos os seus autovalores são maiores ou iguais a zero.

Resultado 3: A distância de Mahalanobis de um ponto é definida como  $d_S^2(x;0) = x' S^{-1} x$ , o que implica que todos os autovalores da matriz simétrica  $S^{-1}$  têm de ser positivos.

## Produtos Kronecker

Seja  $A = \{a_{ij}\}$  uma matriz  $m \times n$  e  $B = \{b_{kl}\}$  uma matriz  $p \times q$ , então, o produto Kronecker de  $A$  e  $B$ , denotado  $A \otimes B$ , é dado pela matriz  $mp \times nq$  seguinte:

$$\begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}$$

### Aplicação na expansão da matriz de variância para efeitos G x E

Em geral, todos os modelos para o estudo da interação G x E têm uma matriz de variância para os efeitos G x E da forma:

$Var(\eta) = A \otimes B$ , em que  $A$  é uma matriz de variância  $t \times t$  para a dimensão ambiente e  $B$  é uma matriz de variância  $m \times m$  para a dimensão genótipo. Considere  $m=4$  genótipos e  $t=2$  ambientes, então:

$$Var \begin{pmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{31} \\ \eta_{41} \\ \eta_{12} \\ \eta_{22} \\ \eta_{32} \\ \eta_{42} \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix}$$

Isto significa que a variância do efeito G x E para o genótipo  $i$  e ambiente  $j$  é dada por

$Var(\eta_{ij}) = a_{jj}b_{ii}$ , e a covariância entre efeitos para genótipo  $i$ , ambiente  $j$  e genótipo  $k$ , ambiente  $l$  é dada por:

$$Cov(\eta_{ij}, \eta_{kl}) = a_{jl}b_{ik}$$

### Aplicação na expansão da equação de modelos mistos para componente de variância

Na abordagem por ANOVA, clássica para a análise de dados de experimentos multi-ambientes (EMA), o modelo para o efeito de genótipo  $i$  no ambiente  $j$  é dado por

$\eta_{ij} = \mu + \alpha_i + \theta_j + \delta_{ij}$  (1), em que  $\mu$  é um efeito médio geral,  $\alpha_i$  é o efeito principal para o genótipo  $i$ ,  $\theta_j$  é o efeito principal para o ambiente  $j$  e  $\delta_{ij}$  é o efeito da interação entre o genótipo  $i$  e o ambiente  $j$ . O modelo na Equação (1) pode ser escrito na notação de vetor considerando o conjunto completo de efeitos genótipos por ambiente, ou seja, o vetor  $\eta$   $mtx1$ . Desta forma:

$\eta = 1_{mt} \mu + (1_t \otimes I_m) \alpha + (I_t \otimes 1_m) \theta + \delta$  (2), em que  $\alpha = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_m)'$  é o vetor  $mx1$  de efeitos principais de genótipos,  $\theta = (\theta_1 \ \theta_2 \ \dots \ \theta_t)'$  é o vetor  $tx1$  de efeitos principais de ambientes e  $\delta$  é o vetor  $mtx1$  de efeitos interação G x E (ordenado como para  $\eta$ ). É usada a notação padrão para vetores unidade e matrizes identidade, de forma que, para o exemplo, o vetor  $1_t$  denota o vetor unidade de comprimento  $t$  e a matriz  $I_t$  denota a matriz identidade  $txt$ .

O símbolo  $\otimes$  representa o produto Kronecker de duas matrizes ou vetores.

Para a expansão da equação do modelo, considere o exemplo específico da Equação (2) e assuma  $m=4$  genótipos e  $t=2$  ambientes, então:

$$\begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{31} \\ \eta_{41} \\ \eta_{12} \\ \eta_{22} \\ \eta_{32} \\ \eta_{42} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}$$

$$\begin{aligned}
& + \left( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \delta_{11} \\ \delta_{21} \\ \delta_{31} \\ \delta_{41} \\ \delta_{12} \\ \delta_{22} \\ \delta_{32} \\ \delta_{42} \end{bmatrix} \\
= & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \delta_{11} \\ \delta_{21} \\ \delta_{31} \\ \delta_{41} \\ \delta_{12} \\ \delta_{22} \\ \delta_{32} \\ \delta_{42} \end{bmatrix}
\end{aligned}$$

# CAPITULO 4

## Componentes de Variância

Componentes de variância são as variâncias associadas aos fatores de efeito aleatório.

### Fatores fixos e aleatórios e componentes de variância

Por exemplo, se os níveis do fator temperatura forem considerados como *fixos*, então, admite-se que os níveis de temperatura utilizados são os únicos de interesse para a pesquisa. O modelo empregado para níveis fixos é chamado de *modelo fixo*.

Quando os níveis de um fator são *aleatórios*, tais como lotes de sementes, operadores de máquinas, amostras de solo, linhagens de soja ou de milho, onde os níveis no experimento tenham sido escolhidos ao acaso a partir de um grande número de níveis possíveis, o modelo é chamado de *modelo aleatório*, e as inferências poderão ser estendidas a todos os níveis da população.

*Então se, um fator é considerado de efeito fixo significa que seus níveis num experimento são os únicos níveis de interesse.*

*Então, níveis aleatórios são escolhidos ao acaso de um conjunto grande ou infinito de níveis.*

No caso do uso de um modelo aleatório, o pesquisador frequentemente está interessado na estimativa de *componentes de variância*.

### Métodos de estimação de componentes de variância

Componentes de variância são as variâncias associadas aos efeitos aleatórios de um modelo. Nos modelos mistos, a solução das equações de modelos mistos (MME) depende do conhecimento da matriz de variâncias-covariâncias  $V$ , cuja estrutura é conhecida, mas seus componentes não o são. Assim, torna-se necessário substituí-los por suas estimativas.

Existem vários métodos de estimação de componentes de variância, sendo os principais os seguintes: Método da Análise de Variância, Métodos de Henderson, MINQUEO, MIVQUE, Máxima Verossimilhança (ML) e Máxima Verossimilhança Restrita (REML).

### Dados balanceados

Nesse caso utiliza-se o método da análise de variância (ou método dos momentos), que consiste em obter os estimadores igualando-se os quadrados médios de um quadro de análise de variância aos seus respectivos valores esperados, que são combinações lineares dos componentes de variância. Portanto, esse método produz equações lineares dos componentes de variância, cujas soluções são tomadas como os estimadores dos referidos componentes.

A aplicação do método da análise de variância é direta e os cálculos exigidos são fáceis. Além disso, apenas as suposições básicas sobre a matriz variâncias-covariâncias são exigidas. Os estimadores obtidos são sempre não-viesados e de variância mínima, mas tem a desvantagem de não excluir a ocorrência de estimativas negativas. Este método já foi estudado com detalhes anteriormente.

### Dados desbalanceados

A principal dificuldade nesse caso é que existem muitos métodos de estimação disponíveis e a escolha do mais adequado a cada situação não é uma tarefa muito simples. Devido às facilidades

computacionais, atualmente a escolha tem sido feita entre os dois métodos baseados na máxima verossimilhança. Os principais métodos disponíveis são os seguintes:

Método de Análise de Variância, Método de Henderson I, Método de Henderson II, Método de Henderson III, Máxima Verossimilhança (ML), Máxima Verossimilhança Restrita (REML), Estimador Quadrático Não-Viesado de Norma Mínima (MINQUE), Estimador Quadrático Não-Viesado de Variância Mínima (MIVQUE), Estimador Quadrático Não-Viesado de Norma Mínima Iterativo (I-MINQUE).

### Métodos de Análise de Variância

O princípio do método de análise de variância pode ser generalizado para dados desbalanceados por meio do uso de qualquer outra forma quadrática em lugar das somas de quadrados.

Considere o vetor de componentes de variância que serão estimados e seja  $q$  um vetor da mesma ordem de  $\sigma^2$ , de qualquer forma quadrática linearmente independente das observações. Suponha que o vetor  $q$  seja tal que  $E(q) = C\sigma^2$ , para alguma matriz  $C$  não-singular, então  $\sigma^2 = C^{-1}q$  é um estimador não-viesado de  $\sigma^2$ . A matriz de dispersão de  $\hat{\sigma}^2$  é dada por

$Var(\hat{\sigma}^2) = C^{-1}[Var(q)](C^{-1})'$ , em que os elementos de  $Var(q)$  são variâncias e covariâncias das formas quadráticas usadas como elementos de  $q$ .

### Método I de Henderson

Os três métodos propostos por Henderson (1953) são três formas de uso do método de análise de variância geral. Eles diferem apenas nas formas quadráticas utilizadas e também podem produzir estimativas negativas.

No método I, as formas quadráticas são análogas às somas de quadrados usadas em dados balanceados, de tal forma que nem sempre são não-negativas devido à estrutura não-balanceada dos dados. Desta forma, para o modelo:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \text{ com}$$

$i=1,2,\dots,I; j=1,2,\dots,J; k=1,2,\dots,n$ , a soma de quadrados para dados desbalanceados é dada por:

$\sum_i n_{i..} (\bar{y}_{i..} - \bar{y}_{...})^2 = \sum_i n_{i..} \bar{y}_{i..}^2 - n_{...} \bar{y}_{...}^2$ . A soma de quadrados para a interação, para dados desbalanceados é dada por:

$$\sum_i \sum_j n_{ij..} (\bar{y}_{ij..} - \bar{y}_{i..} - \bar{y}_{.j..} + \bar{y}_{...})^2 =$$

$$= \sum_i \sum_j n_{ij} \bar{y}_{ij}^2 - \sum_i n_{i..} \bar{y}_{i..}^2 - \sum_j n_{.j} \bar{y}_{.j}^2 + n_{...} \bar{y}_{...}^2.$$

O método I de Henderson consiste em igualar os quadrados médios às suas respectivas esperanças matemáticas, fornecendo estimativas não-viesadas, com variância mínima quando os dados são não-balanceados ou o modelo é aleatório e os efeitos não-correlacionados. Esse método não pode ser usado para modelos mistos.

### Método II de Henderson

Esse método foi proposto para ampliar o uso do Método I removendo sua limitação de não poder ser usado para modelos mistos. O método consta de duas partes. Primeiro supõe, temporariamente, que os efeitos aleatórios são fixados, e com base no modelo  $y = X\beta + Z\gamma + e$ , resolve as equações normais:

$$\begin{bmatrix} XX & XZ \\ ZX & ZZ \end{bmatrix} \begin{bmatrix} \hat{\beta}^0 \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}, \text{ para } \beta^0. \text{ Em seguida considera o vetor de dados ajustados para } \beta^0, \text{ ou seja, } z = y - X\beta^0 \text{ e então se aplica o Método I para } z.$$

Dessa forma, o Método II de Henderson consiste em primeiro estimar os efeitos fixos, e então aplicar o Método I para os resíduos restantes. Para que os estimadores resultantes sejam não-viesados é necessário que os resíduos dependam apenas dos fatores aleatórios. Esse método apresenta os inconvenientes de não haver uma solução única e de não poder adotar modelos com interações entre os efeitos fixos e aleatórios.

### Método III de Henderson

O Método III, também chamado de método de ajuste de constantes, usa as reduções nas somas de quadrados do modelo completo e de submodelos para estimar os componentes de variância. Considere o modelo seguinte:

$y = X\beta^0 + Z\gamma + e = W\theta + e$ . A matriz  $W$  pode ser particionada como  $[W_1 | W_2]$ , e  $\theta'$  pode ser particionada como  $[\theta'_1 | \theta'_2]$  de acordo com  $W$ , ou seja, o modelo pode ser reescrito como:

$y = W_1\theta_1 + W_2\theta_2 + e$ . Note que nenhuma suposição é feita sobre o particionamento de  $W$  e  $\theta$  no que se refere aos efeitos fixos ou aleatórios.

Denominando  $R(\theta_1, \theta_2)$  e  $R(\theta_1)$ , respectivamente, de reduções nas somas de quadrados do modelo completo e do submodelo,  $y = W_1\theta_1 + e$ , tem-se:

$$R(\theta_1 | \theta_2) = R(\theta_1, \theta_2) - R(\theta_1) \text{ e } R(\theta_2 | \theta_1) = R(\theta_1, \theta_2) - R(\theta_2), \text{ sendo}$$

$$E[R(\theta_2|\theta_1)] = \text{tr} \left\{ W_2' \left[ I - W_1 (W_1' W_1)^{-1} W_1' \right] W_2 E(\theta_2 \theta_2') \right\} + \\ + \sigma_e^2 [r(W_1) - r(W_2)], \text{ onde } r(W_1) \text{ é o posto da submatriz } W_1 \text{ e } r(W_2) \text{ o posto de } W_2.$$

Observe que  $R(\theta_2|\theta_1)$  não envolve  $\theta_1$  e portanto  $E[R(\theta_2|\theta_1)]$  não depende do vetor

de efeitos  $\theta_1$ , sejam eles fixos ou aleatórios. Portanto, o Método III consiste em obter os estimadores para os componentes de variância, construindo um sistema de equações a partir das diferenças entre as reduções do modelo completo e um submodelo, e assim igualando-as às suas respectivas esperanças.

Esse método é particularmente vantajoso para modelos mistos, porque se tomando  $\theta_1$  como o vetor dos efeitos fixos e  $\theta_2$  como vetor dos efeitos aleatórios, a  $E[R(\theta_2|\theta_1)]$  não conterá termos devido a esses efeitos fixos, ou seja, a esperança é função apenas de  $\sigma_e^2$  e das variâncias dos efeitos aleatórios em  $\theta_2$ , que são os próprios componentes que se deseja estimar.

Para ilustrar o Método III, considere o modelo a seguir:

$$y = \mu 1 + X_1 \alpha + X_2 \beta + X_3 \gamma + e, \text{ em que:}$$

$\mu$ : constante;

$\alpha$ : vetor de efeitos fixos;

$\beta$  e  $\gamma$ : vetores de efeitos aleatórios.

Nesse caso, a matriz  $W$  pode ser escrita como:

$$W = \begin{bmatrix} 1 & X_1 & X_2 & X_3 \end{bmatrix} \text{ e} \\ R(\mu, \alpha, \beta, \gamma) = y' W (W' W)^{-1} W' y, \text{ com } r(W) = r.$$

Considere os submodelos a seguir:

$$y = \mu 1 + e$$

$$y = \mu 1 + X_1 \alpha + e$$

$$y = \mu 1 + X_1 \alpha + X_2 \beta + e$$

Sejam as reduções correspondentes dadas por:

$$R(\mu) = y' \mathbf{1} (\mathbf{1}' \mathbf{1})^{-1} \mathbf{1}' y = y' \mathbf{1} (n)^{-1} \mathbf{1}' y = \frac{1}{n} y' J y, \text{ com } r(W_1) = r(J) = 1$$

$$R(\mu, \alpha) = y' W_1 (W_1' W_1)^{-1} W_1' y, \text{ com } W_1 = \begin{bmatrix} 1 & X_1 \end{bmatrix} \text{ e } r(W_1) = q$$

$$R(\mu, \alpha, \beta) = y' W_1 (W_1' W_1)^{-1} W_1' y, \text{ com } W_1 = \begin{bmatrix} 1 & X_1 & X_2 \end{bmatrix} \text{ e } r(W_1) = s.$$

Então, os componentes de variância são obtidos sucessivamente com base no conjunto de equações seguintes:

Somas de Quadrados	Esperanças Matemáticas
$SQE = \sum y^2 - R(\mu, \alpha, \beta, \gamma)$	$(n-r)\sigma_e^2$
$R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta)$	$h_1 \sigma_\gamma^2 + (r-s)\sigma_e^2$
$R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha)$	$h_2 \sigma_\beta^2 + h_3 \sigma_\gamma^2 + (r-q)\sigma_e^2$

A partir dessas três equações são calculados os valores  $\hat{\sigma}_e^2$ ,  $\hat{\sigma}_\beta^2$  e  $\hat{\sigma}_\gamma^2$ . Os fatores  $h_1$ ,  $h_2$  e  $h_3$  são obtidos com base na expressão:

$$\begin{aligned} E[R(\theta_2 | \theta_1)] &= tr \left\{ W_2' \left[ I - W_1 (W_1' W_1)^{-1} W_1' \right] W_2 E(\theta_2 \theta_2') \right\} + \\ &+ I \hat{\sigma}_e^2 [r(W) - r(W_1)], \text{ em que as matrizes } W_1 \text{ e } W_2 \text{ são especificadas para cada equação.} \end{aligned}$$

Observe que não é necessário usar a quarta equação,  $R(\mu, \alpha, \beta, \gamma) - R(\mu)$ , cuja esperança seria  $h_4 \sigma_\alpha^2 + h_5 \sigma_\beta^2 + h_6 \sigma_\gamma^2 + (n-1)\sigma_e^2$ , uma vez que, supondo-se  $\alpha$  como efeito fixo, não se considera a existência de  $\sigma_\alpha^2$ .

O Método III pode ser usado para qualquer modelo misto e produz estimadores não-viesados.

### Método da Máxima Verossimilhança (ML)

O método da máxima verossimilhança (ML) consiste em maximizar a função densidade de probabilidade das observações, em relação aos efeitos fixos e aos componentes de variância. Considere o modelo linear misto seguinte:

$y = X\beta + Z\gamma + e$ . Admitindo que os efeitos aleatórios  $\gamma_i$ ,  $i=1, \dots, r$  e o termo erro  $e$  têm distribuição normal, com média zero e matrizes de variâncias-covariâncias  $\sigma_i^2 I_m$  e  $\sigma_e^2 I_n$  respectivamente, o vetor  $y$  terá distribuição normal multivariada, com média  $X\beta$  e matriz de variâncias-covariâncias  $V$ , ou seja,  $y \sim N(X\beta, V)$ , sendo:

$$V = \sum_{i=1}^r Z_i Z_i' \sigma_i^2 + \sigma_e^2 I = \sum_{i=0}^r Z_i Z_i' \sigma_i^2, \text{ com } \sigma_0^2 = \sigma_e^2 \text{ e } Z_0 = I.$$

A função de verossimilhança é expressa como:

$$L = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\beta)' V^{-1} (y - X\beta)\right], \text{ sendo } |V| \text{ o determinante da matriz } V.$$

Maximizando  $L$  em relação aos elementos de  $\beta$  e aos componentes de variância, os  $\sigma_i^2$  que ocorrem em  $V$ , obtém-se um sistema de equações que, uma vez resolvido, produz os estimadores de máxima verossimilhança de  $\beta$  e de  $\sigma^2 = \{\sigma_i^2\}_{i=0}^{i=r}$ . Então, o modelo pode ser reescrito como:

$$X \tilde{V}^{-1} X \tilde{\beta} = X \tilde{V}^{-1} y, \text{ e as equações são:}$$

$$\operatorname{tr}\left(\tilde{V}^{-1} Z_i Z_i'\right) = (y - X \tilde{\beta})' \tilde{V}^{-1} Z_i Z_i' \tilde{V}^{-1} (y - X \tilde{\beta}), \text{ para } i=1, \dots, r.$$

As equações anteriores devem ser resolvidas para  $\tilde{\beta}$  e  $\tilde{\sigma}^2$ , os elementos implícitos em  $\tilde{V}$ . Obviamente que essas equações não são lineares nos elementos  $\tilde{\sigma}^2$ , entretanto, uma vez obtidos os valores  $\tilde{\sigma}_i^2$ , eles podem ser usados para obter  $\tilde{\beta}$ . Essas equações são resolvidas numericamente por iteração. Por conveniência pode-se escrever:

$$P = V^{-1} - V^{-1} X \left( X V^{-1} X \right) X V^{-1} \text{ e}$$

$$I = V^{-1} V = V^{-1} \sum_{i=0}^r Z_i Z_i' \sigma_i^2. \text{ Desta forma, o conjunto das } r+1 \text{ equações anteriores pode ser descrito como:}$$

$tr\left(\tilde{V}^{-1}Z_i Z'_i \tilde{V}^{-1}Z_j Z'_j\right) \tilde{\sigma}_i^2 = \left(y' \tilde{P} Z_i Z'_i \tilde{P} y\right)$ , o que fornece uma visualização mais fácil de

um processo iterativo. Pode-se utilizar um valor inicial para  $\tilde{\sigma}^2$  em  $\tilde{V}$  e  $\tilde{P}$ , resolver as equações anteriores e repetir o processo até que o critério de convergência seja satisfeito.

O Método ML é iterativo e sempre fornece estimativas não-negativas de componentes de variância, mas estas são viesadas porque o método não considera a perda de graus de liberdade resultante da estimação dos efeitos fixos do modelo.

### Método da Máxima Verossimilhança Restrita (REML)

Os estimadores REML são obtidos maximizando a parte da função de verossimilhança que é invariante para o parâmetro de locação. Neste caso, o modelo misto linear geral  $y = X\beta + Z\gamma + e$  é invariante para  $X\beta$ . Outro conceito, os estimadores REML maximizam a função de verossimilhança de um vetor de combinações lineares das observações que são invariantes para  $X\beta$ . Seja  $Ly$  esse vetor, então  $Ly = LX\beta + LZ\gamma + Le$  é invariante para  $X\beta$ , se e somente se,  $LX = 0$ . Mas,  $LX = 0$ , se e somente se,  $L = TM$ , para

$M = I - X(X'X)^{-1}X'$  e algum  $T$ . Claramente que  $L$  deve ser de posto-linha completo e assim também  $T$ . Portanto,  $r_L = r_T$  e  $r_L \leq r_M$ , com  $r_M = n - r_X$ .

As equações para a estimação REML de  $\sigma^2$ , para  $i, j = 0, 1, \dots, r$  são dadas por:

$tr\left(\tilde{P} Z_i Z'_i \tilde{P} Z_j Z'_j\right) \sigma_i^2 = \left(y' \tilde{P} Z_i Z'_i \tilde{P} y\right)$ . Observe que essas equações são similares às equações ML, exceto para  $\tilde{P}$  em vez de  $\tilde{V}^{-1}$ .

No Método REML, cada observação é dividida em duas partes independentes, uma referente aos efeitos fixos e outra aos efeitos aleatórios, de forma que a função densidade de probabilidade das observações é dada pela soma das funções densidade de probabilidade de cada parte. A maximização da função densidade de probabilidade da parte referente aos efeitos aleatórios, em relação aos componentes de variância, elimina o viés resultante da perda de graus de liberdade na estimação dos efeitos fixos do modelo.

As equações REML para dados balanceados são idênticas aos estimadores de análise de variância, que são não-viesados e de variância mínima. O estimador REML leva em conta os graus de liberdade envolvidos nas estimativas dos efeitos fixos, enquanto que os estimadores ML não. Para dados desbalanceados os estimadores ML e os estimadores REML são viesados.

Os estimadores ML e REML dos componentes de variância não são formas explícitas, ou seja, o estimador de cada componente está em função dos estimadores dos outros componentes, e podem ser encontrados somente por métodos numéricos iterativos.

### Método do Estimador Quadrático Não-Viesado de Norma Mínima (MINQUE)

Esse método de estimação é derivado de forma que o estimador minimize a norma euclidiana da matriz núcleo, que seja uma forma quadrática das observações e que seja não-viesado. Utiliza valores escolhidos previamente para os componentes de variância desconhecidos. A estimação dos componentes de variância é feita com base na equação MINQUE, descrita a seguir:

$$tr(P_w V_i P_w V_j) \hat{\sigma}_i^2 = (y' P_w V_i P_w y), \text{ em que:}$$

$\hat{\sigma}$ : vetor de componentes de variância;

$$P_w = V_w^{-1} - V_w^{-1} \left( X V_w^{-1} X \right)^{-1} X V_w^{-1};$$

$V_w$ : estimativa a priori da matriz de variâncias-covariâncias.

Esse método tem as vantagens de não envolver a suposição de normalidade como o ML e REML e das equações de MINQUE terem soluções explícitas, ou seja, não terem de ser resolvidas iterativamente. Entretanto, a solução depende do conhecimento a priori dos valores dos componentes de variância a serem estimados, usados em  $V_w$ . Assim, diferentes valores de  $V_w$  podem levar a obtenção de diferentes estimativas para um mesmo conjunto de dados. Portanto, é obtido “um” estimador MINQUE e não “o” estimador MINQUE.

Um relacionamento importante entre o REML e o MINQUE é que se o valor inicial no processo iterativo REML é  $V_w$ , então a primeira solução é uma estimativa MINQUE.

#### Método do Estimador Quadrático Não-Viesado de Variância Mínima (MIVQUE)

Se a suposição usual de normalidade da variável aleatória  $y$  é satisfeita, o estimador MINQUE apresenta a propriedade de ser uma forma quadrática não-viesada com variância mínima das observações, ou seja, é um estimador não-viesado de variância mínima ou MIVQUE.

O estimador MIVQUE(A) usa as equações REML tomando as estimativas de análise de variância como valores a priori. Embora a teoria MIVQUE especifique que os valores a priori devam ser independentes, o uso das estimativas de análise de variância pode ser justificado pela facilidade de obtenção. O estimador MIVQUEO é o MIVQUE com a suposição a priori de que a matriz de variâncias-covariâncias é a matriz identidade.

#### Método MINQUE Iterativo (I-MINQUE)

O estimador MINQUE utiliza uma estimativa a priori para  $V$  (matriz variâncias-covariâncias) e nenhuma iteração é envolvida. Entretanto, obtida uma solução, por exemplo,  $\tilde{V}_1$ , pode-se usá-la como uma nova estimativa em  $V_w$  a partir da qual um novo conjunto de equações pode ser estabelecido e resolvido, produzindo  $\tilde{V}_2$  e assim sucessivamente. Isto leva ao uso das equações MINQUE de forma iterativa.

### Estimação de componentes de variância

Exemplo 1-A seguir será considerado um exemplo que analisa e interpreta componentes de variância em um modelo aleatório.

*Os dados para o exemplo são:* Considere uma empresa de sementes que abastece uma revendedora com um grande número de lotes de sementes. A revendedora faz determinações em três amostras

de cada cinco lotes selecionados aleatoriamente, para controlar a qualidade das sementes recebidas. O modelo para análise dos dados obtidos é:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ , com  $i=1,2,\dots,k$ , em que os  $k$  níveis (no exemplo, os lotes) são escolhidos ao acaso (aleatórios) de uma população com variância  $\sigma_\tau^2$ . Os dados são apresentados a seguir:

Lotes				
1	2	3	4	5
74	68	75	72	79
76	71	77	74	81
75	72	77	73	79

Uma análise de variância de um fator é realizada com dados. *O quadro de Resumo da Análise de Variância para o exemplo é o seguinte:*

Fonte	SQ	GL	QM	E (QM)
Lotes	147,74	4	36,935	$\sigma_\varepsilon^2 + 3\sigma_\tau^2$
Resíduo	17,99	10	1,799	$\sigma_\varepsilon^2$
Total	165,73	14		

A interpretação do quadro de Análise de Variância é como a seguir: as computações que produzem as SQ são as mesmas, tanto para modelo efeitos fixos quanto para modelo efeitos aleatórios.

Entretanto, para o modelo aleatório, o quadrado médio de lotes é uma estimativa de  $\sigma_\varepsilon^2 + 3\sigma_\tau^2$ . Isto é mostrado na coluna Esperança de Quadrados Médios [E (QM)] do quadro de Análise de Variância. A estatística  $F = 36,94 / 1,80 = 20,5 > F_{0,05(4;10)} = 5,99$ .

*Método de Análise de Variância:* uma vez que a estatística do teste é maior que o valor crítico, rejeita-se a hipótese de médias iguais para lotes. Uma vez que estes lotes foram escolhidos por meio de um processo de seleção aleatória, pode-se estar interessado em encontrar quanto da variância no experimento pode ser atribuída às diferenças entre lotes e quanto a erro aleatório. A estimativa da variância devida a erro (resíduo),  $\sigma_\varepsilon^2$ , é 1,80 e o quadrado médio de lote computado 36,94 é uma

estimativa de  $\sigma_\varepsilon^2 + 3\sigma_\tau^2$ . Estabelecendo a igualdade entre os valores de QM e as expressões de E (QM), o que é chamado de método de momentos ou da análise de variância, se obtém:

$s_\varepsilon^2 = 1,80$  e  $s_\varepsilon^2 + 3s_\tau^2 = 36,94$ , onde os  $s_\varepsilon^2$  são estimadores de variância residual  $\sigma^2$ .

A computação do componente de variância para lote é feita resolvendo a expressão seguinte:

$s_{\tau}^2 = \frac{36,94 - 1,80}{3} = 11,71$ . A variância total pode ser estimada como:

$$s_{total}^2 = s_{\tau}^2 + s_{\varepsilon}^2 = 11,71 + 1,80 = 13,51.$$

A interpretação do resultado pode ser como a seguir: em termos de porcentagens, pode-se observar que  $11,71 / 13,51 (100) = 86,7\%$  da variância total é atribuída a diferenças entre lotes e  $13,3\%$  a variância de erro dentro de lotes.

Exemplo 2-Componentes de variância para fatores aleatórios (Extraído de Cruz et al., 2004):

Considere um modelo que descreve as observações obtidas da avaliação de  $g$  genótipos em um experimento blocos casualizados com  $r$  blocos. O modelo pode ser escrito como:

$$Y_{ij} = m + g_i + b_j + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : observação da parcela do  $i$ -ésimo genótipo no  $j$ -ésimo bloco;

$m$ : média geral do experimento;

$g_i$ : efeito do  $i$ -ésimo genótipo, considerado aleatório, sendo  $E(g_i) = 0$ ,  $E(g_i^2) = \sigma_g^2$  e

$$E(g_i, g_{i'}) = 0;$$

$b_j$ : efeito do  $j$ -ésimo bloco, considerado aleatório, sendo  $E(b_j) = 0$ ,  $E(b_j^2) = \sigma_b^2$  e

$$E(b_j, b_{j'}) = 0;$$

$\varepsilon_{ij}$ : efeito do erro aleatório associado à observação  $Y_{ij}$ , sendo  $E(\varepsilon_{ij}) = 0$ ,  $E(\varepsilon_{ij}^2) = \sigma^2$  e

$\varepsilon_{ij} \sim NID(0, \sigma^2)$ . Considera-se ainda que os efeitos aleatórios (genótipos e blocos) sejam independentes entre si.

O esquema de análise de variância é como a seguir:

FV	GL	SQ	QM
Blocos	$r - 1$	SQB	QMB
Genótipos	$g - 1$	SQG	QMG
Resíduo	$(r - 1)(g - 1)$	SQR	QMR
Total	$rg - 1$	SQTotal	

Os dados, de uma variável hipotética, são referentes à avaliação de dez famílias de meios-irmãos no delineamento blocos casualizados com três repetições (blocos):

Família	Bloco 1	Bloco 2	Bloco 3
1	132	127	128
2	137	161	164
3	137	165	121
4	168	166	161
5	140	161	135
6	112	125	113
7	132	141	143
8	139	144	165
9	154	153	185
10	143	167	160

O resultado da análise de variância é o seguinte:

FV	GL	SQ	QM	F
Blocos	2	708,0667	354,0333	2,39 <sup>ns</sup>
Genótipos	9	6307,6333	700,8481	4,73 <sup>**</sup>
Resíduo	18	2663,2667	147,9592	
Total	29	9678,9667		

\*\*: Significativo a 1% de probabilidade pelo teste F.

ns: Não significativo a 5% de probabilidade pelo teste F.

Como nesse caso o *modelo é aleatório*, a hipótese testada é  $H_0: \sigma_g^2 = 0$  e a significância para o teste F indica a rejeição de  $H_0$  e, portanto, a existência de variabilidade genotípica entre as médias das famílias. A potencialidade das famílias pode ser avaliada por meio da média geral,  $m=4379/30=145,96$ , e a precisão experimental por meio do coeficiente de variação,  $CV\% = \frac{100\sqrt{QMR}}{m} = \frac{100\sqrt{147,9592}}{145,96} = 8,33$ .

As esperanças de quadrados médios das fontes de variação são fornecidas por:

A esperança matemática do quadrado médio de blocos é:

$$E(QMB) = E\left(\frac{SQB}{r-1}\right) = \sigma^2 + g\sigma_b^2.$$

A esperança matemática do quadrado médio de tratamentos (genótipos) é:

$$E(QMG) = E\left(\frac{SQG}{g-1}\right) = \sigma^2 + r\sigma_g^2.$$

A esperança matemática do quadrado médio de resíduo é:

$$E(QMR) = E\left(\frac{SQR}{(r-1)(g-1)}\right) = \sigma^2.$$

A utilização conjunta do resultado da análise de variância do experimento e das expressões de esperanças de quadrados médios possibilita a *estimação de componentes de variâncias genéticas e ambientais*, que são informações de grande utilidade nos programas de melhoramento de plantas. Considerando o modelo aleatório, o esquema de análise de variância com as respectivas esperanças dos quadrados médios é o seguinte:

FV	GL	QM	E (QM)
Blocos	$r - 1$	QMB	$\sigma^2 + g\sigma_b^2$
Genótipos	$g - 1$	QMG	$\sigma^2 + r\sigma_g^2$
Resíduo	$(r - 1)(g - 1)$	QMR	$\sigma^2$

A hipótese  $H_0: \sigma_g^2 = 0$  é avaliada por meio da estatística  $F = \frac{QMG}{QMR}$ , associada a  $(g-1)$  e  $(r-1)(g-1)$  graus de liberdade.

O estimador do componente de variância genética entre médias das famílias avaliadas é dado por:

$$\hat{\sigma}_g^2 = \frac{QMG - QMR}{r} = \frac{\sigma^2 + r\sigma_g^2 - \sigma^2}{r}.$$

O estimador do componente de variância ambiental, em nível de parcelas, é dado por:

$$\hat{\sigma}^2 = QMR.$$

O coeficiente de herdabilidade, em nível de médias de famílias, é dado por:

$$h^2 = \frac{\hat{\sigma}_g^2}{QMG/r} = \frac{\hat{\sigma}_g^2}{\frac{\hat{\sigma}^2}{r} + \hat{\sigma}_g^2}.$$

O *coeficiente de correlação intraclass*, que quantifica a repetibilidade do desempenho dos genótipos no experimento, é dado por:

$$r = \frac{\hat{Cov}(Y_{ij}, Y_{ij'})}{\sqrt{\hat{V}(Y_{ij})\hat{V}(Y_{ij'})}}, \text{ sendo } \hat{Cov}(Y_{ij}, Y_{ij'}) = \hat{\sigma}_g^2 \text{ e}$$

$$\hat{V}(Y_{ij}) = \hat{V}(Y_{ij'}) = \hat{\sigma}^2 + \hat{\sigma}_g^2. \text{ Então, } r = \frac{\hat{\sigma}_g^2}{\hat{\sigma}^2 + \hat{\sigma}_g^2}.$$

Alternativamente, o valor da variância genotípica pode ser obtido por meio da média das covariâncias entre os valores obtidos considerando pares de blocos, ou seja:

Pares de Blocos	Covariância	Correlação
1 e 2	159,0000	0,6843
1 e 3	234,5556	0,6980
2 e 3	159,3333	0,4345
Média	184,2963	0,6056

$$\text{Então, } \hat{\sigma}_g^2 = 184,2963 \text{ e } \hat{\sigma}^2 = QMR = 147,9592.$$

O coeficiente de herdabilidade é obtido por meio de:

$$h^2 = \frac{\hat{\sigma}_g^2}{QMG/r} = \frac{184,2963}{700,8481/3} = 0,7889.$$

O coeficiente de correlação intraclass é obtido por meio de:

$$r = \frac{\hat{\sigma}_g^2}{\hat{\sigma}^2 + \hat{\sigma}_g^2} = \frac{184,2963}{147,9592 + 184,2963} = 0,5547.$$

Exemplo 3-Estimativa de componentes de variância para o caso de blocos casualizados com informação de indivíduos dentro de parcela (Extraído de Cruz et al.,2004)

Considere a avaliação de  $g$  genótipos em blocos casualizados com  $r$  repetições. As observações obtidas nos  $n$  indivíduos que compõem a parcela, são modeladas por:

$$Y_{ijk} = m + g_i + b_j + \varepsilon_{ij} + \delta_{ijk}, \text{ em que:}$$

$Y_{ijk}$ : observação obtida no k-ésimo indivíduo do i-ésimo genótipo no j-ésimo bloco;

$m$ : média geral do experimento;

$g_i$ : efeito do i-ésimo genótipo, considerado aleatório, sendo:

$$E(g_i) = 0, E(g_i^2) = \sigma_g^2 \text{ e } E(g_i, g_{i'}) = 0;$$

$b_j$ : efeito do j-ésimo bloco, considerado aleatório, sendo:

$$E(b_j) = 0, E(b_j^2) = \sigma_b^2 \text{ e } E(b_j, b_{j'}) = 0;$$

$\varepsilon_{ij}$ : efeito aleatório da variação entre parcelas experimentais, sendo:

$$E(\varepsilon_{ij}) = 0, E(\varepsilon_{ij}^2) = \sigma^2 \text{ e } E(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0;$$

$\delta_{ijk}$ : efeito aleatório da variação entre plantas dentro da parcela, sendo:

$$E(\delta_{ijk}) = 0, E(\delta_{ijk}^2) = \sigma_d^2 \text{ e } E(\delta_{ijk}, \delta_{i'j'k'}) = 0.$$

O esquema de análise de variância para o modelo estatístico adotado é o seguinte:

FV	GL	SQ	QM
Blocos	$r - 1$	SQB	QMB
Genótipos	$g - 1$	SQG	QMG
Entre parcelas	$(r - 1)(g - 1)$	SQE	QME
Dentro de parcelas	$(n - 1)rg$	SQD	QMD

As somas de quadrados são obtidas por meio dos seguintes estimadores:

$$SQT_{Total} = \sum_i \sum_j \sum_k Y_{ijk}^2 - C; SQB = \frac{1}{ng} \sum_j Y_{.j.}^2 - C;$$

$$SQG = \frac{1}{nr} \sum_i Y_{i..}^2 - C;$$

$$SQE = \frac{1}{n} \sum_i \sum_j Y_{ij.}^2 - \frac{1}{ng} \sum_j Y_{.j.}^2 - \frac{1}{nr} \sum_i Y_{i..}^2 + C;$$

$$SQD = \sum_i \sum_j \sum_k Y_{ijk}^2 - \frac{1}{n} \sum_i \sum_j Y_{ij.}^2, \text{ sendo } C = \frac{Y_{...}^2}{ngr}$$

$$Y_{...} = \sum_i \sum_j \sum_k Y_{ijk}.$$

Considere como ilustração, a avaliação de 10 famílias num experimento em blocos casualizados, com duas repetições, tendo quatro indivíduos dentro da parcela. Os valores obtidos para uma variável hipotética são:

Família	Bloco1				Bloco2			
	P1	P2	P3	P4	P1	P2	P3	P4
1	62,4	61,1	60,1	53,8	67,5	66,3	65,2	59,2
2	59,8	54,5	52,7	49,0	53,4	50,0	47,3	46,5
3	49,2	42,4	41,0	37,4	62,1	61,6	59,1	48,4
4	60,8	58,9	51,2	50,5	47,2	40,1	36,7	36,3
5	56,6	46,7	46,4	43,8	46,8	45,7	44,4	41,2
6	54,5	52,2	48,9	40,3	50,1	46,8	39,9	38,0
7	41,4	39,0	39,1	38,2	63,9	53,9	48,9	44,6
8	53,0	47,4	46,9	43,8	44,3	42,5	40,5	37,8
9	57,0	47,8	45,8	39,6	48,6	46,8	33,9	32,2
10	56,5	50,2	43,9	41,2	39,5	37,6	37,6	32,7

O resultado da análise de variância é o seguinte:

FV	GL	SQ	QM	F
Blocos	1	79,7844	79,7844	
Genótipos	9	2247,1094	249,6788	1,2245 <sup>ns</sup>
Entre parcelas	9	1831,6062	203,5118	
Dentro de parcelas	60	1675,2344	27,9206	

ns: não significativo ( $p > 0,05$ ).

Não foi detectada variabilidade genética entre famílias, o que indica que a seleção de famílias não terá êxito. Entretanto, é necessário avaliar a variação genética existente em nível de plantas dentro das famílias. Se a herdabilidade, em nível de plantas, for elevada pode-se praticar a seleção dentro de família e obter consideráveis ganhos genéticos.

As esperanças de quadrados médios podem ser obtidas por:

Esperança matemática de Quadrado Médio de Blocos:

$$E(QMB) = E\left(\frac{SQB}{b-1}\right) = \sigma_d^2 + n\sigma_e^2 + ng\sigma_b^2.$$

Esperança matemática de Quadrado Médio de Genótipos:

$$E(QMG) = E\left(\frac{SQG}{g-1}\right) = \sigma_d^2 + n\sigma_e^2 + nr\sigma_g^2.$$

Esperança matemática de Quadrado Médio de Erro Entre Parcelas:

$$E(QME) = E\left(\frac{SQE}{(r-1)(g-1)}\right) = \sigma_d^2 + n\sigma_e^2$$

Esperança matemática de Quadrado Médio de Erro Dentro de Parcelas:

$$E(QMD) = E\left(\frac{SQD}{(n-1)gr}\right) = \sigma_d^2.$$

A *estimação de componentes de variâncias genéticas e ambientais* pode ser realizada com base no esquema de análise de variância seguinte:

FV	GL	QM	E (QM)
Blocos	$r - 1$	QMB	$\sigma_d^2 + n\sigma_e^2 + ng\sigma_b^2$
Genótipos	$g - 1$	QMG	$\sigma_d^2 + n\sigma_e^2 + nr\sigma_g^2$
Entre parcelas	$(r - 1)(g - 1)$	QME	$\sigma_d^2 + n\sigma_e^2$
Dentro de parcelas	$(n - 1)gr$	QMD	$\sigma_d^2$

Nesse estudo devem ser obtidas as *estimativas de componentes de variância* seguintes:

$\sigma_b^2$ : *componente de variância ambiental*, que mede as variações entre blocos, estimado por:

$$\hat{\sigma}_b^2 = \frac{QMB - QME}{ng} = \frac{\sigma_d^2 + n\sigma_e^2 + ng\sigma_b^2 - \sigma_d^2 + n\sigma_e^2}{ng}, \text{ então,}$$

$$\hat{\sigma}_b^2 = \frac{79,7844 - 203,5118}{40} = -3,0931;$$

$\sigma_g^2$ : componente de variância genética entre médias de genótipos (famílias), estimado por:

$$\hat{\sigma}_g^2 = \frac{QMG - QME}{nr} = \frac{\sigma_d^2 + n\sigma_e^2 + nr\sigma_g^2 - \sigma_d^2 + n\sigma_e^2}{nr}, \text{ então,}$$

$$\hat{\sigma}_g^2 = \frac{249,6788 - 203,5118}{8} = 5,7709;$$

$\sigma_e^2$ : componente de variância ambiental, que mede as variações entre parcelas, estimado por:

$$\hat{\sigma}_e^2 = \frac{QME - QMD}{n} = \frac{\sigma_d^2 + n\sigma_e^2 - \sigma_d^2}{n}, \text{ então,}$$

$$\hat{\sigma}_e^2 = \frac{203,5118 - 27,9206}{4} = 43,8978;$$

$\sigma_d^2$ : componente de variância fenotípica (ambiental e genética), que mede as variações entre plantas dentro da parcela, estimado por:

$$\hat{\sigma}_d^2 = QMD, \text{ então, } \hat{\sigma}_d^2 = 27,9206.$$

No caso de melhoramento de plantas, são estimadas também as herdabilidades para a seleção tendo como base as médias entre famílias ou os valores dos indivíduos dentro de parcelas. Desta forma, tem-se:

$$h_{entre}^2 = \frac{\hat{\sigma}_g^2}{QMG/nr}, \text{ logo}$$

$$h_{entre}^2 = \frac{5,7709}{249,6788/8} = 0,1849. \text{ Para o caso de famílias de meios-irmãos tem-se:}$$

$$h_{dentro}^2 = \frac{3\hat{\sigma}_g^2}{QMD}, \text{ logo}$$

$$h_{dentro}^2 = \frac{3(5,7709)}{27,9206} = 0,6201.$$

Assim, existe possibilidade de se obter ganhos pela seleção de plantas dentro de parcela (família), uma vez que a herdabilidade estimada entre plantas é cerca de três vezes a estimada entre médias de famílias.

Inferências relativas aos componentes de variância

*Variância de uma estimativa de variância:*

Considere um quadrado médio,  $QM$ , associado a  $f$  graus de liberdade (gl). Tem-se que a razão  $\frac{(f)QM}{E(QM)}$  segue a distribuição qui-quadrado com  $f$  graus de liberdade  $\left(\chi_f^2\right)$ . Tem-se ainda que na distribuição qui-quadrado a média é igual a  $gl$  ( $f$ ) e a variância é igual a  $2 \times gl$  ( $2f$ ). Desta forma,  $Var\left[\frac{(f)QM}{E(QM)}\right] = 2f$ , e usando as propriedades de variância obtém-se:

$$\frac{f^2}{[E(QM)]^2} Var(QM) = 2f, \text{ logo}$$

$$Var(QM) = \frac{2[E(QM)]^2}{f}.$$

Uma estimativa não-viesada para  $Var(QM)$  é dada por:

$$\hat{Var}(QM) = \frac{2(QM)^2}{f+2}$$

Por exemplo, a variância do componente de variância  $\hat{\sigma}^2$  é obtida por meio da seguinte expressão:

$$Var(\hat{\sigma}^2) = \frac{2(\sigma^2)^2}{f+2} = \frac{2(QMR)^2}{f+2}.$$

Sob a pressuposição de que os quadrados médios de uma análise de variância sejam independentes, pode-se obter a variância de qualquer estimador de componentes de variância com base na expressão anterior.

Considere, por exemplo, o componente  $\hat{\sigma}_g^2$ , dado por  $\hat{\sigma}_g^2 = \frac{QMG - QMR}{r}$ , em que:

**$QMG$** : quadrado médio de genótipos, associado a  $g-1$  graus de liberdade;

**$QMR$** : quadrado médio do resíduo, associado a  $(r-1)(g-1)$  graus de liberdade.

Então, tem-se:

$$\hat{Var}(\hat{\sigma}_g^2) = \hat{Var}\left(\frac{QMG - QMR}{r}\right)$$

$$\hat{Var}(\hat{\sigma}_g^2) = \frac{1}{r^2} [Var(QMG) + Var(QMR)]$$

$$\hat{Var}(\hat{\sigma}_g^2) = \frac{1}{r^2} \left[ \frac{2QMG^2}{g-1+2} + \frac{2QMR^2}{(r-1)(g-1)+2} \right]$$

$$\hat{Var}(\hat{\sigma}_g^2) = \frac{2}{r^2} \left[ \frac{QMG^2}{g+1} + \frac{QMR^2}{(r-1)(g-1)+2} \right]$$

*Graus de liberdade associados às estimativas de variâncias de componentes de variância:*

Sejam  $\hat{\sigma}_1^2$  e  $\hat{\sigma}_2^2$  duas estimativas de diferentes componentes de variâncias ou dois **QMs**, aos quais estão associados  $v_1$  e  $v_2$  graus de liberdade, respectivamente. Considere agora uma função linear dessas estimativas:

$$\hat{\sigma}_0^2 = a_1 \hat{\sigma}_1^2 + a_2 \hat{\sigma}_2^2, \text{ então}$$

$Var(\hat{\sigma}_0^2) = a_1^2 Var(\hat{\sigma}_1^2) + a_2^2 Var(\hat{\sigma}_2^2)$ . Admitindo que o grau de liberdade, a ser obtido, associado à  $Var(\hat{\sigma}_0^2)$  seja  $n$ , tem-se:

$$\frac{2(\hat{\sigma}_0^2)^2}{n} = \frac{2a_1^2(\hat{\sigma}_1^2)^2}{n_1} + \frac{2a_2^2(\hat{\sigma}_2^2)^2}{n_2}, \text{ ou}$$

$$n = \frac{\left(\hat{\sigma}_0^2\right)^2}{\frac{a_1^2(\hat{\sigma}_1^2)^2}{n_1} + \frac{a_2^2(\hat{\sigma}_2^2)^2}{n_2}}.$$

Para os valores dos respectivos estimadores, tem-se:

$$n' = \frac{\left(\hat{\sigma}_0^2\right)^2}{\frac{a_1^2(\hat{\sigma}_1^2)^2}{n_1+2} + \frac{a_2^2(\hat{\sigma}_2^2)^2}{n_2+2}}.$$

Por exemplo, os graus de liberdade associados a  $\hat{Var}(\hat{\sigma}_g^2)$  podem ser estimados por meio de:

$$n' = \frac{\frac{(QMG - QMR)^2}{(QMG)^2 + (QMR)^2}}{g+1 + (r-1)(g-1)+2}.$$

Deve ser salientado que esta é uma aproximação grosseira do número de graus de liberdade, uma vez que a diferença de dois quadrados médios não tem distribuição conhecida e, para se obter a estimativa de  $n$  pelo método de Satterthwaite, é assumida a distribuição de  $\chi^2$ . Existem métodos mais precisos para estimação do número de graus de liberdade associados às variâncias de estimativas de componentes de variância, como por exemplo, o proposto por Kenward e Roger (1997).

Considere um quadrado médio  $QM$  associado a  $f$  graus de liberdade (gl). Tem-se que a razão  $\frac{f \cdot QM}{E(QM)}$  segue a distribuição qui-quadrado com  $f$  graus de liberdade ( $\chi_f^2$ ). Tem-se ainda que na distribuição qui-quadrado a média é igual a  $gl$  ( $f$ ) e a variância é igual a  $2 \cdot gl$  ( $2f$ ). Desta forma,  $Var\left[\frac{f \cdot QM}{E(QM)}\right] = 2f$ , e usando as propriedades de variância, obtém-se

$$\frac{f^2}{[E(QM)]^2} Var(QM) = 2f, \text{ logo } Var(QM) = \frac{2[E(QM)]^2}{f}.$$

Uma estimativa não-viesada para  $Var(QM)$  é dada por  $V\hat{a}r(QM) = \frac{2(QM)^2}{f+2}$

Por exemplo, a variância do componente de variância  $\hat{\sigma}^2$  é dada por

$$Var(\hat{\sigma}^2) = \frac{2(\sigma^2)^2}{f} = \frac{2(QMR)^2}{f}.$$

Sob a pressuposição de que os quadrados médios de uma análise de variância sejam independentes, pode-se obter a variância de qualquer estimador de componentes de variância com base nas expressões anteriores.

Considere, por exemplo, o componente  $\hat{\sigma}_g^2$ , dado por  $\hat{\sigma}_g^2 = \frac{QMG - QMR}{r}$ , em que

$QMG$  é o quadrado médio de genótipos, associado a  $g-1$  graus de liberdade e  $QMR$  é o quadrado médio do resíduo, associado a  $(r-1)(g-1)$  graus de liberdade, então, têm-se:

$$V\hat{a}r(\hat{\sigma}_g^2) = V\hat{a}r\left(\frac{QMG - QMR}{r}\right)$$

$$V\hat{a}r(\hat{\sigma}_g^2) = \frac{1}{r^2} [V\hat{a}r(QMG) + V\hat{a}r(QMR)]$$

$$V\hat{a}r(\hat{\sigma}_g^2) = \frac{1}{r^2} \left[ \frac{2QMG^2}{g-1+2} + \frac{2QMR^2}{(r-1)(g-1)+2} \right]$$

$$V\hat{a}r(\hat{\sigma}_g^2) = \frac{2}{r^2} \left[ \frac{QMG^2}{g+1} + \frac{QMR^2}{(r-1)(g-1)+2} \right]$$

Graus de liberdade associados às estimativas de variâncias de componentes de variância

Sejam  $\hat{\sigma}_1^2$  e  $\hat{\sigma}_2^2$  duas estimativas de diferentes componentes de variâncias ou dois *QMs*, aos quais estão associados  $v_1$  e  $v_2$  graus de liberdade, respectivamente. Considere agora uma função linear dessas estimativas:

$$\hat{\sigma}_0^2 = a_1 \hat{\sigma}_1^2 + a_2 \hat{\sigma}_2^2, \text{ então}$$

$V\hat{a}r(\hat{\sigma}_0^2) = a_1^2 V\hat{a}r(\hat{\sigma}_1^2) + a_2^2 V\hat{a}r(\hat{\sigma}_2^2)$ . Admitindo que o grau de liberdade, a ser obtido, associado a  $V\hat{a}r(\hat{\sigma}_0^2)$  seja  $n$ , tem-se:

$$\frac{2(\hat{\sigma}_0^2)^2}{n} = \frac{2a_1^2(\hat{\sigma}_1^2)^2}{n_1} + \frac{2a_2^2(\hat{\sigma}_2^2)^2}{n_2}, \text{ ou } n = \frac{(\hat{\sigma}_0^2)^2}{\frac{a_1^2(\hat{\sigma}_1^2)^2}{n_1} + \frac{a_2^2(\hat{\sigma}_2^2)^2}{n_2}}.$$

Para os valores dos respectivos estimadores, tem-se:

$$n' = \frac{\left(\hat{\sigma}_0^2\right)^2}{\frac{a_1^2(\hat{\sigma}_1^2)^2}{n_1+2} + \frac{a_2^2(\hat{\sigma}_2^2)^2}{n_2+2}}.$$

Por exemplo, os graus de liberdade associados à  $V\hat{a}r(\hat{\sigma}_g^2)$  podem ser estimados por meio de

$$n' = \frac{\frac{(QMG - QMR)^2}{(QMG)^2} + \frac{(QMR)^2}{(r-1)(g-1)+2}}{g+1}.$$

Deve ser salientado que esta é uma aproximação grosseira do número de graus de liberdade, uma vez que a diferença de dois quadrados médios não tem distribuição conhecida e, para se obter a estimativa de  $n$  pelo método de Satherthwaite, é assumida a distribuição de  $\chi^2$ . Existem métodos mais precisos para estimação do número de graus de liberdade associados às variâncias de estimativas de componentes de variância, como, por exemplo, o proposto por Kenward e Roger (1997).

# CAPITULO 5

## Análise de Covariância

### Conceito de análise de covariância

A análise de covariância está relacionada com duas ou mais variáveis, sendo uma considerada dependente (resposta) e a outra (ou outras) considerada (s) independente (s) e faz uso dos princípios e métodos de análise de variância e análise de regressão.

As principais utilizações da análise de covariância são:

**1º) Controlar o erro experimental** - A variância de uma média de tratamento é  $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$ . Para

diminuir esta variância, temos apenas duas alternativas: aumentar o tamanho da amostra ou controlar a variância na população amostrada; o controle de  $\sigma^2$  é fornecido pelo delineamento experimental ou por meio de uma ou mais covariáveis.

O uso de covariáveis aumenta a precisão com que os efeitos de tratamentos são medidos, o que é conseguido através da remoção, por regressão, de certos efeitos que não podem ser controlados ou que não foram controlados pelo delineamento experimental.

Um exemplo clássico é a correlação entre peso inicial e ganho de peso em experimentos de nutrição animal; ou, stand final e produtividade de plantas.

**2º) Ajustar médias de tratamentos** - Quando a variação observada em Y é parcialmente atribuível a uma variação em X (covariável), tem-se que a variação nas médias de tratamentos  $\bar{Y}'s$  deve ser também afetada pela variação entre as médias da covariável  $\bar{X}'s$ .

Para serem comparáveis, as médias de tratamentos devem ser ajustadas (para terem a mesma qualidade de estimativa que teriam se todas as médias  $\bar{X}'s$  tivessem sido iguais). Portanto, é necessário controlar a variância do erro pelo uso de uma covariável, por meio do uso de regressão.

**3º) Interpretar dados** - Sabe-se que certos tratamentos produzem efeitos reais tanto nas variáveis dependentes quanto nas independentes; a covariância como um meio de controlar erro e ajustar médias de tratamentos, deve ser usada primariamente quando a variável independente mede efeitos ambientais e não é influenciada pelos tratamentos.

O ajustamento de médias por meio de análise de covariância remove parte dos efeitos de tratamentos, quando as médias da covariável (variável independente) são afetadas pelos tratamentos. As médias de tratamento ajustadas estimam os valores esperados apenas quando as médias de tratamentos para a covariável são iguais.

Por exemplo, quando a covariável “stand” é influenciada pelos tratamentos, a análise de produtividade ajustada para diferenças de “stand” remove parte dos efeitos dos tratamentos e o pesquisador pode enganar-se na interpretação dos dados.

### Modelo para Análise de Covariância

As pressuposições para análise de covariância são uma combinação daquelas para análise de variância e de regressão linear. O modelo linear aditivo de análise de covariância para qualquer

delineamento é aquele para análise de variância mais um termo adicional para a covariável (variável independente).

Por exemplo, para a avaliação de produtividade de genótipos no delineamento experimental blocos completos casualizados, com medição do número de plantas por parcela (estande), o modelo estatístico é expresso por:

$$Y_{ij} = \mu + \tau_i + \delta_j + \beta_i (X_{ij} - \bar{X}_{..}) + \varepsilon_{ij} \quad (1)$$

, em que:

$Y_{ij}$ : produtividade do  $i$ -ésimo genótipo no  $j$ -ésimo bloco;

$X_{ij}$ : estande (covariável) na parcela do  $i$ -ésimo genótipo no  $j$ -ésimo bloco;

$\mu$ : média geral de produtividade;

$\tau_i$ : efeito do  $i$ -ésimo genótipo;

$\delta_j$ : efeito do  $j$ -ésimo bloco;

$\beta_i$ : inclinação da reta de regressão de produtividade ( $Y$ ) sobre estande ( $X$ ) para o  $i$ -ésimo genótipo;

$\bar{X}_{..}$ : média geral de estande;

$\varepsilon_{ij}$ : erro experimental associado à observação  $Y_{ij}$ ;

Se os estandes não fossem medidos, então a variação de  $Y_{ij}$  devida a  $\beta_i (X_{ij} - \bar{X}_{..})$  não poderia ser determinada e seria incluída no erro. Neste caso, o modelo anterior (1) torna-se um modelo de análise de variância:

$$Y_{ij} = \mu + \tau_i + \delta_j + e_{ij} \quad (2)$$

Uma vez que se sabe que  $Y$  e  $X$  são estreitamente correlacionadas, o modelo (1) deve ajustar os valores dos dados melhor que o modelo (2) e os resíduos  $\varepsilon_{ij}$  devem ser menores que os  $e_{ij}$ . Se

todos os genótipos tiverem a mesma resposta para estande, ou seja,

$\beta_1 = \beta_2 = \dots = \beta_t = \beta$ , então o modelo (1) é simplificado e torna-se:

$$Y_{ij} = \mu + \tau_i + \delta_j + \beta (X_{ij} - \bar{X}_{..}) + \varepsilon_{ij} \quad (3)$$

, que é o modelo básico para a análise de covariância.

A variável que está sendo analisada, que é a variável dependente, é denotada por Y, enquanto a variável usada no controle do erro e no ajustamento de médias, a variável independente ou covariável, é denotada por X.

É interessante reescrever a expressão anterior nas formas seguintes:

$$Y_{ij} - \beta(X_{ij} - \bar{X}_{..}) = \mu + \tau_i + \delta_j + \varepsilon_{ij} \text{ e}$$

$$Y_{ij} - \tau_i - \delta_j = \mu + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$

No primeiro caso enfatiza-se o delineamento experimental, ou seja, análise de variância de valores da variável dependente que foram ajustados por regressão sobre a variável independente.

No segundo caso, enfatiza-se a estimativa de regressão (regressão de Y sobre X) sem a interferência dos efeitos de genótipos e blocos. Observe então que o termo  $\beta(X_{ij} - \bar{X}_{..})$  estaria sendo incluído no erro ou resíduo.

As pressuposições para o uso válido de análise de covariância são:

1º) Os  $\bar{X}'s$  são fixos, medidos sem erro e independentes dos genótipos;

2º) A regressão de Y sobre X, após remoção das diferenças entre blocos e entre genótipos, é linear e independente de blocos e genótipos;

3º) Os resíduos são normal e independentemente distribuídos com média zero e uma variância comum.

A variância residual é estimada com base em estimadores de quadrados mínimos para  $\mu$ ,  $\tau_i$ ,  $\delta_j$  e  $\beta$ , ou seja,

$$\sum \left[ Y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\delta}_j - \hat{\beta}(X_{ij} - \bar{X}_{..}) \right]^2 = \min .$$

Os estimadores de quadrados mínimos para os parâmetros do modelo são:

$$\hat{\mu} = \bar{Y}$$

$$\hat{\tau}_i = t_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} - b(\bar{X}_{i\cdot} - \bar{X}_{..})$$

$$\hat{\delta}_j = r_j = \bar{Y}_{\cdot j} - \bar{Y}_{..} - b(\bar{X}_{\cdot j} - \bar{X}_{..})$$

$$\hat{\beta} = b = \frac{E_{XY}}{E_{XX}}$$

$$\hat{\sigma}_{Y.X}^2 = S_{Y.X}^2 = \frac{E_{YY} - (E_{XY})^2 / E_{XX}}{f_e}$$

Onde  $E_{XX}$ ,  $E_{YY}$  e  $E_{XY}$  são as somas de quadrados e de produtos do erro e  $f_e$  são os graus de liberdade do erro.

Observa-se em  $\hat{\tau}_i = t_i = \bar{Y}_{i..} - \bar{Y}_{..} - b(\bar{X}_{i..} - \bar{X}_{..})$  que para estimar o efeito de genótipo  $\hat{\tau}_i$ , o desvio de qualquer média de genótipo em relação à média geral deve ser ajustado pela quantidade  $b(\bar{X}_{i..} - \bar{X}_{..})$ . Este ajustamento remove qualquer efeito que é atribuível à variável X, o que torna as médias de genótipos ajustadas efetivamente comparáveis.

Esquema de análise de covariância par o delineamento blocos completos casualizados:

FV	GL	Somas de Quadrados e de Produtos*			GL	SQ**	QM***
		SQ <sub>X</sub>	SP <sub>XY</sub>	SQ <sub>Y</sub>			
Total	$rt-1$						
Blocos	$r-1$	$R_{XX}$	$R_{XY}$	$R_{YY}$			
Genótipo	$t-1$	$T_{XX}$	$T_{XY}$	$T_{YY}$			
Erro	$(r-1)(t-1)$	$E_{XX}$	$E_{XY}$	$E_{YY}$	$(r-1)(t-1)-1$	SQErro	$S_{Y.X}^2$
Genótipo + Erro	$r(t-1)$	$S_{XX}$	$S_{XY}$	$S_{YY}$	$r(t-1)-1$	SQ Genótipo+ Erro	
Genótipo Ajustado					$t-1$	SQ Genótipo Ajustado	QM Genótipo Ajustado

$$*: SQ_X = (X - \bar{X})^2; SQ_Y = (Y - \bar{Y})^2; SP_{XY} = \sum [(X - \bar{X})(Y - \bar{Y})]$$

$$**: SQEro = E_{YY} - \frac{(E_{XY})^2}{E_{XX}}; E_{XX} = \Sigma(X_{ij} - \bar{X}_{..})^2 - R_{XX} - T_{XX};$$

$$E_{YY} = \Sigma(Y_{ij} - \bar{Y}_{..})^2 - R_{YY} - T_{YY};$$

$$E_{XY} = \Sigma(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) - R_{XY} - T_{XY}$$

$$SQGenótipo + Erro = S_{YY} - \frac{(S_{XY})^2}{S_{XX}};$$

$$S_{XX} = T_{XX} + E_{XX}; S_{YY} = T_{YY} + E_{YY}$$

$$S_{XY} = T_{XY} + E_{XY};$$

$$SQGenótipoAjustado = \left[ S_{YY} - \frac{(S_{XY})^2}{S_{XX}} \right] - \left[ E_{YY} - \frac{(E_{XY})^2}{E_{XX}} \right]$$

\*\*\*:  $S_{Y.X}^2$  = Quadrado médio do erro ajustado.

Exemplo de Análise de Covariância no Delineamento Blocos Casualizados:

No experimento de Bowers & Permenter (citado por Stell & Torrie, 1981) estudou-se 11 variedades de feijão em relação ao conteúdo de ácido ascórbico. Com base em experiências prévias sabe-se que o aumento de maturidade reduz o conteúdo de ácido ascórbico.

Uma vez que nem todas as variedades apresentavam o mesmo grau de maturação na colheita, não foi possível colher todas as parcelas com mesmo estágio de maturação. Desta forma, a porcentagem de matéria seca do feijão recém-colhido foi tomada como índice de maturação e usada como uma covariável. Os dados do experimento estão na Tabela a seguir:

Tabela. Conteúdo de ácido ascórbico em mg/100g de matéria fresca (Y) e percentagem de matéria seca (X) em variedades de feijão.

Variedade	Blocos										Totais de Variedade	
	1		2		3		4		5			
	X	Y	X	Y	X	Y	X	Y	X	Y	X <sub>i</sub>	Y <sub>i</sub>
1	34,0	93,0	33,4	94,8	34,7	91,7	38,9	80,8	36,1	80,2	177,1	440,5
2	39,6	47,3	39,8	51,5	51,2	33,3	52,0	27,2	56,2	20,6	238,8	179,9
3	31,7	81,4	30,1	109,0	33,8	71,6	39,6	57,5	47,8	30,1	183,0	349,6
4	37,7	66,9	38,2	74,1	40,3	64,7	39,4	69,3	41,3	63,2	196,9	338,2
5	24,9	119,5	24,0	128,5	24,9	125,6	23,5	129,0	25,1	126,2	122,4	628,8
6	30,3	106,6	29,1	111,4	31,7	99,0	28,3	126,1	34,2	95,6	153,6	538,7
7	32,7	106,1	33,8	107,2	34,8	97,5	35,4	86,0	37,8	88,8	174,5	485,6
8	34,5	61,5	31,5	83,4	31,1	93,9	36,1	69,0	38,5	46,9	171,7	354,7
9	31,4	80,5	30,5	106,5	34,6	76,7	30,9	91,8	36,8	68,2	164,2	423,7
10	21,2	149,2	25,3	151,6	23,5	170,1	24,8	155,2	24,6	146,1	119,4	772,2
11	30,8	78,7	26,4	116,9	33,2	71,8	33,5	70,3	43,8	40,9	167,7	378,6
Totais de Blocos (X <sub>j</sub> ; Y <sub>j</sub> )	348,8	990,7	342,1	1134,9	373,8	995,9	382,4	962,2	422,2	806,8	1869,3	4890,5

Os cálculos de somas de quadrados e somas de produtos para a análise de covariância no delineamento blocos casualizados são dados basicamente por:

$$C = \frac{\left( \sum Y_{ij} \right)^2}{rt}; \quad SQ\ Total = \sum Y_{ij}^2 - C;$$

$$SQ\ Blocos = \frac{\sum Y_j^2}{t} - C; \quad SQ\ Variedade = \frac{\sum Y_i^2}{r} - C;$$

$$SQ\ Res = SQ\ Total - SQ\ Blocos - SQ\ Variedade$$

$$SP = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$$

Para os dados da Tabela anterior as Somas de Quadrados e de Produtos Total são:

$$\Sigma(X_{ij} - \bar{X}_{..})^2 = \Sigma X_{ij}^2 - \frac{X^2}{rt} = 2916,22$$

$$\Sigma(Y_{ij} - \bar{Y}_{..})^2 = \Sigma Y_{ij}^2 - \frac{Y^2}{rt} = 61934,42$$

$$\Sigma(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \Sigma X_{ij}Y_{ij} - \frac{(X_{..})(Y_{..})}{rt} = -12226,14$$

As Somas de Quadrados e de Produtos de Blocos são:

$$R_{XX} = \frac{\Sigma X_{\cdot j}^2}{t} - \frac{X^2}{rt} = 367,85$$

$$R_{YY} = \frac{\Sigma Y_{\cdot j}^2}{t} - \frac{Y^2}{rt} = 4968,84$$

$$R_{XY} = \frac{\Sigma(X_{\cdot j})(Y_{\cdot j})}{t} - \frac{(X_{..})(Y_{..})}{rt} = -1246,66$$

As Somas de Quadrados e de Produtos de Tratamentos (Variedades) são:

$$T_{XX} = \frac{\Sigma X_{i\cdot}^2}{r} - \frac{X^2}{rt} = 2166,71$$

$$T_{YY} = \frac{\Sigma Y_{i\cdot}^2}{r} - \frac{Y^2}{rt} = 51018,18$$

$$T_{XY} = \frac{\Sigma(X_{i\cdot})(Y_{i\cdot})}{r} - \frac{(X_{..})(Y_{..})}{rt} = -9784,14$$

As Somas de Quadrados e de Produtos de Erro são obtidas por subtração e são:

$$E_{XX} = \Sigma(X_{ij} - \bar{X}_{..})^2 - R_{XX} - T_{XX}$$

$$E_{XX} = 2916,22 - 367,85 - 2166,71 = 381,66$$

$$E_{YY} = \Sigma(Y_{ij} - \bar{Y}_{..})^2 - R_{YY} - T_{YY} = 5947,30$$

$$E_{XY} = \Sigma(X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) - R_{XY} - T_{XY}$$

$$E_{XY} = -12226,14 - (-1246,66) - (-9784,14)$$

$$E_{XY} = -1195,34$$

As somas de quadrados e de produtos anteriores são utilizadas nas análises de variância de X e Y bem como para a análise de covariância. Assim, para testar a hipótese de nenhuma diferença entre as médias não ajustadas de teor de ácido ascórbico (Y) para variedades, tem-se:

$$F = \frac{T_{YY} / (t-1)}{E_{YY} / (r-1)(t-1)} = \frac{51018,18 / 10}{5947,30 / 40} = 34,31 (p < 0,01)$$

, com 10 e 40 g.l.

Conclui-se que existem diferenças reais entre as médias de variedades não ajustadas para a característica conteúdo de ácido ascórbico.

Para testar a hipótese de nenhuma diferença entre médias de variedades, para percentagem de matéria seca (X) na colheita, tem-se:

$$F = \frac{T_{XX} / (t-1)}{E_{XX} / (r-1)(t-1)} = \frac{2166,71 / 10}{381,66 / 40} = 22,71 (p < 0,01)$$

, com 10 e 40 g.l.

Conclui-se que existem diferenças reais entre médias de variedades para percentagem de matéria seca na colheita. Portanto, a necessidade de ajustamento das médias de variedades da variável dependente (ácido ascórbico) está justificada, admitindo-se que essas diferenças não são devidas aos efeitos das variedades.

Desta forma, devem-se ajustar as médias de ácido ascórbico das variedades para a mesma percentagem de matéria seca ou maturidade, uma vez que esta é a condição lógica para comparar as variedades em relação ao conteúdo de ácido ascórbico.

Para testar a hipótese de que não existe nenhuma diferença entre as médias de variedades ajustadas é necessário calcular as somas de quadrados de erro e de erro mais variedades da variável Y.

A soma de quadrados para testar a hipótese de nenhuma diferença entre as médias de variedades ajustadas é a diferença entre as duas somas de quadrados citadas anteriormente.

O procedimento é o seguinte:

- a) Para o erro, o coeficiente de regressão é obtido como:

$$b_{YX} = \frac{E_{XY}}{E_{XX}} = \frac{-1195,34}{381,66} = -3,13 \text{ mg de ácido ascórbico para cada } 1\% \text{ de matéria seca.}$$

- b) A soma de quadrados de Y atribuível à regressão sobre X é:

$$b_{YX} E_{XY} = \frac{(E_{XY})^2}{E_{XX}} = \frac{(-1195,34)^2}{381,66} = 3743,85, \text{ com } 1 \text{ g.l.}$$

c) A soma de quadrados do erro ajustada é dada por:

$$E_{YY} - \frac{(E_{XY})^2}{E_{XX}} = 5947,30 - 3743,85 = 2203,45,$$

com  $(r-1)(t-1)-1=39$  g.l.;

O quadrado médio do erro ajustado é então:

$$S_{Y.X}^2 = \frac{2203,45}{39} = 56,50$$

d) Para variedades mais erro, a soma de quadrados ajustada é:

$$S_{YY} - \frac{(S_{XY})^2}{S_{XX}} = 56965,48 - \frac{(10979,48)^2}{2548,37} = 9661,13$$

, com  $r(t-1)-1=49$  g.l.;  $S_{YY}$  = Variedades + Erro, para Y

e) Para variedades ajustados, a soma de quadrados é a diferença entre as somas de quadrados de variedades mais erro e a soma de quadrados de erro, ou seja:

$$T_{YY \text{ ajustado}} = 9661,13 - 2203,45 = 7457,62, \text{ com } t-1=10 \text{ g.l.}$$

f) Para testar a hipótese de nenhuma diferença entre médias de variedades ajustados (Y ajustado), tem-se:

$$F = \frac{QM (\text{Médias Variedades Ajustadas})}{S_{Y.X}^2} = \frac{745,76}{56,50} = 13,20 (p < 0,01), \text{ com } 10 \text{ e } 39 \text{ g.l.}$$

Conclui-se que existem diferenças reais entre as médias de variedades para Y quando ajustadas em relação a X. Se as médias não ajustadas tivessem sido significativas e as ajustadas não, poderia indicar que as diferenças entre as médias não ajustadas refletem simplesmente diferenças em maturidade e não em conteúdo de ácido ascórbico.

g) Os resultados da análise de covariância estão resumidos na Tabela a seguir:

Tabela. Análise de Covariância dos dados de teor de ácido ascórbico em feijão.

Fonte de Variação	Somas de Quadrados e Produtos				Y ajustado por X			
	G.L.	X, X	X, Y	Y, Y	G.L.	S.Q.	Q.M.	F
Total	54	2916,22	-12226,14	61934,42				
Blocos	4	367,85	-1246,66	4968,94				
Variedade	10	2166,71	-9784,14	51018,18				
Erro	40	381,66	-1195,34	5947,30	39	2203,45	56,50	
Variedade + Erro	50	2548,37	-10979,48	56965,48	49	9661,13		
Variedade Ajustado					10	7457,62	745,76	13,2**

Ajuste de Médias de Variedades:

A expressão para ajustar as médias de variedades é dada essencialmente por  $\hat{\tau}_i = \bar{t}_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} - b(\bar{X}_{i\cdot} - \bar{X}_{..})$ , o que é uma aplicação do princípio expresso por:

$$Y_{Ajustado} = \bar{Y} + e_{Y.X} = Y - b(X - \bar{X}).$$

Numa notação mais comum, a equação para uma média de variedade ajustada é dada por

$$\hat{Y}_{i\cdot} = \bar{Y}_{i\cdot} - b_{Y.X} (\bar{X}_{i\cdot} - \bar{X}_{..}), \text{ onde } b_{Y.X} = b \text{ é o coeficiente de regressão do erro.}$$

As médias de variedade ajustadas são estimativas do que as médias de variedade seriam se todos os  $\bar{X}_{i\cdot}$  tendessem para  $\bar{X}_{..}$ .

As médias de variedades ajustadas dos dados de teor de ácido ascórbico em feijoeiro são as seguintes:

Variedade	% Média de Matéria Seca $(\bar{X}_{i\cdot})$	Desvio $(\bar{X}_{i\cdot} - \bar{X}_{..})$	Ajuste $b_{YX} (\bar{X}_{i\cdot} - \bar{X}_{..})$	Média obs. do teor de Ácido Ascórbico $(\bar{Y}_{i\cdot})$	Média ajustada do teor de Ácido Ascórbico $\hat{Y}_{i\cdot} = \bar{Y}_{i\cdot} - b_{YX} (\bar{X}_{i\cdot} - \bar{X}_{..})$
1	35,42	1,43	-4,48	88,10(5)**	92,59(5)
2	47,76	13,77	-43,10	35,98(11)	79,12(8)
3	36,60	2,61	-8,17	69,98(9)	78,10(9)
4	39,38	5,39	-16,87	67,64(10)	84,53(6)
5	24,48	-9,51	29,77	125,76(2)	95,98(4)
6	30,72	-3,27	10,24	107,74(3)	97,51(3)

7	34,90	0,91	-2,85	97,12(4)	99,98(2)
8	34,34	0,35	-1,10	70,94(8)	72,04(11)
9	32,84	-1,15	3,60	94,74(6)	81,15(7)
10	23,88	-10,11	31,64	154,44(1)	122,78(1)
11	33,54	-0,45	1,41	75,72(7)	74,32(10)

$$\bar{X}_{..} = 33,99 \quad \Sigma = -0,03^* \quad \Sigma = +0,09^* \quad \bar{Y}_{..} = 88,92 \quad \hat{\bar{Y}}_{..} = 88,92$$

\* Teoricamente  $\Sigma(\bar{X}_i - \bar{X}_{..}) = 0$ , logo,  $\Sigma b_{YX}(\bar{X}_i - \bar{X}_{..}) = 0$  e desta forma  $\Sigma \bar{Y}_i = \Sigma \hat{\bar{Y}}_i$ .

\*\* Entre parênteses estão as ordens de teor de ácido ascórbico.

Inferências sobre as médias ajustadas temos:

- a) O erro padrão de uma média de variedade ajustada é expressa como:

$$S_{\hat{Y}_i} = S_{Y.X}^2 \sqrt{\frac{1}{r} + \frac{(\bar{X}_i - \bar{X}_{..})^2}{E_{XX}}}$$

- b) O erro padrão da diferença entre duas médias de variedade ajustadas é obtido por meio de:

$$S_{\hat{Y}_i - \hat{Y}_{i'}} = \sqrt{S_{Y.X}^2 \left[ \frac{2}{r} + \frac{(\bar{X}_i - \bar{X}_{i'})^2}{E_{XX}} \right]}$$

Por exemplo, para comparar as médias das variedades 7 e 10, precisa-se de:

$$S_{\hat{Y}_7 - \hat{Y}_{10}} = \sqrt{56,50 \left[ \frac{2}{5} + \frac{(34,90 - 23,88)^2}{381,66} \right]} = 6,38 \text{ mg de ácido ascórbico.}$$

A comparação é entre as médias ajustadas  $\hat{Y}_7 = 99,97$  e  $\hat{Y}_{10} = 122,78$ , sendo a diferença significativa.

Na equação anterior é mostrada a necessidade de um cálculo separado para cada comparação. Provavelmente, isto é justificado se o experimento tomou uma grande quantidade de tempo ou foi algo mais que um simples experimento preliminar.

Entretanto, pode-se utilizar uma aproximação para  $S_{\hat{Y}_i - \hat{Y}_{i'}}$ , que utiliza uma média no lugar dos termos  $(\bar{X}_{i.} - \bar{X}_{i'})$  individuais, que são requeridos na equação anterior. A expressão é a seguinte:

$$S_{\hat{Y}_i - \hat{Y}_{i'}} = \sqrt{\frac{2S_{Y.X}^2}{r} \left[ 1 + \frac{T_{XX}}{(t-1)E_{XX}} \right]}$$

Então, tem-se que:

$$S_{\hat{Y}_7 - \hat{Y}_{10}} = \sqrt{\frac{2(56,50)^2}{5} \left[ 1 + \frac{2166,71}{(10)(381,66)} \right]} = 5,95$$

mg de ácido ascórbico.

Para médias de tratamentos com número diferente de repetições deve-se substituir  $s/r$  nas expressões anteriores por  $(1/r_1 + 1/r_2)$ .

A aproximação anterior geralmente é bastante próxima da situação ideal se os graus de liberdade do erro excedem 20, uma vez que a contribuição para o erro de amostragem em  $b_{Y.X}$  (fator de ajustamento) é pequeno neste caso.

Se a variação entre os  $\bar{X}_i$  é significativa, por exemplo, devido a variedades, a expressão de aproximação pode levar a sérios erros e não deveria ser usada, como é o caso do exemplo aqui apresentado.

Eficiência Relativa da Análise de Covariância:

Para avaliar a eficiência da análise de covariância, como um meio de controle do erro, é feita uma comparação da variância de uma média de variedade antes e depois do ajuste para a variável independente X ou covariável (teor de ácido ascórbico). O quadrado médio do erro antes do ajuste é

$E_{YY}/(r-1)(t-1) = 5947,30/40 = 148,68$ , com 40 graus de liberdade e após o ajuste é

$$S_{Y.X}^2 = \left[ E_{YY} - \frac{(E_{XY})^2}{E_{XX}} \right] / (r-1)(t-1)-1, \text{ então}$$

$S_{Y.X}^2 = 2203,45/39 = 56,50$ , com 39 graus de liberdade.

É necessário reajustar o último valor para levar em conta o erro de amostragem na determinação do coeficiente de regressão usado no ajuste. Então, tem-se que o quadrado médio do erro efetivo após o ajuste para X é dado por:

$$QMERroEfectivo = S_{Y.X}^2 \left[ 1 + \frac{T_{XX}}{(t-1) E_{XX}} \right].$$

$$\text{Então, } QMERroEfectivo = 56,50 \left[ 1 + \frac{2166,71}{(10)(381,66)} \right] = 88,58$$

Uma estimativa da eficiência relativa da análise de covariância comparada com a análise de variância é obtida como

$$ER\% = \left[ \left( E_{YY} / (r-1)(t-1) \right) / QMERroEfectivo \right] (100). \quad \text{Então, neste caso tem-se}$$

$$ER\% = (148,68 / 88,58) 100 = 167,85.$$

Este resultado indica que 100 repetições usando covariância são tão efetivas quanto 168 repetições sem covariância, o que dá uma razão de aproximadamente 3:5.

# CAPITULO 6

## Delineamento Experimental Blocos Completos Casualizados

Para melhorar a precisão do experimento pode-se partitionar o material experimental em grupos homogêneos denominados de blocos. Os blocos representam o controle local e cada um deve incluir todos os tratamentos. Para que o delineamento seja eficiente, cada bloco deve ser o mais uniforme possível, mas os blocos podem diferir bastante uns dos outros.

Nos experimentos fitotécnicos, cada bloco deverá ser constituído de uma área de solo bem uniforme. Nos experimentos zootécnicos, cada bloco deverá ser constituído de animais com características semelhantes. Dentro de cada bloco os tratamentos são alocados às parcelas por meio de sorteio, ou seja, por meio de casualizações independentes.

A vantagem deste delineamento é que ele extrai da variação total a variação devida a blocos, além da variação devida a tratamento.

As desvantagens são:

- (a) redução do número de graus de liberdade do resíduo e
- (b) se o número de tratamentos é muito grande, é difícil conseguir um bom agrupamento das parcelas em blocos homogêneos.

O objetivo da formação de blocos é diminuir o erro experimental por meio de controle local.

O modelo linear aditivo associado ao DBC é expresso como:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : observação do tratamento de ordem  $i$  no bloco de ordem  $j$ ,  $i=1, 2, \dots, t$  e

$j=1, 2, \dots, r$

$\mu$ : média geral do experimento

$\tau_i$ : efeito do tratamento de ordem  $i$

$\beta_j$ : efeito do bloco de ordem  $j$

$\varepsilon_{ij}$ : erro associado a observação  $Y_{ij}$

Este delineamento permite fracionar a variação total em três componentes, ou seja, em variações devidas a blocos, a tratamentos e ao erro experimental. As somas de quadrados, os graus de liberdade, os quadrados médios e o F são calculados e apresentados como a seguir:

$$SQBlocs = \sum_{j=1}^r Y_{\cdot j}^2 / t - C; SQTrat = \sum_{i=1}^t Y_{i \cdot}^2 / r - C;$$

$$SQRes = \sum_{i=1}^t \sum_{j=1}^r Y_{ij}^2 - \sum_{j=1}^r Y_{\cdot j}^2 / t - \sum_{i=1}^t Y_{i \cdot}^2 / r + C;$$

$$SQTotal = \sum_{i=1}^t \sum_{j=1}^r Y_{ij}^2 - C; C = \left( \sum_{i=1}^t \sum_{j=1}^r Y_{ij} \right)^2 / rt$$

O coeficiente de variação é expresso como:  $CV\% = \frac{\sqrt{QM\ Res}}{\bar{Y}_{..}} (100)$

### Exemplo de Aplicação

Considere o conjunto de dados de produtividade de grãos (t/ha) apresentado a seguir referente a um experimento de avaliação de 10 cultivares de arroz, conduzido no delineamento experimental blocos completos casualizados:

Cultivares (i)	Blocos (j)			Totais de cultivares ( $\bar{Y}_{..}$ )
	1	2	3	
1	5,427	5,979	5,396	16,802
2	3,719	3,583	4,771	12,073
3	4,375	3,792	3,750	11,917
4	4,594	5,313	5,760	15,667
5	4,740	4,000	4,813	13,553
6	3,229	2,927	3,000	9,156
7	4,323	4,167	4,792	13,282
8	4,385	3,740	3,646	11,771
9	6,375	5,083	5,292	16,75
10	4,844	4,198	4,750	13,792
Totais de Blocos ( $\bar{Y}_{..}$ )	46,011	42,782	45,97	134,763

A análise estatística dos dados experimentais é feita por meio da análise de variância que permite avaliar o efeito do fator cultivar e a qualidade do experimento:

a) Efeito de cultivares

$$SQTotal = (5,427)^2 + (5,979)^2 + \dots + (4,750)^2 - \frac{(134,763)^2}{3 \times 10} = 21,7565$$

$$SQBlocos = (46,001)^2 / 10 + (42,782)^2 / 10 + (45,97)^2 / 10 - (134,763)^2 / 30 = 0,68638$$

$$SQCultivares = (16,802)^2 / 3 + (12,073)^2 / 3 + \dots + (13,792)^2 / 3 - (134,763)^2 / 30 = 17,5645$$

$$SQResíduo = SQ Total - SQ Tratamento - SQ Blocos$$

$$SQResíduo = 21,7565 - 0,68638 - 17,5645 = 3,50562$$

FV	GL	SQ	QM	F
Bloco	3-1=2	0,68638	0,34319	
Cultivar	10-1=9	17,5645	1,9516	10,02*
Resíduo	18	3,50562	0,1947	
Total	(3.10)-1=29	21,7565	-	

$$F_{0,05} (9;18) = 2,46$$

Como  $F = 10,02 > F_{0,05} (9;18) = 2,46$ , logo rejeita-se  $H_0$ , e concluímos que pelo menos 2 médias são diferentes entre si.

b) Qualidade do experimento:

$$CV\% = \frac{\sqrt{QM \text{ Resíduo}}}{\bar{Y}_{..}} (100) = \frac{\sqrt{0,1947}}{4,492} (100) = 9,82$$

Como o CV% é inferior a 10%, podemos dizer que a precisão e qualidade do experimento são altas.

Teste de Scott-Knott para agrupamento de médias:

As médias de cultivares podem ser avaliadas aplicando o teste para comparações de médias de Scott-Knott. Este teste permite a formação de grupos de médias sem a sobreposição de médias utilizando várias informações estatísticas obtidas do experimento.

O procedimento do teste consiste em colocar as médias em ordem crescente e calcular a estatística de teste  $\lambda = \pi B_0 / [2\hat{\sigma}_0^2(\pi-2)]$ , que tem distribuição aproximada por uma distribuição  $\chi^2$  com  $v_0 = t/(\pi-2)$  graus de liberdade. Então se tem:

Ordenamento das médias de cultivares:

Médias em ordem crescente		
Ordem das Cultivares (i)	Identificação das médias ordenadas	Médias de Cultivares ( $\bar{Y}_{i.}$ )
6	1	3,052
8	2	3,924
3	3	3,972
2	4	4,024
7	5	4,427
5	6	4,518
10	7	4,597
4	8	5,222
9	9	5,583
1	10	5,601

Total Geral $(Y_{..})$	44,92
Média $(\bar{Y}_{..})$	4,4921

Partições em dois grupos:

Partições possíveis em dois grupos	Valores de SQEntre Grupos (SQEG)
1 / 2345678910	2,303
12 / 345678910	2,5214
123 / 45678910	3,0440
1234 / 5678910	3,7402
12345 / 678910	3,7475
123456 / 78910	3,8387
1234567 / 8910	4,0881
12345678 / 910	3,0245
123456789 / 10	1,3655

Obtenção das Soma de Quadrados Entre Grupos (SQEG):

$$FC = (3,052 + 3,924 + \dots + 5,601)^2 / 10 = (44,92)^2 / 10 = 201,79$$

$$SQEG1 = [(3,052)^2 + (3,924 + 3,972 + \dots + 5,601)^2 / 9] - FC = 2,303$$

$$SQEG2 = [(3,052 + 3,924)^2 / 2 + (3,972 + \dots + 5,601)^2 / 8] - FC = 2,5214$$

$$SQEG3 = [(3,052 + 3,924 + 3,972)^2 / 3 + (4,024 + \dots + 5,601)^2 / 7] - FC = 3,0440$$

$$SGEG4 = [(3,052 + 3,924 + 3,972 + 4,024)^2 / 4 + (4,427 + \dots + 5,601)^2 / 6] - FC = 3,7402$$

$$SGEG5 = [(3,052 + \dots + 4,427)^2 / 5 + (4,518 + \dots + 5,601)^2 / 5] - FC = 3,7475$$

$$SGEG6 = [(3,052 + \dots + 4,518)^2 / 6 + (4,597 + \dots + 5,601)^2 / 4] - FC = 3,8387$$

$$SGEG7 = [(3,052 + \dots + 4,597)^2 / 7 (5,222 + \dots + 5,601)^2 / 3] - FC = 4,0881$$

$$SGEG8 = [(3,052 + \dots + 5,2223)^2 / 8 (5,583 + 5,601)^2 / 2] - FC = 3,0245$$

$$SGEG9 = [(3,052 + \dots + 5,583)^2 / 9 + (5,601)^2] - FC = 1,3655$$

O valor de  $B_0$  utilizado no cálculo da estatística de teste  $\lambda$  é igual ao maior valor da SQEG, ou seja,

$$B_0 = \max SQEG.$$

Logo o valor de  $B_0 = 4,0881$

O valor da variância da média utilizado no cálculo de  $\sigma_0^2$  que é um componente da equação da estatística de teste  $\lambda$  é:

$$S\frac{2}{Y} = \frac{QM\ Residuo}{r} = \frac{0,1947}{3} = 0,06492$$

A Soma de Quadrados Total obtida com base nas médias, que também é utilizada no cálculo de  $\sigma_0^2$  é obtida como:

$$SQTotal = \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 = \sum_i \bar{Y}_{i\cdot}^2 - \frac{\left(\sum_i \bar{Y}_{i\cdot}\right)^2}{t}, \text{ ou seja,}$$

$$SQTotal = \left[ (3,052)^2 + (3,924)^2 + \dots + (5,601)^2 \right] - \frac{(44,92)^2}{10} = 5,8548$$

$$\sigma_0^2 = \left[ \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + v S_Y^2 \right] / (t+v)$$

$$\sigma_0^2 = [5,8548 + 18 \times 0,06492] / (10+18)$$

$$\sigma_0^2 = 0,2443$$

Então o valor da estatística de teste  $\lambda$  é:

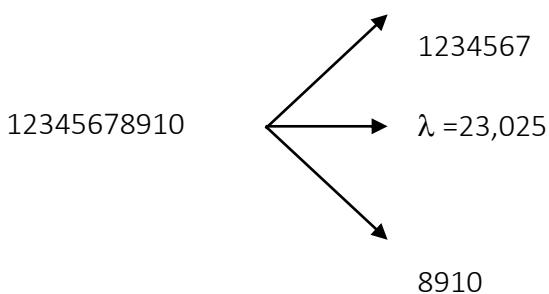
$$\lambda = \pi B_0 / \left[ 2 \hat{\sigma}_0^2 (\pi - 2) \right] = \pi \times 4,0881 / [2 \times 0,2443 - (\pi - 2)]$$

$$\lambda = 23,025$$

$$v_0 = t / (\pi - 2) = 10/\pi = 8,7596 \cong 9,0 \text{ gl}$$

O valor  $\chi_9^2$  para  $\alpha = 5\%$  é 16,9 e para  $\alpha = 1\%$  é 21,7

Como  $\lambda = 23,025 > \chi_9^2 (\alpha = 1\%) = 21,7$ , conclui-se que  $\lambda = 23,025$  é significativo ( $p < 0,01$ ) e que os dois grupos G1=1234 e G2=5678910 são significativamente diferentes a 1% de probabilidade.  
Então temos:



Agora os dois grupos formados devem ser testados separadamente utilizando o mesmo procedimento descrito anteriormente.

Partição em dois grupos para o Grupo 1234567:

Partições possíveis em dois grupos	Valores de SQEntre Grupos (SQEG)
1 / 234567	1,217
12 / 34567	0,9606
123 / 4567	0,9449
1234 / 567	1,019
12345 / 67	0,6577
123456 / 7	0,3198

$$Y_{..} = 28,514$$

$$\bar{Y}_{..} = 4,0734$$

$$FC = (3,052 + \dots + 4,597)^2 / 7 = 116,1571$$

$$SQEG1 = [(3,052)^2 + (3,924 + 3,972 + 4,024 + 4,427 + 4,518 + 4,597)^2 / 6] - FC = 1,217$$

$$SQEG2 = [(3,052 + 3,924)^2 / 2 + (3,972 + 4,024 + 4,427 + 4,518 + 4,597)^2 / 5] - FC = 0,9606$$

$$SQEG3 = [(3,052 + 3,924 + 3,972)^2 / 3 + (4,024 + 4,427 + 4,518 + 4,597)^2 / 4] - FC = 0,9449$$

$$SQEG4 = [(3,052 + 3,924 + 3,972 + 4,024)^2 / 4 + (4,427 + 4,518 + 4,597)^2 / 3] - FC = 1,019$$

$$SQEG5 = [(3,052 + 3,924 + 3,972 + 4,024 + 4,427)^2 / 5 + (4,518 + 4,597)^2 / 2] - FC = 0,6577$$

$$SQEG6 = [(3,052 + 3,924 + 3,972 + 4,024 + 4,427 + 4,518)^2 / 6 + (4,597)^2] - FC = 0,3198$$

Como  $B_0 = \max SQEG$ , então o valor de  $B_0 = 1,217$

$$S\frac{\sum}{Y} = \frac{QM \text{ Residuo}}{r} = \frac{0,1947}{3} = 0,06492$$

$$SQT_{Total} = \sum_i (\bar{Y}_i - \bar{Y}_{..})^2 = 1,6756$$

$$\sigma_0^2 = \left[ \sum_i (\bar{Y}_i - \bar{Y}_{..})^2 + v S\frac{\sum}{Y}^2 \right] / (t+v)$$

$$\sigma_0^2 = [1,6756 + 18 \times 0,06492] / (7+18)$$

$$\sigma_0^2 = 0,1137$$

$$\lambda = \pi B_0 / \left[ 2 \hat{\sigma}_0^2 (\pi - 2) \right]$$

$$\lambda = 14,727$$

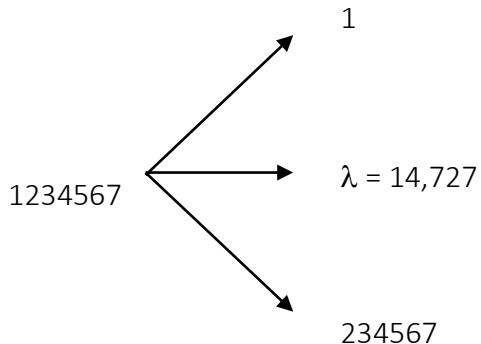
$$v_0 = t / (\pi - 2)$$

$$v_0 = 7 / (\pi - 2) = 6,13 \cong 6,0$$

O valor  $\chi_6^2$  para  $\alpha = 5\%$  é 12,6 e para  $\alpha = 1\%$  é 16,8

Como  $\lambda = 14,727 < \chi_6^2(0,01) = 16,8$ , os dois grupos não são significativamente diferentes a 1% de probabilidade.

Por outro lado, como  $\lambda = 14,727 > \chi^2_6(0,05) = 12,6$ , conclui-se que  $\lambda = 14,727$  é significativo ( $p < 0,05$ ) e que os dois grupos G1=1 e G2=234567 são significativamente diferentes a 5% de probabilidade. Então temos o seguinte:



Partição em dois grupos para o Grupo 8910:

Partições possíveis em dois grupos	Valores de SQEntre Grupos
8 / 910	0,093
89 / 10	0,028

$$Y_{..} = 16,406$$

$$\bar{Y}_{..} = 5,4687$$

$$FC = (5,222 + 5,583 + 5,601)^2/3 = 89,72$$

$$SQEG1 = [(5,222)^2 + (5,583)^2 + (5,601)^2]/2 - FC = 0,093$$

$$SQEG2 = [(5,222 + 5,583)^2/2 + (5,601)^2] - FC = 0,028$$

Logo o valor de  $B_0 = 0,093$

$$SQT_{Total} = \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 = 0,0599$$

$$\sigma_0^2 = \left[ \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 + v S_{\bar{Y}}^2 \right] / (t+v)$$

$$\sigma_0^2 = [0,0599 + 18 \times 0,06492] / (3+18) = 0,0599$$

$$\lambda = \pi B_0 / \left[ 2 \hat{\sigma}_0^2 (\pi - 2) \right]$$

$$\lambda = 2,1331$$

$$v_0 = t / (\pi - 2)$$

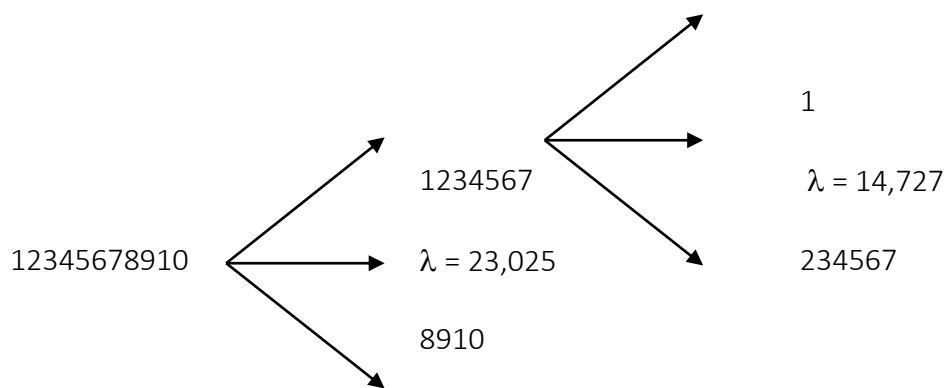
$$v_0 = 3 / (\pi - 2) = 2,6279 \cong 3$$

O valor  $\chi^2_3$  para  $\alpha = 5\%$  é 7,81 e para  $\alpha = 1\%$  é 11,3

Como  $\lambda = 2,1331 < \chi^2_3$  com  $\alpha 5\% \text{ e } 1\%$ , conclui-se que  $\lambda = 2,1331$  é não significativo

$(p > 0,05)$  e que não há diferença entre os grupos tanto a 5% de probabilidade.

Então se tem para o teste de Scott-Knott com o seguinte resultado final:



Outra forma de apresentação do resultado do teste:

Cultivares	Médias
6	3,052 c
8	3,924 b
3	3,972 b
2	4,024 b
7	4,427 b
5	4,518 b
10	4,597 b
4	5,222 a
9	5,583 a
1	5,601 a

A interpretação geral dos resultados da análise estatística dos dados é a seguinte:

- Como  $F = 10,02 > F_{0,05}(9;18) = 2,46$ , logo rejeita-se  $H_0$ , e concluímos que pelo menos 2 médias são diferentes entre si.
- Como o CV% é inferior a 10% (9,82%) podemos dizer que a precisão e a qualidade do experimento são altas.
- As cultivares foram agrupadas em três grupos, sendo que no primeiro grupo ficou apenas a cultivar 6 (a de menor produtividade de grãos), no segundo grupo ficou enquadrado as cultivares 8, 3, 2, 7, 5 e 10, no terceiro grupo ficaram as cultivares 4, 9 e 1 (as de maiores produtividades de grãos).

# CAPITULO 7

## Delineamentos Experimentais Blocos Incompletos

Os princípios e procedimentos envolvidos na construção e aplicação de delineamentos experimentais blocos incompletos são os seguintes:

### Formação de Blocos

Na experimentação, a formação de blocos tem os seguintes objetivos:

- a) Reduzir o erro e aumentar a precisão experimental;
- b) Comparar os tratamentos/genótipos sob condições ambientais mais uniformes.

Entretanto, na prática o uso de blocos completos (contendo todos os tratamentos) apresenta algumas limitações:

- a) No melhoramento de plantas o número de tratamentos pode ser grande e quando os blocos se tornam maiores, as condições ambientais dentro dos blocos tornam-se mais heterogêneas, o que diminui a eficiência dos blocos;
- b) Outros fatores, tais como disponibilidade de área no campo, tamanho da casa de vegetação e número de amostras que podem ser processadas de uma vez, também pode limitar o número de parcelas num bloco.

### Características do Delineamentos Blocos Incompletos

Nos delineamentos blocos incompletos, as parcelas são agrupadas em blocos menores de tamanho não o suficiente para conter todos os tratamentos/genótipos. Blocos menores, contendo diferentes grupos de tratamentos são mais homogêneos. As características que distinguem os blocos incompletos são:

- a) Uso de critério para definir o número de blocos;
- b) Delineamento balanceado ou não balanceado;
- c) Delineamento resolvível ou não resolvível;
- d) Delineamento no formato quadrado ou retangular;
- e) Uso de processo matemático para gerar o esquema do delineamento.

### Delineamentos Blocos Incompletos Balanceados

Para ocorrer balanceamento deve-se ter  $r = \lambda(t-1)/(k-1)$ , em que:

$r$ : número de repetições de cada tratamento/genótipo;

$\lambda$ : número de vezes que dois tratamentos ocorrem juntos no mesmo bloco;  $\lambda$  deve ser inteiro;

$t$ : número de tratamentos/genótipos;

$k$ : número de unidades experimentais/parcelas por bloco ou tamanho do bloco.

Então, tem-se:

Número total de parcelas:  $N = rt = bk$ , em que:

$N$ : número total de parcelas no experimento;

$b$ : número de blocos no experimento.

Cada tratamento ocorre junto com cada um de todos os outros no mesmo bloco um número igual de vezes:

O número de vezes é  $\lambda = r(k-1)/(t-1)$ , em que:

$t$  : número de tratamentos;

$k$  : número de parcelas por bloco;

$r$  : número de repetições de cada tratamento.

Todos os pares de tratamentos são comparados com o mesmo grau de precisão, invariável mesmo que as diferenças entre blocos possam se tornar grande.

Por exemplo:  $t = 45$  tratamentos com blocos de tamanho  $k = 12$  e número de vezes que dois tratamentos ocorrem juntos num mesmo bloco  $\lambda = 2$ , implica em número de repetições  $r = 2(45-1)/(12-1) = 8$ , número total de parcelas  $N = (8)(45) = 360$  e número total de blocos no experimento  $b = 360/12 = 30$ .

No melhoramento de plantas o número mínimo de repetições requeridas para ter balanceamento, em geral é muito grande para ser prático. Então, os blocos incompletos平衡ados são utilizados apenas quando o número de tratamentos não é muito grande.

### Delineamentos Blocos Incompletos Parcialmente Balanceados

Neste delineamento, diferentes pares de tratamentos ocorrem juntos no mesmo bloco um número igual de vezes ou alguns pares nunca ocorrem juntos no mesmo bloco, o que resulta nas seguintes características:

- Comparações entre médias de tratamentos têm diferentes níveis de precisão;
- Maior precisão quando os tratamentos que estão sendo comparados ocorrem no mesmo bloco;
- Análise de variância mais complexa.

Estes delineamentos são comumente utilizados no melhoramento de plantas devido ao grande número de genótipos que geralmente são avaliados nos experimentos de campo.

### Critérios de Formação de Delineamentos Blocos Incompletos

Critério de um fator para formação de blocos: baseado em blocos incompletos casualizados

Critério de dois fatores para formação de blocos:

- Baseado no Quadrado Latino, que é um tipo de delineamento blocos completos que requer  $N = t^2$ . Pode ser impraticável para um grande número de tratamentos;
- Delineamentos Linha-Coluna – tanto linhas quanto colunas ou ambos são blocos incompletos.

### Delineamentos Experimentais Látice

Características dos Delineamentos Látice Quadrado:

- Número de tratamentos deve ser um quadrado perfeito, sendo  $t = k^2$
- Número de blocos por repetição ( $s$ ) é igual a número de parcelas por bloco ( $k$ ), ou seja  $s = k$  e são iguais à raiz quadrada do número de tratamentos ( $t$ )
- Para um balanceamento completo, o número de repetições é  $r = k + 1$

Características dos Delineamentos Látice Retangular:

- a) Número de tratamentos é  $t = s(s-1)$  e o número de parcelas por bloco é  $k = s-1$ , em que:  
 $s$ : número de blocos por repetição;  
b) Por exemplo: látice  $4 \times 5$  tem 4 parcelas por bloco, 5 blocos e 20 tratamentos por repetição.

Características dos Delineamentos Alfa Látice:

- a) Número de tratamentos é  $t = (s)(k)$ , em que:  
 $s$ : número de blocos por repetição  
 $k$ : número de parcelas por bloco;  
b) Mais flexibilidade na escolha de  $s$  e  $k$ .

Esquema Básico para Látice Quadrado  $3 \times 3$

Bloco	Rep I	Rep II	Rep III	Rep IV
1	1 2 3	1 4 7	1 5 9	1 6 8
2	4 5 6	2 5 8	2 6 7	2 4 9
3	7 8 9	3 6 9	3 4 8	3 5 7

Balanceamento: cada tratamento ocorre junto com cada um dos outros tratamentos um número igual de vezes. Uma vez neste caso, então  $\lambda = 1$ .

### Delineamentos Blocos Incompletos Resolvíveis

Nos delineamentos resolvíveis, os blocos são agrupados de forma que cada grupo de blocos constitui uma repetição completa dos tratamentos. Então tem-se:

- a) Bloco = bloco incompleto  
b) Repetição = grupo de blocos incompletos

Neste caso, os experimentos podem ser manejados no campo numa base de repetição por repetição. Repetições completas podem ser perdidas sem perder o experimento todo. Se tiver duas ou mais repetições completas pode-se analisar como delineamento blocos completos casualizados se a formação de blocos incompletos se tornou não efetiva.

### Delineamentos Látice – Resolvíveis

Delineamentos látice são um tipo bem conhecido de delineamento blocos incompletos resolvíveis, com esquema básico como a seguir:

	REP I				
Bloco 3					
Bloco 2					
Bloco 1					

	REP II				
Bloco 6					
Bloco 5					
Bloco 4					

$$b = (s)(r)$$

$t$  : número de tratamentos = 15

$k$  : número de parcelas por bloco = 5

$s$  : número de blocos em cada repetição = 3

$r$  : número de repetições de cada tratamento = 2

$b$  : número total de blocos no experimento = 6

### Tipos de Látices Parcialmente Balanceado

Látices Simples:

- a) Duas repetições – usar as duas primeiras repetições do esquema básico
- b) Látices  $3 \times 3$  e  $4 \times 4$  não são mais precisos que DBC porque  $g_l$  do erro é muito pequeno

Látices Triplo:

- a) Três repetições – usar as três primeiras repetições do esquema básico
- b) Possível para quadrados menores,  $3 \times 3$  e  $4 \times 4$

Látices Quádruplo

- a) Quatro repetições – usar as quatro primeiras repetições do esquema básico
- b) Pode ser repetido látice simples, mas com análise diferente

### Modelo Linear para Delineamento Experimental Látice

Exemplo para um fator de formação de bloco:

$$Y_{ijl} = \mu + \tau_i + \gamma_j + \rho_{l(j)} + \varepsilon_{ijl}, \text{ em que:}$$

$Y_{ijl}$  : observação do  $i$ -ésimo tratamento no  $l$ -ésimo bloco dentro da  $j$ -ésima

repetição;

$\mu$  : média geral;

$\tau_i$  : efeito de tratamento,  $i = 1, 2, \dots, t$ ;

$\gamma_j$  : efeito de repetição,  $j = 1, 2, \dots, r$ ;

$\rho_{l(j)}$  : efeito de bloco dentro de repetição,

$l=1,2,\dots,s$

### Casualização do Delineamento Experimental Látice

Arranjo de campo:

- a) Blocos compostos de parcelas que devem ser as mais homogêneas possíveis
- b) Blocos agrupados em repetições

Casualização de um Esquema Básico:

- a) Casualize a ordem de blocos dentro de repetições
- b) Casualize a ordem de tratamentos dentro de blocos

Exemplo de casualização de um látice  $3 \times 3$  balanceado ( $t=9$ ,  $s=k=3$ ,  $r=4$ ):

1-Atribua  $r$  números aleatórios

Aleatório	Sequência	Ordem
372	1	2
217	2	1
963	3	4
404	4	3

2-Considere o esquema básico

Bloco	Rep I	Rep II	Rep III	Rep IV
1	1 2 3	1 4 7	1 5 9	1 6 8
2	4 5 6	2 5 8	2 6 7	2 4 9
3	7 8 9	3 6 9	3 4 8	3 5 7

3-Casualize a ordem de repetições

Bloco	Rep I	Rep II	Rep III	Rep IV
1	1 4 7	1 2 3	1 6 8	1 5 9
2	2 5 8	4 5 6	2 4 9	2 6 7
3	3 6 9	7 8 9	3 5 7	3 4 8

4-Casualize a ordem de blocos dentro de repetições

	Rep I	Rep II	Rep III	Rep IV
	3	2	3	1
	2	1	1	3
	1	3	2	2

5-Esquema de campo resultante

Bloco	Rep I	Rep II	Rep III	Rep IV
1	3 6 9	4 5 6	3 5 7	1 5 9

2	2 5 8	1 2 3	1 6 8	3 4 8
3	1 4 7	7 8 9	2 4 9	2 6 7

### Análise de Variância de Látice Balanceado

#### Esquema de Análise de Variância

É o mesmo para os latices simples, triplo e quádruplo:

Fonte de Variação	GL	SQ	QM
Total	$rk^2 - 1$	SQTotal	
Repetições	$r - 1$	SQR	
Bloco/Repetição (ajustado)	$k^2 - 1$	SQB	
Tratamentos (não-ajustado)	$r(k-1)$	SQT	$E_b$
Erro Intrabloco	$(k-1)(rk-k-1)$	SQE	$E_e$

Observe que são computados dois termos de erro, ou seja:

$E_b$ : Erro para bloco,  $E_b = SQB / r(k-1)$

$E_e$ : Erro experimental,  $E_e = SQE / (k-1)(rk-k-1)$

#### Cálculos de Somas de Quadrados

$SQTotal = \sum_{i,j,l} Y_{ijl}^2 - \left( G^2 / rk^2 \right)$ , em que:

$G$ : soma das observações de todas as parcelas do experimento;

$SQR = \left( 1/k^2 \right) \sum_j R_j^2 - \left( G^2 / rk^2 \right)$ , em que:

$R_j$ : soma de todas das observações de todas as parcelas na  $j$ -ésima repetição;

$SQB(\text{ajustado}) = [1/kr(r-1)] \sum_{j,l} C_{jl}^2 - [1/k^2 r(r-1)] \sum_j C_j^2$ , em que:

$C_{jl}$ : soma das observações de todos tratamentos no  $l$ -ésimo bloco da  $j$ -ésima repetição

menos  $rB_{jl}$ , sobre todas as repetições;

$B_{jl}$ : soma das observações das  $k$  parcelas no  $l$ -ésimo bloco da  $j$ -ésima repetição;

$C_j$ : soma de todos os valores  $C_{jl}$  na  $j$ -ésima repetição

$$SQT(\text{não-ajustado}) = (1/r) \sum_i T_i^2 - \left( G^2 / rk^2 \right), \text{ em que:}$$

$T_i$ : soma das observações de todas as repetições do  $i$ -ésimo tratamento;

$$SQE = SQT_{\text{Total}} - SQR - SQB - SQT$$

#### Fator de Ajustamento

Deve-se comparar  $E_b$  com  $E_e$ . Se  $E_b \leq E_e$  tem-se que:

- a) O ajuste das médias de tratamentos não terá nenhum efeito
- b) Analise como se o delineamento fosse DBC usando as repetições como blocos completos

Se  $E_b > E_e$ , então utilize um fator de ajuste/ponderação  $A$  da seguinte forma:

$$\text{Cálculo de: } A = (E_b - E_e) / [k(r-1)E_b]$$

Valor de  $A$  é usado para calcular médias de tratamentos ajustadas, por meio da expressão seguinte:

$$\bar{Y}_{i(\text{ajustada})} = \left( T_i + \sum_{j,l} AC_{jl} \right) / r, \text{ para todos os blocos nos quais o } i\text{-ésimo}$$

tratamento ocorre.

#### Teste diferenças entre médias de tratamentos

Para o teste de significância entre médias de tratamentos ajustadas, calcule um quadrado médio ajustado:

$$SQB_{\text{não-ajustado}} = SQB_{naj} = (1/k) \sum_{j,l} B_{jl}^2 - \left( G^2 / rk^2 \right) - SQR$$

$$SQT_{\text{ajustado}} = SQT_{aj} = SQT - Ak(r-1) \left[ \left( rSQB_{naj} \right) / (r-1)(1+kA) - SQB \right]$$

$$QMT_{aj} = SQT_{aj} / (k^2 - 1)$$

Calcule a estatística  $F$  para testar as diferenças entre médias de tratamentos ajustadas como:

$$F = QMT_{aj} / E_e, \text{ com } k^2 - 1 \text{ e } (k-1)(rk-k-1) \text{ graus de liberdade.}$$

# CAPITULO 8

## Delineamentos Experimentais Látice Quadrado

O delineamento blocos completos casualizados (DBC) torna-se menos eficiente à medida que aumentamos o número de tratamentos porque se torna difícil manter a homogeneidade das unidades experimentais, quando o tamanho do bloco aumenta.

Como alternativa para um grande número de tratamentos temos os delineamentos blocos incompletos casualizados, nos quais cada bloco não contém todos os tratamentos, o que possibilita manter um bloco razoavelmente pequeno, mesmo com um grande número de tratamentos, e dentre os quais o látice é o mais comumente utilizado na pesquisa em melhoramento de plantas. Entretanto, o delineamento látice (DL) apresenta as desvantagens seguintes: a) inflexibilidade para o número de tratamentos e/ou repetições; b) diferentes graus de precisão na comparação dos tratamentos; e c) análise de variância complexa.

Uma regra prática útil é considerar a possibilidade de uso do látice apenas quando tornar-se difícil manter um nível razoável de uniformidade das unidades experimentais dentro de um mesmo bloco. Os dois tipos mais comuns de látice são o látice balanceado e o látice parcialmente balanceado; ambos os delineamentos requerem que o número de tratamentos seja um quadrado perfeito.

### Látice Quadrado Balanceado

As características deste delineamento são as seguintes:

- a) O número de tratamentos ( $t$ ) deve ser um quadrado perfeito ( $t = k^2$ );
- b) O tamanho do bloco ( $k$ ) é igual à raiz quadrada do número de tratamentos ( $k = t^{1/2}$ );
- c) O número de repetições ( $r$ ) é um mais o tamanho do bloco ( $r = k + 1$ );
- d) O número de blocos por repetição é  $k = t^{1/2}$ .

A casualização e o croqui serão ilustrados para um delineamento látice quadrado  $3 \times 3$  balanceado envolvendo nove tratamentos. Portanto, temos quatro repetições, cada uma consistindo de três blocos contendo três parcelas.

Os passos a seguir são:

- 1º) Divida a área experimental em  $r = (k + 1)$  repetições, cada uma contendo  $t = k^2$  parcelas, sendo  $k = t^{1/2}$ . Para o exemplo considerado, a área experimental é dividida em  $r = 4$  repetições, cada uma contendo  $t = 9$  parcelas;

- 2º) Divida cada repetição em  $k$  blocos, cada uma contendo  $k$  parcelas. Para nosso exemplo, cada repetição é dividida em  $k=3$  blocos, cada um contendo  $k=3$  parcelas, como mostrado a seguir:

Bloco 1	1   2   3	10   11   12	19   20   21	28   29   30
Bloco 2	4   5   6	13   14   15	22   23   24	31   32   33
Bloco 3	7   8   9	16   17   18	25   26   27	34   35   36
	Repetição I	Repetição II	Repetição III	Repetição IV

Figura 1. Croqui da área experimental, que consiste de 36 parcelas agrupadas em 4 repetições com 3 blocos cada e 3 parcelas por bloco, para o delineamento látice  $3 \times 3$ .

- 3º) Selecionar em tabelas um esquema básico de látice balanceado que corresponda ao número de tratamentos a ser avaliado. Para o nosso exemplo, o esquema adequado é o do látice  $3 \times 3$  balanceado, mostrado a seguir:

Esquema básico de um DLQ balanceado, que envolve 9 tratamentos em blocos de 3 unidades experimentais e 4 repetições

Numeração de Blocos	Numeração de Rep / Tratamentos			
	Rep. I	Rep. II	Rep. III	Rep. IV
1	1 2 3	1 4 7	1 5 9	1 6 8
2	4 5 6	2 5 8	2 6 7	2 4 9
3	7 8 9	3 6 9	3 4 8	3 5 7

- 4º) Casualizar o arranjo de repetições e de blocos dentro de repetições do esquema básico utilizado, seguindo, por exemplo, uma tabela de números aleatórios.

Então o novo esquema fica:

Numeração de Blocos	Numeração de Rep / Tratamentos			
	Rep. I	Rep. II	Rep. III	Rep. IV
1	6 9 3	5 4 6	7 5 3	1 9 5
2	8 5 2	3 2 1	8 1 6	8 3 4
3	4 7 1	9 8 7	2 4 9	6 7 2

- 5º) Aplique o resultado final do processo de casualização anterior ao croqui de campo da Figura 1, o que resulta na Figura 2, mostrada a seguir:

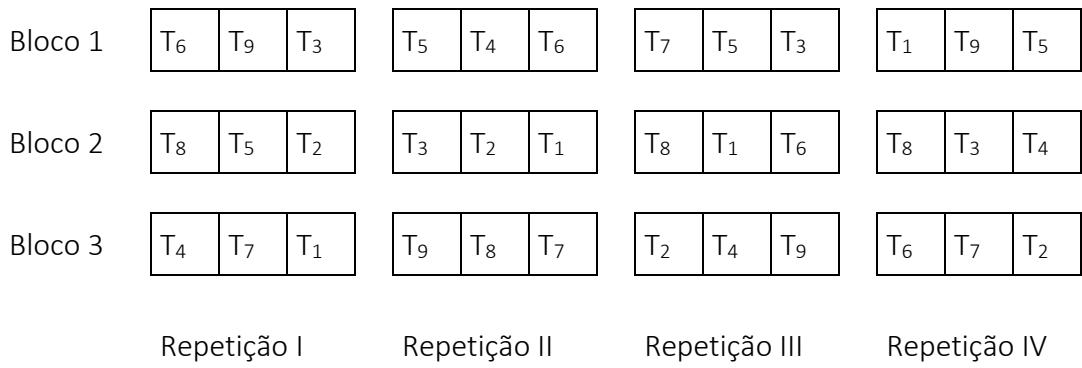


Figura 2. Croqui de campo de um látice balanceado, envolvendo nove tratamentos.

Uma característica importante no delineamento de um látice quadrado balanceado é que todos os pares de tratamentos ocorrem juntos apenas uma vez, no mesmo bloco. Por exemplo, o tratamento 1 aparece apenas uma vez: com os tratamentos 4 e 7 no bloco 3 da repetição I; com os tratamentos 2 e 3 no bloco 2 da repetição II; com os tratamentos 6 e 8 no bloco 2 da repetição III e com os tratamentos 5 e 9 no bloco 1 da repetição IV. A consequência disto é que o grau de precisão para comparar cada par de tratamentos num delineamento látice quadrado balanceado é o mesmo para todos os pares.

Para a análise de variância deve-se considerar quatro fontes de variações básicas que são: repetição, bloco dentro de repetição, tratamento e erro experimental. Em relação ao DBC, a fonte bloco dentro de repetição constitui uma fonte de variação adicional que reflete as diferenças entre blocos da mesma repetição.

### Exemplo de aplicação

O procedimento de cálculo da análise de variância do delineamento látice balanceado será ilustrado por um experimento envolvendo 9 variedades de milho num látice 3 x 3 (Extraído de Leclerc et al., 1962). Os dados de produtividade em kg/parcela estão apresentados a seguir:

Bloco	Kg/parcela	Total do Bloco			Kg/parcela	Total do Bloco
		Bloco	(B)	Rep. I		
1	(2) 15,6	(3) 12,9	(1) 10,4	38,9	4	(6) 14,5
						(9) 15,2
2	(7) 11,4	(9) 13,4	(8) 14,3	39,1	5	(1) 10,7
						(7) 13,7
3	(6) 13,2	(4) 12,8	(5) 15,3	41,3	6	(8) 15,0
						(5) 13,9
						(2) 14,7
						44,0
						37,3
						43,6

Total de Rep. I			119,3			Total de Rep. II			124,9		
Rep. III						Rep. IV					
7	(1)	(9)	(5)	32,9	10	(8)	(1)	(6)	39,5		
	7,7	13,3	11,9			14,5	10,9	14,1			
8	(8)	(4)	(3)	35,9	11	(3)	(7)	(5)	39,7		
	12,6	11,4	11,9			12,7	13,3	13,7			
9	(2)	(6)	(7)	39,0	12	(9)	(2)	(4)	45,4		
	12,1	15,6	11,3			13,5	16,6	15,3			
Total de Rep. III			107,8			Total de Rep. IV			124,6		

O procedimento analítico para a análise de variância é o seguinte:

- 1º) Calcule os totais de blocos (B) e de repetições (R);
- 2º) Calcule os totais de tratamentos (T) e o total geral (G);
- 3º) Para cada tratamento, calcule o valor  $B_i$  como sendo a soma dos totais de blocos para todos os blocos nos quais um determinado tratamento aparece;
- 4º) Para cada tratamento, calcule o valor  $W_i$ , que é dado por  $W_i = kT_i - (k+1)B_i + G$ . Observe que a soma dos valores  $W_i$ , para todos os tratamentos, deve ser zero;
- 5º) Construa um esquema de análise de variância, como:

Fontes de Variação	Graus de Liberdade
Repetição	$k=3$
Tratamento (não-ajustado)	$k^2-1=8$
Bloco (ajustado)	$k^2-1=8$
Erro Intrabloco	$(k-1)(k^2-1)=16$
Tratamento (ajustado)	$[(k^2-1)=8]$
Erro Efetivo	$[(k-1)(k^2-1)=16]$
Total	$k^2(k+1)-1=35$

A tabela a seguir fornece os valores calculados de  $T_i$ ,  $B_i$  e  $W_i$ , para cada tratamento:

Tabela -Valores de  $T_i$ ,  $B_i$  e  $W_i$ , para os dados do látice balanceado  $3 \times 3$ .

Numeração de Tratamento	Total de $(T_i)$	Total de Bloco $(B_i)$	$W_i = 3T_i - 4B_i + G$
1	39,7	148,6	1,3
2	59,0	166,9	-14,0
3	51,8	158,5	-2,0
4	52,4	159,9	-5,8
5	54,8	157,5	11,0
6	57,4	163,8	-6,4
7	49,7	155,1	5,3
8	56,4	158,1	13,4
9	55,4	161,4	-2,8
Total	476,6 (G)	1429,8	=0,0

6º) Calcule as somas de quadrados total, de repetição e de tratamento não ajustado como:

$$C = \frac{G^2}{(k^2)(k+1)}$$

$$C = \frac{(476,6)^2}{(9)(4)}$$

$$C = 6309,6544$$

$$SQTotal = \sum Y^2 - C$$

$$SQTotal = (15,6)^2 + (12,9)^2 + \dots + (15,3)^2 - 6309,6544$$

$$SQTotal = 6424,0000 - 6309,6544$$

$$SQTotal = 114,3456$$

$$SQRe\ p = \sum \frac{R^2}{k^2} - C$$

$$SQRe\ p = \frac{(119,3)^2 + (124,9)^2 + (107,8)^2 + (124,6)^2}{9} - C$$

$$SQRe\ p = 6330,9444 - 6309,6544$$

$$SQRe\ p = 21,29004$$

$$SQTrat\ (não-ajust.) = \sum \frac{T^2}{(k+1)} - C$$

$$SQTrat\ (não-ajust.) = \frac{(39,7)^2 + (59,0)^2 + \dots + (55,4)^2}{4} - C$$

$$SQTrat\ (não-ajust.) = 6376,0250 - 6309,6544$$

$$SQTrat\ (não-ajust.) = 66,3706$$

- 7º) Calcule a soma de quadrados de blocos ajustados, isto é, a soma de quadrados para bloco dentro de repetição, ajustada para efeitos de tratamentos, que é dada por:

$$SQBlocos\ (ajust.) = \frac{\sum_{i=1}^t W_i^2}{(k^3)(k+1)}$$

$$SQBlocos\ (ajust.) = \frac{(1,3)^2 + (-14,0)^2 + \dots + (-2,5)^2}{(27)(4)}$$

$$SQBlocos\ (ajust.) = 5,6739$$

- 8º) Calcule a soma de quadrado do Erro Intrabloco como:

$$SQErro\ Intrablocos = SQTotal - SQRe\ p -$$

$$- SQTrat\ (não-ajust.) - SQBlocos\ (ajust.)$$

$$SQErro\ Intrablocos = 114,34560 - 21,29004 - 66,3706 - 5,65917$$

$$SQErro\ Intrablocos = 21,01106$$

- 9º) Calcule o quadrado médio de blocos ajustados e o quadrado médio do Erro Intrabloco como:

$$QMBlocos \text{ (ajust)} = \frac{SQB locos \text{ (ajust)}}{k^2 - 1}$$

$$QMBlocos \text{ (ajust)} = \frac{5,6739}{8}$$

$$QMBlocos \text{ (ajust)} = 0,7092375$$

$$QMErro \text{ Intrabloco} = \frac{SQErro \text{ Intrabloco}}{(k-1)(k^2-1)}$$

$$QMErro \text{ Intrabloco} = \frac{21,01106}{(2)(8)}$$

$$QMErro \text{ Intrabloco} = 1,3131913$$

10º) Para cada tratamento, calcule os totais de tratamentos ajustados ( $T'$ ) como:

$$T'_i = T_i + \mu \quad W_i, \text{ onde}$$

$$\mu = \frac{QMBlocos \text{ (ajust.)} - QMErro \text{ Intrabloco}}{k^2 [QMBlocos \text{ (ajust.)}]}$$

Deve ser observado que se o  $QMErro \text{ Intrabloco}$  for maior que o  $QMBlocos \text{ (ajust.)}$ , o  $\mu$  é tomado como sendo zero e não é necessário nenhum ajuste para tratamentos. Neste caso, o teste de significância F para tratamentos é feito da maneira usual, ou seja, como a razão entre QM Tratamentos (não-ajustado) e QM do Erro Intrabloco.

No exemplo considerado, o fator de ajustamento  $\mu$  é:

$$\mu = \frac{0,70736 - 1,314112}{9 (0,70736)}$$

$$\mu = -0,94617$$

Como o QM Erro Intrabloco é maior que o QM Blocos (ajustado), considera-se  $\mu = 0$ . Se o fator de ajustamento é zero, então  $T = T'$ .

11º) Para cada tratamento, calcule a média de tratamento ajustado ( $M'$ ) como:

$$M' = \frac{T'}{k+1}$$

Os totais e as médias de tratamentos ajustadas, que neste caso são iguais as médias de tratamentos não ajustadas, constam do quadro a seguir:

Quadro. Totais e médias de tratamentos ajustados.

Número do Tratamento	$T' = T + \mu$	$W_i$	$M' = \frac{T'}{4}$
1	39,7		9,93
2	59,0		14,75
3	51,8		12,95
4	52,4		13,10
5	54,8		13,70
6	57,4		14,35
7	49,7		12,43
8	56,4		14,10
9	55,4		13,85

12º) Calcule o quadrado médio de tratamentos ajustado como:

$$QMTrat\ (ajust.) = \left[ \frac{1}{(k+1)(k^2-1)} \right] \left[ \sum T'^2 - \frac{G^2}{k^2} \right]$$

$$QMTrat\ (ajust.) = \left[ \frac{1}{(4)(8)} \right] \left\{ \left[ (39,7)^2 + (59,0)^2 + \dots + (55,4)^2 \right] - \frac{(476,6)^2}{9} \right\}$$

$$QMTrat\ (ajust.) = \left[ \frac{1}{32} \right] \left\{ [25504,100 - 25238,6178] \right\}$$

$$QMTrat\ (ajust.) = 8,29632$$

13º) Calcule o quadrado médio do erro efetivo como:

$$QMErro\ Efetivo = (QMErro\ Intrabloco)(1+k\mu)$$

$$QMErro\ Efetivo = (1,3131913)(1)$$

$$QMErro\ Efetivo = 1,3131913$$

14º) Calcule o valor F para testar a significância de tratamentos como:

$$F = \frac{QMT_{\text{Tratamentos (ajust.)}}}{QME_{\text{erro Efetivo}}}$$

$$F = \frac{8,29632}{1,314112}$$

$$F = 6,32 (p < 0,01)$$

15º) Compare o valor F calculado com o F tabelado com  $n_1 = (k^2 - 1) = 8$  e

$n_2 = (k-1)(k^2 - 1) = 16$  graus de liberdade ( $F_{1\%} = 3,89$ ). Como o valor de F calculado é maior que o valor de F tabelado, a uma probabilidade de 1%, deve existir pelo menos uma diferença significativa entre médias de tratamento.

16º) Construa o quadro resumo da Análise de Variância como:

Quadro. Análise de Variância de um látice  $3 \times 3$  para produção (kg/parcela) de variedades de milho.

F.V.	G.L.	S.Q.	Q.M.	F
Repetição	3	21,2900		
Tratamentos (n-ajust.)	8	66,3706		
Blocos (ajust.)	8	5,6592	0,707396	
Erro Intrabloco	16	21,0258	1,314112	
Tratamento (ajust.)	(8)	-	8,29632	6,31**
Erro Efetivo	(16)	-	1,314112	
Total	35	114,3456		

17º) Calcule o valor do coeficiente de variação como:

$$CV = \sqrt{\frac{QME_{\text{erro Efetivo}}}{Média Geral}} (100)$$

$$CV = \sqrt{\frac{1,314112}{13,23889}} (100) = 8,66\%$$

18º) Estima-se o ganho em precisão de um delineamento látice quadrado balanceado em relação ao DBC como:

$$ER = \frac{100 [SQBlocos (ajust.) + SQErro Intrabloco]}{k (k^2 - 1)(QMErro Efetivo)}$$

$$ER = \frac{100 (5,6592 + 21,0258)}{(3)(8)(1,314112)} = 84,61\%$$

Este resultado indica que o uso do látice balanceado 3 x 3 é estimado como sendo responsável por um decréscimo de 15,4% na precisão experimental que teria sido obtida com o DBC.

### Látice Quadrado Parcialmente Balanceado

Este delineamento também exige que o número de tratamentos seja um quadrado perfeito e que o tamanho do bloco seja igual à raiz quadrada do número de tratamentos, mas não exige que o número de repetições seja uma função do número de tratamentos. Desta forma, qualquer número de repetições pode ser usado num delineamento látice parcialmente balanceado.

Diz-se que o látice quadrado parcialmente balanceado é simples quando tem duas repetições; triplo com três repetições; quádruplo com quatro repetições e assim sucessivamente. Esta flexibilidade na escolha do número de repetições tem como consequência a perda de simetria no arranjo dos tratamentos sobre os blocos, ou seja, alguns pares de tratamentos nunca vão ocorrer juntos num mesmo bloco.

Desta forma, os pares de tratamentos que são testados no mesmo bloco são comparados com um nível de precisão que é superior àquele dos tratamentos que não são testados no mesmo bloco. Como existe mais de um nível de precisão para comparar os tratamentos, a análise dos dados torna-se mais complexa.

Os pares de tratamentos que ocorrem juntos no mesmo bloco são chamados de primeiros associados e testados com o QMErro1. Os pares de tratamentos que não ocorrem juntos em nenhum bloco são chamados de segundos associados e são testados com o QMErro2.

Os procedimentos para casualização e construção de croquis de campo de delineamentos látice quadrado parcialmente balanceado são semelhantes àqueles para látice quadrado balanceado, exceto para a mudança no número de repetições.

Quando o número de repetições (r) de um látice quadrado parcialmente balanceado é maior que três e par, o esquema básico deste látice pode ser obtido como: a) as primeiras repetições do esquema básico de um látice quadrado balanceado que tem o mesmo número de tratamentos; ou b) as primeiras r repetições do esquema básico do látice quadrado balanceado que tem o mesmo número de tratamentos, repetido p vezes, com p casualizações separadas e independentes. A opção (a) é dita ser sem repetição do esquema básico e a (b) com repetição do esquema básico.

Por exemplo, para um látice 5 x 5 quádruplo (látice quadrado parcialmente balanceado com quatro repetições) o esquema básico pode ser obtido tanto como as primeiras quatro repetições do látice 5 x 5 balanceado ou como látice 5 x 5 simples repetido duas vezes (p=2) e com duas casualizações separadas, como se fosse dois látices 5 x 5 simples.

Para o látice triplo (látice quadrado parcialmente balanceado com três repetições) as três primeiras repetições do esquema básico do látice quadrado balanceado correspondente podem ser utilizadas. Geralmente, o procedimento que usa o esquema básico sem repetições do mesmo é preferido porque ele fornece uma simetria de tratamentos mais próxima daquela encontrada num látice balanceado.

Uma vez escolhido o esquema básico, o processo de casualização do croqui de campo é semelhante ao visto para o látice quadrado balanceado. Os procedimentos para a análise de variância de um

látice quadrado parcialmente balanceado são ligeiramente diferentes para os casos com repetição e sem repetição de um esquema básico.

### Exemplo de Aplicação

A análise de variância apresentada a seguir é de um látice quadrado parcialmente balanceado sem repetição do esquema básico. Trata-se de um látice triplo 9 x 9 que avaliou o desempenho de 81 variedades de arroz. Os dados de produtividade de grãos estão no quadro a seguir (Extraído de Gómez & Gómez, 1984).

Quadro. Produtividade de grãos de um experimento com 81 variedades de arroz conduzido no delineamento látice triplo 9 x 9

Blocos	Produtividade de grãos (t/ha)									Total de Bloco (B)
	Rep. I									
	(1) 2,70	(2) 1,60	(3) 4,45	(4) 2,91	(5) 2,78	(6) 3,32	(7) 1,70	(8) 4,72	(9) 4,79	28,97
1	(10) 4,20	(11) 5,22	(12) 3,96	(13) 1,51	(14) 3,48	(15) 4,69	(16) 1,57	(17) 2,61	(18) 3,16	30,40
2	(19) 4,63	(20) 3,33	(21) 6,31	(22) 6,08	(23) 1,86	(24) 4,10	(25) 5,72	(26) 5,87	(27) 4,20	42,10
3	(28) 3,74	(29) 3,05	(30) 5,16	(31) 4,76	(32) 3,75	(33) 3,66	(34) 4,52	(35) 4,64	(36) 5,36	38,64
4	(37) 4,76	(38) 4,43	(39) 5,36	(40) 4,73	(41) 5,30	(42) 3,93	(43) 3,37	(44) 3,74	(45) 4,06	39,68
5	(46) 3,45	(47) 2,56	(48) 2,39	(49) 2,30	(50) 3,54	(51) 3,66	(52) 1,20	(53) 3,34	(54) 4,04	26,48
6	(55) 3,99	(56) 4,48	(57) 2,69	(58) 3,95	(59) 2,59	(60) 3,99	(61) 4,37	(62) 4,24	(63) 3,70	34,00
7	(64) 5,29	(65) 3,58	(66) 2,14	(67) 5,54	(68) 5,14	(69) 5,73	(70) 3,38	(71) 3,63	(72) 5,08	39,51
8	(73) 3,76	(74) 6,45	(75) 3,96	(76) 3,64	(77) 4,42	(78) 6,57	(79) 6,39	(80) 3,39	(81) 4,89	43,47
	Total de Repetição (R <sub>I</sub> )									323,25

Continua...

Continuação...

	Produção de grãos (t/ha)										Total de Blocos (B)
	Rep. II										
1	(1) 3,06	(10) 2,08	(19) 2,95	(28) 3,75	(37) 4,08	(46) 3,88	(55) 2,14	(64) 3,68	(73) 2,85		28,47
2	(2) 1,61	(11) 5,30	(20) 2,75	(29) 4,06	(38) 3,89	(47) 2,60	(56) 4,19	(65) 3,14	(74) 4,82		32,36
3	(3) 4,19	(12) 3,33	(21) 4,76	(30) 4,99	(39) 4,58	(48) 3,17	(57) 2,69	(66) 2,57	(75) 3,82		34,01
4	(4) 1,61	(13) 5,30	(22) 2,75	(31) 4,06	(40) 3,89	(49) 2,60	(58) 4,19	(67) 3,14	(76) 4,82		34,18
5	(5) 3,81	(14) 3,48	(23) 1,87	(32) 4,34	(41) 4,36	(50) 3,24	(59) 3,62	(68) 4,49	(77) 3,62		32,83
6	(6) 3,34	(15) 3,30	(24) 3,68	(33) 3,84	(42) 4,25	(51) 3,90	(60) 3,64	(69) 5,09	(78) 6,10		37,14
7	(7) 2,98	(16) 2,69	(25) 5,55	(34) 3,52	(43) 4,03	(52) 1,20	(61) 4,36	(70) 3,18	(79) 6,77		34,28
8	(8) 4,20	(17) 2,69	(26) 5,14	(35) 4,32	(44) 3,47	(53) 3,41	(62) 3,74	(71) 3,67	(80) 2,27		32,91
9	(9) 4,75	(18) 2,59	(27) 3,94	(36) 4,51	(45) 3,10	(54) 3,59	(63) 2,70	(72) 4,40	(81) 4,86		34,44
Total de Repetição (R <sub>II</sub> )											300,62

Rep. III

1	(1) 3,52	(12) 2,18	(20) 3,50	(34) 3,30	(45) 3,88	(53) 2,45	(58) 3,75	(70) 4,45	(77) 4,14		31,17
2	(2) 0,79	(10) 3,58	(21) 4,83	(35) 3,63	(43) 3,02	(54) 4,20	(59) 3,59	(67) 5,06	(78) 6,51		35,21

	(3)	(11)	(19)	(36)	(44)	(52)	(60)	(68)	(76)	
3	4,69	5,33	4,43	5,31	4,13	1,98	4,66	4,50	4,50	39,53
4	(4)	(15)	(23)	(28)	(39)	(47)	(61)	(72)	(80)	33,18
5	3,79	0,88	3,40	4,92	2,12	1,89	3,73	3,51	3,50	27,74
6	(6)	(14)	(22)	(30)	(38)	(46)	(63)	(71)	(79)	37,13
7	2,35	2,87	5,50	2,72	4,20	2,87	2,99	1,62	5,33	30,45
8	(8)	(16)	(27)	(32)	(40)	(51)	(56)	(64)	(75)	32,78
9	4,51	1,26	4,20	3,19	4,76	3,35	3,61	4,52	3,38	
	(9)	(17)	(25)	(33)	(41)	(49)	(57)	(65)	(73)	33,99
	4,21	3,17	5,03	3,34	5,31	3,05	3,19	2,63	4,06	
Total de Repetição ( $R_{III}$ )										301,18

Para computar a Análise de Variância devem ser feitos os cálculos seguintes:

1º) Calcule os totais dos blocos (B), totais de repetição (R) e o total geral (G):

$$G = R_I + R_{II} + R_{III}$$

$$G = 323,25 + 300,62 + 301,18 = 925,05$$

2º) Calcule os totais de tratamentos (T):

Tratamento									
Nº	Total (T)								
1	9,28	2	4,00	3	13,33	4	8,96	5	10,38
6	10,00	7	7,03	8	13,43	9	13,75	10	9,86
11	15,85	12	9,47	13	4,89	14	10,90	15	12,29
16	5,52	17	8,47	18	8,62	19	12,01	20	9,58
21	15,81	22	16,67	23	5,75	24	11,18	25	16,30
26	16,51	27	12,34	28	11,06	29	12,03	30	15,49
31	11,19	32	11,28	33	10,84	34	11,34	35	12,59
36	15,18	37	10,96	38	12,79	39	15,74	40	14,34
41	14,97	42	12,38	43	10,42	44	11,34	45	11,04
46	11,51	47	7,74	48	7,45	49	8,22	50	9,65
51	10,91	52	4,38	53	9,20	54	11,83	55	9,12
56	12,28	57	8,57	58	11,49	59	9,80	60	12,29
61	13,00	62	11,71	63	9,10	64	13,49	65	9,35

66	6,33	67	15,88	68	14,13	69	15,24	70	10,07
71	11,26	72	14,32	73	10,67	74	16,60	75	11,16
76	11,46	77	12,18	78	19,18	79	16,64	80	8,40
81	13,25								

3º) Construa o esquema de Análise de Variância do látice triplo 9 x 9 como:

Fontes de Variação	Graus de Liberdade
Repetição	$r-1=2$
Bloco (ajustado)	$r(k-1)=24$
Tratamento (não-ajustado)	$k^2-1=80$
Erro Intrabloco	$(k-1)(rk-k-1)=136$
Tratamento (ajustado)	$\left[ k^2-1=80 \right]$
Total	$rk^2-1=242$

Obs.: r é o número de repetições e k é o tamanho do bloco, ou seja, o número de parcelas por bloco.

4º) Calcule as somas de quadrados total, de repetição e de tratamento (não-ajustado) da maneira usual:

$$C = \frac{G^2}{(r)(k^2)} = \frac{(925,05)^2}{(3)(81)} = 3521,4712$$

$$SQTotal = \sum Y^2 - C$$

$$SQTotal = \left[ (2,70)^2 + (1,60)^2 + \dots + (4,06)^2 \right] - 3521,4712$$

$$SQTotal = 308,9883$$

$$SQR = \frac{\sum R^2}{k^2} - C$$

$$SQR = \frac{(323,25)^2 + (300,62)^2 + (301,18)^2}{81} - 3521,4712$$

$$SQR = 4,1132$$

$$SQTrat(Não-Ajustado) = \frac{\sum T^2}{r} - C$$

$$SQTrat(Não-Ajustado) = \frac{(9,28)^2 + (4,00)^2 + \dots + (13,25)^2}{3} - \\ - 3521,4712$$

$$SQTrat(NãoAjustado) = 256,7386$$

- 5º) Em cada repetição, para cada bloco, calcule o valor  $C_b = M_b - rB$ , onde  $M_b$  é a soma dos totais de tratamentos para todos os tratamentos que aparecem naquele determinado bloco e  $B$  é o total deste bloco na repetição considerada. Por exemplo, na repetição II o bloco 2 contém os tratamentos 2, 11, 20, 29, 38, 47, 56, 65 e 74; portanto, o valor  $M$  para o bloco 2 da repetição II é:

$$M_b = T_2 + T_{11} + T_{20} + T_{29} + T_{38} + T_{47} + T_{56} + T_{65} + T_{74}$$

$$M_b = 4,00 + 15,85 + \dots + 16,60 = 100,22$$

O total do bloco 2 na repetição 2 é 32,36. Então, tem-se:

$$C_b = 100,22 - 3(32,36) = 3,14$$

Os valores  $C_b$ , para os 27 blocos são:

Rep. I		Rep. II		Rep. III	
Bloco	$C_b$	Bloco	$C_b$	Bloco	$C_b$
1	3,25	1	12,55	1	5,34
2	-5,33	2	3,14	2	3,74
3	-10,15	3	1,32	3	-8,62
4	-4,92	4	0,56	4	-2,28
5	-5,06	5	0,55	5	8,70
6	1,45	6	2,92	6	2,97
7	-4,64	7	-8,14	7	6,08
8	-8,43	8	4,18	8	6,41
9	-10,87	9	6,11	9	-0,83
Total	-44,70	Total	23,19	Total	21,51

- 6º) Para cada repetição, calcule a soma de valores  $C_b$  sobre todos os blocos:

$$R_c(I) = 3,25 + (-5,33) + \dots + (-10,87) = -44,70$$

$$R_c(II) = 12,55 + 3,14 + \dots + 6,11 = 23,19$$

$$R_c(III) = 5,34 + 3,74 + \dots + (-0,83) = 21,51$$

Observe que os valores  $R_c$  devem ter soma igual a zero.

- 7º) Calcule a SQBlocos (ajustado) como:

$$SQB(\text{ajustado}) = \frac{\sum C_b^2}{(k)(r)(r-1)} - \frac{\sum R_c^2}{\binom{k^2}{2}(r)(r-1)}$$

$$SQB(\text{ajustado}) = \frac{(3,25)^2 + (-5,33)^2 + \dots + (-0,83)^2}{(9)(3)(3-1)} - \frac{(-44,70)^2 + (23,19)^2 + (21,15)^2}{(81)(3)(3-1)}$$

$$SQB(\text{ajustado}) = 12,1492$$

- 8º) Calcule a SQErro Intrabloco como:

$$SQErroIntrabloco = SQT_{\text{Total}} - SQR - SQT_{\text{(não ajustado)}} - SQB(\text{ajustado})$$

$$SQErroIntrabloco = 308,9883 - 4,1132 - 256,7386 - 12,1492$$

$$SQErroIntrabloco = 35,9873$$

- 9º) Calcule o Quadrado Médio de Bloco (ajustado) e o Quadrado Médio do Erro Intrabloco como:

$$QMB(\text{ajustado}) = \frac{SQB(\text{ajustado})}{r(k-1)}$$

$$QMB(\text{ajustado}) = \frac{12,1492}{3(9-1)}$$

$$QMB(\text{ajustado}) = 0,5062$$

$$QMErroIntrabloco = \frac{SQErroIntrabloco}{(k-1)(rk-k-1)}$$

$$QMErroIntrabloco = \frac{35,9873}{(9-1)[(3)(9)-9-1]}$$

$$QMErroIntrabloco = 0,2646$$

10º) Calcule o fator de ajustamento  $\mu$  como:

$$\mu = \frac{\left( \frac{1}{QMErro} - \frac{2}{3QMBloco - QMErro} \right)}{k \left( \frac{2}{QMErro} + \frac{2}{3QMBloco - QMErro} \right)}, \text{ onde}$$

QMErro = Quadrado Médio do Erro Intrabloco; e

QMBloco = Quadrado Médio de Blocos (ajustado).

Observe que se QMBlocos é menor que QMErro,  $\mu$  é tomado como sendo zero e nenhum ajustamento é feito. Neste caso, o teste de significância F para o efeito de tratamento é feito da maneira usual como a razão entre QMTratamentos (não-ajustado) e QMErro Intrabloco. Para o exemplo considerado tem-se que:

$$\mu = \frac{\left( \frac{1}{0,2646} - \frac{2}{3(0,5062) - 0,2646} \right)}{9 \left( \frac{2}{0,2646} + \frac{2}{3(0,5062) - 0,2646} \right)}$$

$$\mu = 0,0265$$

11º) Para cada tratamento, calcule os totais de tratamentos ajustados  $(T'_i)$  como:

$$T'_i = T_i + \mu \sum C_b$$

Onde  $\sum C_b$  envolve o  $C_b$  de todos os blocos nos quais aparece um determinado tratamento.

Por exemplo, o total de tratamento ajustado para o tratamento número 2 é calculado como:

$$T'_2 = 4,00 + 0,0265(3,25 + 3,14 + 3,74) = 4,27$$

Para as comparações de médias deve-se utilizar as médias de tratamento ajustadas, que são computadas simplesmente dividindo cada total de tratamento ajustado pelo número de repetições.

12º) Calcule a Soma de Quadrados de Tratamento Ajustada e o valor F do teste de significância.

$$SQT(\text{ajustado}) = SQT(\text{não-ajustado}) - A, \text{ onde}$$

$$A = \left( \frac{1}{QMErro} - \frac{2}{3QMBloco - QMErro} \right) \times \\ \left[ (QMErro)B_{\mu} - (k-1)(QMErro)(rQMBloco - QMErro) \right] \\ e B_{\mu} = \frac{\sum B^2}{k} - \frac{\sum R^2}{k^2}.$$

Então tem-se que:

$$B_{\mu} = \frac{(28,97)^2 + (30,40)^2 + \dots + (33,99)^2}{9} - \\ - \frac{(323,25)^2 + (300,62)^2 + (301,18)^2}{81} \\ B_{\mu} = 49,4653$$

$$A = \left( \frac{1}{0,2646} - \frac{2}{3(0,5062) - 0,2646} \right) \times \\ \{(0,2646)(49,4653) - 8(0,2646)[3(0,5062) - 0,2646]\} \\ A = 22,7921$$

$$SQT(ajustado) = 256,7386 - 22,7921$$

$$SQT(ajustado) = 233,9465$$

$$QMT(ajustado) = \frac{SQT(ajustado)}{k^2 - 1}$$

$$QMT(ajustado) = \frac{233,9465}{80} = 2,9243$$

$$F = \frac{QMT(ajustado)}{QMErro} = \frac{2,9243}{0,2646} = 11,05$$

O coeficiente de variação é computado como:

$$CV(\%) = \frac{\sqrt{QMErro}}{MédiaGeral} (100)$$

$$CV(\%) = \frac{\sqrt{0,2646}}{3,81} (100) = 13,50$$

13º) Construa o quadro de Análise de Variância como:

F.V.	G.L.	S.Q.	Q.M.	F
Repetição	2	4,1132		
Blocos (ajust.)	24	12,1492	0,5062	
Tratamentos (não-ajust.)	80	256,7386		
Erro Intrabloco	136	35,9873	0,2646	
Tratamento (ajust.)	(80)	233,9465	2,9243	11,05*
Total	242	308,9883		

14º) Calcule o quadrado médio do erro efetivo.

Para um látice parcialmente balanceado, existem dois erros: um para comparações entre tratamentos que aparecem no mesmo bloco ( $QMErroEfetivo_1$ ) e outro para comparações entre tratamentos que não aparecem no mesmo bloco ( $QMErroEfetivo_2$ ), que são dados por:

$$QMErroEfetivo_1 = \frac{QMErro}{k} \left[ \frac{\frac{6}{\overline{QMErro}}}{\frac{2}{QMErro} + \frac{2}{3QMBloco - QMErro}} + (k-2) \right]$$

$$QMErroEfetivo_2 = \frac{QMErro}{k} \left[ \frac{\frac{9}{\overline{QMErro}}}{\frac{2}{QMErro} + \frac{2}{3QMBloco - QMErro}} + (k-3) \right]$$

Para um experimento grande, os valores destes dois erros podem não diferir muito, e por simplicidade, o QMErro Efetivo Médio pode ser usado para comparar qualquer par de médias:

$$QMERroEfectivoMédio = \frac{QMERro}{k+1} \left[ \frac{\frac{9}{QMERro}}{\frac{2}{QMERro} + \frac{2}{3QMBloco - QMERro}} + (k-2) \right]$$

Para o exemplo considerado tem-se:

$$QMERroEfectivo_1 = \frac{0,2646}{9} \left[ \frac{\frac{6}{0,2646}}{\frac{2}{0,2646} + \frac{2}{3(0,5062) - 0,2646}} + 7 \right]$$

$$QMERroEfectivo_1 = 0,2786$$

$$QMERroEfectivo_2 = \frac{0,2646}{9} \left[ \frac{\frac{9}{0,2646}}{\frac{2}{0,2646} + \frac{2}{3(0,5062) - 0,2646}} + 6 \right]$$

$$QMERroEfectivo_2 = 0,2856$$

$$QMERroEfectivoMédio = \frac{0,2646}{10} \left[ \frac{\frac{9}{0,2646}}{\frac{2}{0,2646} + \frac{2}{3(0,5062) - 0,2646}} + 7 \right]$$

$$QMERroEfectivoMédio = 0,2835$$

- 15º) Calcule a eficiência relativa média do delineamento látice parcialmente balanceado em relação ao DBC, como:

$$ERMédia = \left[ \frac{SQB(ajustado) + SQErroIntrabloco}{r(k-1) + (k-1)(rk-k-1)} \right]$$

$$\times \left( \frac{100}{QMERroEfetivoMédio} \right)$$

$$ERMédia = \left( \frac{12,1492 + 35,9873}{24 + 136} \right) \left( \frac{100}{0,2835} \right) = 106,10$$

# CAPITULO 9

## Delineamento Experimental Látice Retangular

O látice retangular permite a comparação de  $k(k+1)$  tratamentos/genótipos em  $k+1$  blocos de  $k$  parcelas, os quais formam repetições completas. Os látices retangulares podem ser simples com esquema básico de grupos X e Y ou triplos com esquema básico de grupos X, Y e Z. Alguns valores de  $k$ , permitem a obtenção de um tipo de látice retangular em que todas as comparações entre tratamentos têm um mesmo erro padrão que é chamado de látice retangular quase balanceado (Harshbarger, 1951). Os esquemas de grupos básicos X e Y e X, Y e Z podem ou não serem repetidos no delineamento de um experimento.

O látice retangular vem preencher algumas lacunas dos outros tipos de blocos incompletos. Senão observe a seguir o número de tratamentos/genótipos que podem ser comparados por diferentes tipos de blocos incompletos:

Blocos incompletos em geral

Número de genótipos: $t$	13	13	16	16	21	25	...
Número de blocos: $b$	13	26	20	16	21	30	...
Número de repetições: $r$	4	6	5	6	5	6	...
Número de parcelas: $k$	4	3	4	6	5	6	...

Látice  $(t=k^2)$

$t$	16	25	36	49	64	81	100	121	...
$k$	4	5	6	7	8	9	10	11	...

Látice retangular  $[t=k(k+1)]$

$t$	12	20	30	42	56	72	90	110	...
$k$	3	4	5	6	7	8	9	10	...

O látice retangular simples é construído usando o esquema básico de grupos X e Y, sem ou com repetições dos grupos. O látice retangular simples com 2 repetições é obtido sem repetição dos grupos e com 4 repetições é obtido com duas repetições dos grupos X e Y. O látice retangular triplo é construído usando o esquema básico de grupos X, Y e Z, sem ou com repetições dos grupos.

Sem repetição dos grupos X, Y e Z tem-se o látice retangular triplo com 3 repetições e com duas repetições dos grupos X, Y e Z com 6 repetições.

### Exemplo de Aplicação

Considere um experimento de avaliação de 56 genótipos de algodoeiro, desenvolvido usando o delineamento experimental látice retangular simples  $7 \times 8$  com 2 repetições. Então, tem-se um látice retangular simples, com duas repetições do grupo X e duas repetições do grupo Y.

O número de parcelas (de uma linha com 12 plantas) por bloco é  $k=7$ , o número de blocos por repetição é  $b=8$  e o número de repetições é  $r=4$ . Os blocos do grupo X são:  $X_1, X_2, \dots, X_8$  e os blocos do grupo Y são:  $Y_1, Y_2, \dots, Y_8$ .

O bloco  $X_1$  não tem nenhum tratamento em comum com o bloco  $Y_1$ , o bloco  $X_2$  não tem nenhum tratamento em comum com o bloco  $Y_2$  e assim sucessivamente. Os blocos com o mesmo índice são chamados de blocos parceiros.

Quadro 1-Dados de produtividade de algodão em caroço, em gramas por parcela, para 56 genótipos de algodoeiro (Extraído de Conagin, 1954):

Rep I									
Bloco	Genótipos / Produtividade								Total de Bloco
$X_6$	(36)	(37)	(39)	(38)	(40)	(42)	(41)		
	735	455	545	389	496	131	332		3083
$X_7$	(44)	(49)	(43)	(46)	(48)	(45)	(47)		
	607	375	740	304	206	261	174		2667
$X_3$	(19)	(21)	(15)	(17)	(20)	(16)	(18)		
	342	405	418	369	419	478	343		2774
$X_4$	(22)	(25)	(26)	(24)	(27)	(23)	(28)		
	476	523	640	467	433	396	557		3492
$X_2$	(14)	(08)	(12)	(11)	(09)	(10)	(13)		
	198	250	270	330	281	410	518		2257
$X_5$	(35)	(32)	(31)	(29)	(34)	(33)	(30)		
	251	271	269	315	210	310	420		2046
$X_1$	(03)	(02)	(01)	(04)	(05)	(07)	(06)		
	316	322	584	281	336	337	343		2519
$X_8$	(50)	(51)	(52)	(56)	(54)	(55)	(53)		
	194	521	566	535	784	794	487		3881
	Total de Rep I $(R_I)$								22719

Rep II

Bloco	Genótipos / Produtividade								Total de Bloco
$Y_4$	(46) (03) (53) (39) (17) (10) (32) 221 343 546 319 540 476 601								3046
$Y_2$	(30) (16) (44) (51) (37) (23) (01) 375 312 239 458 490 383 643								2900
$Y_6$	(48) (33) (05) (26) (55) (12) (19) 394 372 249 473 514 380 456								2838
$Y_7$	(27) (06) (41) (34) (13) (20) (56) 446 385 482 227 271 288 327								2426
$Y_5$	(40) (04) (11) (18) (25) (54) (47) 492 596 369 335 077 578 364								2811
$Y_3$	(52) (02) (24) (09) (45) (31) (38) 676 467 324 412 382 415 603								3279
$Y_8$	(35) (07) (14) (21) (28) (49) (42) 433 325 350 241 368 392 217								2326
$Y_1$	(43) (15) (22) (29) (50) (36) (08) 559 178 372 410 526 765 325								3135
	Total de Rep II $(R_{II})$								22761

Rep III

Bloco	Genótipos / Produtividade								Total de Bloco
$X_4$	(23) (27) (26) (25) (28) (22) (24) 440 452 857 351 425 437 383								3345
$X_2$	(13) (12) (11) (08) (10) (14) (09) 752 487 782 472 310 329 488								3611
$X_1$	(07) (01) (05) (03) (06) (04) (02) 459 752 470 509 419 203 458								3270
$X_3$	(15) (16) (20) (21) (19) (17) (18) 273 416 601 351 307 386 436								2770
$X_5$	(29) (31) (30) (32) (35) (33) (34) 417 654 311 705 568 433 300								3388
$X_6$	(36) (38) (41) (39) (40) (42) (37) 885 359 731 319 538 545 741								4118
$X_8$	(56) (52) (54) (55) (53) (51) (50) 541 736 639 313 434 402 583								3648
$X_7$	(49) (43) (48) (45) (46) (44) (47) 517 680 423 333 474 390 433								3250
	Total de Rep III $(R_{III})$								27400

Rep IV

Bloco	Genótipos / Produtividade							Totais de Bloco
$Y_1$	(43) 342	(50) 517	(08) 325	(29) 233	(22) 295	(36) 414	(15) 221	2347
$Y_4$	(46) 431	(53) 275	(32) 251	(03) 320	(17) 206	(10) 234	(39) 253	1970
$Y_5$	(25) 335	(40) 343	(54) 178	(11) 411	(18) 183	(47) 184	(04) 198	1832
$Y_2$	(44) 280	(51) 238	(23) 201	(16) 248	(01) 210	(37) 422	(30) 321	1920
$Y_7$	(34) 510	(20) 680	(06) 447	(56) 509	(13) 236	(27) 334	(41) 430	3146
$Y_6$	(05) 310	(33) 298	(19) 362	(48) 325	(55) 364	(26) 400	(12) 351	2410
$Y_3$	(52) 520	(02) 358	(45) 327	(31) 219	(24) 385	(38) 213	(09) 290	2312
$Y_8$	(21) 450	(28) 479	(42) 474	(35) 518	(49) 633	(07) 444	(14) 400	3398
	Total de Rep IV ( $R_{IV}$ )							19335

$$Total Geral = 22719 + 22761 + 27400 + 19335 = 92215$$

A análise de variância é realizada segundo método descrito em Cochran e Cox (1950), o qual envolve os seguintes passos:

1-Obtenção dos totais de blocos (B), de repetições (R) e o total geral (G), que já foram apresentados no Quadro 1. Os totais de tratamentos são apresentados no Quadro 2 a seguir:

Quadro 2-Totais de tratamentos

Tratamento ( $i$ )	Total $(T_i)$	Tratamento ( $i$ )	Total $(T_i)$
01	2189	29	1375
02	1605	30	1427
03	1488	31	1557
04	1278	32	1828
05	1365	33	1413
06	1594	34	1247
07	1565	35	1770
08	1372	36	2799
09	1471	37	2108
10	1430	38	1564
11	1892	39	1436
12	1479	40	1869
13	1777	41	1975
14	1277	42	1367
15	1090	43	2321
16	1454	44	1516
17	1501	45	1303
18	1297	46	1430
19	1467	47	1155
20	1988	48	1348
21	1447	49	1917
22	1580	50	1820
23	1420	51	1619
24	1559	52	2498
25	1286	53	1742
26	2370	54	2179
27	1665	55	1985
28	1829	56	1912

2-Reorganização dos dados de acordo com o esquema básico de grupos X e Y, como apresentado a seguir nos Quadros 3 e 4:

Quadro 3-Totais de tratamentos e de blocos nos grupos X e Y

Grupo X (RepI+RepIII)

Bloco								B
$X_1$	(01) 1336	(02) 780	(03) 825	(04) 484	(05) 806	(06) 762	(07) 796	5789
$X_2$	(08) 722	(09) 769	(10) 720	(11) 1112	(12) 748	(13) 1270	(14) 527	5868
$X_3$	(15) 691	(16) 894	(17) 755	(18) 779	(19) 649	(20) 1020	(21) 756	5544
$X_4$	(22) 913	(23) 836	(24) 850	(25) 874	(26) 1497	(27) 885	(28) 982	6837
$X_5$	(29) 732	(30) 731	(31) 923	(32) 976	(33) 743	(34) 510	(35) 819	5434
$X_6$	(36) 1620	(37) 1196	(38) 748	(39) 864	(40) 1034	(41) 1063	(42) 676	7201
$X_7$	(43) 1420	(44) 997	(45) 594	(46) 778	(47) 607	(48) 629	(49) 892	5917
$X_8$	(50) 777	(51) 923	(52) 1302	(53) 921	(54) 1423	(55) 1107	(56) 1076	7529

B: Total de Bloco

Grupo Y (RepII+RepIV)

Bloco								B
$Y_1$	(08) 650	(15) 399	(22) 667	(29) 643	(36) 1179	(43) 901	(50) 1043	5482
$Y_2$	(01) 853	(16) 560	(23) 583	(30) 696	(37) 912	(44) 519	(51) 696	4820
$Y_3$	(02) 825	(09) 702	(24) 709	(31) 634	(38) 816	(45) 709	(52) 1196	5591
$Y_4$	(03) 663	(10) 710	(17) 746	(32) 852	(39) 572	(46) 652	(53) 821	5016
$Y_5$	(04) 794	(11) 780	(18) 518	(25) 412	(40) 835	(47) 548	(54) 756	4643
$Y_6$	(05) 559	(12) 731	(19) 818	(26) 873	(33) 670	(48) 719	(55) 878	5248
$Y_7$	(06) 832	(13) 507	(20) 968	(27) 780	(34) 737	(41) 912	(56) 836	5572
$Y_8$	(07) 769	(14) 750	(21) 691	(28) 847	(35) 951	(42) 691	(49) 1025	5724

B: Total de Bloco

Quadro 4-Totais de tratamentos (grupo X + grupo Y)

Bloco	$Y_1$	$Y_2$	$Y_3$	$Y_4$	$Y_5$	$Y_6$	$Y_7$	$Y_8$
$X_1$	-	(1) 2189	(2) 1605	(3) 1488	(4) 1278	(5) 1365	(6) 1594	(7) 1565
$X_2$	(8) 1372	-	(9) 1471	(10) 1430	(11) 1892	(12) 1479	(13) 1777	(14) 1277
$X_3$	(15) 1090	(16) 1454	-	(17) 1501	(18) 1297	(19) 1467	(20) 1988	(21) 1447
$X_4$	(22) 1580	(23) 1420	(24) 1559	-	(25) 1286	(26) 2370	(27) 1665	(28) 1829
$X_5$	(29) 1375	(30) 1427	(31) 1557	(32) 1828	-	(33) 1413	(34) 1247	(35) 1770
$X_6$	(36) 2799	(37) 2108	(38) 1564	(39) 1436	(40) 1869	-	(41) 1975	(42) 1367
$X_7$	(43) 2321	(44) 1516	(45) 1303	(46) 1430	(47) 1155	(48) 1348	-	(49) 1917
$X_8$	(50) 1820	(51) 1619	(52) 2498	(53) 1742	(54) 2179	(55) 1985	(56) 1912	-

Cálculos de Somas de Quadrados Gerais:

$$C = 37962527,80$$

$$SQT_{Total} = 42972255,00 - 37962527,80 = 5009727,20$$

$$SQR = 38550326,90 - 37962527,80 = 587799,10$$

$$SQT_{Não-Ajustado} = 39678554,7 - 37962527,80$$

$$SQT_{Não-Ajustado} = 1716026,90$$

3-Organizar os blocos do mesmo grupo de acordo com o número do bloco e com a repetição do esquema básico de grupos X e Y, como apresentado no Quadro5 a seguir:

Quadro 5-Arranjo de totais de blocos por grupos X e Y e por repetições, de forma que os blocos parceiros aparecem na mesma linha

Bloco	Blocos X		Totais de Blocos X	Blocos Y		Totais de Blocos Y
	I	III		II	IV	
1	2519	3270	5789	3135	2347	5482
2	2257	3611	5868	2900	1920	4820
3	2774	2770	5544	3279	2312	5591
4	3492	3345	6837	3046	1970	5016
5	2046	3388	5434	2811	1832	4643
6	3083	4118	7201	2838	2410	5248
7	2667	3250	5917	2426	3146	5572
8	3881	3648	7529	2326	3398	5724
Totais	22719	27400	50119	22761	19335	42096

Dentro de cada grupo (X e Y) para cada bloco, calcular os valores de  $C = T - n \sum B$ , onde  $T$  é o total de todos os tratamentos no bloco e  $B$  o total do bloco. Por exemplo, o valor de  $C$  para o bloco 6 dentro do grupo X, ou seja,  $X_6$  é calculado como:

$$C_{X_6} = (2799 + 2108 + 1564 + 1436 + 1869 + 1975 + 1367) - \\ - 2(3083 + 4118) = -1284$$

A soma de todos os valores de  $C$  é igual a zero, ou seja  $\sum C_X + \sum C_Y = 0$

Organizar os valores de  $C$  num quadro de tal forma que os parceiros apareçam na mesma linha. O total de cada linha fornece a soma dos valores de  $C$  dos parceiros, sendo denotado por  $S_l$ , ou seja:

$$S_l = C_{X_l} + C_{Y_l}, l=1,2,\dots,8. \text{ As somas dos valores de } C_{X_l} \text{ e de } C_{Y_l} \text{ são denotadas por } R_c, c=1,2. \text{ Então, } R_1 = C_{X_1} + C_{X_2} + \dots + C_{X_8} \text{ e } R_2 = C_{Y_1} + C_{Y_2} + \dots + C_{Y_8}$$

Estes resultados estão apresentados no Quadro 6 a seguir:

Quadro 6-Valores de  $C_X$ ,  $C_Y$ ,  $S_l$  e  $R_c$  para serem utilizados na análise de variância

Bloco	$C_X$	$C_Y$	$S_l$
1	- 494	1393	899
2	-1038	2093	1055
3	- 844	375	- 469
4	-1965	823	- 1142
5	- 251	1670	1419
6	-1284	931	- 353
7	- 844	1014	170
8	- 1303	- 276	- 1579
Totais $R_c$	- 8023	8023	0

Para a análise de variância, a soma de quadrados de blocos ajustada para tratamentos é calculada a partir dos denominados componentes  $a$  e  $b$ , o que é feito pelo fato de ter repetição do esquema básico de grupos X e Y. No caso de não haver repetição do esquema básico de grupos X e Y obtém-se apenas o componente  $b$ . Todas as outras somas de quadrados são calculadas da maneira usual.

4-Cálculo da soma de quadrados do **Componente(a)**: este cálculo é feito a partir das somas de quadrados das interações bloco x grupo X e bloco x grupo Y, as quais podem ser obtidas com base no Quadro 5.

$$BX_l^2 = \frac{1}{7} \left[ (2519)^2 + (2257)^2 + \dots + (3648)^2 \right] - \\ - \frac{1}{56} \left[ (22719)^2 + (27400)^2 \right] - \\ - \left\{ \frac{1}{14} \left[ (5789)^2 + (5868)^2 + \dots + (7529)^2 \right] + \frac{1}{112} \left[ (50119)^2 \right] \right\}$$

$$BX_l^2 = 23156517,6 - 22623445,7 -$$

$$-(22750424,1 + 22427805,0)$$

$$BX_l^2 = 210452,8$$

O valor de  $BY_l^2 = 358302,5$  é obtido de forma análoga, dentro do grupo Y. Então tem-se:

$$SQComponente(a) = BX_l^2 + BY_l^2$$

$$SQComponente(a) = 210452,8 + 358302,5$$

$$SQComponente(a) = 568755,3$$

5-Cálculo da soma de quadrados do **Componente(b)**, para blocos ajustados para tratamento é realizado usando a expressão seguinte:

$$SQComponente(b) = \frac{\sum C^2}{r(nk-k-1)} - \frac{\sum R_c^2}{r(k+1)(nk-k-1)} - \\ - \frac{\sum S_l^2}{r(n-1)(k+1)(nk-k-1)}$$

, em que:

$n$  : número de grupos,  $n=2$

$p$  : número de repetições dos grupos,  $p=2$

$k$  : número de parcelas por bloco,  $k=7$

$r=np$  : número total de repetições,  $r=4$

$$\sum C^2 = (-494)^2 + (1393)^2 + \dots + (-276)^2 = 21915928$$

$$\sum R_c^2 = (-8023)^2 + (8023)^2 = 128737058$$

$$\sum S_l^2 = (899)^2 + (1055)^2 + \dots + (-1579)^2 = 8105662$$

Então, tem-se:

$$SQComponente(b) = \frac{21915928}{24} - \frac{128737058}{192} - \\ - \frac{8105662}{192}$$

$$SQComponente(b) = 200441,20$$

6-Cálculos de soma de quadrados de bloco ajustado e soma de quadrados de erro intrabloco e resumo da análise de variância (Quadro 7):

$$\begin{aligned}
 SQBAjustado &= SQComponente(a) + \\
 &+ SQComponente(b) \\
 SQBAjustado &= 568755,30 + 200441,20 \\
 SQBAjustado &= 769196,50 \\
 SQErroIntrabloco &= SQTTotal - \\
 &- (SQR + SQBAjustado + SQTNAjustado) \\
 SQErroIntrabloco &= 5009727,20 - \\
 &(587799,10 + 769196,50 + 1716026,90) \\
 SQErroIntrabloco &= 1936704,70
 \end{aligned}$$

O resumo da análise de variância é apresentado no Quadro 7 a seguir:

Quadro 7-Resumo da análise de variância de produtividade (g/parcela) de genótipos de algodoeiro

FV	SQ	GL	QM
Total	5009727,20	223	-
Repetição	587799,10	3	195933,00
Tratamento (naj)	1716026,90	55	31200,50
Bloco (aj)	769196,50	28	$27471,30 (E_b)$
Componente(a)	568755,30	14	-
Componente(b)	200441,20	14	-
Erro Intrabloco	1936704,70	137	$14136,50 (E_e)$
CV%	28,88		

7-Calcular a soma de quadrados de tratamento ajustado para o efeito de blocos. Então, deve-se fazer o ajuste dos totais de tratamentos, o que é feito usando os valores de  $\lambda$  e  $\mu$  que são obtidos usando as seguintes expressões:

$$\lambda = \frac{r(E_b - E_e)}{r(k-1)E_b + (rk-2k+r)E_e} \quad \text{e}$$

$$\mu = \frac{\lambda r(E_b - E_e)}{r(k+1)E_b + (rk-2k-r)E_e} . \text{ Então tem-se:}$$

$$\lambda = \frac{4(27471,30 - 14136,50)}{4(7-1)27471,30 + [(4)(7)-2(7)+4]14136,50}$$

$$\lambda = 0,058370$$

$$\mu = \frac{(0,05837)4(27471,30 - 14136,50)}{4(7+1)27471,30 + [(4)(7)-2(7)-4]14136,50}$$

$$\mu = 0,003051$$

A seguir é apresentado um quadro com os valores de  $\lambda C_X$ ,  $\lambda C_Y$  e  $\mu S_l$  que serão utilizados para o ajustamento dos totais de tratamentos:

Quadro 8-Valores de  $\lambda C_X$ ,  $\lambda C_Y$  e  $\mu S_l$  para serem utilizados no ajuste de totais de blocos e de tratamentos

Bloco	$\lambda C_X$	$\lambda C_Y$	$\mu S_l$
1	- 28,8	81,3	2,7
2	-60,6	122,2	3,2
3	- 49,3	21,9	- 1,4
4	-114,7	48,0	- 3,5
5	- 14,7	97,5	4,3
6	-74,9	54,3	- 1,1
7	- 49,3	59,2	0,5
8	- 76,1	- 16,1	- 4,8
Totais	- 468,4	468,3	- 0,1

O ajuste para os blocos é obtido por meio do cálculo dos valores  $\lambda C - \mu S_l$ , onde  $C$  é  $C_X$  para ajuste de blocos do grupo X e  $C_Y$  para ajuste de blocos do grupo Y. Por exemplo, os ajustes para os blocos  $X_1$  do grupo X e  $Y_2$  do grupo Y são os seguintes:

$$Ajuste B_{X_1} = -28,8 - 2,7 = -31,5$$

$$Ajuste B_{Y_2} = 122,2 - 3,2 = 119,0$$

Os valores dos ajustes para todos os blocos estão apresentados no quadro a seguir:

Quadro 9-Fatores de ajuste para blocos

Grupo X (RepI+RepIII)	Grupo Y (RepII+RepIV)						
Bloco	Total de Bloco	$C_X$	Fator de Ajuste	Bloco	Total de Bloco	$C_Y$	Fator de Ajuste
$X_1$	5789	-494	-31,5	$Y_1$	5482	1392	78,6
$X_2$	5868	-1038	-63,8	$Y_2$	4820	2093	119,0
$X_3$	5544	-844	-47,9	$Y_3$	5591	375	23,3
$X_4$	6837	-1965	-111,2	$Y_4$	5016	823	51,5
$X_5$	5434	-251	-19,0	$Y_5$	4643	1670	93,2
$X_6$	7201	-1284	-73,8	$Y_6$	5248	931	55,4
$X_7$	5917	-844	-49,8	$Y_7$	5572	1014	58,7
$X_8$	7529	-1303	-71,3	$Y_8$	5724	-276	-11,3

Cada total de tratamento é ajustado pela adição do ajuste de cada bloco no qual o tratamento está contido. Por exemplo, para o total do tratamento 1 tem-se:

$2189 + (-31,5) + (119,0) = 2276,5$  (Veja no Quadro 4 e Quadro 9). Para o tratamento 2 tem-se:

$$1605 + (-31,5) + (23,3) = 1596,8$$

Os totais de tratamentos ajustados para todos os tratamentos estão apresentados no quadro a seguir:

Quadro 10-Totais de tratamentos ajustados e não ajustados

Trat( <i>i</i> )	Total $(T_i)$ Não-Ajustado	Total $(T_i)$ Ajustado	Trat( <i>i</i> )	Total $(T_i)$ Não-Ajustado	Total $(T_i)$ Ajustado
01	2189	2276,5	29	1375	1434,6
02	1605	1596,8	30	1427	1527,0
03	1488	1508,0	31	1557	1561,3
04	1278	1339,7	32	1828	1860,5
05	1365	1388,9	33	1413	1449,4
06	1594	1621,2	34	1247	1286,7
07	1565	1522,2	35	1770	1739,7
08	1372	1386,8	36	2799	2803,8
09	1471	1430,5	37	2108	2153,2
10	1430	1417,7	38	1564	1513,5
11	1892	1921,4	39	1436	1413,7
12	1479	1470,6	40	1869	1888,4
13	1777	1771,9	41	1975	1959,9
14	1277	1201,9	42	1367	1281,9
15	1090	1120,7	43	2321	2349,8
16	1454	1525,1	44	1516	1585,2
17	1501	1504,6	45	1303	1276,5
18	1297	1342,3	46	1430	1431,7
19	1467	1474,5	47	1155	1198,4
20	1988	1998,8	48	1348	1353,6
21	1447	1387,8	49	1917	1855,9
22	1580	1547,4	50	1820	1827,3
23	1420	1427,8	51	1619	1666,7
24	1559	1471,1	52	2498	2450,0
25	1286	1268,0	53	1742	1722,2
26	2370	2314,2	54	2179	2200,9
27	1665	1612,5	55	1985	1969,1
28	1829	1706,5	56	1912	1899,4

## 8-Comparações entre tratamentos ajustados aos pares

Para a comparação entre dois totais de tratamentos ajustados são necessários três erros padrões, o que resulta em três diferenças mínimas significativas. Por exemplo, para comparações com base no teste, as diferenças mínimas significativas (dms) são as seguintes:

### 8.1-Para dois tratamentos que ocorrem juntos no mesmo bloco

$$dms_1 = t \sqrt{2rE_e(1+\lambda-\mu)}$$

$$dms_1 = 1,98 \sqrt{2(4)(14136,5)(1+0,05837-0,003051)}$$

$$dms_1 = 648,0$$

8.2-Para dois tratamentos que não ocorrem juntos no mesmo bloco

$$dms_2 = t \sqrt{2rE_e(1+2\lambda-\mu)}$$

$$dms_2 = 1,98 \sqrt{2(4)(14136,5)(1+(2)0,05837-0,003051)}$$

$$dms_2 = 702,9$$

8.3-Para dois tratamentos quaisquer

$$dms_3 = t \sqrt{2rE_e \left( 1 + \frac{2k^2 \lambda}{k^2 + k - 1} - \mu \right)}$$

$$dms_3 = 1,98 \sqrt{2(4)(14136,5) \left( 1 + \frac{2(7^2)0,05837}{7^2 + 7 - 1} - 0,003051 \right)}$$

$$dms_3 = 698,6$$

9-Eficiência relativa do látice retangular simples com 4 repetições

Deve-se avaliar o ganho em precisão decorrente do uso do delineamento experimental látice retangular simples ao invés do delineamento experimental blocos completos casualizados da seguinte forma:

A variância do erro considerando a análise como blocos completos casualizados é calculada como:

$$QMERro_{DBC} = \frac{(SQBAjustado + SQerroIntrabloco)}{(GLBAjustado + GLEerroIntrabloco)}$$

$$QMERro_{DBC} = \frac{(769196,5 + 1936704,7)}{(28 + 137)} = 16399,4$$

A variância do erro para a análise como látice retangular simples com 4 repetições é calculada como:

$$QMERro_{DLR} = E_e \left( 1 + \frac{2k^2 \lambda}{k^2 + k - 1} - \mu \right)$$

$$QMERro_{DLR} = (14136,5) \left( 1 + \frac{2(7^2)0,05837}{7^2+7-1} - 0,003051 \right)$$

$$QMERro_{DLR} = 15563,6$$

A eficiência relativa do látice retangular simples em relação a blocos casualizados é expressa como:

$$ER\% = \frac{QMERro_{DBC}}{QMERro_{DLR}} (100)$$

$$ER\% = \frac{16399,4}{15563,6} = 105,4$$

Este resultado significa que houve um ganho em precisão de 5,4% com o uso do látice retangular simples. Significa também que 15 repetições do látice simples retangular equivalem a 16 repetições de blocos casualizados.

# CAPITULO 10

## Delineamentos Experimentais Alfa

Os delineamentos-alfa são essencialmente delineamentos blocos resolvíveis. Nos delineamentos blocos resolvíveis, a variação entre blocos dentro de repetição contribui para a redução no erro experimental, aumentando assim a precisão da estimação dos contrastes entre tratamentos.

Os delineamentos bloco resolvíveis permitem realizar um experimento na base de uma repetição por vez. Por exemplo, nos experimentos agronômicos o campo pode ser dividido num certo número de grandes áreas, que correspondem a repetições e então cada área é subdividida em blocos. No melhoramento de plantas, estes delineamentos são úteis para experimentos de campo com grande número de tratamentos/genótipos que não podem sempre serem executados num único local ou estação de cultivo. Então, é necessário que a variação devida à local ou época possa ser controlada dentro de local ou época. Isto pode então ser manejado usando delineamentos blocos resolvíveis. Neste caso, local ou época podem ser tomados como repetições e a variação dentro de um local ou época pode ser cuidada pela formação de blocos.

### Construção de delineamentos- $\alpha$

Para construir delineamentos- $\alpha$  para  $t = ks$  tratamentos em  $r$  repetições inicia-se com um arranjo  $k \times r$  chamado arranjo- $\alpha$  com elementos inteiros módulo  $S$ . Um delineamento- $\alpha$  é determinado completamente por um arranjo- $\alpha$   $k \times r$ . Para cada coluna deste arranjo desenvolve-se uma repetição contendo  $S$  blocos, cada um com tamanho  $k$ . Portanto, obtém-se  $r$  repetições e  $rs$  blocos de tamanho  $k$  a partir de um arranjo- $\alpha$   $k \times r$ . Para fim de complementação define-se os seguintes termos:

Arranjo: Disposição de elementos em linhas e colunas. O arranjo que é utilizado para desenvolver delineamentos é chamado de arranjo gerador.

Arranjo reduzido: Arranjo gerador com todos os elementos da primeira linha e primeira coluna iguais a zero.

Sob uma adequada permutação de elementos, todos os arranjos podem ser representados como arranjos reduzidos, restringindo com isso a construção de delineamentos apenas a arranjos reduzidos. Por exemplo, considere a obtenção do arranjo reduzido a partir do arranjo gerador a seguir, para  $s = 3$ :

1	0	1
2	2	1
0	1	1
1	0	2

Primeiro, adicione 2 aos elementos das colunas 1 e 3, reduzindo a soma ao módulo 3 quando necessário. Obtém-se o seguinte arranjo:

0	0	0
1	2	0
2	1	0
0	0	1

Segundo, adicione 2 aos elementos da linha 2 e 1 aos elementos da linha 3, reduzindo a soma ao módulo 3 quando necessário. O arranjo obtido é o arranjo reduzido:

0	0	0
0	1	2
0	2	1
0	0	1

Na medida do possível todas as  $r-1$  colunas de tamanho  $k-1$  cada devem ser distintas. O delineamento possível de ser obtido a partir deste arranjo deve ser então checado para eficiência-A. Se a eficiência-A for fraca, então pode-se selecionar outro arranjo reduzido, até que o delineamento com a eficiência-A desejada seja obtido. Existem algoritmos computacionais disponíveis com procedimentos para gerar delineamentos- $\alpha$  eficientes. O procedimento para a geração delineamentos- $\alpha$  a partir de arranjo- $\alpha$  reduzido e a terminologia envolvida são descritos a seguir:

1-Arranjo- $\alpha$ : Arranjo  $k \times r$  de elementos que consistem de classe de resíduo de  $S$ , podendo estes elementos ser  $0, 1, 2, \dots, s-1$ . Se a primeira linha e a primeira coluna deste arranjo consistem de zeros como elementos, então o arranjo é chamado de arranjo reduzido. A construção de um delineamento- $\alpha$  é realizada com base no arranjo reduzido. Pode-se determinar a confluência dos diferentes pares de tratamentos por meio do arranjo- $\alpha$ .

2-Arranjo intermediário: Arranjo  $k \times sr$  obtido da seguinte forma: Obter  $(s-1)$  novas colunas a partir de cada uma das colunas do arranjo- $\alpha$ , desenvolvendo ciclicamente cada coluna módulo  $S$ . Desta forma, pode-se construir  $r(s-1)$  novas colunas aplicando o processo a cada coluna do arranjo- $\alpha$  módulo  $S$ .

3-Delineamento- $\alpha$ : Delineamento com os parâmetros  $t=ks$ ,  $b=rs$ ,  $r$  e  $k$  obtido da seguinte forma: Tomar o arranjo intermediário e adicionar  $[1+(j-1)s]$  aos elementos da

$j$ -ésima linha ( $j=1, 2, \dots, k$ ) deste arranjo. Agora, toma-se a correspondência um a um entre os tratamentos e os elementos do arranjo final obtendo-se então o delineamento- $\alpha$  requerido.

Os passos para construção de um delineamento- $\alpha$  são descritos a seguir com a ajuda de um exemplo:

Exemplo 1: Considere a construção de um delineamento- $\alpha$  para  $t=12$ ,  $b=12$ ,  $r=3$ ,  $k=3$  e  $s=4$ . O arranjo- $\alpha$  gerador para este delineamento é:

0	0	0
0	2	3
0	3	1

Dado este arranjo gerador, desenvolve-se o arranjo- $\alpha^*$  intermediário desenvolvendo ciclicamente cada coluna módulo  $S$ :

0	1	2	3	0	1	2	3	0	1	2	3
0	1	2	3	2	3	0	1	3	0	1	2
0	1	2	3	3	0	1	2	1	2	3	0

Agora, adiciona-se  $[1+(j-1)s]$ , ( $j=1,2,\dots,k$ ), aos elementos da  $j$ -ésima linha deste arranjo. Então, o delineamento- $\alpha$  com colunas como blocos é o seguinte:

Repetição I				Repetição II				Repetição III			
Blocos				Blocos				Blocos			
1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	1	2	3	4	1	2	3	4
5	6	7	8	7	8	5	6	8	5	6	7
9	10	11	12	12	9	10	11	10	11	12	9

A construção de um delineamento- $\alpha$  a partir de um arranjo- $\alpha$  desenvolvendo ciclicamente módulo  $S$  aos blocos iniciais que correspondem a uma coluna do arranjo- $\alpha$  e adicionando  $[1+(j-1)s]$  à  $j$ -ésima posição de cada um dos blocos desenvolvidos, implica que algumas

das confluências, no mínimo  $k \binom{s}{2}$ , são zero. Para melhor compreensão, considere o exemplo

a seguir:

Exemplo 2-Considere um arranjo- $\alpha$   $3 \times 4$  com  $k=3$ ,  $r=4$  e  $s=3$ :

0	0	0	0
0	1	2	2
0	2	1	2

Tome a terceira coluna como conjunto inicial e então desenvolvendo ciclicamente módulo 3 obtém-se:

0	1	2
2	0	1
1	2	0

Agora adicione  $1+(j-1)s=1+(j-1)3$  aos elementos na  $j$ -ésima linha, para  $j=1,2,\dots,k$  e tem-se:

1	2	3
6	4	5
8	9	7

Os grupos de tratamentos 1,2,3; 4,5,6 e 7,8,9 aparecerão em diferentes blocos em qualquer uma das repetições e, portanto, nunca ocorrerão juntos num bloco.

Observe que diferentes arranjos geradores fornecem diferentes delineamentos- $\alpha$  para os parâmetros  $t$ ,  $k$  e  $r$  definidos. Então, a escolha de um delineamento apropriado deve ser baseada no limite inferior da eficiência média dos delineamentos.

Os delineamentos blocos incompletos resolvíveis têm uma estrutura de blocos aninhada, ou seja, os blocos incompletos dentro de cada repetição são casualizados separadamente para cada repetição. Este processo de casualização é apropriado para um esquema experimental onde as repetições são entidades separadas. Desta forma, os blocos de uma repetição não têm nenhuma relação com os blocos de outra repetição.

#### Casualização de delineamentos- $\alpha$

Existem quatro estágios de casualização para um delineamento- $\alpha$ , que são os seguintes:

1-Casualizar os tratamentos, ou seja, alocar de forma aleatória os tratamentos do experimento aos números dos tratamentos do delineamento;

2-Casualizar as repetições;

3-Casualizar os blocos dentro de cada repetição, separadamente para cada repetição;

4-Casualizar os tratamentos nas parcelas dentro de cada bloco, separadamente para cada bloco.

A casualização do delineamento- $\alpha$  é feita por meio da geração de números aleatórios ou utilizando tabelas de números aleatórios. Considere como exemplo os passos para casualização do delineamento- $\alpha$  do Exemplo 1:

1-Assumir que os tratamentos tenham sido numerados aleatoriamente;

2-Casualizar as repetições

Gerar números aleatórios de 1 a 3. Seja, por exemplo, os números  $\{2,1,3\}$ . Então, deve-se arranjar a repetição 2 como repetição 1, a repetição 1 como repetição 2 e a repetição 3 como repetição 3. A disposição do delineamento após este passo é:

Repetição I				Repetição II				Repetição III			
Blocos				Blocos				Blocos			
1	2	3	4	5	6	7	8	9	10	11	12
1	2	3	4	1	2	3	4	1	2	3	4
7	8	5	6	5	6	7	8	8	5	6	7
12	9	10	11	9	10	11	12	10	11	12	9

### 3-Casualizar os blocos

Gerar 3 conjuntos de 4 números aleatórios distintos menor ou igual a 4. Seja, por exemplo, os conjuntos de números  $\{3,1,2,4\}$ ,  $\{3,2,4,1\}$  e  $\{4,1,2,3\}$ . Casualizar os blocos por meio destes conjuntos de números aleatórios. Usar o primeiro conjunto para casualizar os blocos dentro da repetição 1, o segundo conjunto para casualizar os blocos da repetição 2 e o terceiro conjunto para casualizar os blocos dentro da repetição 3. A disposição resultante deste passo é:

Repetição I				Repetição II				Repetição III			
Blocos				Blocos				Blocos			
1	2	3	4	5	6	7	8	9	10	11	12
3	1	2	4	3	2	4	1	4	1	2	3
5	7	8	6	7	6	8	5	7	8	5	6
10	12	9	11	11	10	12	9	9	10	11	12

### 4-Casualizar os tratamentos dentro de cada bloco

Gerar 12 conjuntos de 3 números aleatórios cada. Seja, por exemplo os conjuntos  $\{1,3,2\}$ ,  $\{3,1,2\}$ ,  $\{3,2,1\}$ ,  $\{2,1,3\}$ ,  $\{1,2,3\}$ ,  $\{3,1,2\}$ ,  $\{2,3,1\}$ ,  $\{3,2,1\}$ ,  $\{3,2,1\}$ ,  $\{2,1,3\}$ ,  $\{1,3,2\}$  e  $\{3,1,2\}$ . Usar estes conjuntos para casualizar os tratamentos dentro de blocos, obtendo-se a seguinte disposição:

Repetição I				Repetição II				Repetição III			
Blocos				Blocos				Blocos			
1	2	3	4	5	6	7	8	9	10	11	12
3	12	9	6	3	10	8	9	9	8	2	12
10	1	8	4	7	2	12	5	7	1	11	3
5	7	2	11	11	6	4	1	4	10	5	6

Ao final do passo 4 obtém o esquema da disposição de repetições, blocos e tratamentos do delineamento- $\alpha$ . Este esquema pode então ser utilizado para experimentação.

### Esquema de análise de variância

A análise de variância de delineamentos- $\alpha$  pode ser realizada como o procedimento geral para delineamentos em blocos. A única mudança nesta análise é que a soma de quadrados de bloco será bifurcada em soma de quadrados devido a repetições e soma de quadrados entre blocos dentro de repetições.

Um esquema geral para a análise de variância é o seguinte:

Fonte de Variação	Graus de Liberdade
Repetições	$r-1$
Blocos dentro de Repetições	$r(s-1)$
Tratamentos	$t-1$ ou $sk-1$
Erro	$rsk-sk-rs+1$
Total	$tr-1$ ou $rsk-1$

### Exemplo de Aplicação

Considere um experimento de avaliação inicial de desempenho de 21 genótipos de toria (*Brassica campestris* var Toria) mais 3 genótipos controle usando um delineamento alfa com três repetições. Foi medida a produtividade de grãos em kg/ha. Os detalhes sobre os genótipos sob avaliação, os genótipos controle/testemunha e o delineamento utilizado são os seguintes:

Trat.	Nº Trat.	Trat.	Nº Trat.	Trat.	Nº Trat.
Rau dt-01-03	1	Tk-06-1	9	Rh-0304	17
Rau dt-01-02	2	Tk-06-2	10	Th-0302	18
Bausm-92-24	3	Tl-2027	11	Jmt-05	19
Rgn 186	4	Tl-2013	12	Pt-303(controle)	20
Ej-17	5	Jmt-02-6	13	Zonal(testemunha)	21
Npj-112	6	Ndt 05-5	14	Ptc-99-14	22
Vlt-4	7	Ndre 200216	15	JD-6(testemunha)	23
Rrn-612	8	Pt-2004-3	16	Ort 17-6-16	24

Os dados obtidos estão apresentados a seguir:

Repetição I						
Bloco1	(1) 1555,6	(5) 1160,5	(9) 1308,6	(13) 1382,7	(17) 987,7	(21) 1135,8
Bloco2	(2) 1284,0	(6) 1086,4	(10) 1284,0	(14) 1111,1	(18) 938,3	(22) 1308,6
Bloco3	(3) 1234,6	(7) 419,8	(11) 1308,6	(15) 963,0	(19) 963,0	(23) 987,7
Bloco4	(4) 1234,6	(8) 987,7	(12) 1284,0	(16) 913,6	(20) 1160,5	(24) 790,1

Repetição II						
Bloco1	(1) 1481,5	(6) 1086,4	(11) 1308,6	(16) 1284,0	(19) 1111,1	(22) 1185,2
Bloco2	(2) 987,7	(7) 308,6	(12) 1234,6	(13) 1308,6	(20) 765,4	(23) 938,3
Bloco3	(3) 1012,3	(8) 864,2	(9) 1234,6	(14) 938,3	(17) 913,6	(24) 864,2
Bloco4	(4) 1135,8	(5) 987,7	(10) 987,7	(15) 740,7	(18) 963,0	(21) 1135,8

Repetição III						
Bloco1	(1) 1284,0	(7) 333,3	(12) 1135,8	(15) 839,5	(18) 814,8	(24) 888,9
Bloco2	(2) 1135,8	(8) 913,6	(9) 1456,8	(16) 1037,0	(19) 938,3	(21) 1037,0
Bloco3	(3) 963,0	(5) 1209,9	(10) 1259,3	(13) 1234,6	(20) 963,0	(22) 1111,1
Bloco4	(4) 1086,4	(6) 765,4	(11) 1111,1	(14) 1037,0	(17) 938,3	(23) 938,3

Pede-se:

1-Realizar a análise de variância dos dados para testar se existe diferença entre os efeitos de tratamentos;

2-Obter as médias de tratamentos ajustadas.

Resolução: Análise usando o programa SAS

Entrada de dados: Os genótipos são codificados como tratamento (trat), repetição (rep), bloco (blc) e produtividade de grãos (prod).

O programa SAS utilizado é o seguinte:

```
DATA toria;
INPUT rep blc trat prod;
DATALINES;
1     1     1     1555.6
1     1     5     1160.5
1     1     9     1308.6
1     1     13    1382.7
1     1     17    987.7
1     1     21    1135.8
1     2     2     1284.0
1     2     6     1086.4
1     2     10    1284.0
1     2     14    1111.1
1     2     18    938.3
1     2     22    1308.6
1     3     3     1234.6
1     3     7     419.8
1     3     11    1308.6
1     3     15    963.0
1     3     19    963.0
1     3     23    987.7
1     4     4     1234.6
1     4     8     987.7
1     4     12    1284.0
1     4     16    913.6
1     4     20    1160.5
1     4     24    790.1
2     1     1     1481.5
2     1     6     1086.4
2     1     11    1308.6
2     1     16    1284.0
2     1     19    1111.1
2     1     22    1185.2
2     2     2     987.7
2     2     7     308.6
```

2	2	12	1234.6
2	2	13	1308.6
2	2	20	765.4
2	2	23	938.3
2	3	3	1012.3
2	3	8	864.2
2	3	9	1234.6
2	3	14	938.3
2	3	17	913.6
2	3	24	864.2
2	4	4	1135.8
2	4	5	987.7
2	4	10	987.7
2	4	15	740.7
2	4	18	963.0
2	4	21	1135.8
3	1	1	1284.0
3	1	7	333.3
3	1	12	1135.8
3	1	15	839.5
3	1	18	814.8
3	1	24	888.9
3	2	2	1135.8
3	2	8	913.6
3	2	9	1456.8
3	2	16	1037.0
3	2	19	938.3
3	2	21	1037.0
3	3	3	963.0
3	3	5	1209.9
3	3	10	1259.3
3	3	13	1234.6
3	3	20	963.0
3	3	22	1111.1
3	4	4	1086.4
3	4	6	765.4
3	4	11	1111.1
3	4	14	1037.0
3	4	17	938.3
3	4	23	938.3

;  
PROC GLM;  
CLASS rep blc trat;  
MODEL prod = trat rep blc (rep);

```

MEANS trat;
LSMEANS trat/PDIFF ADJUST=TUKEY;
RUN;

```

Principais resultados da análise usando o PROC GLM:

The GLM Procedure

Dependent Variable: prod

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	34	3535274.909	103978.674	12.79	<.0001
Error	37	300877.637	8131.828		
Corrected Total	71	3836152.546			

R-Square	Coeff Var	Root MSE	prod Mean
0.921568	8.542984	90.17665	1055.564

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trat	23	3205798.159	139382.529	17.14	<.0001
rep	2	135161.752	67580.876	8.31	0.0010
blc(rep)	9	194314.998	21590.555	2.66	0.0175

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trat	23	2555476.216	111107.662	13.66	<.0001
rep	2	135161.752	67580.876	8.31	0.0010
blc(rep)	9	194314.998	21590.555	2.66	0.0175

Level of	-----prod-----		
trat	N	Mean	Std Dev

1	3	1440.36667	140.394456
---	---	------------	------------

2	3	1135.83333	148.150003
3	3	1069.96667	144.691822
4	3	1152.26667	75.459746
5	3	1119.36667	116.671219
6	3	979.40000	185.329436
7	3	353.90000	58.392037
8	3	921.83333	62.160304
9	3	1333.33333	113.145982
10	3	1177.00000	164.403133
11	3	1242.76667	114.026678
12	3	1218.13333	75.459746
13	3	1308.63333	74.050006
14	3	1028.80000	86.691349
15	3	847.73333	111.378469
16	3	1078.20000	188.605726
17	3	946.53333	37.729873
18	3	905.36667	79.399391
19	3	1004.13333	93.455462
20	3	962.96667	197.550002
21	3	1102.86667	57.042207
22	3	1201.63333	99.770253
23	3	954.76667	28.521103
24	3	847.73333	51.417150

The GLM Procedure

Least Squares Means

Adjustment for Multiple Comparisons: Tukey-Kramer

		LSMEAN	
trat	prod	LSMEAN	Number
1		1403.80119	1
2		1121.76052	2
3		1074.43274	3
4		1198.43889	4
5		1136.00327	5
6		912.96482	6
7		384.23923	7
8		941.29268	8
9		1343.10923	9
10		1164.34601	10
11		1193.55661	11
12		1270.22149	12

13	1330.60677	13
14	1035.89312	14
15	872.73573	15
16	1024.13105	16
17	982.91704	17
18	913.14395	18
19	928.31546	19
20	994.62355	20
21	1112.74762	21
22	1115.34599	22
23	984.52679	23
24	894.37960	24

Least Squares Means for effect trat

Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: prod

i/j	1	2	3	4	5	6	7	8
1	0.1441	0.0371	0.6568	0.1766	<.0001	<.0001	<.0001	0.0004
2	0.1441		1.0000	1.0000	1.0000	0.5879	<.0001	0.8124
3	0.0371	1.0000		0.9959	1.0000	0.9370	<.0001	0.9877
4	0.6568	1.0000	0.9959		1.0000	0.1101	<.0001	0.2297
5	0.1766	1.0000	1.0000	1.0000		0.5525	<.0001	0.7741
6	<.0001	0.5879	0.9370	0.1101	0.5525		<.0001	1.0000
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
8	0.0004	0.8124	0.9877	0.2297	0.7741	1.0000	<.0001	
9	1.0000	0.4862	0.1727	0.9781	0.6134	0.0017	<.0001	0.0013
10	0.3969	1.0000	0.9999	1.0000	1.0000	0.2725	<.0001	0.5632
11	0.5755	1.0000	0.9967	1.0000	1.0000	0.0928	<.0001	0.3314
12	0.9873	0.9609	0.7461	1.0000	0.9919	0.0212	<.0001	0.0316
13	1.0000	0.5811	0.2264	0.9919	0.6440	0.0026	<.0001	0.0064
14	0.0117	1.0000	1.0000	0.9083	0.9998	0.9922	<.0001	0.9999

i/j	1	2	3	4	5	6	7	8
15	<.0001	0.3274	0.6465	0.0316	0.2055	1.0000	<.0001	1.0000
16	0.0050	0.9998	1.0000	0.8507	0.9992	0.9988	<.0001	1.0000
17	0.0012	0.9857	0.9999	0.5239	0.9506	1.0000	<.0001	1.0000
18	<.0001	0.5876	0.9371	0.1054	0.4677	1.0000	<.0001	1.0000
19	0.0001	0.7106	0.9661	0.2049	0.6767	1.0000	<.0001	1.0000
20	0.0029	0.9920	1.0000	0.6286	0.9772	1.0000	<.0001	1.0000
21	0.0895	1.0000	1.0000	1.0000	1.0000	0.7330	<.0001	0.8707

22	0.1006	1.0000	1.0000	1.0000	1.0000	0.5757	<.0001	0.8962
23	0.0019	0.9822	0.9999	0.5428	0.9698	1.0000	<.0001	1.0000
24	<.0001	0.4852	0.8143	0.0609	0.4010	1.0000	<.0001	1.0000

i/j	9	10	11	12	13	14	15	16	
1	1.0000	0.3969	0.5755	0.9873	1.0000	0.0117	<.0001	0.0050	
2	0.4862	1.0000	1.0000	0.9609	0.5811	1.0000	0.3274	0.9998	
3	0.1727	0.9999	0.9967	0.7461	0.2264	1.0000	0.6465	1.0000	
4	0.9781	1.0000	1.0000	1.0000	0.9919	0.9083	0.0316	0.8507	
5	0.6134	1.0000	1.0000	0.9919	0.6440	0.9998	0.2055	0.9992	
6	0.0017	0.2725	0.0928	0.0212	0.0026	0.9922	1.0000	0.9988	
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
8	0.0013	0.5632	0.3314	0.0316	0.0064	0.9999	1.0000	1.0000	
9		0.8726	0.9735	1.0000	1.0000	0.0613	0.0004	0.0403	
10	0.8726		1.0000	0.9997	0.8949	0.9922	0.0967	0.9866	
11	0.9735	1.0000		1.0000	0.9896	0.9320	0.0406	0.8841	
12	1.0000	0.9997	1.0000		1.0000	0.4620	0.0027	0.3054	
13	1.0000	0.8949	0.9896	1.0000		0.1172	0.0005	0.0851	
14	0.0613	0.9922	0.9320	0.4620	0.1172		0.9356	1.0000	
15	0.0004	0.0967	0.0406	0.0027	0.0005	0.9356		0.9678	
16	0.0403	0.9866	0.8841	0.3054	0.0851	1.0000	0.9678		
17	0.0064	0.8520	0.5825	0.1487	0.0169	1.0000	0.9993	1.0000	
18	0.0017	0.2093	0.1698	0.0124	0.0022	0.9955	1.0000	0.9993	
19	0.0016	0.4479	0.1442	0.0300	0.0041	0.9995	1.0000	0.9997	
20	0.0245	0.8830	0.7431	0.1074	0.0152	1.0000	0.9971	1.0000	
21	0.3423	1.0000	1.0000	0.9546	0.5088	1.0000	0.3475	1.0000	
22	0.5086	1.0000	1.0000	0.9624	0.5373	1.0000	0.3821	0.9999	
23	0.0191	0.8641	0.5169	0.1117	0.0180	1.0000	0.9986	1.0000	
24	0.0004	0.2268	0.1070	0.0036	0.0012	0.9772	1.0000	0.9905	

i/j	17	18	19	20	21	22	23	24	
1	0.0012	<.0001	0.0001	0.0029	0.0895	0.1006	0.0019	<.0001	
2	0.9857	0.5876	0.7106	0.9920	1.0000	1.0000	0.9822	0.4852	
3	0.9999	0.9371	0.9661	1.0000	1.0000	1.0000	0.9999	0.8143	
4	0.5239	0.1054	0.2049	0.6286	1.0000	1.0000	0.5428	0.0609	
5	0.9506	0.4677	0.6767	0.9772	1.0000	1.0000	0.9698	0.4010	
6	1.0000	1.0000	1.0000	1.0000	0.7330	0.5757	1.0000	1.0000	
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
8	1.0000	1.0000	1.0000	1.0000	0.8707	0.8962	1.0000	1.0000	
9	0.0064	0.0017	0.0016	0.0245	0.3423	0.5086	0.0191	0.0004	
10	0.8520	0.2093	0.4479	0.8830	1.0000	1.0000	0.8641	0.2268	
11	0.5825	0.1698	0.1442	0.7431	1.0000	1.0000	0.5169	0.1070	

12	0.1487	0.0124	0.0300	0.1074	0.9546	0.9624	0.1117	0.0036
13	0.0169	0.0022	0.0041	0.0152	0.5088	0.5373	0.0180	0.0012
14	1.0000	0.9955	0.9995	1.0000	1.0000	1.0000	1.0000	0.9772
15	0.9993	1.0000	1.0000	0.9971	0.3475	0.3821	0.9986	1.0000
16	1.0000	0.9993	0.9997	1.0000	1.0000	0.9999	1.0000	0.9905
17		1.0000	1.0000	1.0000	0.9904	0.9922	1.0000	1.0000
18	1.0000		1.0000	1.0000	0.6700	0.6422	1.0000	1.0000
19	1.0000	1.0000		1.0000	0.7927	0.7693	1.0000	1.0000
20	1.0000	1.0000	1.0000		0.9982	0.9964	1.0000	0.9997
21	0.9904	0.6700	0.7927	0.9982		1.0000	0.9953	0.5695
22	0.9922	0.6422	0.7693	0.9964	1.0000		0.9931	0.5685
23	1.0000	1.0000	1.0000	1.0000	0.9953	0.9931		1.0000
24	1.0000	1.0000	1.0000	0.9997	0.5695	0.5685	1.0000	

A análise anterior pode ser feita também por meio da metodologia de modelos mistos, usando os códigos SAS a seguir:

```
PROC MIXED METHOD = REML;
CLASS rep blc trat;
MODEL prod = trat rep blc(rep);
RANDOM blc(rep);
LSMEANS trat/PDIFF ADJUST = TUKEY;
RUN;
```

Principais resultados da análise usando o PROC MIXED:

#### Type 3 Tests of Fixed Effects

Effect	Num		Den		F Value	Pr > F
	DF	DF	DF	DF		
trat	23	37	13.66		<.0001	
rep	2	0	8.31		.	
blc(rep)	9	0	2.66		.	

#### Least Squares Means

Effect	trat	Standard					
		Estimate	Error	DF	t Value	Pr >  t	
trat	1	1403.80	56.3067	37	24.93	<.0001	
trat	2	1121.76	56.3067	37	19.92	<.0001	
trat	3	1074.43	56.3067	37	19.08	<.0001	
trat	4	1198.44	56.3067	37	21.28	<.0001	
trat	5	1136.00	57.3571	37	19.81	<.0001	
trat	6	912.96	57.3571	37	15.92	<.0001	
trat	7	384.24	57.3571	37	6.70	<.0001	
trat	8	941.29	57.3571	37	16.41	<.0001	
trat	9	1343.11	57.3571	37	23.42	<.0001	
trat	10	1164.35	57.3571	37	20.30	<.0001	
trat	11	1193.56	57.3571	37	20.81	<.0001	
trat	12	1270.22	57.3571	37	22.15	<.0001	
trat	13	1330.61	56.9620	37	23.36	<.0001	
trat	14	1035.89	56.9620	37	18.19	<.0001	
trat	15	872.74	56.9620	37	15.32	<.0001	
trat	16	1024.13	56.9620	37	17.98	<.0001	

trat	17	982.92	56.9620	37	17.26	<.0001
trat	18	913.14	56.9620	37	16.03	<.0001
trat	19	928.32	56.9620	37	16.30	<.0001
trat	20	994.62	56.9620	37	17.46	<.0001
trat	21	1112.75	56.9903	37	19.53	<.0001
trat	22	1115.35	56.9903	37	19.57	<.0001
trat	23	984.53	56.9903	37	17.28	<.0001
trat	24	894.38	56.9903	37	15.69	<.0001

### Differences of Least Squares Means

Effect	trat	_trat	Standard		DF	t Value	Pr >  t	Adjustment	Adj P
			Estimate	Error					
trat	1	2	282.04	81.5319	37	3.46	0.0014	Tukey-Kramer	0.1441
trat	1	3	329.37	81.5319	37	4.04	0.0003	Tukey-Kramer	0.0371
trat	1	4	205.36	81.5319	37	2.52	0.0162	Tukey-Kramer	0.6568
trat	1	5	267.80	79.6702	37	3.36	0.0018	Tukey-Kramer	0.1766
trat	1	6	490.84	79.8725	37	6.15	<.0001	Tukey-Kramer	<.0001
trat	1	7	1019.56	79.6702	37	12.80	<.0001	Tukey-Kramer	<.0001
trat	1	8	462.51	82.2608	37	5.62	<.0001	Tukey-Kramer	0.0004
trat	1	9	60.6920	79.6702	37	0.76	0.4510	Tukey-Kramer	1.0000
trat	1	10	239.46	82.2608	37	2.91	0.0061	Tukey-Kramer	0.3969
trat	1	11	210.24	79.6702	37	2.64	0.0121	Tukey-Kramer	0.5755
trat	1	12	133.58	79.8725	37	1.67	0.1029	Tukey-Kramer	0.9873
trat	1	13	73.1944	79.5892	37	0.92	0.3637	Tukey-Kramer	1.0000
trat	1	14	367.91	82.1823	37	4.48	<.0001	Tukey-Kramer	0.0117
trat	1	15	531.07	79.1827	37	6.71	<.0001	Tukey-Kramer	<.0001
trat	1	16	379.67	79.3862	37	4.78	<.0001	Tukey-Kramer	0.0050
trat	1	17	420.88	79.5892	37	5.29	<.0001	Tukey-Kramer	0.0012
trat	1	18	490.66	79.3862	37	6.18	<.0001	Tukey-Kramer	<.0001
trat	1	19	475.49	79.1827	37	6.00	<.0001	Tukey-Kramer	0.0001
trat	1	20	409.18	82.1823	37	4.98	<.0001	Tukey-Kramer	0.0029
trat	1	21	291.05	79.2031	37	3.67	0.0007	Tukey-Kramer	0.0895
trat	1	22	288.46	79.6094	37	3.62	0.0009	Tukey-Kramer	0.1006
trat	1	23	419.27	82.0054	37	5.11	<.0001	Tukey-Kramer	0.0019
trat	1	24	509.42	79.6094	37	6.40	<.0001	Tukey-Kramer	<.0001
trat	2	3	47.3278	81.5319	37	0.58	0.5651	Tukey-Kramer	1.0000
trat	2	4	-76.6784	81.5319	37	-0.94	0.3531	Tukey-Kramer	1.0000
trat	2	5	-14.2428	82.2608	37	-0.17	0.8635	Tukey-Kramer	1.0000
trat	2	6	208.80	79.6702	37	2.62	0.0127	Tukey-Kramer	0.5879
trat	2	7	737.52	79.8725	37	9.23	<.0001	Tukey-Kramer	<.0001
trat	2	8	180.47	79.6702	37	2.27	0.0294	Tukey-Kramer	0.8124

trat	2	9	-221.35	79.8725	37	-2.77	0.0087	Tukey-Kramer	0.4862
trat	2	10	-42.5855	79.6702	37	-0.53	0.5962	Tukey-Kramer	1.0000
trat	2	11	-71.7961	82.2608	37	-0.87	0.3884	Tukey-Kramer	1.0000
trat	2	12	-148.46	79.6702	37	-1.86	0.0704	Tukey-Kramer	0.9609
trat	2	13	-208.85	79.3862	37	-2.63	0.0123	Tukey-Kramer	0.5811
trat	2	14	85.8674	79.5892	37	1.08	0.2876	Tukey-Kramer	1.0000
trat	2	15	249.02	82.1823	37	3.03	0.0044	Tukey-Kramer	0.3274
trat	2	16	97.6295	79.1827	37	1.23	0.2254	Tukey-Kramer	0.9998
trat	2	17	138.84	82.1823	37	1.69	0.0995	Tukey-Kramer	0.9857
trat	2	18	208.62	79.5892	37	2.62	0.0126	Tukey-Kramer	0.5876
trat	2	19	193.45	79.3862	37	2.44	0.0197	Tukey-Kramer	0.7106
trat	2	20	127.14	79.1827	37	1.61	0.1169	Tukey-Kramer	0.9920
trat	2	21	9.0129	79.6094	37	0.11	0.9105	Tukey-Kramer	1.0000
trat	2	22	6.4145	79.2031	37	0.08	0.9359	Tukey-Kramer	1.0000
trat	2	23	137.23	79.6094	37	1.72	0.0931	Tukey-Kramer	0.9822
trat	2	24	227.38	82.0054	37	2.77	0.0087	Tukey-Kramer	0.4852
trat	3	4	-124.01	81.5319	37	-1.52	0.1368	Tukey-Kramer	0.9959
trat	3	5	-61.5705	79.6702	37	-0.77	0.4445	Tukey-Kramer	1.0000
trat	3	6	161.47	82.2608	37	1.96	0.0572	Tukey-Kramer	0.9370
trat	3	7	690.19	79.6702	37	8.66	<.0001	Tukey-Kramer	<.0001
trat	3	8	133.14	79.8725	37	1.67	0.1040	Tukey-Kramer	0.9877
trat	3	9	-268.68	79.6702	37	-3.37	0.0018	Tukey-Kramer	0.1727
trat	3	10	-89.9133	79.8725	37	-1.13	0.2675	Tukey-Kramer	0.9999
trat	3	11	-119.12	79.6702	37	-1.50	0.1433	Tukey-Kramer	0.9967
trat	3	12	-195.79	82.2608	37	-2.38	0.0226	Tukey-Kramer	0.7461
trat	3	13	-256.17	79.1827	37	-3.24	0.0026	Tukey-Kramer	0.2264
trat	3	14	38.5396	79.3862	37	0.49	0.6302	Tukey-Kramer	1.0000
trat	3	15	201.70	79.5892	37	2.53	0.0156	Tukey-Kramer	0.6465
trat	3	16	50.3017	82.1823	37	0.61	0.5442	Tukey-Kramer	1.0000
trat	3	17	91.5157	79.1827	37	1.16	0.2552	Tukey-Kramer	0.9999
trat	3	18	161.29	82.1823	37	1.96	0.0572	Tukey-Kramer	0.9371
trat	3	19	146.12	79.5892	37	1.84	0.0744	Tukey-Kramer	0.9661
trat	3	20	79.8092	79.3862	37	1.01	0.3213	Tukey-Kramer	1.0000
trat	3	21	-38.3149	82.0054	37	-0.47	0.6431	Tukey-Kramer	1.0000
trat	3	22	-40.9133	79.6094	37	-0.51	0.6104	Tukey-Kramer	1.0000
trat	3	23	89.9060	79.2031	37	1.14	0.2636	Tukey-Kramer	0.9999
trat	3	24	180.05	79.6094	37	2.26	0.0297	Tukey-Kramer	0.8143
trat	4	5	62.4356	79.8725	37	0.78	0.4394	Tukey-Kramer	1.0000
trat	4	6	285.47	79.6702	37	3.58	0.0010	Tukey-Kramer	0.1101
trat	4	7	814.20	82.2608	37	9.90	<.0001	Tukey-Kramer	<.0001
trat	4	8	257.15	79.6702	37	3.23	0.0026	Tukey-Kramer	0.2297
trat	4	9	-144.67	82.2608	37	-1.76	0.0869	Tukey-Kramer	0.9781
trat	4	10	34.0929	79.6702	37	0.43	0.6712	Tukey-Kramer	1.0000
trat	4	11	4.8823	79.8725	37	0.06	0.9516	Tukey-Kramer	1.0000

trat	4	12	-71.7826	79.6702	37	-0.90	0.3734	Tukey-Kramer	1.0000
trat	4	13	-132.17	82.1823	37	-1.61	0.1163	Tukey-Kramer	0.9919
trat	4	14	162.55	79.1827	37	2.05	0.0472	Tukey-Kramer	0.9083
trat	4	15	325.70	79.3862	37	4.10	0.0002	Tukey-Kramer	0.0316
trat	4	16	174.31	79.5892	37	2.19	0.0349	Tukey-Kramer	0.8507
trat	4	17	215.52	79.3862	37	2.71	0.0100	Tukey-Kramer	0.5239
trat	4	18	285.29	79.1827	37	3.60	0.0009	Tukey-Kramer	0.1054
trat	4	19	270.12	82.1823	37	3.29	0.0022	Tukey-Kramer	0.2049
trat	4	20	203.82	79.5892	37	2.56	0.0147	Tukey-Kramer	0.6286
trat	4	21	85.6913	79.6094	37	1.08	0.2887	Tukey-Kramer	1.0000
trat	4	22	83.0929	82.0054	37	1.01	0.3175	Tukey-Kramer	1.0000
trat	4	23	213.91	79.6094	37	2.69	0.0107	Tukey-Kramer	0.5428
trat	4	24	304.06	79.2031	37	3.84	0.0005	Tukey-Kramer	0.0609
trat	5	6	223.04	83.4487	37	2.67	0.0111	Tukey-Kramer	0.5525
trat	5	7	751.76	83.4874	37	9.00	<.0001	Tukey-Kramer	<.0001
trat	5	8	194.71	83.4487	37	2.33	0.0252	Tukey-Kramer	0.7741
trat	5	9	-207.11	80.1749	37	-2.58	0.0139	Tukey-Kramer	0.6134
trat	5	10	-28.3427	76.6775	37	-0.37	0.7138	Tukey-Kramer	1.0000
trat	5	11	-57.5533	83.9499	37	-0.69	0.4973	Tukey-Kramer	1.0000
trat	5	12	-134.22	83.4487	37	-1.61	0.1163	Tukey-Kramer	0.9919
trat	5	13	-194.60	76.6775	37	-2.54	0.0155	Tukey-Kramer	0.6440
trat	5	14	100.11	82.9445	37	1.21	0.2351	Tukey-Kramer	0.9998
trat	5	15	263.27	80.1347	37	3.29	0.0022	Tukey-Kramer	0.2055
trat	5	16	111.87	83.4100	37	1.34	0.1880	Tukey-Kramer	0.9992
trat	5	17	153.09	80.1347	37	1.91	0.0639	Tukey-Kramer	0.9506
trat	5	18	222.86	79.6094	37	2.80	0.0081	Tukey-Kramer	0.4677
trat	5	19	207.69	83.4487	37	2.49	0.0174	Tukey-Kramer	0.6767
trat	5	20	141.38	80.0944	37	1.77	0.0858	Tukey-Kramer	0.9772
trat	5	21	23.2557	76.6986	37	0.30	0.7634	Tukey-Kramer	1.0000
trat	5	22	20.6573	79.8725	37	0.26	0.7974	Tukey-Kramer	1.0000
trat	5	23	151.48	83.4680	37	1.81	0.0777	Tukey-Kramer	0.9698
trat	5	24	241.62	83.1970	37	2.90	0.0062	Tukey-Kramer	0.4010
trat	6	7	528.73	83.4487	37	6.34	<.0001	Tukey-Kramer	<.0001
trat	6	8	-28.3279	83.4874	37	-0.34	0.7363	Tukey-Kramer	1.0000
trat	6	9	-430.14	83.4487	37	-5.15	<.0001	Tukey-Kramer	0.0017
trat	6	10	-251.38	80.1749	37	-3.14	0.0034	Tukey-Kramer	0.2725
trat	6	11	-280.59	76.6775	37	-3.66	0.0008	Tukey-Kramer	0.0928
trat	6	12	-357.26	83.9499	37	-4.26	0.0001	Tukey-Kramer	0.0212
trat	6	13	-417.64	83.4100	37	-5.01	<.0001	Tukey-Kramer	0.0026
trat	6	14	-122.93	76.6775	37	-1.60	0.1174	Tukey-Kramer	0.9922
trat	6	15	40.2291	82.9445	37	0.49	0.6305	Tukey-Kramer	1.0000
trat	6	16	-111.17	80.1347	37	-1.39	0.1737	Tukey-Kramer	0.9988
trat	6	17	-69.9522	80.0944	37	-0.87	0.3881	Tukey-Kramer	1.0000
trat	6	18	-0.1791	80.1347	37	-0.00	0.9982	Tukey-Kramer	1.0000

trat	6	19	-15.3506	79.6094	37	-0.19	0.8482	Tukey-Kramer	1.0000
trat	6	20	-81.6587	83.4487	37	-0.98	0.3342	Tukey-Kramer	1.0000
trat	6	21	-199.78	83.1970	37	-2.40	0.0215	Tukey-Kramer	0.7330
trat	6	22	-202.38	76.6986	37	-2.64	0.0121	Tukey-Kramer	0.5757
trat	6	23	-71.5620	79.8725	37	-0.90	0.3761	Tukey-Kramer	1.0000
trat	6	24	18.5852	83.4680	37	0.22	0.8250	Tukey-Kramer	1.0000
trat	7	8	-557.05	83.4487	37	-6.68	<.0001	Tukey-Kramer	<.0001
trat	7	9	-958.87	83.9499	37	-11.42	<.0001	Tukey-Kramer	<.0001
trat	7	10	-780.11	83.4487	37	-9.35	<.0001	Tukey-Kramer	<.0001
trat	7	11	-809.32	80.1749	37	-10.09	<.0001	Tukey-Kramer	<.0001
trat	7	12	-885.98	76.6775	37	-11.55	<.0001	Tukey-Kramer	<.0001
trat	7	13	-946.37	80.1347	37	-11.81	<.0001	Tukey-Kramer	<.0001
trat	7	14	-651.65	83.4100	37	-7.81	<.0001	Tukey-Kramer	<.0001
trat	7	15	-488.50	76.6775	37	-6.37	<.0001	Tukey-Kramer	<.0001
trat	7	16	-639.89	82.9445	37	-7.71	<.0001	Tukey-Kramer	<.0001
trat	7	17	-598.68	83.4487	37	-7.17	<.0001	Tukey-Kramer	<.0001
trat	7	18	-528.90	80.0944	37	-6.60	<.0001	Tukey-Kramer	<.0001
trat	7	19	-544.08	80.1347	37	-6.79	<.0001	Tukey-Kramer	<.0001
trat	7	20	-610.38	79.6094	37	-7.67	<.0001	Tukey-Kramer	<.0001
trat	7	21	-728.51	83.4680	37	-8.73	<.0001	Tukey-Kramer	<.0001
trat	7	22	-731.11	83.1970	37	-8.79	<.0001	Tukey-Kramer	<.0001
trat	7	23	-600.29	76.6986	37	-7.83	<.0001	Tukey-Kramer	<.0001
trat	7	24	-510.14	79.8725	37	-6.39	<.0001	Tukey-Kramer	<.0001
trat	8	9	-401.82	76.6775	37	-5.24	<.0001	Tukey-Kramer	0.0013
trat	8	10	-223.05	83.9499	37	-2.66	0.0116	Tukey-Kramer	0.5632
trat	8	11	-252.26	83.4487	37	-3.02	0.0045	Tukey-Kramer	0.3314
trat	8	12	-328.93	80.1749	37	-4.10	0.0002	Tukey-Kramer	0.0316
trat	8	13	-389.31	82.9445	37	-4.69	<.0001	Tukey-Kramer	0.0064
trat	8	14	-94.6004	80.1347	37	-1.18	0.2453	Tukey-Kramer	0.9999
trat	8	15	68.5569	83.4100	37	0.82	0.4164	Tukey-Kramer	1.0000
trat	8	16	-82.8384	76.6775	37	-1.08	0.2870	Tukey-Kramer	1.0000
trat	8	17	-41.6244	79.6094	37	-0.52	0.6042	Tukey-Kramer	1.0000
trat	8	18	28.1487	83.4487	37	0.34	0.7378	Tukey-Kramer	1.0000
trat	8	19	12.9772	80.0944	37	0.16	0.8722	Tukey-Kramer	1.0000
trat	8	20	-53.3309	80.1347	37	-0.67	0.5098	Tukey-Kramer	1.0000
trat	8	21	-171.45	79.8725	37	-2.15	0.0384	Tukey-Kramer	0.8707
trat	8	22	-174.05	83.4680	37	-2.09	0.0440	Tukey-Kramer	0.8962
trat	8	23	-43.2341	83.1970	37	-0.52	0.6064	Tukey-Kramer	1.0000
trat	8	24	46.9131	76.6986	37	0.61	0.5445	Tukey-Kramer	1.0000
trat	9	10	178.76	83.4487	37	2.14	0.0388	Tukey-Kramer	0.8726
trat	9	11	149.55	83.4874	37	1.79	0.0814	Tukey-Kramer	0.9735
trat	9	12	72.8877	83.4487	37	0.87	0.3881	Tukey-Kramer	1.0000
trat	9	13	12.5025	80.1347	37	0.16	0.8769	Tukey-Kramer	1.0000
trat	9	14	307.22	80.0944	37	3.84	0.0005	Tukey-Kramer	0.0613

trat	9	15	470.37	83.4487	37	5.64	<.0001	Tukey-Kramer	0.0004
trat	9	16	318.98	79.6094	37	4.01	0.0003	Tukey-Kramer	0.0403
trat	9	17	360.19	76.6775	37	4.70	<.0001	Tukey-Kramer	0.0064
trat	9	18	429.97	83.4100	37	5.15	<.0001	Tukey-Kramer	0.0017
trat	9	19	414.79	80.1347	37	5.18	<.0001	Tukey-Kramer	0.0016
trat	9	20	348.49	82.9445	37	4.20	0.0002	Tukey-Kramer	0.0245
trat	9	21	230.36	76.6986	37	3.00	0.0048	Tukey-Kramer	0.3423
trat	9	22	227.76	83.1970	37	2.74	0.0095	Tukey-Kramer	0.5086
trat	9	23	358.58	83.4680	37	4.30	0.0001	Tukey-Kramer	0.0191
trat	9	24	448.73	79.8725	37	5.62	<.0001	Tukey-Kramer	0.0004
trat	10	11	-29.2106	83.4487	37	-0.35	0.7283	Tukey-Kramer	1.0000
trat	10	12	-105.88	83.4874	37	-1.27	0.2127	Tukey-Kramer	0.9997
trat	10	13	-166.26	79.6094	37	-2.09	0.0437	Tukey-Kramer	0.8949
trat	10	14	128.45	80.1347	37	1.60	0.1174	Tukey-Kramer	0.9922
trat	10	15	291.61	80.0944	37	3.64	0.0008	Tukey-Kramer	0.0967
trat	10	16	140.21	83.4487	37	1.68	0.1013	Tukey-Kramer	0.9866
trat	10	17	181.43	82.9445	37	2.19	0.0351	Tukey-Kramer	0.8520
trat	10	18	251.20	76.6775	37	3.28	0.0023	Tukey-Kramer	0.2093
trat	10	19	236.03	83.4100	37	2.83	0.0075	Tukey-Kramer	0.4479
trat	10	20	169.72	80.1347	37	2.12	0.0410	Tukey-Kramer	0.8830
trat	10	21	51.5984	79.8725	37	0.65	0.5223	Tukey-Kramer	1.0000
trat	10	22	49.0000	76.6986	37	0.64	0.5268	Tukey-Kramer	1.0000
trat	10	23	179.82	83.1970	37	2.16	0.0372	Tukey-Kramer	0.8641
trat	10	24	269.97	83.4680	37	3.23	0.0026	Tukey-Kramer	0.2268
trat	11	12	-76.6649	83.4487	37	-0.92	0.3642	Tukey-Kramer	1.0000
trat	11	13	-137.05	83.4487	37	-1.64	0.1090	Tukey-Kramer	0.9896
trat	11	14	157.66	79.6094	37	1.98	0.0551	Tukey-Kramer	0.9320
trat	11	15	320.82	80.1347	37	4.00	0.0003	Tukey-Kramer	0.0406
trat	11	16	169.43	80.0944	37	2.12	0.0412	Tukey-Kramer	0.8841
trat	11	17	210.64	80.1347	37	2.63	0.0124	Tukey-Kramer	0.5825
trat	11	18	280.41	82.9445	37	3.38	0.0017	Tukey-Kramer	0.1698
trat	11	19	265.24	76.6775	37	3.46	0.0014	Tukey-Kramer	0.1442
trat	11	20	198.93	83.4100	37	2.39	0.0223	Tukey-Kramer	0.7431
trat	11	21	80.8090	83.4680	37	0.97	0.3393	Tukey-Kramer	1.0000
trat	11	22	78.2106	79.8725	37	0.98	0.3338	Tukey-Kramer	1.0000
trat	11	23	209.03	76.6986	37	2.73	0.0098	Tukey-Kramer	0.5169
trat	11	24	299.18	83.1970	37	3.60	0.0009	Tukey-Kramer	0.1070
trat	12	13	-60.3853	80.0944	37	-0.75	0.4557	Tukey-Kramer	1.0000
trat	12	14	234.33	83.4487	37	2.81	0.0079	Tukey-Kramer	0.4620
trat	12	15	397.49	79.6094	37	4.99	<.0001	Tukey-Kramer	0.0027
trat	12	16	246.09	80.1347	37	3.07	0.0040	Tukey-Kramer	0.3054
trat	12	17	287.30	83.4100	37	3.44	0.0014	Tukey-Kramer	0.1487
trat	12	18	357.08	80.1347	37	4.46	<.0001	Tukey-Kramer	0.0124
trat	12	19	341.91	82.9445	37	4.12	0.0002	Tukey-Kramer	0.0300

trat	12	20	275.60	76.6775	37	3.59	0.0009	Tukey-Kramer	0.1074
trat	12	21	157.47	83.1970	37	1.89	0.0662	Tukey-Kramer	0.9546
trat	12	22	154.88	83.4680	37	1.86	0.0715	Tukey-Kramer	0.9624
trat	12	23	285.69	79.8725	37	3.58	0.0010	Tukey-Kramer	0.1117
trat	12	24	375.84	76.6986	37	4.90	<.0001	Tukey-Kramer	0.0036
trat	13	14	294.71	82.9056	37	3.55	0.0011	Tukey-Kramer	0.1172
trat	13	15	457.87	82.3980	37	5.56	<.0001	Tukey-Kramer	0.0005
trat	13	16	306.48	82.9056	37	3.70	0.0007	Tukey-Kramer	0.0851
trat	13	17	347.69	80.0944	37	4.34	0.0001	Tukey-Kramer	0.0169
trat	13	18	417.46	82.3588	37	5.07	<.0001	Tukey-Kramer	0.0022
trat	13	19	402.29	82.9445	37	4.85	<.0001	Tukey-Kramer	0.0041
trat	13	20	335.98	76.6775	37	4.38	<.0001	Tukey-Kramer	0.0152
trat	13	21	217.86	79.5892	37	2.74	0.0095	Tukey-Kramer	0.5088
trat	13	22	215.26	79.8725	37	2.70	0.0105	Tukey-Kramer	0.5373
trat	13	23	346.08	80.1548	37	4.32	0.0001	Tukey-Kramer	0.0180
trat	13	24	436.23	82.6522	37	5.28	<.0001	Tukey-Kramer	0.0012
trat	14	15	163.16	82.9056	37	1.97	0.0566	Tukey-Kramer	0.9356
trat	14	16	11.7621	82.3980	37	0.14	0.8873	Tukey-Kramer	1.0000
trat	14	17	52.9761	76.6775	37	0.69	0.4939	Tukey-Kramer	1.000
trat	14	18	122.75	80.0944	37	1.53	0.1339	Tukey-Kramer	0.9955
trat	14	19	107.58	82.3588	37	1.31	0.1995	Tukey-Kramer	0.9995
trat	14	20	41.2696	82.9445	37	0.50	0.6217	Tukey-Kramer	1.0000
trat	14	21	-76.8545	82.6522	37	-0.93	0.3585	Tukey-Kramer	1.0000
trat	14	22	-79.4529	79.5892	37	-1.00	0.3246	Tukey-Kramer	1.0000
trat	14	23	51.3663	79.8725	37	0.64	0.5241	Tukey-Kramer	1.0000
trat	14	24	141.51	80.1548	37	1.77	0.0857	Tukey-Kramer	0.9772
trat	15	16	-151.40	82.9056	37	-1.83	0.0759	Tukey-Kramer	0.9678
trat	15	17	-110.18	82.9445	37	-1.33	0.1922	Tukey-Kramer	0.9993
trat	15	18	-40.4082	76.6775	37	-0.53	0.6013	Tukey-Kramer	1.0000
trat	15	19	-55.5797	80.0944	37	-0.69	0.4921	Tukey-Kramer	1.0000
trat	15	20	-121.89	82.3588	37	-1.48	0.1473	Tukey-Kramer	0.9971
trat	15	21	-240.01	80.1548	37	-2.99	0.0049	Tukey-Kramer	0.3475
trat	15	22	-242.61	82.6522	37	-2.94	0.0057	Tukey-Kramer	0.3821
trat	15	23	-111.79	79.5892	37	-1.40	0.1685	Tukey-Kramer	0.9986
trat	15	24	-21.6439	79.8725	37	-0.27	0.7879	Tukey-Kramer	1.0000
trat	16	17	41.2140	82.3588	37	0.50	0.6197	Tukey-Kramer	1.0000
trat	16	18	110.99	82.9445	37	1.34	0.1890	Tukey-Kramer	0.9993
trat	16	19	95.8156	76.6775	37	1.25	0.2193	Tukey-Kramer	0.9997
trat	16	20	29.5075	80.0944	37	0.37	0.7147	Tukey-Kramer	1.0000
trat	16	21	-88.6166	79.8725	37	-1.11	0.2744	Tukey-Kramer	1.0000
trat	16	22	-91.2149	80.1548	37	-1.14	0.2624	Tukey-Kramer	0.9999
trat	16	23	39.6043	82.6522	37	0.48	0.6346	Tukey-Kramer	1.0000
trat	16	24	129.75	79.5892	37	1.63	0.1115	Tukey-Kramer	0.9905
trat	17	18	69.7731	82.9056	37	0.84	0.4054	Tukey-Kramer	1.0000

trat	17	19	54.6016	82.3980	37	0.66	0.5117	Tukey-Kramer	1.0000
trat	17	20	-11.7065	82.9056	37	-0.14	0.8885	Tukey-Kramer	1.0000
trat	17	21	-129.83	79.5892	37	-1.63	0.1113	Tukey-Kramer	0.9904
trat	17	22	-132.43	82.6522	37	-1.60	0.1176	Tukey-Kramer	0.9922
trat	17	23	-1.6097	80.1548	37	-0.02	0.9841	Tukey-Kramer	1.0000
trat	17	24	88.5374	79.8725	37	1.11	0.2748	Tukey-Kramer	1.0000
trat	18	19	-15.1715	82.9056	37	-0.18	0.8558	Tukey-Kramer	1.0000
trat	18	20	-81.4796	82.3980	37	-0.99	0.3292	Tukey-Kramer	1.0000
trat	18	21	-199.60	79.8725	37	-2.50	0.0170	Tukey-Kramer	0.6700
trat	18	22	-202.20	79.5892	37	-2.54	0.0154	Tukey-Kramer	0.6422
trat	18	23	-71.3828	82.6522	37	-0.86	0.3933	Tukey-Kramer	1.0000
trat	18	24	18.7643	80.1548	37	0.23	0.8162	Tukey-Kramer	1.0000
trat	19	20	-66.3081	82.9056	37	-0.80	0.4289	Tukey-Kramer	1.0000
trat	19	21	-184.43	80.1548	37	-2.30	0.0271	Tukey-Kramer	0.7927
trat	19	22	-187.03	79.8725	37	-2.34	0.0247	Tukey-Kramer	0.7693
trat	19	23	-56.2113	79.5892	37	-0.71	0.4844	Tukey-Kramer	1.0000
trat	19	24	33.9359	82.6522	37	0.41	0.6837	Tukey-Kramer	1.0000
trat	20	21	-118.12	82.6522	37	-1.43	0.1613	Tukey-Kramer	0.9982
trat	20	22	-120.72	80.1548	37	-1.51	0.1405	Tukey-Kramer	0.9964
trat	20	23	10.0968	79.8725	37	0.13	0.9001	Tukey-Kramer	1.0000
trat	20	24	100.24	79.5892	37	1.26	0.2157	Tukey-Kramer	0.9997
trat	21	22	-2.5984	82.4763	37	-0.03	0.9750	Tukey-Kramer	1.0000
trat	21	23	128.22	83.4100	37	1.54	0.1327	Tukey-Kramer	0.9953
trat	21	24	218.37	82.4763	37	2.65	0.0118	Tukey-Kramer	0.5695
trat	22	23	130.82	82.4763	37	1.59	0.1212	Tukey-Kramer	0.9931
trat	22	24	220.97	83.4100	37	2.65	0.0118	Tukey-Kramer	0.5685
trat	23	24	90.1472	82.4763	37	1.09	0.2815	Tukey-Kramer	1.0000

# CAPITULO 11

## Delineamentos Experimentais Aumentados

Nestes delineamentos, as testemunhas (tratamentos comuns) são repetidas num determinado delineamento experimental básico. Os tratamentos novos (tratamentos regulares) não são repetidos no delineamento experimental; eles aumentam o delineamento.

### Características dos Delineamentos Experimentais Aumentados

- a) Fornecem estimativas de erro padrão que podem ser usadas para as comparações seguintes: entre os tratamentos regulares, entre os tratamentos comuns e entre os tratamentos regulares e os tratamentos comuns.
- b) As observações sobre os tratamentos regulares podem ser ajustadas para a heterogeneidade da área experimental por meio da formação de blocos.
- c) São delineamentos sem repetições dos tratamentos regulares (genótipos a serem avaliados), que podem ser bem utilizados em condições de escassez de recursos materiais ou financeiros.
- d) São delineamentos flexíveis, ou seja, os blocos podem ser de tamanhos diferentes.

### Aplicações dos Delineamentos Experimentais Aumentados

- a) Nos estágios iniciais em um programa de melhoramento, porque pode-se ter quantidade insuficiente de sementes para repetição e o uso de uma única repetição permite que mais genótipos possam ser avaliados.
- b) No melhoramento de plantas participatório, porque os produtores podem preferir cultivar uma única repetição quando existem muitos genótipos para serem avaliados.
- c) Na pesquisa de sistemas de produção para produtores, que precisa avaliar genótipos promissores num maior número de ambientes possíveis.

### Delineamentos Experimentais Aumentados no DBC

A área experimental é dividida em blocos, que são blocos incompletos porque eles contêm apenas um subconjunto dos genótipos a serem avaliados.

Dois ou mais tratamentos comuns (testemunhas) são designados ao acaso às parcelas dentro dos blocos, de forma que os mesmos tratamentos comuns aparecem em cada bloco.

Este delineamento é mais eficiente quando o tamanho de bloco é constante.

Os tratamentos comuns são repetidos, mas os tratamentos regulares não são repetidos.

### Número de blocos necessários

É necessário ter no mínimo 10 graus de liberdade para erro na análise de variância de, considerando apenas os tratamentos comuns. Neste caso, o número de graus de liberdade para erro pode ser calculado por meio da seguinte expressão:

$$gl = (r-1)(c-1), \text{ em que:}$$

$c$ : número de tratamentos comuns por bloco;

$r$ : número de repetições de um tratamento comum (número de blocos).

Então, o número mínimo de blocos deve ser:

$$r \geq \lceil (10)/(c-1) \rceil + 1$$

Por exemplo, para 4 tratamentos comuns (testemunhas) o número mínimo de blocos necessário é:

$$r \geq \lceil (10)/(4-1) \rceil + 1$$

$$r \geq 10/3 + 1$$

$$r \geq 4,33, \text{ ou seja, } r \geq 5 \text{ blocos.}$$

Desta forma, cada bloco tem  $c+t$  parcelas, sendo  $t$  o número de tratamentos regulares (genótipos a serem avaliados) e  $c$  o número de tratamentos comuns (testemunhas).

### Análise de Variância para Delineamentos Experimentais Blocos Aumentados

O erro experimental é estimado tratando os tratamentos comuns como se eles fossem os tratamentos em um delineamento experimental blocos casualizados. O quadrado médio do erro é então usado para construir erros padrões para os diferentes tipos de comparações entre médias de tratamentos.

### Ajustamento de tratamentos para as diferenças entre blocos

O ajuste é feito com base na diferença entre as médias de bloco de cada tratamento comum e a média geral de tratamentos comuns. Este cálculo assume que os blocos são de efeitos fixos. Então o modelo linear utilizado é o seguinte:

$$Y_{ij} = \mu + T_i + B_j + e_{ij}, \text{ com } B_j = Y_{.j} - \bar{Y}_{..} \text{ e } \sum_j B_j = 0.$$

Passos na análise de variância:

1-Construa uma tabela de dupla entrada de médias de tratamento comum x bloco;

2-Calcule a média geral e a média dos tratamentos comuns em cada bloco;

3-Calcule o ajuste de bloco como  $b_j = \bar{Y}_{.j} - \bar{Y}_{..}$ ;

4-Ajuste as observações dos tratamentos regulares como  $\hat{Y}_{ij} = Y_{ij} - b_j$ ;

5-Complete a análise de variância padrão para o delineamento blocos casualizados usando as observações dos tratamentos comuns;

6-Faça o esquema de análise de variância:

FV	GL	SQ	QM
----	----	----	----

---

Blocos	$r-1$	$SQB = t \sum_j (\bar{Y}_{\cdot j} - \bar{Y}_{..})^2$	QMB
Tratamentos Comuns	$c-1$	$SQC = r \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2$	QMC
Erro	$(r-1)(c-1)$	$SQE = SQT - SQB - SQC$	QME
Total	$rc-1$	$SQT = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	

---

7-Obtenha os erros padrões para os diferentes tipos de comparações de médias, utilizando as seguintes expressões:

$$\text{Diferença entre dois tratamentos comuns} - s_c = \sqrt{2QME / r};$$

$$\text{Diferença entre médias ajustadas de dois tratamentos regulares no mesmo bloco} - s_d = \sqrt{2QME};$$

$$\text{Diferença entre médias ajustadas de dois tratamentos regulares em diferentes blocos} - s_v = \sqrt{2(c+1)QME / c};$$

$$\text{Diferença entre a média ajustada de um tratamento regular e a média de um tratamento comum} - s_{vc} = \sqrt{[(r+1)(c+1)QME] / rc}, \text{ em que:}$$

$c$ : número de diferentes tratamentos comuns por bloco;

$r$ : número de blocos ou número de repetições de um tratamento comum.

Ilustração de Delineamento Aumentado em Blocos Casualizados

Considere a avaliação de 30 novas cultivares (tratamentos regulares) usando 3 variedades comerciais como padrão (tratamentos comuns). Então tem-se:

Número de blocos:

$$r \geq \lceil (10) / (c-1) \rceil + 1, \text{ ou } r = [(10) / (3-1)] + 1 = 6$$

Número de tratamentos regulares por bloco:

$$30 / 6 = 5$$

Os 5 tratamentos regulares são casualizados dentro de cada um dos 6 blocos, separadamente.

Número total de parcelas:

$$(5+3)(6) = 48$$

Croqui de campo:

Blocos					
1	2	3	4	5	6
C1	C1	C1	C1	C1	C1
R14	C2	R18	R09	R02	R29
R26	R04	R27	R06	R21	R07
C2	R15	C2	C2	C3	C2
R17	R30	R25	C3	C2	R01
C3	R03	R28	R20	R10	C3
R22	C3	R05	R11	R08	R12
R13	R24	C3	R23	R16	R19

C: Tratamentos comuns (variedades comerciais padrão)

R: Tratamentos regulares (novos genótipos)

Observe que C1 é colocado sistematicamente primeiro em cada bloco como um marcador.

### Variações nos Delineamentos Aumentados

- 1 – Os tratamentos regulares podem ser considerados de efeitos fixo ou aleatório. No caso de efeitos aleatórios é melhor utilizar a metodologia de modelos mistos para analisar os dados;
- 2 – Pode ser feito o ajuste para duas fontes de heterogeneidade usando linhas e colunas;
- 3 – Os delineamentos modificados utilizam a alocação sistemática dos tratamentos comuns;
- 4 – Os delineamentos de tratamentos fatorial e parcelas subdivididas podem ser utilizados;
- 5 – Os delineamentos aumentados parcialmente repetidos usam os tratamentos regulares ao invés dos tratamentos comuns para estimar o erro e fazer os ajustamentos para os efeitos de heterogeneidade da área experimental.

### Delineamento Experimental Blocos Aumentados (DBA)

O delineamento blocos aumentados, proposto por Federer (1956) especificamente para o melhoramento de plantas, consiste na formação de blocos completos com os tratamentos comuns (testemunhas ou controle) e a inclusão nesses blocos dos tratamentos regulares (tratamentos a serem avaliados), sendo que cada tratamento regular aparece em somente um bloco.

Os tratamentos comuns são então utilizados para a estimativa do erro experimental.

A casualização dos tratamentos nos blocos pode ser feita de duas formas:

- 1 – Sorteio inicial dos tratamentos comuns em cada bloco e em seguida o sorteio dos tratamentos regulares nas parcelas remanescentes de cada bloco. Por exemplo, considere o esquema de casualização com quatro blocos para quatro testemunhas (C1 a C4) e 12 tratamentos regulares a serem avaliados (T1 a T12):

Bloco1	C1	T12	C4	C3	T6	T2	C2
Bloco2	T4	C2	C3	T5	C1	T1	C4
Bloco3	C4	C3	T8	T3	C2	C1	T11
Bloco4	T10	C1	C2	T7	C4	C3	T9

2 – Fazer a casualização do experimento como blocos completos casualizados e, então, substituir os tratamentos não comuns pelos tratamentos regulares de forma ordenada. Considere o mesmo exemplo anterior.

Neste caso, dentro de cada bloco, os tratamentos de 1 a 4 representam as testemunhas (C1 a C4) e os tratamentos de 5 a 7 são substituídos de forma ordenada pelos tratamentos regulares (T1 a T12):

Bloco1	5 (T1)	1 (C1)	3 (C3)	4 (C4)	6 (T2)	2 (C2)	7 (T3)
Bloco2	1 (C1)	5 (T4)	4 (C4)	6 (T5)	7 (T6)	3 (C3)	2 (C2)
Bloco3	6 (T8)	3 (C3)	2 (C2)	7 (T9)	4 (C4)	1 (C1)	5 (T7)
Bloco4	4 (C4)	7 (T12)	1 (C1)	5 (T10)	2 (C2)	3 (C3)	6 (T11)

### Utilização do Delineamento Blocos Aumentados (DBA)

O DBA é útil quando a disponibilidade de material experimental (sementes, por exemplo) ou de área é pequena, ou ainda quando o número de tratamentos é grande, principalmente nas etapas iniciais de uma pesquisa.

O inconveniente do DBA é que a perda de uma parcela contendo um tratamento regular implica na falta de informação sobre este tratamento.

As vantagens do DBA são:

- (a) Os blocos têm seus tamanhos reduzidos e, portanto, com maior possibilidade de homogeneidade dentro dos blocos;
- (b) A casualização de duas ou mais testemunhas nos blocos permite uma estimativa não viesada de seus efeitos;
- (c) Tem-se um número de graus de liberdade adequado para a estimativa do erro experimental.

A limitação do DBA é: apresenta variâncias diferentes para as quatro formas diferentes de comparação entre médias, que são:

- (1) Entre testemunhas;
- (2) Entre tratamentos regulares do mesmo bloco;
- (3) Entre tratamentos regulares de blocos diferentes;
- (4) Entre testemunhas e tratamentos regulares.

Entretanto, se o objetivo for comparar os tratamentos regulares com as testemunhas, esta desvantagem deixa de existir.

O esquema para análise de variância de dados obtidos com base no DBA e que está associado ao seu modelo estatístico é o seguinte:

Fontes de Variação	Graus de Liberdade
Blocos (Não Ajustado)	$b-1$
Tratamentos (Ajustados)	$c+t-1$
Testemunhas (Test)	$c-1$
Tratamentos Regulares (TratReg)	$t-1$
Test x TratReg	1
Erro	$(b-1)(c-1)$
Total	$bc+t-1$

Os cálculos das somas de quadrados são realizados da seguinte forma:

1 – Cálculo de soma de quadrados total e soma de quadrados de blocos (não ajustado) da forma usual, ou seja,

$$SQTotal = \sum_{i,j} Y_{ij}^2 - FC, \text{ em que}$$

$$FC = \frac{\left( \sum_{i,j} Y_{ij} \right)^2}{IJ} \text{ com } i=1,2,\dots,I \text{ e } j=1,2,\dots,J,$$

, onde  $I$ : número de tratamentos (testemunhas mais tratamentos regulares, ou seja,  $(c+t)$ ) e  $J$ : número de blocos.

$$SQBlocos = \frac{1}{I} \left( \sum_j B_j^2 \right) - FC, \text{ onde } B_j \text{ é o total de cada bloco e } I: \text{número de tratamentos } (c+t).$$

2 – Cálculo de soma de quadrados de tratamentos ajustados: (1) primeiro faz-se o ajuste dos valores obtido para os tratamentos regulares, subtraindo-se de cada valor de tratamento obtido o valor dado pela diferença,

$MC_j - MC$ , que é o fator de correção de tratamentos, onde  $MC_j$ : média das testemunhas no bloco  $j$  onde se encontra o tratamento regular a ser ajustado e  $MC$ : média geral das

testemunhas no experimento; (2) denominando-se  $C_i$  o total da testemunha  $i$  e de  $T_k$ , o valor ajustado de cada tratamento regular, a soma de quadrados de tratamentos ajustada é dada por:

$$SQTrat.Ajustado = \sum_i C_i^2 / J + \sum_k T_k^2 - FC$$

3 – Cálculos das somas de quadrados dos desdobramentos da soma de quadrados de tratamentos ajustada:

$$SQTestemunhas = \frac{1}{b} \left( \sum_i C_i^2 \right) - \frac{1}{(b)(c)} \left( \sum_i C_i \right)^2, \text{ onde } C_i: \text{total de cada uma das}$$

testemunhas,  $b$ : número de repetições (blocos) para cada testemunha e  $c$ : número de testemunhas;

$$SQTrat Regulares = \left( \sum_l T_l^2 \right) - \frac{1}{bt} \left( \sum_l T_l \right)^2, \text{ onde } T_l: \text{total ajustado de cada um dos}$$

tratamentos regulares,  $t$ : número de tratamentos regulares e  $b$ : número de blocos;

$$SQTestxTrat Reg = \frac{1}{bc} \left( \sum_i C_i \right)^2 + \frac{1}{bt} \left( \sum_l T_l \right)^2 - \\ - \frac{1}{bc+bt} \left( \sum_i C_i + \sum_l T_l \right)^2$$

### Comparações entre médias de tratamentos

As comparações entre médias de tratamentos regulares e entre médias de testemunhas têm variâncias diferentes, sendo a variância das testemunhas sempre a menor de todas. Os estimadores das variâncias de contrastes utilizados nas comparações entre médias são:

1 – Para comparação entre médias de testemunhas tem-se:

$$\hat{Vár}(\hat{m}_i - \hat{m}_{i'}) = \frac{2QM\ Res}{b}, \text{ sendo } b \text{ o número de blocos e}$$

$QM\ Res$  o quadrado médio do resíduo;

2 – Para comparação entre médias de tratamentos regulares que ocorrem juntos em um mesmo bloco tem-se:

$$\hat{Vár}(\hat{m}_i - \hat{m}_{i'}) = 2QM\ Res;$$

3 – Para comparação entre médias de dois tratamentos regulares que ocorrem em blocos diferentes tem-se:

$$\hat{Vár}(\hat{m}_i - \hat{m}_{i'}) = 2QM\ Res \left( 1 + \frac{1}{c} \right), \text{ sendo } c \text{ o número de testemunhas;}$$

4 – Para comparação entre a média de uma testemunha e a média de um tratamento regular tem-se:

$$\hat{Var}(\hat{m}_i - \hat{m}_{i'}) = QM \operatorname{Res} \left( 1 + \frac{1}{b} + \frac{1}{c} - \frac{1}{bc} \right);$$

5 – Para comparação entre a média de um tratamento regular e a média das testemunhas tem-se:

$$\hat{Var}(\hat{m}_i - \hat{m}_{i'}) = QM \operatorname{Res} \left( 1 + \frac{1}{(c)(b)} \right), \text{ sendo } c \text{ o número de testemunhas e } b \text{ o}$$

número de blocos.

### Exemplo de Aplicação

Considere um experimento de avaliação de linhagens de feijoeiro, delineado em blocos aumentados de Federer, com 10 blocos contendo nove tratamentos (quatro testemunhas e cinco linhagens) cada um. Os dados são de produtividade de grãos em kg/ha (Extraído de Zimmermann, 2004) são apresentados no quadro a seguir:

Quadro. Dados de produtividade de grãos de feijoeiro de um experimento desenvolvido no DBA

Bloco1	Trat.	9	5	8	3	4	1	2	6	7	13673
	Prod.	761	1516	1692	1563	2019	2003	1671	1245	1203	
Bloco2	Trat.	21	20	2	3	4	1	17	18	19	16735
	Prod.	1536	1888	1838	1818	2210	1866	1786	1756	2037	
Bloco3	Trat.	30	1	29	31	2	4	3	32	33	12922
	Prod.	1052	1695	1080	1443	1472	1745	1680	1123	1632	
Bloco4	Trat.	2	41	42	45	43	44	3	1	4	12454
	Prod.	1113	1357	1219	1135	1076	1384	1516	1773	1881	
Bloco5	Trat.	3	57	1	2	4	53	55	54	56	14675
	Prod.	1959	1503	2085	1454	1837	1873	1243	1212	1509	
Bloco6	Trat.	67	65	4	68	1	69	3	2	66	15207
	Prod.	1414	1279	2077	1586	1846	1956	1775	1728	1546	
Bloco7	Trat.	80	2	79	3	78	81	1	77	4	18592
	Prod.	2213	1930	2653	2054	2361	1812	1849	1885	1835	
Bloco8	Trat.	90	91	89	2	3	92	1	4	93	13721
	Prod.	1145	1194	1920	1415	1589	1138	1840	2062	1418	
Bloco9	Trat.	3	1	102	101	105	104	4	103	2	15690
	Prod.	1894	1695	1885	1433	1803	1451	1975	1879	1675	
Bloco10	Trat.	115	114	4	117	2	113	1	3	116	17456
	Prod.	2356	1450	2191	2351	1687	2180	1757	1879	1605	

A soma de quadrados total e a soma de quadrados de blocos são obtidas da forma usual, ou seja:

$$FC = \frac{\left( \sum_{i,j} Y_{ij} \right)^2}{IJ}, \text{ então}$$

$$FC = \frac{(151125)^2}{(9)(10)} = 253764062,5$$

$$SQT_{otal} = \sum_{i,j} Y_{ij}^2 - FC, \text{ logo}$$

$$\begin{aligned} SQT_{otal} &= (761)^2 + \dots + (1605)^2 - 253764062,5 \\ &= 11147428,50 \end{aligned}$$

$$SQB_{locos} = \frac{1}{I} \left( \sum_j B_j^2 \right) - FC, \text{ então,}$$

$$SQB_{locos} = \frac{1}{9} \left[ (13673)^2 + \dots + (17456)^2 \right] -$$

$$-253764062,5 = 4071067,39$$

A soma de quadrados de tratamentos ajustada é calculada da seguinte forma:

- 1) Calcular a média geral de todas as testemunhas para o experimento e para cada bloco;
- 2) Obter o fator de correção dos tratamentos regulares por meio da expressão  $MC_j - MC$ , onde  $MC_j$ : média das testemunhas no bloco  $j$  onde se encontra o tratamento regular a ser ajustado e  $MC$ : média geral das testemunhas no experimento.

Observe que o fator de correção dos dados de tratamentos regulares (linhagens) é a diferença entre a média das testemunhas no bloco em que o tratamento regular se encontra e a média das testemunhas no experimento.

Para o exemplo considerado, o quadro a seguir resume todos os valores calculados:

Quadro. Fator de correção dos tratamentos regulares em função das médias das testemunhas em cada bloco e da média geral das testemunhas

Blocos	Média Geral das Testemunhas no Experimento	Média das Testemunhas no Bloco	Fator de correção dos Tratamentos
1	1798,775	1814,000	15,225
2	1798,775	1933,000	134,225
3	1798,775	1648,000	-150,775
4	1798,775	1570,750	-228,025
5	1798,775	1833,750	34,975
6	1798,775	1856,500	57,725
7	1798,775	1917,000	118,225
8	1798,775	1726,500	-72,275
9	1798,775	1809,750	10,975
10	1798,775	1878,500	79,725

O ajuste dos dados dos tratamentos regulares é feito da seguinte forma:

Por exemplo, a tratamento número 5, no bloco 1, tem produtividade ajustada igual a  $1516 - 15,225 = 1500,775$ ; o tratamento número 30, no bloco 3, tem produtividade ajustada igual a  $1052 - (-150,775) = 1202,775$ .

O quadro a seguir contém os totais das testemunhas e o valor ajustado de cada tratamento.

Quadro. Valores de totais de produtividade das testemunhas e de produtividade ajustada dos tratamentos (linhagens) num experimento em blocos aumentados.

Bloco1	Linhagem	5	6	7	8	9
	Prod.Ajustada	1500,775	1229,775	1187,775	1676,775	745,775
Bloco2	Linhagem	17	18	19	20	21
	Prod.Ajustada	1651,775	1621,775	1902,775	1753,775	1401,775
Bloco3	Linhagem	29	30	31	32	33
	Prod.Ajustada	1230,775	1202,775	1593,775	1273,775	1782,775
Bloco4	Linhagem	41	42	43	44	45
	Prod.Ajustada	1585,025	1447,025	1304,025	1612,025	1363,025
Bloco5	Linhagem	53	54	55	56	57
	Prod.Ajustada	1838,025	1177,025	1208,025	1474,025	1468,025
Bloco6	Linhagem	65	66	67	68	69
	Prod.Ajustada	1221,275	1488,275	1356,275	1528,275	1898,275
Bloco7	Linhagem	77	78	79	80	81
	Prod.Ajustada	1766,775	2242,775	2534,775	2094,775	1693,275
Bloco8	Linhagem	89	90	91	92	93
	Prod.Ajustada	1992,275	1217,275	1266,275	1210,275	1490,275
Bloco9	Linhagem	101	102	103	104	105
	Prod.Ajustada	1422,025	1874,025	1868,025	1440,025	1792,025
Bloco10	Linhagem	113	114	115	116	117
	Prod.Ajustada	2100,275	1370,275	2276,275	1525,275	2271,275
	Testemunhas	1	2	3	4	
	Prod.Ajustada	18409	15983	17727	19832	

A soma de quadrados de tratamentos ajustada é obtida por meio da seguinte expressão:

$$SQTrat.Ajustado = \sum_i C_i^2 / J + \sum_k T_k^2 - FC, \text{ logo,}$$

$$\begin{aligned} SQTrat.Ajustado &= \frac{1}{10} \left[ (18409)^2 + \dots + (19832)^2 \right] + \\ &+ \left[ (1500,775)^2 + \dots + (2271,275)^2 \right] - FC \\ &= 6403550,136 \end{aligned}$$

Os desdobramentos da soma de quadrados de tratamentos ajustada são feitos da seguinte forma:

$$SQTestemunhas = \frac{1}{b} \left( \sum_i C_i^2 \right) - \frac{1}{(b)(c)} \left( \sum_i C_i \right)^2, \text{ então tem-se,}$$

$$SQTestemunhas = \frac{1}{10} \left[ (18409)^2 + \dots + (19832)^2 \right] - \\ - \frac{1}{(10)(4)} (18409 + \dots + 19832)^2 = 766572,275$$

$$SQTrat\text{Re gulares} = \left( \sum_l T_l^2 \right) - \frac{1}{bt} \left( \sum_l T_l \right)^2, \text{ logo tem-se,} \\ SQTrat.\text{Re gulares} = \left[ (1500,775)^2 + \dots + (1525,275)^2 \right] \\ - \frac{1}{(10)(5)} (79174)^2 = 4606934,816$$

$$SQTestxTrat\text{Re g} = \frac{1}{bc} \left( \sum_i C_i \right)^2 + \frac{1}{bt} \left( \sum_l T_l \right)^2, \text{ então tem-se,} \\ - \frac{1}{bc+bt} \left( \sum_i C_i + \sum_l T_l \right)^2$$

$$SQTest.xTrat.\text{Re g} = \frac{1}{(10)(4)} (71951)^2 +$$

$$- \frac{1}{(10)(5)} (79174)^2 - \frac{1}{40+50} (151125)^2$$

$$= 1030043,045$$

$$SQ\text{Residuo} = SQTotal - \\ - (SQBlocs + SQTrat.\text{Ajustado}), \text{ logo tem-se,}$$

$$SQ\text{Residuo} = 11147428,50 -$$

$$- (4071067,389 + 6403550,136)$$

$$= 672810,97$$

O resumo da análise de variância é apresentado no quadro a seguir:

Quadro. Resumo da análise de variância de produtividade de grãos (kg/ha) para um experimento em blocos aumentados, com quatro testemunhas e 50 linhagens de feijoeiro

FV	GL	QM	F (Valor p)
Blocos (não ajust)	9	452340,821	18,15(<,0001)
Tratamentos(ajust)	53	120821,701	4,85(<,0001)
Testemunhas	3	255524,092	10,25(<,0001)
Trat. Regulares	49	94019,078	3,77(<,0002)
Test.xTrat. Reg.	1	1030043,045	41,34(<,0001)
Resíduo	27	24918,920	-
Total	89	-	-

O valor da média geral é obtido por meio da expressão  $MG = \frac{TG}{IJ}$ , sendo  $TG$ : total geral;  $I$ : número de tratamentos ( $c+t$ );  $J$ : número de blocos. Então tem-se,

$$MG = \frac{TG}{IJ} = \frac{151125}{(9)(10)} = 1679,167.$$

O valor do coeficiente de variação experimental é obtido da seguinte forma:

$$CV\% = \sqrt{\frac{QM \text{ Residuo}}{MG}} (100), \text{ ou seja,}$$

$$CV\% = \sqrt{\frac{24918,92}{1679,167}} (100) = 9,40$$

Considere como exemplo, o contraste de médias envolvendo um tratamento regular (linhagem) e a média das testemunhas, que apresenta estimativa de variância obtida da seguinte forma:

$$\hat{Var}(\hat{m}_i - \hat{m}_{i'}) = QM \text{ Res} \left( 1 + \frac{1}{(c)(b)} \right), \text{ então tem-se,}$$

$$\hat{Var}(\hat{m}_i - \hat{m}_{i'}) = 24918,92 \left( 1 + \frac{1}{(4)(10)} \right)$$

$$= 25541,893$$

Considere a diferença mínima significativa (DMS), baseada no teste  $t$  a 5% de probabilidade, que é obtida como:

$$DMS = [t_{0,05}(27gl)] \hat{s}(\hat{m}_i - \hat{m}_{i'}), \text{ sendo}$$

$$\hat{s}(\hat{m}_i - \hat{m}_{i'}) = \sqrt{\text{Var}(\hat{m}_i - \hat{m}_{i'})}. \text{ Então tem-se,}$$

$$DMS = (2,052) \left( \sqrt{25541,893} \right) = 327,95$$

Portanto, qualquer diferença entre uma média de tratamento regular e a média das testemunhas que for maior que **DMS = 327,95** é considerada significativa.

# CAPITULO 12

## Experimentos com Medidas Repetidas

### Introdução

- O termo medidas repetidas refere-se a múltiplas respostas tomadas em seqüência sobre a mesma unidade experimental;
- O motivo para análise como medidas repetidas é a suspeita de que os efeitos de tratamentos se alteram ao longo da seqüência avaliada;
- Geralmente as respostas são tomadas ao longo do tempo, mas podem ser tomadas também numa seqüência espacial;
- Experimentos com medidas repetidas têm objetivos comuns a qualquer experimento fatorial e são semelhantes a parcelas subdivididas;
- Uma característica de experimentos com medidas repetidas que requer atenção especial na análise de dados é o padrão de correlação entre as respostas obtidas sobre uma mesma unidade experimental;
- A existência de correlação entre respostas seqüenciais (não casualizadas) gera a correlação de erros, a qual, se ignorada, pode ou não distorcer as inferências, dependendo do grau de homogeneidade das variâncias e covariâncias dos dados nas diferentes respostas;
- Métodos estatísticos especiais, que modelam a estrutura de covariância, podem ser necessários para uma análise efetiva e eficiente de experimentos com medidas repetidas;
- Nos experimentos com medidas repetidas no tempo, a interação tempo x tratamento é o aspecto mais importante, e quando significativa, indica uma tendência não paralela da variável resposta em relação ao tempo para diferentes níveis do tratamento;

### Abordagens para Análise de Dados com Medidas Repetidas

#### 1. Abordagem por Análise de Variância Univariada

- Historicamente, a análise de variância univariada (ANOVA) é o método comumente aplicado aos dados com medidas repetidas. Nesta análise os dados são tratados como se fossem provenientes de um delineamento parcelas subdivididas, em que os tratamentos são as parcelas e os tempos as subparcelas. Esta abordagem é também referida como análise de parcelas subdivididas no tempo;
- Se as medições tiverem variâncias iguais em todos os tempos ou espaços, e se pares de medições sobre a mesma unidade experimental são igualmente

correlacionados, indiferente da defasagem de tempo ou espaço entre as medições (simetria composta), então a análise de variância univariada é válida, e, de fato constitui um ótimo método de análise;

- A condição requerida para que a análise univariada seja válida é que a matriz variância-covariância do fator tempo ou espaço tenha uma forma estrutural conhecida como Tipo H-F (ou esfericidade), que pressupõe a igualdade de todas as variâncias das diferenças entre pares de medições, e que é estatisticamente menos restritiva que variâncias e covariâncias iguais (Huynh e Feldt, 1970);
- O teste estatístico para a estrutura de covariância Tipo H-F é conhecido como *teste de esfericidade de Mauchly*, realizado por meio da opção *print* no comando *repeated* do *proc glm* do SAS. Quando a hipótese de esfericidade não é rejeitada, então a análise univariada é mais poderosa que a análise multivariada;
- Quando o teste de esfericidade é significativo (a hipótese de esfericidade é rejeitada), pode-se ainda testar a significância dos efeitos dentro de sujeitos (indivíduos, objetos, parcelas) ajustando-se os próprios testes univariados usando dois tipos de ajustes para os graus de liberdade, disponíveis com o comando *repeated* no *proc glm* do SAS: G-G (Greenhouse e Geisser, 1959) e o menos conservativo H-F (Huynh e Feldt, 1976).

## 2. Abordagem por Análise de Variância Multivariada

- Outra estratégia de análise, quando o teste de esfericidade é significativo, envolve quatro testes da análise de variância multivariada (MANOVA) do *proc glm* do SAS: Lambda de Wilks, Traço de Pillai, Traço de Hotelling-Lawley e Raiz Máxima de Roy (Johnson e Wichern, 1998). Todos estes testes são baseados numa matriz variância-covariância dentro de sujeitos completamente sem estrutura; assume uma correlação única (diferente de zero) entre cada par de medições;
- Algumas críticas à abordagem multivariada: a) a MANOVA requer a estimativa de  $k(k+1)/2$  parâmetros da matriz variância-covariância; b) para um modelo mais moderado, os dados em cada ponto no tempo ou espaço terão variância  $\sigma^2$  e as correlações entre os dados em dois pontos no tempo ou espaço dependem somente da diferença entre os tempos ou espaços; c) o ajuste de um modelo super-parametrizado pode levar a estimativas imprecisas dos parâmetros do modelo, menor poder para detectar efeitos e previsões viesadas;
- As estratégias de análises anteriores consideram os seguintes aspectos: a) os dados dentro de sujeitos (objetos, indivíduos, parcelas) devem ser balanceados; b) deve-

se escolher uma modelagem de médias em termos de efeitos fixos entre objetos (são aqueles cujos níveis permanecem constantes dentro de objetos) e dentro de objetos (são aqueles cujos níveis variam dentro de objetos); c) as inferências sobre os efeitos fixos fornecem respostas para questões de pesquisa por meio de testes de hipóteses lineares (teste F Tipo I e Tipo III, comandos *test*, *contrast*, *estimates*, *means* e *lsmeans* do SAS).

1. Experimentos com um fator (delineamento experimental inteiramente casualizados) e medidas repetidas em todas as unidades experimentais

Considere a seguinte estrutura de dados:

		Tratamentos Sequenciais (Medidas Repetidas)				
		1	2	3	...	t
EU	1	$y_{11}$	$y_{12}$	$y_{13}$	...	$y_{1t}$
	2	$y_{21}$	$y_{22}$	$y_{23}$	...	$y_{2t}$
	:	:	:	...		:
	n	$y_{n1}$	$y_{n2}$	$y_{n3}$	...	$y_{nt}$

Neste caso tem-se uma amostra de  $n$  unidades experimentais submetidas a  $t$  tratamentos sequenciais. Então, tem-se  $t$  respostas (observações) para cada unidade experimental.

O modelo associado à análise deste conjunto de dados é o seguinte:

$$y_{ij} = \mu + \rho_i + \alpha_j + (\rho\alpha)_{ij} + \varepsilon_{ij}, \quad i=1,2,\dots,n; \quad j=1,2,\dots,t; \text{ em que:}$$

$\rho_i$ : efeito da unidade experimental de ordem  $i$ ;

$\alpha_j$ : efeito do tratamento sequencial  $j$ ;

$\varepsilon_{ij}$ : erro experimental;

$(\rho\alpha)_{ij}$ : interação da unidade experimental com o tratamento sequencial;

$\varepsilon_{ij} \square N(0, \sigma^2)$ , independentes;

$$\begin{aligned}
\rho_i &\square N\left(0, \sigma_{\rho}^2\right), \text{ independentes; } \\
(\rho\alpha)_{ij} &\square N\left(0, \frac{t-1}{t} \sigma_{\rho\alpha}^2\right); \\
\text{cov}\left[(\rho\alpha)_{ij}, (\rho\alpha)_{ij'}\right] &= -\frac{1}{t} \sigma_{\rho\alpha}^2, \text{ para } j \neq j'; \\
\text{cov}\left[(\rho\alpha)_{ij}, (\rho\alpha)_{i'j}\right] &= 0, \text{ para } i \neq i'; \\
\sum_i (\rho\alpha)_{ij} &= 0; \sum_j (\rho\alpha)_{ij} = 0; \sum_j \alpha_j = 0; \\
\varepsilon_{ij}, \rho_i \text{ e } (\rho\alpha)_{ij} &\text{ são independentes dois a dois.}
\end{aligned}$$

Para este modelo tem-se que:

$$\begin{aligned}
E(y_{ij}) &= \mu + \alpha_j; \quad \text{var}(y_{ij}) = \sigma_Y^2 = \sigma_{\rho}^2 + \frac{t-1}{t} \sigma_{\rho\alpha}^2 + \sigma^2; \\
\text{cov}(y_{ij}, y_{ij'}) &= \sigma_{\rho}^2 - \frac{1}{t} \sigma_{\rho\alpha}^2, \text{ para } j \neq j'; \\
\text{cov}(y_{ij}, y_{i'j}) &= 0, \text{ para } i \neq i'.
\end{aligned}$$

Observe que:

- a) Duas observações pertencentes a unidades experimentais diferentes são independentes;
- b) Duas observações pertencentes à mesma unidade experimental são correlacionadas, sendo que a correlação (que pode ser positiva ou negativa) é a mesma para todas as unidades experimentais;
- c) O coeficiente de correlação entre duas observações quaisquer pertencentes à mesma

$$\text{unidade experimental, denotada por } w, \text{ é dado por } w = \frac{\sigma_{\rho}^2 - \frac{1}{t} \sigma_{\rho\alpha}^2}{\sigma_Y^2}.$$

As somas de quadrados e os graus de liberdade são obtidos como:

SQ	Expressão SQ	GL
SQUE	$t \sum_i (\bar{y}_{i\cdot} - \bar{y}_{..})^2$	$n-1$
SQTR	$n \sum_j (\bar{y}_{\cdot j} - \bar{y}_{..})^2$	$t-1$
SQUE*TR	$\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{..})^2$	$(n-1)(t-1)$
SQT	$\sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{..})^2$	$nt-1$

SQUE: Soma de quadrados de unidade experimental;

SQTR: Soma de quadrados de tratamentos sequenciais;

SQUE\*TR: Soma de quadrados de interação unidade experimental x tratamento sequencial;

SQT: Soma de quadrados de total.

Os valores esperados dos quadrados médios são dados por:

$$E(QMUE) = \sigma^2 + t\sigma_\rho^2;$$

$$E(QMTR) = \sigma^2 + \sigma_\rho^2 \alpha + n \frac{\sum \alpha_j^2}{t-1};$$

$$E(QMUE*TR) = \sigma^2 + \sigma_\rho^2 \alpha.$$

Não há um teste exato para a hipótese  $H_{02}: \sigma_\rho^2 = 0$ . A estatística para testar a hipótese

$H_{01}: \alpha_1 = \alpha_2 = \dots = \alpha_t = 0$  é dada por:

$F = \frac{QMTR}{QMUE*TR}$ , que sob a hipótese  $H_{01}$ , tem distribuição F central com  $(t-1)$  e  $(n-1)(t-1)$  graus de liberdade.

Os efeitos dos tratamentos podem ser analisados da seguinte forma:

Se  $H_{01}:\alpha_1=\alpha_2=\dots=\alpha_t=0$  for rejeitada, em geral, o interesse concentra-se na análise desses efeitos por meio da estimação de um ou mais contrastes entre as respostas médias  $\mu_{\cdot j}$ ,  $j=1,2,\dots,t$ , sob os tratamentos.

Seja  $C = \sum_j c_j \mu_{\cdot j}$ ,  $\sum_j c_j = 0$ , um contraste entre as respostas médias  $\mu_{\cdot j}$  sob os tratamentos. Como  $\hat{\mu}_{\cdot j} = \bar{y}_{\cdot j}$ , obtém-se  $\hat{C} = \sum_j c_j \bar{y}_{\cdot j}$ . Verifica-se também que

$$\text{var}(\hat{C}) = \sum_j c_j^2 \frac{(\sigma^2 + \sigma_{\rho\alpha}^2)}{n}. \text{ Como } QMUE*TR \text{ é um estimador não viesado de } \sigma^2 + \sigma_{\rho\alpha}^2, \text{ obtém-se } \text{var}(\hat{C}) = \frac{QMUE*TR}{n} \sum_j c_j^2.$$

Os limites de confiança para  $C$ , obtidos com um coeficiente de confiança  $\gamma=1-\alpha$ , são dados por  $\hat{C} \pm A \sqrt{\text{var}(\hat{C})}$  em que  $A=t_{[\alpha/2;(n-1)(t-1)]}$  é o percentil de ordem  $\alpha/2$  obtido da distribuição t-Student com  $(n-1)(t-1)$  graus de liberdade.

Para um conjunto de contrastes em que é fixado o coeficiente de confiança global em  $\gamma=1-\alpha$ , podem ser utilizados os procedimentos de Tukey, Scheffé ou Bonferroni, para os quais os valores de  $A$  são obtidos como:

a) Para Tukey,  $A = \frac{1}{\sqrt{2}} q_{[\alpha;t,(n-1)(t-1)]}'$  com

$q_{[\alpha;t,(n-1)(t-1)]}$  igual ao percentil de ordem  $\alpha$  obtido da distribuição amplitude estudentizada com  $t$  e  $(n-1)(t-1)$  graus de liberdade;

b) Para Scheffé,  $A = \sqrt{(t-1) F_{[\alpha;t-1,(n-1)(t-1)]}}$ , com

$F_{[\alpha;t-1,(n-1)(t-1)]}$  igual ao percentil de ordem  $\alpha$  obtido da distribuição  $F$  com  $t-1$  e  $(n-1)(t-1)$  graus de liberdade;

- c) Para Bonferroni,  $A = t_{[\alpha/(2g);(n-1)(t-1)]}$ , sendo  $g$  o número de contrastes a serem estimados e  $t_{[\alpha/(2g);(n-1)(t-1)]}$  igual ao percentil de ordem  $\alpha/(2g)$  obtido da distribuição t-Student com  $(n-1)(t-1)$  graus de liberdade.
2. Experimentos com dois fatores (delineamento experimental blocos casualizados) e medidas repetidas em um fator

Considere a seguinte estrutura de dados:

Grupo	EU	Tratamentos Sequenciais (Medidas Repetidas)			
		1	2	...	$t$
$A_1$	1	$y_{111}$	$y_{121}$	...	$y_{1t1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$N$	$y_{11n}$	$y_{12n}$	...	$y_{1tn}$
$A_a$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
	1	$y_{a11}$	$y_{a21}$	...	$y_{at1}$
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
	$N$	$y_{a1n}$	$y_{a2n}$	...	$y_{atn}$

Neste caso tem-se um fator A com  $a$  níveis, caracterizando  $a$  grupos,  $A_1, A_2, \dots, A_a$ . Suponha que a cada grupo estejam alocadas  $n$  unidades experimentais e que cada uma delas receba  $t$  tratamentos sequenciais. Estes tratamentos caracterizam os  $t$  níveis de um fator B.

O modelo associado à análise deste conjunto de dados é o seguinte:

$$y_{ijk} = \mu + \alpha_i + \rho_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\beta\rho)_{jk(i)} + \varepsilon_{ijk},$$

$i=1,2,\dots,a; j=1,2,\dots,t; k=1,2,\dots,n$ . Sendo que:

$\alpha_i$ : efeito fixo do grupo  $i$ ;  $\beta_j$ : efeito fixo do tratamento sequencial  $j$ ;

$\rho_{k(i)}$ : efeito aleatório da unidade experimental  $k$  dentro do grupo  $i$ ;

$(\alpha\beta)_{ij}$ : efeito da interação entre o grupo  $i$  e o tratamento sequencial  $j$ ;

$(\beta\rho)_{jk(i)}$ : efeito da interação entre o tratamento sequencial  $j$  e a unidade experimental  $k$  dentro do grupo  $i$ ;

$\varepsilon_{ijk}$ : erro experimental;

$\varepsilon_{ijk} \sim N(0, \sigma^2)$ , independentes;

$\rho_{k(i)} \sim N(0, \sigma_\rho^2)$ , independentes;

$(\beta\rho)_{jk(i)} \sim N\left(0, \frac{t-1}{t}\sigma_{\beta\rho}^2\right)$ ;

$\text{cov}\left[(\beta\rho)_{jk(i)}, (\beta\rho)_{j'k(i)}\right] = -\frac{1}{t}\sigma_{\beta\rho}^2$ , para  $j \neq j'$ ;

$\text{cov}\left[(\beta\rho)_{jk(i)}, (\beta\rho)_{jk(i')}\right] = 0$ , para  $i \neq i'$ ;

$\sum_j (\beta\rho)_{jk(i)} = 0$ ,  $k=1,2,\dots,n$ ;  $i=1,2,\dots,a$ ;

$\sum_k (\beta\rho)_{jk(i)} = 0$ ,  $j=1,2,\dots,t$ ;  $i=1,2,\dots,a$ ;

$\sum_i \alpha_i = 0$ ,  $\sum_j \beta_j = 0$ ,  $\sum_i (\alpha\beta)_{ij} = 0$  e  $\sum_j (\alpha\beta)_{ij} = 0$ ;

$\varepsilon_{ijk}$ ,  $\rho_{k(i)}$  e  $(\beta\rho)_{jk(i)}$  são independentes dois a dois.

As somas de quadrados e os graus de liberdade para a análise de variância são obtidos da seguinte forma:

SQ	Expressão SQ	GL
SQG	$tn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$	$a-1$
SQUE(G)	$t \sum_i \sum_k (\bar{y}_{i.k} - \bar{y}_{i..})^2$	$a(n-1)$
SQTR	$an \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$	$t-1$
SQG*TR	$n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$(a-1)(t-1)$
SQTR*EU(G)	$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.} - \bar{y}_{i.k} + \bar{y}_{i..})^2$	$a(t-1)(n-1)$
SQT	$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$	$atn-1$

Os valores esperados dos quadrados médios são obtidos com base nas expressões seguintes:

$$E(QMG) = \sigma^2 + t\sigma_\rho^2 + nt \frac{\sum \alpha_i^2}{a-1}$$

$$E[QMUE(G)] = \sigma^2 + t\sigma_\rho^2$$

$$E(QMTR) = \sigma^2 + \sigma_{\rho\alpha}^2 + na \frac{\sum \beta_j^2}{t-1}$$

$$E(QMG*TR) = \sigma^2 + \sigma_{\rho\alpha}^2 + n \frac{\sum \sum (\alpha\beta)_{ij}^2}{(a-1)(t-1)}$$

$$E[QMTR*UE(G)] = \sigma^2 + \sigma_{\rho\alpha}^2.$$

A estatística de teste para a hipótese  $H_{01}: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0$  é dada por

$F_1 = \frac{QMG}{QMUE(G)}$ , que, sob a hipótese  $H_{01}$ , tem distribuição  $F$  central com  $(a-1)$  e

$a(n-1)$  graus de liberdade. Observe que  $F_1$  tem distribuição  $F$  exata mesmo quando a matriz de variâncias-covariâncias  $\Sigma$  para o vetor de observações de uma unidade experimental qualquer não satisfaz a condição de esfericidade.

A estatística de teste para a hipótese  $H_{02}:(\alpha\beta)_{11}=(\alpha\beta)_{12}=\dots=(\alpha\beta)_{at}=0$  é

dada por  $F_2 = \frac{QMG*TR}{QMTR*UE(G)}$ , que, sob a hipótese  $H_{02}$ , tem distribuição  $F$  central

com  $(a-1)(t-1)$  e  $a(t-1)(n-1)$  graus de liberdade.

A estatística de teste para a hipótese  $H_{03}:\beta_1=\beta_2=\dots=\beta_t=0$  é dada por

$F_3 = \frac{QMTR}{QMTR*UE(G)}$ , que, sob a hipótese  $H_{03}$ , tem distribuição  $F$  central com

$t-1$  e  $a(t-1)(n-1)$  graus de liberdade.

Se as matrizes de variâncias-covariâncias  $\Sigma$ , para o vetor de observações de uma unidade experimental qualquer não satisfizer a condição de esfericidade, as estatísticas  $F_2$  e  $F_3$  não

têm distribuições  $F$  exatas, mas podem ser aproximadas, respectivamente, pelas distribuições  $F_{2[\in(a-1)(t-1), \in a(t-1)(n-1)]}$  e  $F_{3[\in(t-1), \in a(t-1)(n-1)]}$ .

A análise dos efeitos dos fatores deve ser abordada da seguinte forma:

1) Não há efeito de interação entre os fatores A e B

a) Comparar as médias da variável resposta sob os níveis de A, ou seja,  $\mu_{i..}$  com  $\mu_{i'..}$

para  $i \neq i'$ . Um intervalo de confiança para  $\mu_{i..} - \mu_{i'..}$ , com coeficiente de confiança

$\gamma = 1 - \alpha$  é dado por  $\left[ \bar{y}_{i..} - \bar{y}_{i'..} \right] \pm A \sqrt{\frac{2}{nt} QMUE(G)}$ . No caso de ser

utilizado o método de Bonferroni, tem-se  $A=t_{[\alpha/2g;a(n-1)]}$ , sendo  $t_{[\alpha/2g;a(n-1)]}$  o percentil de ordem  $\alpha/2g$  da distribuição t-Student com  $a(n-1)$  graus de liberdade.

- b) Comparar as médias da variável resposta sob os níveis do fator B, ou seja,  $\mu_{.j}$  com  $\mu_{.j'}$ ,  $j \neq j'$ . Um intervalo de confiança para  $\mu_{.j} - \mu_{.j'}$ , com coeficiente de confiança  $\gamma=1-\alpha$  é dado por  $\left[ \bar{y}_{.j} - \bar{y}_{.j'} \right] \pm A \sqrt{\frac{2}{an} QMTR * UE(G)}$ . No caso de se utilizar o método de Bonferroni tem-se  $A=t_{[\alpha/2g;a(n-1)(t-1)]}$ , sendo  $t_{[\alpha/2g;a(n-1)(t-1)]}$  o percentil de ordem  $\alpha/2g$  da distribuição t-Student com  $a(n-1)(t-1)$  graus de liberdade.

## 2. Há efeito de interação entre os fatores A e B

- a) Comparar as médias da variável resposta sob os níveis de A para cada nível de B, ou seja,  $\mu_{ij}$  com  $\mu_{i'j}$ , para  $i < i' = 1, 2, \dots, a$  e  $j = 1, 2, \dots, t$ . Corresponde a realizar  $b$  ANOVAs com um fator.
- b) Comparar as médias da variável resposta sob os níveis de B para cada nível de A, ou seja,  $\mu_{ij}$  com  $\mu_{ij'}$ , para  $j < j' = 1, 2, \dots, t$  e  $i = 1, 2, \dots, a$ . Corresponde a realizar  $a$  ANOVAs com um fator. Um intervalo de confiança para  $\mu_{ij} - \mu_{ij'}$ , com coeficiente de confiança  $\gamma=1-\alpha$  é dado por  $\left[ \bar{y}_{ij} - \bar{y}_{ij'} \right] \pm A \sqrt{\frac{2}{n} QMTR * UE(G)}$ . No caso de ser utilizado o método de Bonferroni, tem-se  $A=t_{[\alpha/2g;a(n-1)(t-1)]}$ , sendo  $t_{[\alpha/2g;a(n-1)(t-1)]}$  o percentil de ordem  $\alpha/2g$  da distribuição t-Student com  $a(n-1)(t-1)$  graus de liberdade.

## Análise de Dados de Experimentos com Medidas Repetidas usando o SAS

1. Exemplo de um experimento com medidas repetidas analisado por abordagem univariada clássica-como parcelas subdivididas

Considere um experimento de avaliação de produtividade (ton/acre) de quatro genótipos de alfafa, testados com cinco repetições num Delineamento Experimental Inteiramente Casualizado (DIC), com medidas repetidas (quatro cortes – set/74, jun/75, ago/75 e set/75). Aqui as medidas repetidas são modeladas como subparcelas. As rotinas SAS para realização desta análise são as seguintes:

```
Options nodate nonumber ls=78;
Data mruniv;
Input genot bloco corte parcela prod;
Datalines;
1   1   1   1   2.8019
1   1   2   1   3.7309
1   1   3   1   3.0986
1   1   4   1   2.5096
1   2   1   2   2.9661
1   2   2   2   4.4355
1   2   3   2   3.1061
1   2   4   2   2.5729
1   3   1   3   2.4323
1   3   2   3   4.3231
1   3   3   3   2.8103
1   3   4   3   2.0797
1   4   1   4   2.9351
1   4   2   4   3.9971
1   4   3   4   2.7797
1   4   4   4   2.4403
1   5   1   5   2.4228
1   5   2   5   3.8566
1   5   3   5   3.2491
1   5   4   5   2.3413
2   1   1   6   2.7621
2   1   2   6   5.4053
2   1   3   6   3.8243
2   1   4   6   2.7299
2   2   1   7   3.0964
2   2   2   7   3.9068
2   2   3   7   3.2623
2   2   4   7   2.5861
```

2	3	1	8	3.0992
2	3	2	8	4.0886
2	3	3	8	3.1315
2	3	4	8	2.6032
2	4	1	9	2.6526
2	4	2	9	5.4288
2	4	3	9	2.7089
2	4	4	9	2.3016
2	5	1	10	2.6367
2	5	2	10	3.7746
2	5	3	10	3.0973
2	5	4	10	2.3008
3	1	1	11	2.2915
3	1	2	11	3.8114
3	1	3	11	2.9258
3	1	4	11	2.3986
3	2	1	12	2.5403
3	2	2	12	3.8272
3	2	3	12	2.8673
3	2	4	12	2.1629
3	3	1	13	2.4119
3	3	2	13	4.0832
3	3	3	13	3.0391
3	3	4	13	2.0708
3	4	1	14	2.3042
3	4	2	14	3.2785
3	4	3	14	2.7271
3	4	4	14	2.0493
3	5	1	15	2.3694
3	5	2	15	3.4484
3	5	3	15	2.5056
3	5	4	15	2.0898
4	1	1	16	2.5663
4	1	2	16	4.9607
4	1	3	16	2.8173
4	1	4	16	2.0575
4	2	1	17	2.3163
4	2	2	17	3.9663
4	2	3	17	2.9146
4	2	4	17	2.1576
4	3	1	18	2.6583
4	3	2	18	3.7186
4	3	3	18	2.9292
4	3	4	18	2.1568

```

4      4      1      19     2.4788
4      4      2      19     3.9205
4      4      3      19     3.0619
4      4      4      19     2.3582
4      5      1      20     2.2359
4      5      2      20     4.0299
4      5      3      20     2.8528
4      5      4      20     1.8574
;
Title "Modelo 1-Abordagem por Análise de Variância Univariada Clássica – Parcelas Subdivididas";
Proc glm data=mruniv;
Class genot bloco corte parcela;
Model prod=bloco genot bloco*genot corte genot*corte;
Test h=genot e=bloco*genot;
Estimate 'med_geral' intercept 1;
Lsmeans genot/stderr cl out=med11;
Run; quit;

Title "Modelo 2-Abordagem por Análise de Variância Univariada – Parcelas Subdivididas";
Proc mixed data=mruniv;
Class genot bloco corte parcela;
Model prod=bloco genot corte genot*corte;
Random bloco*genot;
Estimate 'med geral' intercept 1;
Lsmeans genot/cl;
Make 'estimates' out=est1;
Make 'dimensions' out=npars1 (rename=(value=nparm1));
Make 'fitstatistics' out=crit1 (rename=(value=value1));
Make 'lsmeans' out=med12;
Run; quit;

```

Os resultados obtidos com as análises anteriores são os seguintes:

Modelo 1-Abordagem por Análise de Variância Univariada Clássica - Parcelas Subdivididas

The GLM Procedure

Dependent Variable: prod

		Sum of			
Source	DF	Squares	Mean Square	F Value	Pr > F

Model 31 42.88578030 1.38341227 14.46 <.0001

Error 48 4.59263308 0.09567986

Total 79 47.47841337

R-Square Coeff Var Root MSE prod Mean

0.903269 10.33211 0.309322 2.993790

Source DF Type I SS Mean Square F Value Pr > F

bloco 4 1.04531324 0.26132831 2.73 0.0397  
genot 3 2.84073874 0.94691291 9.90 <.0001  
genot\*bloco 12 1.00336350 0.08361362 0.87 0.5779  
corte 3 37.44862545 12.48287515 130.47 <.0001  
genot\*corte 9 0.54773937 0.06085993 0.64 0.7606

Source DF Type III SS Mean Square F Value Pr > F

bloco 4 1.04531324 0.26132831 2.73 0.0397  
genot 3 2.84073874 0.94691291 9.90 <.0001  
genot\*bloco 12 1.00336350 0.08361362 0.87 0.5779  
corte 3 37.44862545 12.48287515 130.47 <.0001  
genot\*corte 9 0.54773937 0.06085993 0.64 0.7606

Tests of Hypotheses Using the Type III MS for genot\*bloco as an Error Term

Source DF Type III SS Mean Square F Value Pr > F

genot 3 2.84073874 0.94691291 11.32 0.0008

Standard

Parameter Estimate Error t Value Pr > |t|

med\_geral 2.99379000 0.03458321 86.57 <.0001

### Least Squares Means

		Standard	
genot	prod LSMEAN	Error	Pr >  t
1	3.04445000	0.06916641	<.0001
2	3.26985000	0.06916641	<.0001
3	2.76011500	0.06916641	<.0001
4	2.90074500	0.06916641	<.0001

genot prod LSMEAN 95% Confidence Limits

1	3.044450	2.905382	3.183518
2	3.269850	3.130782	3.408918
3	2.760115	2.621047	2.899183
4	2.900745	2.761677	3.039813

### Modelo 2-Abordagem por Análise de Variância Univariada - Parcelas Subdivididas

#### The Mixed Procedure

##### Iteration History

Iteration Evaluations -2 Res Log Like Criterion

0	1	63.16696049	
1	1	63.16696049	0.00000000

Convergence criteria met.

Covariance Parameter  
Estimates

Cov Parm Estimate

genot*bloco	0
Residual	0.09327

##### Fit Statistics

Res Log Likelihood -31.6

Akaike's Information Criterion	-32.6
Schwarz's Bayesian Criterion	-33.1
-2 Res Log Likelihood	63.2

### Type 3 Tests of Fixed Effects

Effect	Num DF		Den DF	F Value	Pr > F
bloco	4	12		2.80	0.0745
genot	3	12		10.15	0.0013
corte	3	48		133.84	<.0001
genot*corte	9	48		0.65	0.7466

### Estimates

Label	Standard				
	Estimate	Error	DF	t Value	Pr >  t
med geral	2.9938	0.03414	12	87.68	<.0001

### Least Squares Means

Effect	genot	Standard					
		Estimate	Error	DF	t Value	Pr >  t	Alpha
	1	3.0445	0.06829	12	44.58	<.0001	0.05
	2	3.2699	0.06829	12	47.88	<.0001	0.05
	3	2.7601	0.06829	12	40.42	<.0001	0.05
	4	2.9007	0.06829	12	42.48	<.0001	0.05

### Least Squares Means

Effect	genot	Lower	Upper
	1	2.8957	3.1932
	2	3.1211	3.4186
	3	2.6113	2.9089
	4	2.7520	3.0495

## 2. Exemplo de experimento com medidas repetidas analisado por abordagem univariada usando o comando *repeated* do SAS

No exemplo anterior, o modelo de parcelas subdivididas especifica que pares de observações (subparcelas) dentro da mesma unidade principal (parcela) são igualmente correlacionados. Com medidas repetidas, pares de observações sobre a mesma unidade não são necessariamente igualmente correlacionados. Medidas próximas no tempo ou espaço podem ser mais altamente correlacionadas do que medidas mais afastadas no tempo ou espaço.

Uma vez que correlação desigual entre medidas repetidas é ignorada na análise por parcelas subdivididas, os testes derivados dessa abordagem podem não serem válidos. Além disso, nesta abordagem a subparcela, que consiste de dados tomados na parcela inteira, difere da subparcela do delineamento de tratamentos parcelas subdivididas, no qual os dados são tomados em partes designadas das parcelas, ou seja, nas subparcelas.

Aqui é considerada a dependência entre as observações, a qual é usada para ajustar os graus de liberdade na análise de variância e fornecer um teste aproximado para as medidas repetidas. Estes cálculos mais precisos podem ser obtidos usando o comando *repeated* do SAS que é colocado especificamente para esta situação.

No exemplo considerado, os quatro cortes são tratados como quatro variáveis respostas e o genótipo é a variável classificação (tratamento). A sintaxe básica do comando é *repeated factor / options*; aqui *factor* é o nome do fator dentro de sujeitos (*subjects*) e não deve ter sido previamente definido na entrada de dados (*input*).

As rotinas SAS para realização desta análise são as seguintes:

```
Options nodate nonumber ls=78;
Data mrmultiv;
Input genot bloco parcela corte1 corte2 corte3 corte4;
Datalines;
1      1      1      2.8019 3.7309 3.0986 2.5096
1      2      2      2.9661 4.4355 3.1061 2.5729
1      3      3      2.4323 4.3231 2.8103 2.0797
1      4      4      2.9351 3.9971 2.7797 2.4403
1      5      5      2.4228 3.8566 3.2491 2.3413
2      1      6      2.7621 5.4053 3.8243 2.7299
2      2      7      3.0964 3.9068 3.2623 2.5861
2      3      8      3.0992 4.0886 3.1315 2.6032
2      4      9      2.6526 5.4288 2.7089 2.3016
2      5     10      2.6367 3.7746 3.0973 2.3008
3      1     11      2.2915 3.8114 2.9258 2.3986
3      2     12      2.5403 3.8272 2.8673 2.1629
```

```

3   3   13   2.4119 4.0832 3.0391 2.0708
3   4   14   2.3042 3.2785 2.7271 2.0493
3   5   15   2.3694 3.4484 2.5056 2.0898
4   1   16   2.5663 4.9607 2.8173 2.0575
4   2   17   2.3163 3.9663 2.9146 2.1576
4   3   18   2.6583 3.7186 2.9292 2.1568
4   4   19   2.4788 3.9205 3.0619 2.3582
4   5   20   2.2359 4.0299 2.8528 1.8574
;

```

```

Title "Modelo 3-Abordagem por Análise de Variância Univariada Usando o Comando Repeated do
SAS – Teste de Esfericidade e Ajuste do Número de Graus de Liberdade";
Proc glm data=mrmultiv;
Class bloco genot;
Model corte1-corte4=bloco genot/nouni;
Repeated corte/nom;
Lsmeans genot/stderr cl out=med21;
Run; quit;

```

Note que os quatro cortes agora estão à esquerda do comando *model* como quatro variáveis dependentes.

Após o comando *repeated*, *corte* é o nome do fator dentro de sujeitos e não foi previamente especificado na entrada de dados. A opção */nouni* informa ao SAS para não realizar uma ANOVA univariada e */nom* para não realizar uma MANOVA. Em geral, os testes univariados são mais capazes de detectar as diferenças existentes que os testes multivariados.

Os resultados obtidos com as análises anteriores são:

Modelo 3-Abordagem por Análise de Variância Univariada Usando o Comando *Repeated* do SAS

The GLM Procedure

Repeated Measures Analysis of Variance

Repeated Measures Level Information

Dependent Variable corte1 corte2 corte3 corte4

Level of corte	1	2	3	4
----------------	---	---	---	---

Repeated Measures Analysis of Variance  
 Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bloco	4	1.04531324	0.26132831	3.13	0.0561
genot	3	2.84073874	0.94691291	11.32	0.0008
Error	12	1.00336350	0.08361362		

Repeated Measures Analysis of Variance  
 Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
corte	3	37.44862545	12.48287515	114.05	<.0001
corte*bloco	12	0.65233146	0.05436095	0.50	0.9030
corte*genot	9	0.54773937	0.06085993	0.56	0.8232
Error(corte)	36	3.94030162	0.10945282		

Source	Adj Pr > F	
	G - G	H - F
corte	<.0001	<.0001
corte*bloco	0.7962	0.8808
corte*genot	0.7095	0.7976
Error(corte)		

Greenhouse-Geisser Epsilon 0.4801  
 Huynh-Feldt Epsilon 0.8461

Least Squares Means

genot	corte1	Standard	
	LSMEAN	Error	Pr >  t
1	2.71164000	0.08642073	<.0001
2	2.84940000	0.08642073	<.0001
3	2.38346000	0.08642073	<.0001
4	2.45112000	0.08642073	<.0001

corte1

genot	LSMEAN	95% Confidence Limits	
1	2.711640	2.523345	2.899935
2	2.849400	2.661105	3.037695
3	2.383460	2.195165	2.571755
4	2.451120	2.262825	2.639415

corte2

genot	LSMEAN	Standard	
		Error	Pr >  t
1	4.06864000	0.23882787	<.0001
2	4.52082000	0.23882787	<.0001
3	3.68974000	0.23882787	<.0001
4	4.11920000	0.23882787	<.0001

corte2

genot	LSMEAN	95% Confidence Limits	
1	4.068640	3.548279	4.589001
2	4.520820	4.000459	5.041181
3	3.689740	3.169379	4.210101
4	4.119200	3.598839	4.639561

corte3

genot	LSMEAN	Standard	
		Error	Pr >  t
1	3.00876000	0.11213021	<.0001
2	3.20486000	0.11213021	<.0001
3	2.81298000	0.11213021	<.0001
4	2.91516000	0.11213021	<.0001

corte3

genot	LSMEAN	95% Confidence Limits	
1	3.008760	2.764449	3.253071
2	3.204860	2.960549	3.449171
3	2.812980	2.568669	3.057291
4	2.915160	2.670849	3.159471

	corte4	Standard		
genot	LSMEAN	Error	Pr >  t	
1	2.38876000	0.07289678	<.0001	
2	2.50432000	0.07289678	<.0001	
3	2.15428000	0.07289678	<.0001	
4	2.11750000	0.07289678	<.0001	

	corte4	95% Confidence Limits		
genot	LSMEAN			
1	2.388760	2.229932	2.547588	
2	2.504320	2.345492	2.663148	
3	2.154280	1.995452	2.313108	
4	2.117500	1.958672	2.276328	

3. Exemplo de experimento com medidas repetidas analisado por abordagem multivariada clássica

```

Options nodate nonumber ls=78;
Data mrmultiv;
Input genot bloco parcela corte1 corte2 corte3 corte4;
Datalines;
1   1   1   2.8019 3.7309 3.0986 2.5096
1   2   2   2.9661 4.4355 3.1061 2.5729
1   3   3   2.4323 4.3231 2.8103 2.0797
1   4   4   2.9351 3.9971 2.7797 2.4403
1   5   5   2.4228 3.8566 3.2491 2.3413
2   1   6   2.7621 5.4053 3.8243 2.7299
2   2   7   3.0964 3.9068 3.2623 2.5861
2   3   8   3.0992 4.0886 3.1315 2.6032
2   4   9   2.6526 5.4288 2.7089 2.3016
2   5   10  2.6367 3.7746 3.0973 2.3008
3   1   11  2.2915 3.8114 2.9258 2.3986
3   2   12  2.5403 3.8272 2.8673 2.1629
3   3   13  2.4119 4.0832 3.0391 2.0708
3   4   14  2.3042 3.2785 2.7271 2.0493
3   5   15  2.3694 3.4484 2.5056 2.0898
4   1   16  2.5663 4.9607 2.8173 2.0575
4   2   17  2.3163 3.9663 2.9146 2.1576
4   3   18  2.6583 3.7186 2.9292 2.1568
4   4   19  2.4788 3.9205 3.0619 2.3582
4   5   20  2.2359 4.0299 2.8528 1.8574
;

```

Title "Modelo 4-Abordagem por Análise de Variância Multivariada Clássica – Teste de Esfericidade e Ajuste do Número de Graus de Liberdade";  
 Proc glm data=mrmultiv;  
 Class bloco genot;  
 Model corte1-corte4=bloco genot/nouni;  
 Manova h=genot/printe;  
 Repeated corte contrast(1)/printe summary; /\* Avalia os cortes em relação ao corte(1)=primeiro corte \*/  
 Lsmeans genot/stderr cl out=med21;  
 Run; quit;

Os resultados obtidos com esta análise são os seguintes:

Modelo 4-Abordagem por Análise de Variância Multivariada Clássica – Teste de Esfericidade e Ajuste do Número de Graus de Liberdade

The GLM Procedure

#### Multivariate Analysis of Variance

E = Error SSCP Matrix

	corte1	corte2	corte3	corte4
corte1	0.448112	-0.414028	-0.022954	0.185116
corte2	-0.414028	3.422325	-0.097161	-0.387348
corte3	-0.022954	-0.097161	0.754391	0.271269
corte4	0.185116	-0.387348	0.271269	0.318836

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 12	corte1	corte2	corte3	corte4
corte1	1.000000	-0.334330	-0.039479	0.489741
		0.2642	0.8981	0.0894
corte2	-0.334330	1.000000	-0.060469	-0.370815
		0.2642	0.8444	0.2123

corte3	-0.039479	-0.060469	1.000000	0.553119	
	0.8981	0.8444		0.0499	

corte4	0.489741	-0.370815	0.553119	1.000000	
	0.0894	0.2123	0.0499		

### Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse \* H, where

H = Type III SSCP Matrix for genot

E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1			
		corte1	corte2	corte3	corte4
3.93919	91.80	0.97956	0.39044	0.21127	0.81875
0.33571	7.82	-0.83458	-0.16736	-1.05985	2.27689
0.01604	0.37	-1.37042	0.18756	0.18846	1.17474
0.00000	0.00	-0.33458	0.37517	-1.08029	0.764021

### MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall genot Effect

H = Type III SSCP Matrix for genot

E = Error SSCP Matrix

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilk	0.14918426	2.11	12	24.103	0.0571
Pillai	1.06465582	1.51	12	33	0.1690
Hotelling-Lawley	4.29093424	3.00	12	12	0.0342
Roy	3.93918883	10.83	4	11	0.0008

### Repeated Measures Analysis of Variance

### Repeated Measures Level Information

Dependent Variable	corte1	corte2	corte3	corte4
Level of corte	1	2	3	4

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|

DF = 12	corte1	corte2	corte3	corte4
corte1	1.000000	-0.334330	-0.039479	0.489741
	0.2642	0.8981	0.0894	
corte2	-0.334330	1.000000	-0.060469	-0.370815
	0.2642	0.8444	0.2123	
corte3	-0.039479	-0.060469	1.000000	0.553119
	0.8981	0.8444		0.0499
corte4	0.489741	-0.370815	0.553119	1.000000
	0.0894	0.2123	0.0499	

E = Error SSCP Matrix

corte\_N represents the contrast between the nth level of corte and the 1st

corte_2	corte_3	corte_4	
corte_2	4.69849	0.78793	0.28968
corte_3	0.78793	1.24841	0.55722
corte_4	0.28968	0.55722	0.39672

Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

DF = 12	corte_2	corte_3	corte_4
corte_2	1.000000	0.325335	0.212174
corte_3	0.325335	1.000000	0.791785
corte_4	0.212174	0.791785	1.000000
	0.2781	0.0013	0.4865

### Repeated Measures Analysis of Variance

#### Sphericity Tests

Variables	Mauchly's			
	DF	Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	5	0.0815955	26.869687	<.0001
OrthogonalComponents	5	0.0851196	26.416318	<.0001

Manova Test Criteria and Exact F Statistics for the Hypothesis of no corte Effect

H = Type III SSCP Matrix for corte

E = Error SSCP Matrix

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks	0.02506727	129.64	3	10	<.0001
Pillai	0.97493273	129.64	3	10	<.0001
Hotelling-Lawley	38.89265200	129.64	3	10	<.0001
Roy	38.89265200	129.64	3	10	<.0001

Manova Test Criteria and F Approximations for the Hypothesis of no corte\*bloco Effect

H = Type III SSCP Matrix for corte\*bloco

E = Error SSCP Matrix

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilk	0.47694200	0.72	12	26.749	0.7199
Pillai	0.62700794	0.79	12	36	0.6548
Hotelling-Lawley	0.88813718	0.69	12	13.692	0.7342
Roy	0.56517621	1.70	4	12	0.2154

Manova Test Criteria and F Approximations for the Hypothesis of no corte\*genot Effect

H = Type III SSCP Matrix for corte\*genot

E = Error SSCP Matrix

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilk	0.67418315	0.48	9	24.488	0.8749
Pillai	0.34700404	0.52	9	36	0.8479
Hotelling-Lawley	0.45212250	0.48	9	12.866	0.8660
Roy	0.37082697	1.48	3	12	0.2688

### Repeated Measures Analysis of Variance Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bloco	4	1.04531324	0.26132831	3.13	0.0561
genot	3	2.84073874	0.94691291	11.32	0.0008
Error	12	1.00336350	0.08361362		

### Univariate Tests of Hypotheses for Within Subject Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F

corte	3	37.44862545	12.48287515	114.05	<.0001
corte*bloco	12	0.65233146	0.05436095	0.50	0.9030
corte*genot	9	0.54773937	0.06085993	0.56	0.8232
Error(corte)	36	3.94030162	0.10945282		

Source	Adj Pr > F	
	G - G	H - F
corte	<.0001	<.0001
corte*bloco	0.7962	0.8808
corte*genot	0.7095	0.7976
Error(corte)		

Greenhouse-Geisser Epsilon	0.4801
Huynh-Feldt Epsilon	0.8461

### Repeated Measures Analysis of Variance Analysis of Variance of Contrast Variables

corte\_N represents the contrast between the nth level of corte and the 1st

Contrast Variable: corte\_2

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	45.04170966	45.04170966	115.04	<.0001
bloco	4	0.83671937	0.20917984	0.53	0.7133
genot	3	0.57805105	0.19268368	0.49	0.6944
Error	12	4.69849312	0.39154109		

Contrast Variable: corte\_3

Source	DF	Type III SS	Mean Square	F Value	Pr > F
--------	----	-------------	-------------	---------	--------

Mean	1	2.98818612	2.98818612	28.72	0.0002
bloco	4	0.32388019	0.08097005	0.78	0.5603
genot	3	0.08407717	0.02802572	0.27	0.8462
Error	12	1.24841135	0.10403428		

Contrast Variable: corte\_4

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Mean	1	1.89346272	1.89346272	57.27	<.0001
bloco	4	0.13359877	0.03339969	1.01	0.4402
genot	3	0.04232467	0.01410822	0.43	0.7375
Error	12	0.39671670	0.03305972		

#### Least Squares Means

genot	corte1		Pr >  t
	LSMEAN	Standard Error	
1	2.71164000	0.08642073	<.0001
2	2.84940000	0.08642073	<.0001
3	2.38346000	0.08642073	<.0001
4	2.45112000	0.08642073	<.0001

genot	corte1		
	LSMEAN	95% Confidence Limits	
1	2.711640	2.523345	2.899935
2	2.849400	2.661105	3.037695
3	2.383460	2.195165	2.571755
4	2.451120	2.262825	2.639415

genot	corte2		Pr >  t
	LSMEAN	Standard Error	

1	4.06864000	0.23882787	<.0001
2	4.52082000	0.23882787	<.0001
3	3.68974000	0.23882787	<.0001
4	4.11920000	0.23882787	<.0001

corte2			
genot	LSMEAN	95% Confidence Limits	
1	4.068640	3.548279	4.589001
2	4.520820	4.000459	5.041181
3	3.689740	3.169379	4.210101
4	4.119200	3.598839	4.639561

corte3			
genot	LSMEAN	Standard	
		Error	Pr >  t
1	3.00876000	0.11213021	<.0001
2	3.20486000	0.11213021	<.0001
3	2.81298000	0.11213021	<.0001
4	2.91516000	0.11213021	<.0001

corte3			
genot	LSMEAN	95% Confidence Limits	
1	3.008760	2.764449	3.253071
2	3.204860	2.960549	3.449171
3	2.812980	2.568669	3.057291
4	2.915160	2.670849	3.159471

corte4			
genot	LSMEAN	Standard	
		Error	Pr >  t
1	2.38876000	0.07289678	<.0001
2	2.50432000	0.07289678	<.0001
3	2.15428000	0.07289678	<.0001
4	2.11750000	0.07289678	<.0001

corte4			
genot	LSMEAN	95% Confidence Limits	
1	2.388760	2.229932	2.547588

2	2.504320	2.345492	2.663148
3	2.154280	1.995452	2.313108
4	2.117500	1.958672	2.276328

## Diagnóstico de Pressuposições para Análise de Medidas Repetidas

### 1. Teste de Esfericidade

A pressuposição de esfericidade é feita em relação à estrutura da matriz de covariância num delineamento medidas repetidas.

A condição de simetria composta ocorre se todas as covariâncias (os elementos fora da diagonal da matriz de covariância) são iguais e todas as variâncias são iguais na população que está sendo amostrada. Observe que as variâncias não têm de ser iguais às covariâncias.

Como ocorre com a pressuposição de homogeneidade de variâncias, raramente se espera que um conjunto de dados reais apresente exatamente uma simetria composta, mas se as covariâncias observadas são aproximadamente iguais em todas as amostras e se as variâncias também são iguais, pode-se assumir que existe simetria composta.

Se ocorrer a simetria composta então a esfericidade também ocorre. Uma vez que simetria composta é um requerimento mais restrito do que esfericidade ainda pode ser necessário checar esfericidade se não ocorre simetria composta.

Teste de Esfericidade de Mauchly:

Testa a hipótese nula de que a matriz populacional de correlações é uma matriz identidade.

A hipótese nula é  $H_0: M' \Sigma M = \sigma^2 I_m$  contra a hipótese alternativa  $H_a: M' \Sigma M \neq \sigma^2 I_m$ , em que:

$\sigma^2 > 0$ : variância não-especificada;

$I$ : matriz identidade  $m \times m$ ;

$M$ : matriz ortonormal  $r \times m$  associada com um efeito dentro de sujeitos.

A matriz  $M$  é gerada usando contrastes polinomiais igualmente espaçados aplicados aos fatores dentro de sujeitos.

A estatística  $W$  do teste de esfericidade de Mauchly é expressa como:

$$W = \frac{|A|}{(trace(A)/m)^m}, \text{ para } trace(A) > 0, \text{ em que:}$$

$A = M' \Sigma M$ ,  $\Sigma = (Y - X\hat{B})' (Y - X\hat{B})$  é a matriz de soma de quadrados e de produtos cruzados dos resíduos,  $M$  é uma matriz ortonormal associada com os efeitos dentro de sujeitos que estão sendo testados.

A aproximação Qui-Quadrado para o teste de Mauchly é a seguinte: quando  $n$  é grande e sob  $n - r_X \geq 1$  e  $m \geq 2$ , tem-se:

$$\rho = 1 - \left( 2m^2 + m + 2 \right) / \left( 6m(n - r_X) \right), \text{ em que } n \text{ é o número de sujeitos e } r_X \text{ é a ordem da matriz de delineamento.}$$

Para efeitos principais,  $m = l - 1$ , onde  $l$  é o número de níveis da variável que está sendo testada. Para efeito de interação simples,  $m = (l_1 - 1)(l_2 - 1)$ , onde  $l_1$  e  $l_2$  são os respectivos números de níveis para as duas variáveis envolvidas na interação.

A estatística Qui-Quadrado é dada por:

$$c = -\rho(n - r_X) \log W, \text{ ou seja,}$$

$$c = \left( \frac{2m^2 + m + 2}{6m} - n - r_X \right) \log(W), \text{ para } W > 0.$$

Os graus de liberdade associados a essa estatística são computados como  $f = \frac{m(m+1)}{2} - 1$ .

A significância é avaliada com base no seguinte critério:

se a estatística Qui-Quadrado calculado for maior que o valor de Qui-Quadrado tabelado afirma-se que o teste é significativo, então se rejeita a hipótese  $H_0: M' \Sigma M = \sigma^2 I_m$ , que é a hipótese de que existe esfericidade.

## 2. Ajuste de Graus de Liberdade

Quando o teste de Mauchly é significativo, ou seja, não existe esfericidade, ainda assim pode-se utilizar a análise de variância univariada com o ajuste dos graus de liberdade, usando os seguintes métodos:

Greenhouse-Geisser (GG):

$$\epsilon_{gg} = \frac{[traço(A)]^2}{m[traço(A'A)]}$$

Huynh-Feldt:

$$\epsilon_{hf} = \min \left[ \frac{(n)(m)(\epsilon_{gg}) - 2}{m(n - r_X) - (m^2)\epsilon_{gg}}, 1 \right]$$

### 3. Teste de Homogeneidade de Variâncias

O teste de Levene é usado para testar se  $k$  amostras têm variâncias iguais. Variâncias iguais ao longo de amostras são denominadas de homogeneidade de variâncias. Alguns testes estatísticos, como por exemplo, a análise de variância, assume que as variâncias são iguais ao longo dos grupos ou amostras.

O teste de Levene é uma alternativa ao teste de Bartlett. O teste de Levene é menos sensível que o teste de Bartlett para afastamentos da normalidade. Quando se tem forte evidência de que os dados de fato provêm de uma distribuição normal, ou aproximadamente normal, então o teste de Bartlett tem um melhor desempenho.

O teste de Levene é definido como:  $H_0: \sigma_1 = \sigma_2 = \dots = \sigma_k$  contra  $H_1: \sigma_i \neq \sigma_j$ , para pelo menos um par  $(i, j)$ .

Dada uma variável  $Y$  com amostra de tamanho  $N$  dividida em  $k$  subgrupos, em que  $N_i$  é o tamanho de amostra do  $i$ -ésimo subgrupo, a estatística do teste de Levene é definida como:

$$W = \frac{(N-k) \sum_{i=1}^k N_i (\bar{Z}_{i\cdot} - \bar{Z}_{..})^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i\cdot})^2}, \text{ em que } Z_{ij} \text{ pode ter uma das três definições seguintes:}$$

1.  $Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$ , em que  $\bar{Y}_{i\cdot}$  é a média do  $i$ -ésimo subgrupo;
2.  $Z_{ij} = |Y_{ij} - \tilde{Y}_{i\cdot}|$ , em que  $\tilde{Y}_{i\cdot}$  é a mediana do  $i$ -ésimo subgrupo;
3.  $Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}^t|$ , em que  $\bar{Y}_{i\cdot}^t$  é a média censurada 10% do  $i$ -ésimo subgrupo.

Os  $\bar{Z}_{i\cdot}$  são as médias de grupos da  $Z_{ij}$  e  $\bar{Z}_{..}$  é a média geral da  $Z_{ij}$ .

As três escolhas para definir  $Z_{ij}$  determinam a robustez e o poder do teste de Levene.

Robustez significa a capacidade do teste em não detectar falsamente variâncias desiguais quando os dados analisados não são normalmente distribuídos e as variâncias de fato são iguais.

Poder significa a capacidade do teste em detectar variâncias desiguais quando as variâncias de fato são desiguais.

O teste de Levene rejeita a hipótese que as variâncias são iguais se  $W > F_{(\alpha, k-1, N-k)}$ , em que  $F_{(\alpha, k-1, N-k)}$  é o valor crítico da distribuição  $F$ , com  $k-1$  e  $N-k$  graus de liberdade a uma probabilidade  $\alpha$  escolhida.

O teste de Bartlett é comumente usado para testar a igualdade (homogeneidade) de variâncias. As hipóteses do teste são as seguintes:

$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$  contra  $H_a: \text{as } \sigma_i^2 \text{ não são todas iguais.}$

Considere que se tenham amostras de tamanho  $n_i$  extraídas da i-ésima população,  $i=1, 2, \dots, k$ , e as estimativas de variâncias usuais de cada amostra:  $S_1^2, S_2^2, \dots, S_k^2$ , em que:

$$S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 / (n_i - 1)$$

Agora, considere a notação seguinte:  $v_i = n_i - 1$  (os  $v_i$  são os graus de liberdade) e  $v = \sum_{i=1}^k v_i$

$$, S^2 = \frac{\sum_{i=1}^k v_i S_i^2}{v}$$

A estatística  $M$  do teste de Bartlett é definida por:

$$M = v \log S^2 - \sum_{i=1}^k v_i \log S_i^2$$

Quando nenhum dos graus de liberdade é pequeno, Bartlett demonstrou que  $M$  é distribuído aproximadamente como  $\chi^2_{k-1}$ .

A aproximação qui-quadrado geralmente é aceitável se todos os  $n_i$  são no mínimo cinco. De acordo com Bartlett, este é um teste ligeiramente viesado. Ele pode ser melhorado dividindo-se  $M$  pelo fator

$$C = 1 + \frac{1}{3(k-1)} \left[ \left( \sum_{i=1}^k \frac{1}{v_i} \right) - \frac{1}{v} \right].$$

Então, é sugerido o uso de  $M/C$  ao invés de  $M$  para o teste estatístico.

Este teste não é robusto, é muito sensível a afastamentos da normalidade. Quando existe suspeita de não-normalidade dos dados, o teste de Levene é uma melhor escolha que o teste de Bartlett.

#### 4. Teste de Normalidade

O teste de Shapiro-Wilk calcula uma estatística  $W$  que testa se uma amostra aleatória,  $x_1, x_2, \dots, x_n$ , é originada de uma distribuição normal. Pequenos valores de  $W$  evidenciam afastamento da normalidade e pontos percentuais para a estatística  $W$  foram reproduzidos por meio de tabela.

Este teste tem se dado muito bem em estudos comparativos com outros testes de qualidade de ajuste.

A estatística  $W$  é calculada como a seguir:

$$W = \frac{\left( \sum_{i=1}^n a_i x_{(i)} \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ em que } x_{(i)} \text{ são valores da amostra ordenados} (x_{(1)} \text{ é o menor}) \text{ e os}$$

$a_i$  são constantes geradas a partir das médias, variâncias e covariâncias das estatísticas ordens de uma amostra de tamanho  $n$  extraída de uma distribuição normal.

# CAPITULO 13

## Interação Genótipos x Ambientes

A interação de genótipos com ambientes (IGE) caracteriza-se pelo fato de os genótipos apresentarem comportamento diferente nos diferentes ambientes em que são avaliados. Os ambientes podem ser representados pelos fatores locais, anos, épocas e adubação ou por combinações destes fatores.

A quantificação dos efeitos da IGE sob as técnicas de melhoramento e a possibilidade de adoção de procedimentos para sua minimização e, ou, aproveitamento devem ser avaliados.

### Quantificação da IGE:

O componente da variação atribuído aos efeitos da interação genótipos x ambientes, representado por  $\sigma_{ge}^2$ , é quantificado por meio de um sistema no qual se igualam os quadrados médios, obtidos em análises de variâncias, aos respectivos estimadores das esperanças de quadrados médios.

Serão considerados experimentos em blocos casualizados, envolvendo  $g$  genótipos avaliados em  $e$  ambientes, em relação a um determinado caráter. Cada observação fenotípica pode ser descrita pelo seguinte modelo estatístico:

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + B/E_{jk} + \varepsilon_{ijk}, \text{ em que:}$$

$\mu$ : média geral;

$G_i$ : efeito do i-ésimo genótipo ( $i = 1, 2, \dots, g$ );

$E_j$ : efeito do j-ésimo ambiente ( $j = 1, 2, \dots, e$ )

$GE_{ij}$ : efeito da interação do i-ésimo genótipo com o j-ésimo ambiente;

$B/E_{jk}$ : efeito do k-ésimo bloco dentro do j-ésimo ambiente ( $k = 1, 2, \dots, r$ );

$\varepsilon_{ijk}$ : erro aleatório, pressupondo que  $\varepsilon_{ijk} \sim NID(0; \sigma^2)$

O esquema da análise de variância conjunta, segundo este modelo estatístico é:

FV	GL	SQ	QM
Blocos/Ambientes	e (r-1)	SQB	QMB
Ambientes (E)	e-1	SQE	QME
Genótipos (G)	g-1	SQG	QMG
G x E	(e-1)(g-1)	SQGE	QMGE
Resíduo	e(r-1)(g-1)	SQR	QMR
Total	egr-1	SQT	

As expressões para obtenção das somas de quadrados são:

$$C = Y_{..}^2 / egr;$$

$$SQB = \frac{1}{g} \sum_j \sum_k Y_{.jk}^2 - \frac{1}{gr} \sum_j Y_{.j.}^2;$$

$$SQE = \frac{1}{gr} \sum_j Y_{.j.}^2 - C;$$

$$SQG = \frac{1}{er} \sum_i Y_{i..}^2 - C;$$

$$SQGE = \frac{1}{r} \sum_i \sum_j Y_{ij.}^2 - \frac{1}{gr} \sum_j Y_{.j.}^2 - \frac{1}{er} \sum_i Y_{i..}^2 + C;$$

$$SQT = \sum_i \sum_j \sum_k Y_{ijk}^2 - C;$$

$$SQR = SQT - (SQB + SQE + SQG + SQGE).$$

As esperanças dos quadrados médios para as fontes de variação descritas no modelo estatístico anterior são:

- (a) Efeitos aleatórios: B, G, E, GE e  $\mathcal{E}$

FV	E (QM)
Blocos/Ambientes	$\sigma^2 + g\sigma_b^2$
Ambientes (E)	$\sigma^2 + r\sigma_{ge}^2 + g\sigma_b^2 + gr\sigma_e^2$
Genótipos (G)	$\sigma^2 + r\sigma_{ge}^2 + er\sigma_g^2$
Interação G x E	$\sigma^2 + r\sigma_{ge}^2$
Resíduo	$\sigma^2$

(b) Efeitos aleatórios: B e  $\mathcal{E}$  e Efeitos fixos: G, E e GE

FV	E (QM)
Blocos/Ambientes	$\sigma^2 + g\sigma_b^2$
Ambientes (E)	$\sigma^2 + g\sigma_b^2 + gr\phi_e$
Genótipos (G)	$\sigma^2 + er\phi_g$
Interação G x E	$\sigma^2 + r\phi_{ge}$
Resíduo	$\sigma^2$

(c) Efeitos aleatórios: B, G, GE e  $\mathcal{E}$  e Efeito fixo: E

FV	E (QM)
Blocos/Ambientes	$\sigma^2 + g\sigma_b^2$
Ambientes (E)	$\sigma^2 + r\ell\sigma_{ge}^2 + g\sigma_b^2 + gr\phi_e$
Genótipos (G)	$\sigma^2 + er\sigma_g^2$
Interação G x E	$\sigma^2 + r\ell\sigma_{ge}^2$
Resíduo	$\sigma^2$

$$\ell = e/(e-1)$$

(d) Efeitos aleatórios: B, E, GE e  $\varepsilon$  e Efeito fixo: G

FV	E (QM)
Blocos/Ambientes	$\sigma^2 + g\sigma_b^2$
Ambientes (E)	$\sigma^2 + g\sigma_b^2 + gr\sigma_e^2$
Genótipos (G)	$\sigma^2 + r\ell\sigma_{ge}^2 + er\phi_g$
Interação G x E	$\sigma^2 + r\ell\sigma_{ge}^2$
Resíduo	$\sigma^2$

$\ell = g/(g-1)$

Nas tabelas de (b) a (d) tem-se:  $\phi_g = \sum_i G_i^2 / (g-1)$ ;  $\phi_e = \sum_j E_j^2 / (e-1)$ ;

$$\phi_{ge} = \sum_i \sum_j GE_{ij}^2 / (e-1)(g-1).$$

Conhecidas as expressões das esperanças de quadrados médios [E (QM)], os componentes de variância associados aos efeitos aleatórios e os componentes quadráticos associados aos efeitos fixos são facilmente estimados. Como exemplo, considere a Tabela (d), da qual se derivam os seguintes estimadores:

$$\hat{\sigma}^2 = QMR; \quad \hat{\sigma}_{ge}^2 = \frac{QMG - QME}{r\ell}; \quad \hat{\sigma}_e^2 = \frac{QME - QMB}{gr};$$

$$\hat{\phi}_g = \frac{QMG - QME}{er}; \quad \hat{\sigma}_b^2 = \frac{QMB - QMR}{g}.$$

Além dos componentes, normalmente tem-se interesse nos testes de significância e então as expressões das E (QM) são de grande utilidade, por indicar, para determinadas hipóteses, qual o denominador apropriado para o teste F. Para a Tabela (d) tem-se:

Hipótese ( $H_0$ )	Estatística F	Graus de liberdade
$H_{01}: \sigma_e^2 = 0$	$F_{01} = QME/QMB$	(e-1), e(r-1)
$H_{02}: G_i = 0, \forall i$	$F_{02} = QMG/QMGE$	(g-1), (e-1)(g-1)
$H_{03}: \sigma_{ge}^2 = 0$	$F_{03} = QMGE/QMR$	(e-1)(g-1), e(r-1)(g-1)

No caso de  $H_{02}$  a restrição  $\sum_i G_i = 0$  é parte integrante do modelo.

*Observação:* Sendo o genótipo um *fator de efeito fixo*, o interesse é estimar e testar a hipótese de funções lineares dos efeitos, e um valor  $\hat{\phi}_g = 10,34$ , usado como exemplo hipotético, é uma estimativa da variabilidade do conjunto de g's fixos.

Por outro lado, se o genótipo é um *fator de efeito aleatório*, o interesse é fazer inferência sobre toda a população de genótipos, e um valor  $\hat{\sigma}_g^2 = 10,34$  é uma estimativa da variância na população de g's.

O F esperado sob a hipótese de nulidade é 1,0. Desta forma, por exemplo, na Tabela (a) tem-se que, para testar  $H_0: \sigma_e^2 = 0$ , o F apropriado é expresso por:

$$F = \frac{QME+QMR}{QMB+QMGE}.$$

Este F está associado a  $n_1$  e  $n_2$  graus de liberdade, obtidos pela equação de Satterthwaite, que, neste caso, tem a seguinte expressão:

$$n_1 = \frac{(QME+QMR)^2}{\frac{(QME)^2}{e-1} + \frac{(QMR)^2}{e(r-1)(g-1)}} \text{ e}$$

$$n_2 = \frac{\frac{(QMB+QMGE)^2}{(QMB)^2 + (QMGE)^2}}{\frac{e(r-1)}{(e-1)(g-1)}}.$$

*Observação:* Dependendo da bibliografia consultada, das suposições e das restrições associadas ao modelo estatístico, particularmente nos modelos mistos, podem-se encontrar expressões para as E (QM) diferentes das apresentadas. A questão sobre qual delas usar permanece aberta e não tem sido considerada.

Exemplo de Aplicação:

Antes de realizar a análise conjunta é importante que se faça a análise de variância em cada ambiente, para avaliar a existência de variabilidade genética entre os genótipos estudados, a precisão relativa de cada experimento e a homogeneidade das variâncias residuais. Considere os resultados seguintes:

Quadro. Resultado da análise de variância conjunta de nove experimentos (ambientes), envolvendo cinco genótipos, avaliados em quatro blocos por experimento

FV	GL	QM	F
Blocos/Ambientes	27		
Genótipos (G)	4	QMG=12,119	QMG/QMGE=
Ambientes (E)	8	QME=42,886	3,00*
Interação G x E	32	QMGE=4,039	QMGE/QMR=
Resíduo	108	QMR=0,456	8,86**

Considerando todos os efeitos, exceto a média, como aleatórios estimam-se os componentes de variância como a seguir:

$$\hat{\sigma}_g^2 = \frac{QMG - QMGE}{er} = \frac{12,119 - 4,039}{(9)(4)} = 0,224 \text{ e}$$

$$\hat{\sigma}_{ge}^2 = \frac{QMGE - QMR}{r} = \frac{4,039 - 0,456}{4} = 0,896$$

Decomposição da interação em partes simples e complexa:

A existência da IGE está associada a dois fatores. O *fator simples* é proporcionado pela diferença de variabilidade entre genótipos nos ambientes e o *fator complexo* é dado pela falta de correlação entre genótipos nos ambientes. Apenas o fator complexo causa dificuldade para o melhoramento de plantas.

A interação complexa indica a inconsistência da superioridade de genótipos com a variação ambiental, o que torna difícil a seleção e, ou, a recomendação dos mesmos.

A partição do quadrado médio da interação genótipos x pares de ambientes em parte simples e complexa é dada pela expressão (Robertson, 1959):

$$QMGxE = S + C, \text{ sendo: } S = \frac{1}{2} \left( \sqrt{Q_1} - \sqrt{Q_2} \right)^2 \text{ e } C = (1-r) \sqrt{Q_1 Q_2}, \text{ em que:}$$

**S** e **C**: partes simples e complexa da interação, respectivamente;

**r**: correlação entre médias de genótipos nos ambientes 1 e 2;

**Q<sub>1</sub>** e **Q<sub>2</sub>**: quadrados médios entre genótipos nos ambientes 1 e 2, respectivamente.

A decomposição apresentada é o resultado da seguinte análise:

Para dois ambientes apenas tem-se:

$$\begin{aligned} QMGxE(g-1) &= \frac{1}{2} \left[ \left( \sum_i Y_{i1}^2 - \frac{Y_{\cdot 1}^2}{g} \right) + \left( \sum_i Y_{i2}^2 - \frac{Y_{\cdot 2}^2}{g} \right) \right] - \\ &- \left( \sum_i Y_{i1} Y_{i2} - \frac{Y_{\cdot 1} Y_{\cdot 2}}{g} \right) \end{aligned}$$

, ou

$$\begin{aligned} QMGxE &= \frac{1}{2} (Q_1 + Q_2) - PM_{12} = \frac{1}{2} (Q_1 + Q_2) - r \sqrt{Q_1 Q_2} \text{ e} \\ QMGxE &= \frac{1}{2} \left( \sqrt{Q_1} - \sqrt{Q_2} \right)^2 + (1-r) \sqrt{Q_1 Q_2}. \end{aligned}$$

Sob o aspecto genético, a interação pode ser decomposta da seguinte forma:

a) Considerando os efeitos de ambientes aleatórios

$$\hat{\sigma}_{ge}^2 = \frac{1}{2} \left( \hat{\sigma}_{g1} - \hat{\sigma}_{g2} \right)^2 + (1 - r_g) \hat{\sigma}_{g1} \hat{\sigma}_{g2}$$

b) Considerando os efeitos de ambientes fixos

$$\hat{\sigma}_{ge}^2 = \frac{1}{4} \left( \hat{\sigma}_{g1} - \hat{\sigma}_{g2} \right)^2 + \frac{1}{2} (1 - r_g) \hat{\sigma}_{g1} \hat{\sigma}_{g2}$$

Nestas equações tem sido utilizada, para  $r_g$ , a estimativa dada pelo quociente entre a covariância entre médias dos genótipos, nos dois ambientes, e o produto dos desvios padrão genéticos em cada ambiente.

A decomposição do quadrado médio da interação proposta por Cruz e Castoldi (1991), dada por:

$\sqrt{(1-r)^3 Q_1 Q_2}$ , pondera de maneira mais eficiente a contribuição da correlação e da diferença de variabilidade genotípica nos ambientes.

Para o exemplo em consideração, estimam-se as seguintes medidas de dissimilaridade entre pares de ambientes:

a) Matriz de Somas de Quadrados entre Genótipos e Pares de Ambientes

A soma de quadrados entre genótipos e pares de ambientes com base em médias de parcelas pode ser estimada por meio da seguinte expressão:

$$SQGE_{jj'} = \theta_{jj'} = \frac{1}{2} \left[ d_{jj'}^2 - \frac{1}{g} \left( Y_{\cdot j} - Y_{\cdot j'} \right)^2 \right], \text{ onde}$$

$d_{jj'}^2 = \sum_i \left( Y_{ij} - Y_{ij'} \right)^2$ , expressa o quadrado da distância euclidiana entre os ambientes  $j$  e  $j'$  com base na média de  $g$  genótipos.

No exemplo sob análise, para os ambientes 1 e 2 tem-se:

$$d_{12}^2 = (2,0 - 6,4)^2 + (3,7 - 6,7)^2 + \dots + (4,9 - 4,9)^2$$

$$d_{12}^2 = 54,3 \text{ e}$$

$$SQGE_{12} = \theta_{12} = \frac{1}{2} \left[ 54,3 - \frac{1}{5} (16,10 - 30,70)^2 \right]$$

$$SQGE_{12} = \theta_{12} = 5,834$$

As demais estimativas são obtidas de forma análoga e estão apresentadas na Tabela 1 a seguir:

Tabela 1-Estimativas das somas de quadrados entre genótipos e pares de ambientes, com base em médias de parcelas

Ambientes	2	3	4	5	6	7	8	9
1	5,834	6,714	2,266	2,390	9,526	5,954	12,054	12,446
2		0,680	1,514	1,604	1,906	0,250	2,360	3,046
3			3,154	1,574	2,566	0,350	5,130	6,026
4				1,246	5,124	1,954	4,114	4,244
5					5,176	1,084	6,904	7,736
6						2,786	3,836	5,480
7							3,580	4,446
8								0,356

b) *Matriz de Quadrados Médios da Interação entre Genótipos e Pares de Ambientes*

Com base nas estimativas das somas de quadrados da interação entre genótipos e pares de ambientes, obtém-se os respectivos quadrados médios:

Tem-se que  $QMGE_{jj'} = \frac{SQGE_{jj'}}{g-1}$ . Então,

$$QMGE_{12} = \frac{SQGE_{12}}{g-1} = \frac{5,834}{5-1} = 1,4585.$$

As demais estimativas estão apresentadas na Tabela 2 a seguir:

Tabela 2-Estimativas dos quadrados médios da interação entre genótipos e pares de ambientes

Ambientes	2	3	4	5	6	7	8	9
1	1,458	1,678	0,566	0,597	2,381	1,488	3,013	3,111
2		0,170	0,378	0,401	0,476	0,062	0,590	0,761
3			0,788	0,393	0,641	0,087	1,282	1,506
4				0,311	1,281	0,488	1,028	1,061
5					1,294	0,271	1,726	1,934
6						0,696	0,959	1,370
7							0,895	1,111
8								0,089

c) *Matriz de Correlações entre Médias dos Genótipos em Cada Par de Ambientes*

Os coeficientes de correlação simples, estimados a partir das médias de produtividade de cinco genótipos avaliados em nove ambientes:

Tabela 3-Médias de produtividade de cinco genótipos avaliados em nove ambientes

	A1	A2	A3	A4	A5	A6	A7	A8	A9	$\bar{Y}_i$
C1	2,0	6,4	7,3	3,8	3,1	5,9	5,4	8,3	7,9	50,1
C2	3,7	6,7	8,4	3,6	4,1	8,1	5,8	6,7	5,5	52,6
C3	3,1	6,6	8,1	3,8	4,7	6,3	6,3	7,1	5,7	51,7
C4	2,4	5,1	8,6	2,8	4,2	5,3	5,9	5,2	4,5	45,0
C5	4,9	4,9	6,3	3,8	4,0	3,8	4,3	4,4	3,8	40,2

, são descritos na Tabela 4 a seguir:

Tabela 4-Estimativas de coeficientes de correlações entre médias de cinco genótipos avaliados em nove ambientes

Ambientes	2	3	4	5	6	7	8	9
1	-0,645	-0,535	0,365	0,338	-0,284	-0,622	-0,654	-0,701
2		0,787	-0,057	0,080	0,886	0,889	0,776	0,630
3			-0,605	0,404	0,695	0,901	0,249	0,103
4				-0,195	0,050	-0,311	0,397	0,362
5					0,105	0,426	-0,396	-0,614
6						0,105	0,604	0,437
7							0,505	0,329
8								0,963

d) Matriz da Estimativa da Parte Complexa do Quadrado Médio da Interação entre Genótipos e Pares de Ambientes

Com base na equação de Cruz e Castoldi (1991), obtém-se as estimativas da parte complexa da interação. Por exemplo, em relação aos ambientes 1 e 2, tem-se:

$$r_{12} = -0,6452; Q_1 = 1,3070; Q_2 = 0,5330, \text{ então}$$

$$C = \sqrt{[1 - (-0,6452)^3] \times 1,3070 \times 0,5330} = 1,7612.$$

O valor percentual de C corresponde a:

$$\%C = \frac{C}{QMG \times E_{12}} \times 100 = \frac{1,7612}{1,4585} \times 100 = 120,75$$

Tabela 5-Estimativas da parte complexa (acima da diagonal) resultantes da decomposição da interação entre genótipos e pares de ambientes e respectivos valores percentuais (abaixo da diagonal)

Ambiente	1	2	3	4	5	6	7	8	9
1		1,76	2,05	0,25	0,36	2,59	1,80	3,77	3,95
2	120,8		0,07	0,34	0,37	0,04	0,02	0,12	0,26
3	122,5	39,9		0,83	0,25	0,25	0,02	0,95	1,25
4	44,2	90,9	105,7		0,33	0,63	0,49	0,31	0,34
5	59,8	93,2	64,2	105,5		0,77	0,19	1,48	1,85
6	109,1	9,2	38,8	48,9	59,3		0,21	0,60	1,03
7	121,2	33,0	25,5	101,7	71,0	30,4		0,41	0,65
8	125,1	20,3	74,3	30,6	85,9	62,9	46,0		0,02
9	126,8	33,5	82,9	32,4	95,8	75,0	58,7	19,2	

Observação: valores de %C menores que 50 indicam os pares de ambientes cuja interação com os genótipos é predominantemente do tipo simples. Nesta tabela a parte abaixo da diagonal representa a percentagem da parte complexa em relação ao total da interação e valores acima de 100% ocorrem nos casos em que a correlação das médias dos genótipos em dois ambientes é negativa.

### Estratificação de Ambientes

Um aspecto importante na análise de interação genótipos x ambientes é identificar subgrupos de ambientes nos quais a interação seja não significativa para o conjunto de genótipos sob avaliação, o que é chamado de estratificação de ambientes.

Estes subgrupos representam ambientes com similaridade de resposta dos genótipos. Esta análise possibilita o descarte de ambientes, em condições de escassez de recursos.

O método de agrupamento de ambientes consiste em estimar a soma de quadrados da interação entre genótipos e pares de ambientes e em seguida agrupar aqueles ambientes cuja interação é não significativa. No próximo passo é realizada a estimação da soma de quadrados entre genótipos e grupos de três ambientes, empregando-se o teste F para avaliar a possibilidade de formação do novo grupo (Lin, 1982).

A soma de quadrados entre genótipos e pares de ambientes, com base em médias de parcelas, pode ser estimada por meio da seguinte expressão:

$$SQGE_{jj'} = \theta_{jj'} = \frac{1}{2} \left[ d_{jj'}^2 - \frac{1}{g} \left( Y_{\cdot j} - Y_{\cdot j'} \right)^2 \right], \text{ onde}$$

$d_{jj'}^2 = \sum_i \left( Y_{ij} - Y_{ij'} \right)^2$ , expressa o quadrado da distância euclidiana entre os ambientes  $j$  e  $j'$  com base na média de  $g$  genótipos.

Para obter a soma de quadrados da interação entre genótipos e ambientes (IGE) com base em parcela deve-se multiplicar a expressão anterior pelo número de repetições.

A soma de quadrados entre genótipos e três ambientes pode ser obtida por meio da seguinte expressão:

$$\theta_{(jj'k)} = \frac{2}{\eta} \left( \theta_{jj'} + \theta_{jk} + \theta_{j'k} \right) = \frac{2}{\eta} S_{jj'k}, \text{ em que:}$$

$\eta = 3$ : número de ambientes considerados na interação;

$S_{jj'k}$ : somatório das somas de quadrados da interação entre genótipos e combinações, dois a dois, dos ambientes  $j$ ,  $j'$  e  $k$ .

A soma de quadrados entre genótipos e quatro ambientes pode ser obtida por meio da seguinte expressão:

$$\theta_{(jj'kl)} = \frac{2}{\eta} \left( S_{jj'k} + \theta_{jl} + \theta_{j'l} + \theta_{kl} \right), \text{ em que:}$$

$\eta = 4$ : número de ambientes considerados na interação;

Para o exemplo sob análise tem-se:

**1) Estimação de Somas de Quadrados entre Genótipos e Pares de Ambientes**

Para os ambientes 1 e 2 obtém-se:

$$d_{12}^2 = (2,0 - 6,4)^2 + (3,7 - 6,7)^2 + \dots + (4,9 - 4,9)^2$$

$$d_{12}^2 = 54,30$$

$$SQGE_{12} = \theta_{12} = \frac{1}{2} \left[ 54,30 - \frac{1}{5} (16,10 - 30,70)^2 \right]$$

$$SQGE_{12} = \theta_{12} = 5,834$$

Para os ambientes 1 e 3 obtém-se:

$$d_{13}^2 = (2,0 - 7,3)^2 + (3,7 - 8,4)^2 + \dots + (4,9 - 6,3)^2$$

$$d_{13}^2 = 115,58$$

$$SQGE_{13} = \theta_{13} = \frac{1}{2} \left[ 115,58 - \frac{1}{5} (16,10 - 38,70)^2 \right]$$

$$SQGE_{13} = \theta_{13} = 6,714$$

As demais estimativas são obtidas de forma análoga e estão apresentadas no Tabela 6 a seguir.

Tabela 6- Estimativas das somas de quadrados entre genótipos e pares de ambientes, com base em médias de parcelas  $(\theta_{jj'})$

Ambientes	2	3	4	5	6	7	8	9
1	5,834	6,714	2,266	2,390	9,526	5,954	12,054	12,446
2		0,680	1,514	1,604	1,906	0,250	2,360	3,046
3			3,154	1,574	2,566	0,350	5,130	6,026
4				1,246	5,124	1,954	4,114	4,244
5					5,176	1,084	6,904	7,736
6						2,786	3,836	5,480
7							3,580	4,446
8								0,356

## 2) Formação do Grupo I

Na formação do primeiro grupo são agrupados os ambientes com menor soma de quadrados da interação entre genótipos e pares de ambientes, que no exemplo sob análise corresponde aos ambientes 2 e 7, com  $SQGE_{27} = \theta_{27} = 0,250$ . A significância desta interação é testada por meio do teste F, da seguinte forma:

$$F = \frac{\theta_{jj'}/(\eta-1)(g-1)}{(QMR/r)}, \text{ sendo}$$

$$\theta_{jj'}/(\eta-1)(g-1) = QMGE_{jj'}. \text{ Então,}$$

$$F = \frac{(0,250/4)}{0,456/4} = 0,548ns, \text{ associado a 4 e 108 graus de liberdade.}$$

Portanto, os ambientes 2 e 7 podem ser agrupados porque apresentam interação G x E não significativa, pelo teste F a 5% de probabilidade.

Em seguida, avalia-se a possibilidade de inclusão de outro ambiente neste grupo inicial. Então, deve-se obter as estimativas das somas de quadrados entre genótipos e três ambientes, o que é feito com base na expressão  $\theta_{(jj'k)} = \frac{2}{\eta} S_{jj'k}$ , sendo  $S_{jj'k} = \theta_{jj'} + \theta_{jk} + \theta_{j'k}$ .

Então tem-se:

$$\theta_{(27)1} = \frac{2}{\eta} (\theta_{27} + \theta_{21} + \theta_{71})$$

$$\theta_{(27)1} = \frac{2}{3}(0,250 + 5,834 + 5,954) = 8,026$$

$$\theta_{(27)3} = \frac{2}{\eta}(\theta_{27} + \theta_{23} + \theta_{73})$$

$$\theta_{(27)3} = \frac{2}{3}(0,250 + 0,680 + 0,350) = \frac{2}{3}S_{237} = 0,853$$

$$\theta_{(27)4} = \frac{2}{3}(0,250 + 1,514 + 1,954) = 2,478$$

$$\theta_{(27)5} = \frac{2}{3}(0,250 + 1,604 + 1,084) = 1,958$$

$$\theta_{(27)6} = \frac{2}{3}(0,250 + 1,906 + 2,786) = 3,294$$

$$\theta_{(27)8} = \frac{2}{3}(0,250 + 2,360 + 3,580) = 4,126$$

$$\theta_{(27)9} = \frac{2}{3}(0,250 + 3,046 + 4,446) = 5,162$$

A menor  $SQGE_{(jj')k}$  é  $\theta_{(27)3} = 0,853$  e

$$QMGE_{(27)3} = \frac{\theta_{(27)3}}{(3-1)(5-1)} = \frac{0,853}{8} = 0,107$$

O teste de significância desta interação é

$$F = \frac{QMGE_{(27)3}}{(QMR/r)} = \frac{0,107}{0,114} = 0,938^{ns}, \text{ associado a } 8 \text{ e } 108 \text{ graus de liberdade.}$$

Como o valor de F é não significativo pode-se incluir o ambiente 3 no grupo inicial formado pelos ambientes 2 e 7.

A seguir avalia-se a possibilidade de inclusão de um quarto ambiente no grupo formado pelos ambientes 2, 3 e 7. Então, deve-se realizar a estimativa das somas de quadrados entre genótipos

$$\theta_{(jj'kl)} = \frac{2}{4} (S_{jj'k} + \theta_{jl} + \theta_{j'l} + \theta_{jk}),$$

sendo  $S_{jj'k} = \theta_{jj'} + \theta_{jk} + \theta_{j'k}$ . Então, tem-se:

$$\theta_{(237)1} = \frac{2}{4} (S_{237} + \theta_{12} + \theta_{13} + \theta_{17}), \text{ sendo}$$

$$S_{237} = 0,250 + 0,680 + 0,350 = 1,280$$

$$\theta_{(237)1} = \frac{2}{4} (1,280 + 5,834 + 6,714 + 5,954) = 9,891$$

$$\theta_{(237)4} = \frac{2}{4} (1,280 + 1,514 + 3,154 + 1,954) = 3,951$$

$$\theta_{(237)5} = \frac{2}{4} (S_{237} + \theta_{25} + \theta_{35} + \theta_{57})$$

$$\theta_{(237)5} = \frac{2}{4} (1,280 + 1,604 + 1,574 + 1,084) = \frac{2}{4} S_{2357} = 2,771$$

$$\theta_{(237)6} = \frac{2}{4} (S_{237} + \theta_{26} + \theta_{36} + \theta_{67})$$

$$\theta_{(237)6} = \frac{2}{4} (1,280 + 1,906 + 2,566 + 2,786) = 4,269$$

$$\theta_{(237)8} = \frac{2}{4} (S_{237} + \theta_{28} + \theta_{38} + \theta_{78})$$

$$\theta_{(237)8} = \frac{2}{4} (1,280 + 2,360 + 5,130 + 3,580) = 6,175$$

$$\theta_{(237)9} = \frac{2}{4} (S_{237} + \theta_{29} + \theta_{39} + \theta_{79})$$

$$\theta_{(237)9} = \frac{2}{4} (1,280 + 3,046 + 6,026 + 4,446) = 7,399$$

A menor  $SQGE_{(jj'k)l}$  é  $\theta_{(237)5} = 2,771$

Então, o ambiente 5 é o que proporciona menor interação G x E quando incorporado ao grupo formado pelos ambientes 2,3 e 7. O quadrado médio desta interação é:

$$QMGE_{(237)5} = \frac{\theta_{(237)5}}{(4-1)(5-1)} = \frac{2,771}{12} = 0,231$$

O teste de significância desta interação é

$$F = \frac{QMGE_{(237)5}}{(QMR/r)} = \frac{0,231}{0,114} = 2,026 (p < 0,05), \text{ associado a } 12 \text{ e } 108 \text{ graus de}$$

liberdade.

Como o valor de F é significativo não se deve incluir o ambiente 5 no grupo formado pelos ambientes 2, 3 e 7 porque a interação G x E torna-se significativa.

Pode-se formar novos grupos com os ambientes restantes, repetindo-se o procedimento anterior.

### 3) Formação do Grupo II

Na Tabela 7 são apresentadas as somas de quadrados entre genótipos e pares de ambientes, referentes aos ambientes que ainda não foram agrupados.

Tabela 7-Estimativas de somas de quadrados entre genótipos e pares de ambientes  $(\theta_{jj'})$ , com base em médias de parcelas, para os ambientes que não foram incluídos no primeiro grupo

Ambientes	4	5	6	8	9
1	2,266	2,390	9,526	12,054	12,446
4		1,246	5,124	4,114	4,244
5			5,176	6,904	7,736
6				3,836	5,480
8					0,356

Os ambientes 8 e 9 são os mais similares, com  $SQGE_{89} = \theta_{89} = 0,356$ . A significância desta interação é testada por meio da seguinte expressão:

$$F = \frac{\left[ \theta_{jj'}/(\eta-1)(g-1) \right]}{(QMR/r)} . \text{ Então,}$$

$$F = \frac{\left[ 0,356/(2-1)(5-1) \right]}{0,114} = \frac{0,089}{0,114} = 0,781 (p > 0,05), \text{ associado a 4 e 108 graus de liberdade.}$$

Como o valor de F é não significativo, pode ser então formado o Grupo II. Em seguida é avaliada a possibilidade de inclusão de outro ambiente no novo grupo formado. Para isto, deve-se obter as estimativas das somas de quadrados entre genótipos e grupos de três ambientes, as quais são:

$$\theta_{(89)1} = \frac{2}{\eta} (\theta_{89} + \theta_{18} + \theta_{19})$$

$$\theta_{(89)1} = \frac{2}{3} (0,356 + 12,054 + 12,446) = 16,571$$

$$\theta_{(89)4} = \frac{2}{\eta} (\theta_{89} + \theta_{48} + \theta_{49})$$

$$\theta_{(89)4} = \frac{2}{3}(0,356 + 4,114 + 4,244) = 5,809$$

$$\theta_{(89)5} = \frac{2}{\eta}(\theta_{89} + \theta_{58} + \theta_{59})$$

$$\theta_{(89)5} = \frac{2}{3}(0,356 + 6,904 + 7,736) = 9,997$$

$$\theta_{(89)6} = \frac{2}{\eta}(\theta_{89} + \theta_{68} + \theta_{69})$$

$$\theta_{(89)6} = \frac{2}{3}(0,356 + 3,836 + 5,480) = 6,448$$

O menor valor de soma de quadrados entre genótipos e três ambientes é

$$SQGE_{(89)4} = \theta_{(89)4} = 5,809 \text{ e o quadrado médio correspondente é}$$

$$QMGE_{(89)4} = \frac{5,809}{(3-1)(5-1)} = 0,726.$$

O valor da estatística F para testar a significância desta interação é

$$F = \frac{QMGE_{(89)4}}{(QMR/r)} = \frac{0,726}{0,114} = 6,37 (p < 0,05), \text{ associado com } 8 \text{ e } 108 \text{ graus de}$$

liberdade.

Como o teste F é significativo não se deve incluir o ambiente 4 no Grupo II porque a interação torna-se significativa.

#### 4) Formação do Grupo III

Na Tabela 8 estão apresentadas as somas de quadrados entre genótipos e pares de ambientes, para os ambientes que ainda não foram agrupados.

Tabela 8- Estimativas de somas de quadrados entre genótipos e pares de ambientes  $(\theta_{jj'})$ , com base em médias de parcelas, para os ambientes que não foram incluídos nos grupos anteriores (Grupo I e Grupo II)

Ambientes	4	5	6
1	2,266	2,390	9,526
4		1,246	5,124
5			5,176

O par de ambientes 4 e 5 é o que proporciona menor interação, com  $SQGE_{45} = \theta_{45} = 1,246$  e

$$QMGE_{45} = \frac{1,246}{(2-1)(5-1)} = 0,3115.$$

O teste de significância desta interação é

$$F = \frac{QMGE_{45}}{(QMR/r)} = \frac{0,3115}{0,114} = 2,73 (p < 0,05)$$

Como o teste F é significativo não se forma mais novos grupos.

Na Tabela 9 é apresentado o resumo da estratificação de ambientes.

Tabela 9-Grupos de ambientes que apresentam interação entre genótipos e ambientes não significativa

Grupo	Ambientes	QMGE	F
I	2, 3 e 7	0,107	0,938
II	8 e 9	0,089	0,781

Apesar de haver interação G x E para todos os genótipos e ambientes em conjunto, existem grupos de ambientes que não apresentam interação G x E significativa.

O processo apresentado estabelece grupos individuais que não apresentam ambientes em comum. O par de ambientes 5 e 7 apresenta interação com genótipos não significativa e poderia

ser considerado um outro grupo (Grupo III). Entretanto, neste caso, os grupos I e III apresentam o ambiente 7 em comum.

# CAPITULO 14

## Análise de Experimentos Multiambientes

### Visão geral de experimentos multiambiente

Experimentos multiambiente (EMA) são simplesmente experimentos desenvolvidos em múltiplos ambientes. Na pesquisa agronômica, os EMA são ferramentas padrão de pesquisa, uma vez que a experiência tem mostrado que os resultados obtidos raramente são repetíveis, porque eles dependem do ambiente (solo e clima).

Os EMA são ferramentas úteis para identificar o que é repetível ao longo de um conjunto de ambientes e para ajudar a entender as razões e os mecanismos da não repetibilidade ou interação com o ambiente. Entretanto, o delineamento de EMA efetivos ainda apresenta desafios.

As questões básicas estão relacionadas a quantos e quais ambientes. O paradigma de melhoramento participativo pode solucionar estas questões.

Nos EMA típicos, os tratamentos experimentais necessários para atingir os objetivos da pesquisa podem ser definidos e controlados pelo pesquisador. Os ambientes podem ser escolhidos de acordo com os objetivos, mas geralmente não podem ser controlados. Pode-se escolher locais com tipo de solo contrastante ou épocas de semeadura com clima contrastante, mas não se pode alocar aleatoriamente o tipo de solo ou o clima. Isto coloca limites no delineamento e interpretação.

O conceito de interação genótipos por ambientes ( $G \times E$ ) tem levado à definição de objetivos, delineamentos e análises de EMA. A interação entre o efeito de tratamentos e ambientes significa que a magnitude e a qualidade do efeito não são as mesmas em diferentes ambientes. Os objetivos da pesquisa determinam o delineamento, a análise de dados e a interpretação de resultados de um grupo de EMA.

No melhoramento de plantas, os objetivos da maioria dos EMA estão centrados nas interações  $G \times E$ . A avaliação e discussão da interação requer que variáveis respondam quantitativas, tais como produtividade, eficiência no uso de nutrientes minerais e severidade de doenças, sejam medidas para descrever o desempenho dos genótipos.

### Então, os objetivos dos EMA são:

- 1) Identificar quais genótipos são estáveis, ou seja, desempenham de forma consistente sobre uma faixa de ambientes.
- 2) Identificar genótipos localmente adaptados, ou seja, que apresentam bom desempenho em determinados ambientes.
- 3) Definir a faixa de ambientes dentro da qual um determinado genótipo está adaptado. Isto é o conceito de mega-ambientes no melhoramento de plantas.
- 4) Coletar evidências para a extração de resultados para além dos ambientes incluídos na pesquisa.

5) Entender as bases e os mecanismos envolvidos nas interações, de forma que possam ser explorados para ganhos futuros em novos ambientes.

6) Avaliar e minimizar os riscos do sistema de produção, principalmente em relação às variações de clima e doenças.

O trabalho envolvido no desenvolvimento de EMA é no mínimo o trabalho de um único experimento multiplicado pelo número de ambientes.

**Os principais passos envolvidos no desenvolvimento de EMA são:**

1) Especificar os objetivos o mais precisamente possível. Estes devem ser baseados na estratégia do programa, na literatura, em resultados já obtidos e nas ideias originais.

2) Formar a equipe. A equipe são os indivíduos e organizações que serão envolvidos no processo de experimentação.

3) Determinar os genótipos e ambientes a serem incluídos.

4) Decidir o que é necessário medir, incluindo dados sobre cada genótipo e cada ambiente.

5) Escolher a forma como a pesquisa será implementada, tal como, quem irá participar de que forma em cada estágio.

6) Escrever todos os detalhes do delineamento e decisões em um protocolo e revisá-lo.

7) Estar seguro de que as equipes, em cada local, têm o completo entendimento de todos os processos para implantação, manejo e medição no experimento e do manuseio de dados, de forma que as abordagens e métodos em cada ambiente sejam os mesmos e verdadeiramente comparáveis.

Observe que vários destes passos serão iterativos, ou seja, em vários pontos será necessário dar um passo atrás e revisitar um estágio anterior.

Os EMA bem delineados e desenvolvidos irão gerar dados de boa qualidade. Tais dados podem ser analisados a partir de diferentes abordagens e os resultados apresentados e interpretados de forma interessante.

A garantia de qualidade ao longo de todo o processo de delineamento, implementação e análise de dados dos EMA depende dos seguintes fatores: uso de procedimentos claros e testados, entendimento das responsabilidades de todos os envolvidos, boa coordenação de todo o empreendimento e boa documentação de todo o processo.

**Princípios básicos de experimentos multiambiente**

A precisão e o poder preditivo de experimentos de campo individuais são muito baixos. A ferramenta mais importante para prever o desempenho de cultivares dentro de uma população de ambientes alvo é o uso de experimentos multiambiente (EMA).

Novas cultivares são testadas em vários locais e ao longo de vários anos nos locais para amostrar a população de ambientes alvo. Médias de cultivares estimadas a partir de EMA são os melhores preditores do futuro desempenho das cultivares, mas elas sempre são estimadas com erro. Este erro pode ser minimizado com o aumento do número de repetições, locais e anos de teste de campo, mas a condução de EMA requer muito do tempo e dinheiro disponível para programas de melhoramento.

Então, é preciso alocar locais, anos e repetições de forma a maximizar a precisão e o poder preditivo do programa de avaliação de cultivares, para determinado recurso financeiro e tempo disponível.

O real propósito de um experimento de avaliação de cultivares é predizer o desempenho de novas cultivares nas condições de cultivo atuais e futuras dentro de uma população de ambientes alvo. A precisão de um experimento de avaliação e cultivares é determinada principalmente pelo número de repetições dentro e ao longo dos ambientes. A precisão relativa de diferentes experimentos pode ser comparada pelo seu erro padrão da média (EPM) ou diferença mínima significativa (DMS).

Modelos lineares que descrevem as medições obtidas de experimentos de campo podem ser descritos da seguinte forma:

Realiza-se a análise conjunta dos experimentos sobre locais e anos para estimar o desempenho médio de cultivares. São necessárias também estimativas da precisão das médias (EPM e DMS).

Para estimar a variância da média de uma cultivar deve-se utilizar um modelo estatístico que descreve os fatores ou fontes que contribuem para esta variância. O modelo mais simples para a análise de um EMA é:

$$Y_{ijkl} = \mu + e_k + b(e)_{j(k)} + g_i + ge_{ik} + \varepsilon_{ijkl}, \text{ em que:}$$

$Y_{ijkl}$  : medição da parcela  $l$  no ambiente  $k$ , bloco  $j$ , que contém o genótipo  $i$ ;

$\mu$  : média geral de todas as parcelas em todos os ambientes;

$e_k$  : efeito do ambiente (experimento)  $k$ ,  $k=1,2,\dots,a$ ;

$b(e)_{j(k)}$  : efeito da repetição  $j$  dentro do ambiente  $k$ ,  $j=1,2,\dots,r$ ;

$g_i$  : efeito do genótipo  $i$ ,  $i=1,2,\dots,g$ ;

$ge_{ik}$  : efeito da interação do genótipo  $i$  com o ambiente  $k$ ;

$\varepsilon_{ijkl}$  : efeito do resíduo de parcela,  $l=1,2,\dots,p$ .

Neste modelo, geralmente o efeito de genótipo é considerado fixo, ou seja, o interesse é estimar o desempenho de genótipos específicos no experimento. Ambientes e blocos são considerados fatores de efeitos aleatórios porque o interesse não está nas médias dos experimentos individuais per se, mas nos experimentos como uma amostra da população de ambientes alvo.

Em algumas situações, genótipos podem ser considerados como de efeito aleatório, se o objetivo do experimento é estimar variâncias genéticas ao invés de predizer o desempenho de cultivares. As interações GE são aleatórias neste modelo, porque a interação entre fatores fixo e aleatório é sempre aleatório. O termo GE contribui para o verdadeiro erro no teste de diferenças entre cultivares e para a variância das médias de cultivares.

De acordo com este modelo, a variância da média de uma cultivar é obtida por meio da seguinte expressão:

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_{ge}^2}{a} + \frac{\sigma_g^2}{ar}, \text{ em que:}$$

$\sigma_{\bar{Y}}^2$  : variância da média de uma cultivar;

$a$  : número de experimentos (ambientes);  
 $r$  : número de repetições por experimento.

Deve-se delinear os experimentos multiambiente procurando minimizar  $\sigma_{\bar{Y}}^2$  em programas de avaliação de cultivares. A minimização de  $\sigma_{\bar{Y}}^2$  leva a melhor predição do desempenho da cultivar no futuro e aumenta o ganho genético com o melhoramento.

A equação  $\sigma_{\bar{Y}}^2 = \frac{\sigma_{ge}^2}{a} + \frac{\sigma_g^2}{ar}$  pode ser usada para determinar o valor mínimo de  $\sigma_{\bar{Y}}^2$  que pode ser obtido com os recursos disponíveis para o melhorista.

Os componentes de variância podem ser estimados a partir de um quadro de análise de variância para um conjunto de experimentos completamente balanceado, ou seja, todas as cultivares testadas em todos os ambientes. Existem métodos para estimar componentes de variância para conjuntos de dados desbalanceados, mas requerem métodos estatísticos mais complexos. As esperanças de quadrados médios da análise de variância de experimentos multiambiente são funções lineares das variâncias dos fatores no modelo linear:

$$Y_{ijkl} = \mu + e_k + b(e)_{j(k)} + g_i + ge_{ik} + \varepsilon_{ijkl}$$

Estas esperanças de quadrados médios são apresentadas no quadro a seguir:

Quadro-Esperanças de quadrados médios (EQM) da análise de variância do modelo genótipos x ambientes, considerando todos os fatores como aleatórios

FV	QM	EQM
Ambientes (E)	-	-
Repetições dentro E [R (E)]	-	-
Genótipos (G)	$QM_G$	$\sigma^2 + r\sigma_{ge}^2 + ra\sigma_g^2$
Interação G x E	$QM_{GE}$	$\sigma^2 + r\sigma_{ge}^2$
Resíduo	$QM_R$	$\sigma^2$

Os componentes de variância podem ser estimados como funções dos quadrados médios da análise de variância, da seguinte forma:

$$\hat{\sigma}_{\varepsilon}^2 = QM_R$$

$$\hat{\sigma}_G^2 = (QM_G - QM_{GE}) / ra$$

$$\sigma_{GE}^2 = (QM_{GE} - QM_R) / r$$

Deve ser observado que estes componentes de variância têm erros padrões muito grandes e que devem ser usados apenas como um guia para o planejamento de experimentos multiambiente. Eles devem ser estimados apenas se o número de graus de liberdade do quadrado médio correspondente for igual ou maior que 50.

Sobre a possibilidade de dividir uma população de ambientes alvo em duas regiões para propósitos de melhoramento:

As análises anteriores são úteis para decidir quantos locais, anos e repetições são necessários para predizer o desempenho de um genótipo com precisão adequada. Existem muitas análises de interação genótipos x ambientes que podem ser utilizadas para agrupar os ambientes em grupos relativamente similares com base no desempenho dos genótipos. Dentre estas estão as análises AMMI, GGE e agrupamento.

### Análise de Experimentos Multiambiente

Considere um experimento genérico, desenvolvido no delineamento experimental blocos casualizados, com  $g$  genótipos e  $r$  repetições. Considere ainda que o referido experimento fosse instalado em  $l$  locais diferentes. O modelo estatístico para a análise individual de cada um dos locais/ambientes é o seguinte:

$$Y_{ij} = \mu + b_j + g_i + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : observação do  $j$ -ésimo bloco, referente ao  $i$ -ésimo genótipo;

$\mu$ : média geral;

$b_j$ : efeito do  $j$ -ésimo bloco, com  $j=1,2,\dots,r$ ;

$g_i$ : efeito do  $i$ -ésimo genótipo, com  $i=1,2,\dots,g$ ;

$\varepsilon_{ij}$ : erro aleatório associado à observação  $Y_{ij}$ , pressupondo que  $\varepsilon_{ij} \sim NID(0; \sigma^2)$ .

O modelo estatístico para a análise conjunta de experimentos multiambiente é o seguinte:

$$Y_{ijk} = \mu + (b/l)_{jk} + g_i + l_k + gl_{ik} + \varepsilon_{ijk}, \text{ em que:}$$

$Y_{ijk}$ : observação do  $k$ -ésimo local e  $j$ -ésimo bloco, referente ao  $i$ -ésimo genótipo;

$\mu$ : média geral;

$(b/l)_{jk}$ : efeito do  $j$ -ésimo bloco dentro do  $k$ -ésimo local, com  $j=1,2,\dots,r$ ;

$g_i$ : efeito do  $i$ -ésimo genótipo, com  $i=1,2,\dots,g$ ;

$l_k$ : efeito do  $k$ -ésimo local, com  $k=1,2,\dots,l$ ;

$gl_{ik}$ : efeito da interação entre o  $i$ -ésimo genótipo e o  $k$ -ésimo local;

$\varepsilon_{ijk}$ : erro aleatório associado à observação  $Y_{ijk}$ , pressupondo que  $\varepsilon_{ijk} \sim NID(0; \sigma^2)$ .

A metodologia de análise conjunta de experimentos multiambiente consta, basicamente, dos passos seguintes:

*Primeiro Passo:* Análise de Variância Individual por Local/Ambiente

O esquema de Análise de Variância é o seguinte:

FV	GL	QM	F
Blocos	$r-1$	$QMB$	
Genótipos	$g-1$	$QMG$	$QMG/QMR$
Resíduo	$(r-1)(g-1)$	$QMR$	
Total	$gr-1$		
Média	$\bar{Y}_{..}$		
CV%		$(100\sqrt{QMR})/\bar{Y}_{..}$	

*Segundo Passo:* Teste de Homogeneidade das Variâncias Residuais

Geralmente utiliza-se o teste de Hartley, cuja equação é:  $H = \frac{s_{\max}^2}{s_{\min}^2}$ .

*Terceiro Passo:* Análise de Variância Conjunta para os Locais/Ambientes

FV	GL	QM	F
Blocos / Local	$l(r-1)$	$QMB$	
Locais (L)	$l-1$	$QML$	*
Genótipos (G)	$g-1$	$QMG$	*
$G \times L$	$(g-1)(l-1)$	$QMGxL$	*
Resíduo	$l[(r-1)(g-1)]$	$QMR$	
Total	$grl-1$		
Média	$\bar{Y}...$		
CV%	$(100\sqrt{QMR})/\bar{Y}...$		

(\*) A estatística F varia de acordo com a natureza do modelo, se fixa, aleatória ou mista e deve ser determinada para cada situação.

*Quarto Passo:* Estimação de Componentes de Variância e de Componentes Quadráticos para os Parâmetros do Modelo da Análise de Variância Conjunta

A obtenção das estimativas de componentes de variância e de componentes quadráticos dos parâmetros do modelo e a realização de testes F dessas estimativas dependem da natureza do modelo estatístico adotado para a pesquisa. Por exemplo, podem ser adotados os seguintes modelos: i) Modelo Aleatório; ii) Modelo Fixo; iii) Modelo Misto, com efeito de local fixo; iv) Modelo Misto, com efeito de genótipo fixo.

Uma das metodologias de estimação dos parâmetros consiste na combinação entre os resultados da análise de variância e as expressões das esperanças de quadrados médios das fontes de variação. As esperanças dos quadrados médios, denotadas por E (QM), e as estatísticas para o teste F em função da natureza do modelo, para o modelo em consideração, são as seguintes:

Modelo 1: Efeitos Aleatórios: B, G, L, GL e R

FV	GL	E (QM)	F
Blocos/ Locais	$l(r-1)$	$\sigma^2 + g\sigma_b^2$	
Locais (L)			
Genótipos (G)	$l-1$	$\sigma^2 + r\sigma_{gl}^2 + g\sigma_b^2 + gr\sigma_l^2$	$\frac{QML+QMR}{QMB+QMGL}$
$G \times L$	$g-1$	$\sigma^2 + r\sigma_{gl}^2 + lr\sigma_g^2$	$QMG/QMGL$
Resíduo	$(g-1)(l-1)$		$QMGL/QMR$
	$[(r-1)(g-1)]l$	$\sigma^2 + r\sigma_{gl}^2$	
		$\sigma^2$	

Modelo 2: Efeitos Aleatórios: B e R; Efeitos Fixos: G, L e GL

FV	GL	E (QM)	F
Blocos/ Locais	$l(r-1)$	$\sigma^2 + g\sigma_b^2$	
Locais (L)			
Genótipos (G)	$l-1$	$\sigma^2 + g\sigma_b^2 + gr\phi_l$	$QML/QMB$
$G \times L$	$g-1$	$\sigma^2 + lr\phi_g$	$QMG/QMR$
Resíduo	$(g-1)(l-1)$	$\sigma^2 + r\phi_{gl}$	$QMGL/QMR$
	$[(r-1)(g-1)]l$		
		$\sigma^2$	

Modelo 3: Efeitos Aleatórios: B, L, GL e R; Efeito Fixo: G

FV	GL	E (QM)	F
Blocos/ Locais	$l(r-1)$	$\sigma^2 + g\sigma_b^2$	
Locais (L)			
Genótipos (G)	$l-1$	$\sigma^2 + g\sigma_b^2 + gr\sigma_l^2$	$QML/QMB$
$G \times L$	$g-1$	$\sigma^2 + r\ell\sigma_{gl}^2 + lr\phi_g$	$QMG/QMGL$
Resíduo	$(g-1)(l-1)$		$QMGL/QMR$
	$[(r-1)(g-1)]l$	$\sigma^2 + r\ell\sigma_{gl}^2$	
		$\sigma^2$	

$$\ell = g / (g-1)$$

Modelo 4: Efeitos Aleatórios: B, G, GL e R; Efeito Fixo: L

FV	GL	E (QM)	F
Blocos/ Locais	$l(r-1)$	$\sigma^2 + g\sigma_b^2$	
Locais (L)			
Genótipos (G)	$l-1$	$\sigma^2 + r\ell\sigma_{gl}^2 + g\sigma_b^2 + gr\phi_l$	$\frac{QML+QMR}{QMB+QMGL}$
$G \times L$	$g-1$	$\sigma^2 + lr\sigma_g^2$	$QMG/QMR$
Resíduo	$(g-1)(l-1)$		$QMGL/QMR$
	$[(r-1)(g-1)]l$	$\sigma^2 + r\ell\sigma_{gl}^2$	
		$\sigma^2$	

$$\ell = l(l-1)$$

As estimativas dos componentes de variância (valores  $\hat{\sigma}^2$ ), obtidas para os fatores de efeitos aleatórios, e dos componentes quadráticos (valores  $\hat{\phi}$ ), obtidos para os fatores de efeitos fixos, para os modelos anteriores, são obtidas por meio das seguintes expressões:

a) Para o fator Genótipo (G)

Modelo 1:  $\hat{\sigma}_g^2 = \frac{QMGL - QMGL}{lr};$

Modelo 2:  $\hat{\phi}_g = \frac{QMGL - QMR}{lr};$

Modelo 3:  $\hat{\phi}_g = \frac{QMGL - QMGL}{lr};$

Modelo 4:  $\hat{\sigma}_g^2 = \frac{QMGL - QMR}{lr};$

b) Para o fator Local (L)

Modelo 1:

$$\hat{\sigma}_l^2 = \frac{(QML + QMR) - (QMGL + QMB)}{gr};$$

Modelo 2:  $\hat{\phi}_l = \frac{QML - QMB}{gr};$

Modelo 3:  $\hat{\sigma}_l^2 = \frac{QML - QMB}{gr};$

Modelo 4:

$$\hat{\phi}_l = \frac{(QML + QMR) - (QMGL + QMB)}{gr};$$

c) Para a Interação Genótipos x Locais (G x L)

Modelo 1:  $\hat{\sigma}_{gl}^2 = \frac{QMGL - QMR}{r};$

Modelo 2:  $\hat{\phi}_{gl} = \frac{QMGL - QMR}{r};$

Modelo 3:  $\hat{\sigma}_{gl}^2 = \frac{(QMGL - QMR)}{r} \left( \frac{g-1}{g} \right);$

Modelo 4:  $\hat{\sigma}_{gl}^2 = \frac{(QMGL - QMR)}{r} \left( \frac{l-1}{l} \right).$

Os números de graus de liberdade associados às estatísticas F, cujas expressões contenham no numerador e, ou, no denominador dois ou mais quadrados médios, são obtidos pelas seguintes expressões:

Para  $F = \frac{QML + QMR}{QMB + QMGL}$ , que testa a significância do fator local no Modelo 1, os números

de graus de liberdade associados ao numerador e ao denominador são, respectivamente:

$$n_1 = \frac{(QML + QMR)^2}{\frac{(QML)^2}{l-1} + \frac{(QMR)^2}{l(r-1)(g-1)}} \text{ e}$$

$$n_2 = \frac{(QMB + QMGL)^2}{\frac{(QMB)^2}{l(r-1)} + \frac{(QMGL)^2}{(l-1)(g-1)}}.$$

As estimativas das correlações intraclassificadas são dadas pelas seguintes expressões:

$$\text{Modelo 1: } \rho = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_{gl}^2 + \hat{\sigma}^2};$$

$$\text{Modelo 2: } \rho = \frac{\hat{\phi}_g}{\hat{\phi}_g + \hat{\sigma}^2};$$

$$\text{Modelo 3: } \rho = \frac{\hat{\phi}_g}{\hat{\phi}_g + \hat{\sigma}_{gl}^2 + \hat{\sigma}^2};$$

$$\text{Modelo 4: } \rho = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}^2}.$$

Na obtenção dessas estimativas tem-se  $\hat{\sigma}^2 = QMR$ .

Observe que, para os fatores de efeitos aleatórios no modelo adotado são estimados os componentes de variância, que fornecem uma estimativa da variabilidade dos fatores, uma vez que os níveis estudados desses fatores são considerados como uma amostra aleatória de uma população de níveis. Por outro lado, para os fatores de efeitos fixos são estimados os componentes quadráticos, que são submetidos a testes de significância para avaliar se existe diferença significativa entre as médias dos níveis dos fatores.

No caso dos componentes de variância não são realizados testes de significância, mas são estimados para os mesmos variâncias e intervalos de confiança. Note que, embora a análise de variância seja realizada da maneira usual, dependendo da natureza do modelo (fixa, aleatória ou mista) as estimativas dos componentes de variância, dos componentes quadráticos e dos valores das estatísticas do teste F podem assumir resultados diferenciados, devido às diferentes composições das expressões das esperanças dos quadrados médios.

Considere agora que os experimentos desenvolvidos em diferentes locais foram realizados durante vários anos. Nesse caso, o modelo estatístico para a análise conjunta do grupo de experimentos é o seguinte:

$$Y_{ijkm} = \mu + G_i + A_k + L_m + (B/A)L_{jkm} + \\ + GA_{ik} + GL_{im} + AL_{km} + GAL_{ikm} + \varepsilon_{ijkm}, \text{ em que:}$$

$\mu$ : média geral;

$G_i$ ,  $A_k$  e  $L_m$ : efeitos de genótipos, anos e locais, respectivamente;

$(B/A)L_{jkm}$ : efeito de blocos dentro de anos e ambos dentro de locais;

$GA_{ik}$ ,  $GL_{im}$  e  $AL_{km}$ : efeitos das interações de primeira ordem entre genótipos e anos, genótipos e locais e anos e locais, respectivamente;

$GAL_{ikm}$ : efeito da interação de segunda ordem entre genótipos, anos e locais;

$\varepsilon_{ijkm}$ : erro aleatório, pressupondo que

$$\varepsilon_{ijkm} \square NID(0; \sigma^2)$$

Neste modelo têm-se:

$$i=1,2,\dots,g; j=1,2,\dots,r; k=1,2,\dots,a; m=1,2,\dots,l$$

A análise é realizada adotando-se os passos seguintes:

*Primeiro Passo:* Análises de Variância Individuais dos  $a \times l$  Ambientes

O esquema de Análise de Variância é o seguinte:

FV	GL	QM	F
Blocos	$r-1$	$QMB$	
Genótipos	$g-1$	$QMG$	$QMG/QMR$
Resíduo	$(r-1)(g-1)$	$QMR$	
Total	$gr-1$		
Média	$\bar{Y}_{..}$		
CV%	$(100\sqrt{QMR})/\bar{Y}_{..}$		

*Segundo Passo:* Teste de Homogeneidade de Variâncias Residuais

Geralmente utiliza-se o teste de Hartley, cuja equação é :

$$H = \frac{S_{\max}^2}{S_{\min}^2}$$

*Terceiro Passo:* Análise de Variância Conjunta do Grupo de Experimentos

O esquema da análise de variância conjunta é o seguinte:

FV	GL	SQ	QM
$(B / A) / L$	$(r-1)al$	SQB	QMB
Anos (A)	$a-1$	SQA	QMA
Locais (L)	$l-1$	SQL	QML
Genótipos (G)	$g-1$	SQG	QMG
$G \times A$		SQGA	QMGA
$G \times L$	$(g-1)(a-1)$	SQGL	QMGL
$A \times L$	$(g-1)(l-1)$	SQAL	QMAL
$G \times A \times L$	$(a-1)(l-1)$	SQGAL	QMGAL
Resíduo	$(g-1)(a-1)(l-1)$	SQR	QMR
	$(r-1)(g-1)al$		

*Quarto Passo:* Estimação dos Componentes de Variância e dos Componentes Quadráticos das Fontes de Variação da Análise de Variância Conjunta

As expressões das esperanças de quadrados médios [E (QM)] são obtidas com base nas pressuposições feitas a respeito da natureza dos efeitos (fixa, aleatória ou mista) dos fatores envolvidos no modelo.

Toda interação envolvendo pelo menos um fator de efeito aleatório é considerada como aleatória. Desta forma, as alternativas de modelos para essas estimativas são as seguintes:

Modelo 1: Efeitos Aleatórios: anos, locais, genótipos, blocos e resíduo

FV	E (QM)
(B / A) / L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + r\sigma_{gal}^2 + g\sigma_b^2 + rg\sigma_{al}^2 + rl\sigma_{ga}^2 + rgl\sigma_a^2$
Locais (L)	$\sigma^2 + r\sigma_{gal}^2 + g\sigma_b^2 + rg\sigma_{al}^2 + ra\sigma_{gl}^2 + rga\sigma_l^2$
Genótipos (G)	
G x A	$\sigma^2 + r\sigma_{gal}^2 + ra\sigma_{gl}^2 + rl\sigma_{ga}^2 + ral\sigma_g^2$
G x L	$\sigma^2 + r\sigma_{gal}^2 + rl\sigma_{ga}^2$
A x L	$\sigma^2 + r\sigma_{gal}^2 + ra\sigma_{gl}^2$
G x A x L	$\sigma^2 + r\sigma_{gal}^2 + g\sigma_b^2 + rg\sigma_{al}^2$
Resíduo	
	$\sigma^2 + r\sigma_{gal}^2$
	$\sigma^2$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	$(QMG+QMGAL)/(QMGA+QMGL)$
Anos (A)	$(QMA+QMGAL)/(QMGA+QMAL)$
Locais (L)	$(QML+QMGAL)/(QMGL+QMAL)$
G x A	$QMGA/QMGAL$
G x L	$QMGL/QMGAL$
A x L	$(QMAL+QMR)/(QMB+QMGAL)$
G x A x L	$QMGAL/QMR$

Modelo 2: Efeitos aleatórios: locais, genótipos e blocos; efeito fixo: anos.

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos(A)	$\sigma^2 + r\alpha\sigma_{gal}^2 + g\sigma_b^2 + rga\sigma_{al}^2 + rl\alpha\sigma_{ga}^2 + rgl\phi_a$
Locais (L)	$\sigma^2 + g\sigma_b^2 + ra\sigma_{gl}^2 + rga\sigma_l^2$
Genótipos (G)	$\sigma^2 + ra\sigma_{gl}^2 + ral\sigma_g^2$
G x A	$\sigma^2 + r\alpha\sigma_{gal}^2 + rl\alpha\sigma_{ga}^2$
G x L	$\sigma^2 + ra\sigma_{gl}^2$
A x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + g\sigma_b^2 + rga\sigma_{al}^2$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$

$$\alpha = a/(a-1)$$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	QMG/QMGL
Anos (A)	(QMA + QMGAL)/(QMGA+QMAL)
Locais(L)	(QML+QMR)/(QMB+QMGL)
G x A	QMGA/QMGAL
G x L	QMGL/QMR
A x L	(QMAL+QMR)/(QMB+QMGAL)
G x A x L	QMGAL/QMR

Modelo 3: Efeitos aleatórios: anos, genótipos e blocos; fixo: locais

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + g\sigma_b^2 + rl\sigma_{ga}^2 + rgl\sigma_a^2$
Locais(L)	$\sigma^2 + r\alpha\sigma_{gal}^2 + g\sigma_b^2 + rg\alpha\sigma_{al}^2 + ra\alpha\sigma_{gl}^2 + rga\phi_l$
Genótipos (G)	$\sigma^2 + rl\sigma_{ga}^2 + ral\sigma_g^2$
G x A	$\sigma^2 + rl\sigma_{ga}^2$
G x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + r\alpha\sigma_{gl}^2$
A x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + g\sigma_b^2 + rg\alpha\sigma_{al}^2$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$

$$\alpha = l/(l-1)$$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	QMG/QMGA
Anos (A)	(QMA+QMR)/(QMB+QMGA)
Locais(L)	(QML+QM GAL)/(QM GL+QM AL)
G x A	QMGA/QMR
G x L	QMGL/QM GAL
A x L	(QM AL+QMR)/QMB+QM GAL)
G x A x L	QM GAL/QMR

Modelo 4: Efeitos aleatórios: genótipos e blocos; efeitos fixos: anos e locais

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + g\sigma_b^2 + rl\lambda\sigma_{ga}^2 + rgl\phi_a$
Locais (L)	$\sigma^2 + g\sigma_b^2 + ra\theta\sigma_{gl}^2 + rga\phi_l$
Tratamentos (G)	$\sigma^2 + ral\sigma_g^2$
G x A	$\sigma^2 + rl\lambda\sigma_{ga}^2$
G x L	$\sigma^2 + ra\theta\sigma_{gl}^2$
A x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + g\sigma_b^2 + rg\phi_{al}$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$
$\alpha = \theta\lambda$	$\theta = l/(l-1)$
	$\lambda = a/(a-1)$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Tratamentos (G)	QMG/QMR
Anos (A)	(QMA+QMR)/(QMB+QMGA)
Locais(L)	(QML+QMR)/(QMGL+QMB)
G x A	QMGA/QMR
G x L	QMGL/QMR
A x L	(QMAL+QMR)/(QMB+QM GAL)
G x A x L	QMGAL/QMR

Modelo 5: Efeitos aleatórios: locais e blocos; efeitos fixos: anos e genótipos

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + g\sigma_b^2 + rg\theta\sigma_{al}^2 + rgl\phi_a$
Locais (L)	$\sigma^2 + g\sigma_b^2 + rga\sigma_l^2$
Genótipos (G)	$\sigma^2 + ra\lambda\sigma_{gl}^2 + ral\phi_g$
G x A	$\sigma^2 + r\alpha\sigma_{gal}^2 + rl\phi_{ga}$
G x L	$\sigma^2 + ra\lambda\sigma_{gl}^2$
A x L	$\sigma^2 + g\sigma_b^2 + rg\theta\sigma_{al}^2$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$

$$\alpha = \theta\lambda \quad \theta = a/(a-1) \quad \lambda = g/(g-1)$$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	QMG/QMGL
Anos (A)	QMA/QMAL
Locais (L)	QML/QMB
G x A	QMGA/QMGAL
G x L	QMGL/QMR
A x L	QMAL/QMB
G x A x L	QMGAL/QMR

Modelo 6: Efeitos aleatórios: anos e blocos; efeitos fixos: locais e genótipos

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + g\sigma_b^2 + rgl\sigma_a^2$
Locais (L)	$\sigma^2 + g\sigma_b^2 + rg\theta\sigma_{al}^2 + rga\phi_l$
Tratamentos (G)	$\sigma^2 + rl\lambda\sigma_{ga}^2 + ral\phi_g$
G x A	$\sigma^2 + rl\lambda\sigma_{ga}^2$
G x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + ra\phi_{gl}$
A x L	$\sigma^2 + g\sigma_b^2 + rg\theta\sigma_{al}^2$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$
$\alpha = \theta\lambda$	$\theta = l/(l-1)$
	$\lambda = g/(g-1)$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Tratamentos (G)	QMG/QMGA
Anos (A)	QMA/QMB
Locais (L)	QML/QMAL
G x A	QMGA/QMR
G x L	QMGL/QMGAL
A x L	QMAL/QMB
G x A x L	QMGAL/QMR

Modelo 7: Efeitos fixos: anos, locais e genótipos; efeito aleatório: blocos

FV	GL	E (QM)
(B/A)/L	(r-1)al	$\sigma^2 + g\sigma_b^2$
Anos (A)	a-1	$\sigma^2 + g\sigma_b^2 + rgl\phi_a$
Locais (L)	l-1	$\sigma^2 + g\sigma_b^2 + rga\phi_l$
Genótipos (G)	g-1	$\sigma^2 + ral\phi_g$
G x A	(g-1)(a-1)	$\sigma^2 + rl\phi_{ga}$
G x L	(g-1)(l-1)	$\sigma^2 + ra\phi_{gl}$
A x L	(a-1)(l-1)	$\sigma^2 + g\sigma_b^2 + rg\phi_{al}$
G x A x L	(g-1)(a-1)(l-1)	$\sigma^2 + r\phi_{gal}$
Resíduo	(r-1)(g-1)al	$\sigma^2$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	QMG/QMR
Anos (A)	QMA/QMB
Locais (L)	QML/QMB
G x A	QMGA/QMR
G x L	QMGL/QMR
A x L	QMAL/QMB
G x A x L	QMGAL/QMR

Modelo 8: Efeitos aleatórios: anos, locais e blocos; efeito fixo: genótipos

FV	E (QM)
(B/A)/L	$\sigma^2 + g\sigma_b^2$
Anos (A)	$\sigma^2 + g\sigma_b^2 + rg\sigma_{al}^2 + rgl\sigma_a^2$
Locais (L)	$\sigma^2 + g\sigma_b^2 + rg\sigma_{al}^2 + rga\sigma_l^2$
Genótipos (G)	$\sigma^2 + r\alpha\sigma_{gal}^2 + ra\alpha\sigma_{gl}^2 + rl\alpha\sigma_{ga}^2 + ral\phi_g$
G x A	$\sigma^2 + r\alpha\sigma_{gal}^2 + rl\alpha\sigma_{ga}^2$
G x L	$\sigma^2 + r\alpha\sigma_{gal}^2 + ra\alpha\sigma_{gl}^2$
A x L	$\sigma^2 + g\sigma_b^2 + rg\sigma_{al}^2$
G x A x L	$\sigma^2 + r\alpha\sigma_{gal}^2$
Resíduo	$\sigma^2$

$$\alpha = g/(g-1)$$

As estatísticas para os testes F das fontes de variação são representadas como a seguir:

FV	F
Genótipos (G)	(QMG+QMGAL)/(QMGL+QMGA)
Anos (A)	QMA/QMAL
Locais (L)	QML/QMAL
G x A	QMGA/QMGAL
G x L	QMGL/QMGAL
A x L	QMAL/QMB
G x A x L	QMGAL/QMR

Exemplo de Aplicação: Considere os dados a seguir, referentes às médias de produtividade de cinco cultivares em nove ambientes, bem como os resultados das estimativas dos quadrados médios de resíduos das análises de variâncias individuais. Os experimentos foram desenvolvidos no delineamento blocos casualizados com quatro repetições (Extraído de Cruz et al., 2014)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	$\bar{Y}_i$
C1	2,0	6,4	7,3	3,8	3,1	5,9	5,4	8,3	7,9	50,1
C2	3,7	6,7	8,4	3,6	4,1	8,1	5,8	6,7	5,5	52,6
C3	3,1	6,6	8,1	3,8	4,7	6,3	6,3	7,1	5,7	51,7
C4	2,4	5,1	8,6	2,8	4,2	5,3	5,9	5,2	4,5	45,0
C5	4,9	4,9	6,3	3,8	4,0	3,8	4,3	4,4	3,8	40,2
$\bar{Y}_{.j}$	16,1	30,7	38,7	17,8	20,1	29,4	27,7	31,7	27,4	239,6
QMR	0,450	0,520	0,730	0,220	0,350	0,580	0,410	0,610	0,234	

1-Fazer a análise de variância individual (em cada ambiente), para avaliar a precisão relativa de cada experimento e a homogeneidade das variâncias residuais. Nesse caso deve ser feita a recomposição da análise de variância, uma vez que se dispõe apenas das médias de cada cultivar em cada ambiente e dos quadrados médios residuais de cada ambiente (experimento).

Neste caso, os quadrados médios de tratamentos (cultivares) são obtidos por meio da expressão seguinte:

$$QM(C/A_j) = \frac{r}{c-1} \left( \sum_i Y_{ij}^2 - \frac{1}{c} \bar{Y}_{.j}^2 \right).$$

Por exemplo, para o ambiente 1 se tem:

$$QM(C/A_1) = \frac{4}{5-1} \left[ (2,0^2 + \dots + 4,9^2) - \frac{(16,1)^2}{5} \right] = 5,228$$

Os cálculos para os outros ambientes são feitos de forma análoga, resultando no quadro seguinte:

FV	GL	Quadrados Médios nos Ambientes								
		A1	A2	A3	A4	A5	A6	A7	A8	A9
Blocos	3	-	-	-	-	-	-	-	-	-
Cultivares	4	5,228**	2,132*	3,572*	0,752*	1,348*	9,768**	2,332**	9,612**	9,688**
Resíduo	12	0,450	0,520	0,730	0,220	0,350	0,580	0,410	0,610	0,234
Média		3,22	6,14	7,74	3,56	4,02	5,88	5,54	6,34	5,48
CV(%)		20,83	11,74	11,04	13,17	14,72	12,95	11,56	12,32	8,83
$\hat{\sigma}_c^2$		1,195	0,403	0,710	0,133	0,250	2,297	0,481	2,250	2,363

\*\* e \* : significativo, pelo teste F, a 1 e 5% de probabilidade, respectivamente.

2-Teste de homogeneidade das variâncias residuais, uma vez que se recomenda fazer a análise conjunta apenas dos ambientes cujas variâncias residuais sejam homogêneas. A homogeneidade das variâncias pode ser avaliada por meio do teste de Pearson e Hartley, cuja expressão é dada por:

$$H = \frac{\hat{\sigma}_{\max}^2}{\hat{\sigma}_{\min}^2}. \text{ O teste é aplicado para } k \text{ variâncias independentes, o valor obtido é}$$

comparado com o tabelado para  $k$  e  $n'$  graus de liberdade, sendo  $n'$  o número de graus de liberdade associado à variância residual de cada experimento. Para o exemplo considerado obtém-se:

$$H = \frac{0,730}{0,220} = 3,318$$

Como  $H = 3,318 < H_{0,01(9;12)} = 9,5$ , não se rejeita a hipótese de que existe homogeneidade entre as variâncias residuais. Então, pode-se realizar a análise conjunta.

3-Realizar a análise de variância conjunta, cujas estimativas de quadrados médios são obtidas da seguinte forma:

$$\begin{aligned} QMC &= \frac{r}{a(c-1)} \left[ \sum_i Y_{i..}^2 - \frac{1}{c} (\bar{Y}_{..})^2 \right] = \\ &= \frac{4}{9 \times 4} \left[ (50,1^2 + \dots + 40,2^2) - \frac{239,6^2}{5} \right] = 12,119 \end{aligned}$$

$$\begin{aligned} QMA &= \frac{r}{c(a-1)} \left[ \sum_j Y_{.j}^2 - \frac{1}{a} (\bar{Y}_{..})^2 \right] = \\ &= \frac{4}{5 \times 8} \left[ (16,1^2 + \dots + 27,4^2) - \frac{239,6^2}{9} \right] = 42,886 \end{aligned}$$

$$\begin{aligned} QMC, A &= \frac{r}{ac-1} \left[ \sum_i \sum_j Y_{ij}^2 - \frac{1}{ac} (\bar{Y}_{..})^2 \right] = \\ &= \frac{4}{44} \left[ (2,0^2 + \dots + 3,8^2) - \frac{239,6^2}{45} \right] = 11,837 \end{aligned}$$

$$\begin{aligned} SQCA &= SQC, A - (SQC + SQA) = \\ &= 44 \times 11,837 - (4 \times 12,119 + 8 \times 42,886) = 129,264 \end{aligned}$$

$$QMCA = \frac{SQCA}{(c-1)(a-1)} = \frac{129,264}{32} = 4,039.$$

O quadrado médio do resíduo pode ser obtido pela média aritmética simples dos quadrados médios residuais em cada ambiente, quando os graus de liberdade são os mesmos em todos os ambientes, sendo dado por:

$$QMR = \frac{1}{a} \sum_j QMR/A_j = \frac{1}{9} (0,450 + \dots + 0,234) = 0,456$$

Duas relações importantes são:

$$\begin{aligned} SQCA + SQC &= \sum_j SQC/A_j = \\ &= 4(5,228 + 2,132 + \dots + 9,688) = 177,728 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_c^2 + \hat{\sigma}_{ca}^2 &= \frac{1}{a} \sum_j \hat{\sigma}_{c/a,j}^2 = \\ &= \frac{1}{9} (1,195 + 0,403 + \dots + 2,363) = 1,120 \end{aligned}$$

O resultado completo da análise de variância conjunta está representado no quadro a seguir:

FV	GL	QM	F
Blocos/Ambientes	27	-	-
Ambientes (A)	8	QMA=42,886	-
Cultivares (C)	4	QMC=12,119	QMC/QMCA=3,00*
C x A	32	QMCA=4,039	QMCA/QMR=8,86**
Resíduo	108	QMR=0,456	

\*,\*\*: significativo, a 5% e 1% de probabilidade, respectivamente.

Considerando, neste caso, todos os efeitos do modelo, exceto a média, como aleatórios devem ser estimados os componentes de variância das fontes de variação de interesse, como a seguir:

$$\hat{\sigma}_c^2 = \frac{QMC - QMCA}{ar} = \frac{12,119 - 4,039}{9 \times 4} = 0,224$$

$$\hat{\sigma}_{ca}^2 = \frac{QMCA - QMR}{r} = \frac{4,039 - 0,456}{4} = 0,896.$$

Para avaliar os erros associados aos componentes de variância é necessário obter as estimativas das variâncias das estimativas de variâncias. A expressão geral para o estimador da variância de um estimador de variância é a seguinte:

$\sigma^2(\hat{\sigma}_X^2) = \frac{2(\sigma_X^2)^2}{v+2}$ , em que  $v$ : número de graus de liberdade associado ao estimador do componente de variância da variável  $X$ .

Para  $X = \frac{Q_1 - Q_2}{r}$ , o estimador de sua variância será:

$$\hat{\sigma}^2(\hat{\sigma}_X^2) = \hat{\sigma}^2\left[\frac{(Q_1 - Q_2)}{r}\right] = \frac{1}{r^2}[\hat{\sigma}^2(Q_1) + \hat{\sigma}^2(Q_2)]$$

Uma vez que  $Q_1$  e  $Q_2$  são independentes, tem-se que:

$$\hat{\sigma}^2(\hat{\sigma}_X^2) = \frac{2}{r^2} \left( \frac{Q_1^2}{v_1+2} + \frac{Q_2^2}{v_2+2} \right).$$

Por exemplo, o componente  $\sigma_c^2$ , que é estimado por  $\hat{\sigma}_c^2 = \frac{QMC - QMCA}{ra}$ , tem o estimador de sua variância obtido por meio de:

$$\hat{\sigma}^2(\hat{\sigma}_c^2) = \frac{2}{r^2 a^2} \left( \frac{QMC^2}{v_c+2} + \frac{QMCA^2}{v_{ca}+2} \right), \text{ logo, a estimativa da}$$

variância da estimativa desse componente é:

$$\hat{\sigma}^2(\hat{\sigma}_c^2) = \frac{2}{4^2 x 9^2} \left[ \frac{(12,119)^2}{4+2} + \frac{(4,039)^2}{32+2} \right] = 0,03816$$

A estimativa do componente  $\sigma_{ca}^2$ , cuja expressão é  $\hat{\sigma}_{ca}^2 = \frac{QMCA - QMR}{r}$ , tem o estimador de sua variância obtido por meio de:

$$\hat{\sigma}^2(\hat{\sigma}_{ca}^2) = \frac{2}{r^2} \left( \frac{QMCA^2}{v_{ca}+2} + \frac{QMR}{v_r+2} \right), \text{ logo a estimativa da variância da estimativa deste}$$

componente é:

$$\hat{\sigma}^2(\hat{\sigma}_{ca}^2) = \frac{2}{4^2} \left[ \frac{(4,039)^2}{32+2} + \frac{(0,456)^2}{108+2} \right] = 0,06021.$$

Os erros padrões das estimativas dos componentes anteriores são:

$$s(\hat{\sigma}_c^2) = \sqrt{\hat{\sigma}^2(\hat{\sigma}_c^2)} = \sqrt{0,03816} = 0,19535 \text{ e}$$

$$s(\hat{\sigma}_{ca}^2) = \sqrt{\hat{\sigma}^2(\hat{\sigma}_{ca}^2)} = \sqrt{0,06021} = 0,24538 .$$

Para avaliar o erro associado às estimativas de variância (componentes de variância), ou seja, para avaliar a precisão dessas estimativas, pode-se obter também o intervalo de confiança por meio da expressão geral seguinte:

$$IC_{(1-\alpha)100\%} : \frac{v_X \hat{\sigma}_X^2}{\chi_{\alpha/2}^2} < \sigma_X^2 < \frac{v_X \hat{\sigma}_X^2}{\chi_{1-\alpha/2}^2}, \text{ em que:}$$

$\hat{\sigma}_X^2$  : estimativa da variância entre os níveis da variável  $X$  ;

$v_X$  : número de graus de liberdade associados à estimativa, que pode ser obtido pelo método de Satterthwaite;

$\chi_{\alpha/2}^2$  e  $\chi_{1-\alpha/2}^2$  : valores do modelo de distribuição de probabilidade  $\chi^2$  para  $v_X$  graus de liberdade.

Admitindo que o número de graus de liberdade associado a  $\hat{\sigma}_X^2$ , a ser obtido, é  $n$ , tem-se, pelo método de Satterthwaite:

$$n = \frac{\left(\hat{\sigma}_X^2\right)^2}{\frac{a_1^2(Q_1)^2}{v_1+2} + \frac{a_2^2(Q_2)^2}{v_2+2}}.$$

Se  $a_1 = a_2 = 1$ , tem-se que  $\hat{\sigma}_X^2$  estará associado a  $v_1$  e  $v_2$  graus de liberdade, de acordo com a distribuição de  $\chi^2$ , e, o número de graus de liberdade a ser obtido será dado por:

$$n' = \frac{\left(\hat{\sigma}_X^2\right)^2}{\frac{\left(Q_1\right)^2 + \left(Q_2\right)^2}{v_1+2 + v_2+2}} = \frac{\left(Q_1 - Q_2\right)^2}{\frac{\left(Q_1\right)^2}{v_1+2} + \frac{\left(Q_2\right)^2}{v_2+2}}.$$

Para o exemplo em consideração, o número de graus de liberdade associado ao componente de variância  $\hat{\sigma}_c^2 = \frac{QMC - QMCA}{ra}$  é obtido por meio da expressão:

$$n' = \frac{\left(QMC - QMCA\right)^2}{\frac{\left(QMC\right)^2 + \left(QMCA\right)^2}{v_c+2 + v_{ca}+2}}, \text{ em que:}$$

$v_c$ : número de graus de liberdade associado à fonte de variação cultivares (C);

$v_{ca}$ : número de graus de liberdade associado à fonte de variação interação cultivares x ambientes (CA).

Então, o número de graus de liberdade associado a  $\hat{\sigma}_c^2$  será:

$$n' = \frac{\left(12,119 - 4,039\right)^2}{\frac{\left(12,119\right)^2 + \left(4,039\right)^2}{4+2 + 32+2}} = 2,62 \approx 3.$$

O intervalo de confiança, com grau de confiança de 95%, para  $\hat{\sigma}_c^2$  é dado por:

$$IC_{(1-0,05)100\%} : \frac{n'\hat{\sigma}_c^2}{\chi_{[0,05/2;3]}^2} < \sigma_c^2 < \frac{n'\hat{\sigma}_c^2}{\chi_{[(1-0,05)/2;3]}^2}$$

$$IC_{95\%} : \frac{3(0,224)}{9,348} < \sigma_c^2 < \frac{3(0,224)}{0,2158}$$

$$IC_{95\%} : 0,07189 < \sigma_c^2 < 3,11399.$$

O número de graus de liberdade associado ao componente de variância

$$\hat{\sigma}_{ca}^2 = \frac{QMCA - QMR}{r}$$

é obtido por meio da expressão:

$$n' = \frac{\frac{(QMCA - QMR)^2}{(QMCA)^2 + (QMR)^2}}{v_{ca} + 2 + v_r + 2}, \text{ em que:}$$

$v_r$ : número de graus de liberdade associado à fonte de variação resíduo (R);

$v_{ca}$ : número de graus de liberdade associado à fonte de variação interação cultivares x ambientes (CA).

Então, o número de graus de liberdade associado a  $\hat{\sigma}_{ca}^2$  será:

$$n' = \frac{\frac{(4,039 - 0,456)^2}{(4,039)^2 + (0,456)^2}}{32+2 + 108+2} = 26,65 \approx 27.$$

O intervalo de confiança, com grau de confiança de 95%, para  $\hat{\sigma}_{ca}^2$  é dado por:

$$IC_{(1-0,05)100\%} : \frac{n'\hat{\sigma}_{ca}^2}{\chi_{[0,05/2;27]}^2} < \sigma_{ca}^2 < \frac{n'\hat{\sigma}_{ca}^2}{\chi_{[(1-0,05)/2;27]}^2}$$

$$IC_{95\%} : \frac{27(0,896)}{43,195} < \sigma_{ca}^2 < \frac{27(0,896)}{14,573}$$

$$IC_{95\%} : 0,56006 < \sigma_{ca}^2 < 1,66006.$$

Note que quanto menor for o valor da estimativa da variância da estimativa de variância obtida, menor será o seu erro padrão e mais estreito será o seu intervalo de confiança. Consequentemente, maior será a confiabilidade na interpretação dos resultados obtidos e nas inferências realizadas.

# CAPITULO 15

## Metodologia de Modelos Mistos

### Modelo linear fixo

Considere o modelo linear fixo geral:

$$y = X\beta + e$$

$y$ : vetor de respostas observadas, de dimensão  $n \times 1$ ;

$X$ : matriz de incidência, de dimensão  $n \times p$ , cuja estrutura relaciona as respostas observadas aos fatores de efeitos fixos do modelo (constante, genótipos, blocos, locais, anos, épocas, etc.);

$\beta$ : vetor de efeitos fixos, de dimensão  $p \times 1$ ;

$e$ : vetor de erros aleatórios, de dimensão  $n \times 1$ .

O modelo deve ser ajustado de forma a minimizar:

$$|e| = |y - X\beta|$$

Para isto utiliza-se o sistema de equações normais do método dos quadrados mínimos ordinários (Ordinary Least Square-OLS):

$$X'X\beta = X'y$$

Tem-se também o estimador de quadrados mínimos generalizado:

$$\beta^0 = (X'X)^{-1} X'y$$

(Infinitas soluções)

Pode-se ainda utilizar o melhor estimador linear não viesado (Best Linear Unbiased Estimator-BLUE):

$$C'\beta^0$$

(Melhor Estimador Linear Não Viesado)

Melhor: a variância é mínima;

Estimador Linear: é uma função linear dos dados;

Não Viesado: o valor esperado do estimador é igual ao parâmetro. Por exemplo,  
 $E[BLUE(g)] = g$ .

Exemplo 1: Considere um experimento no DBC com três genótipos (Efeitos Aleatórios) e duas repetições (Efeitos Fixos)

Genótipos	Bloco 1	Bloco 2
G1	4,45	5,00
G2	4,61	5,82
G3	5,27	5,79

$$y_{ij} = \mu + g_i + b_j + e_{ij}$$

$y_{ij}$ : observação da parcela que recebeu o genótipo  $i$  no bloco  $j$ ;

$\mu$ : média geral, constante;

$g_i$ : efeito do genótipo  $i$ ,  $g_i \sim N(0, \sigma_g^2)$ ;

$b_j$ : efeito do bloco  $j$ ;

$e_{ij}$ : erro experimental (ou resíduo) associado a  $y_{ij}$ ,  $e_{ij} \sim N(0, \sigma_e^2)$ .

Modelo Linear Fixo e Estimação de Efeitos para Exemplo 1:

$$y = X\beta + e$$

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{12} \\ y_{22} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ b_1 \\ b_2 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ e_{31} \\ e_{12} \\ e_{22} \\ e_{32} \end{bmatrix}$$

$$\beta^0 = (X'X)^{-1} X'y$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} y_{..} \\ y_{.1} \\ y_{.2} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix}$$

(Infinitas soluções: restrições  $\sum b_j = 0; \sum g_i = 0$ )

$$BLUE \begin{bmatrix} \hat{\mu} + \bar{b} + \hat{g}_1 \\ \hat{\mu} + \bar{b} + \hat{g}_2 \\ \hat{\mu} + \bar{b} + \hat{g}_3 \end{bmatrix} = \begin{bmatrix} y_{1\cdot}/r \\ y_{2\cdot}/r \\ y_{3\cdot}/r \end{bmatrix} = \begin{bmatrix} 4,725 \\ 5,215 \\ 5,530 \end{bmatrix}$$

(Invariante)

Estimação de Componentes de Variância para Exemplo 1:

FV	GL	QM	E(QM)
Blocos	1	0,8664	
Genótipos	2	0,3291	$\sigma_e^2 + 2\sigma_g^2$
Erro	2	0,0761	$\sigma_e^2$

Método dos Momentos ou da Análise de Variância:

$$\hat{\sigma}_g^2 = \frac{QM_{Genotipos} - QM_{Erro}}{r} = \frac{0,3291 - 0,0761}{2} = 0,1265$$

$$\hat{\sigma}_F^2 = \hat{\sigma}_g^2 + \frac{\hat{\sigma}_e^2}{r} = 0,1265 + \frac{0,0761}{2} = 0,16455$$

**Modelo linear misto**

Considere agora o modelo linear misto geral:

$$Y = X\beta + Zg + e; g \perp N(0, G) e \perp N(0, R)$$

(Observe que é o modelo fixo estendido, para acomodar o termo referente ao fator de efeitos aleatórios,  $Zg$ ). Neste caso, tem-se que:

$$Y \perp N_p \left[ E(Y) = X\beta; Var(Y) = ZGZ' + R = V \right]$$

A estimação e predição por meio de modelo misto é feita da seguinte forma:

Solução dos Efeitos Fixos (Estimação):

$$BLUE(\beta) = \left( X'V^{-1}X \right)^{-1} X'V^{-1}y$$

Distribuição Conjunta de  $Y$  e  $g$ :

$$f(Y, g) = f(y|g)f(g)$$

Solução dos Efeitos Aleatórios (Predição):

$$E(g|Y) = BLUP(g) = GZ'V^{-1}(y - X\beta) \text{ e}$$

$$V_{n \times n} = ZGZ' + R$$

Observe que a predição  $BLUP(g)$ : é a esperança condicional de  $g$ , dado um conjunto de observações fenotípicas.

O melhor preditor linear não viesado (Best Linear Unbiased Predictor-BLUP) é expresso como:

$$BLUP(g) = GZ'V^{-1}(y - X\beta)$$

(Melhor Preditor Linear Não Viesado)

Melhor: minimiza a variância do erro de predição;

Preditor Linear: é uma função linear dos dados;

$$\text{Não Viesado: } E[BLUP(g)] = E(g) = 0;$$

BLUP pressupõe que os componentes de variância são conhecidos. Na prática são utilizados EBLUP (Empirical Best Linear Unbiased Predictor) porque os componentes de variância são desconhecidos e, portanto, devem ser estimados. Entretanto, utiliza-se a sigla BLUP, indiferentemente.

Dificuldade: Inversão da matriz  $V_{n \times n}$ .

Sistema de Equações de Modelos Mistos:

$$\begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} XX & XZ \\ ZX & ZZ + A^{-1} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

A estimação dos efeitos fixos (BLUE) e a predição dos efeitos aleatórios (BLUP) do modelo são obtidas com a solução do Sistema de Equações de Modelos Mistos de Henderson (1953) da seguinte forma:

Estimação dos Efeitos Fixos:

$$BLUE(\beta) = (X'V^{-1}X)^{-1} X'V^{-1}y$$

Predição dos Efeitos Aleatórios:

$$BLUP(g) = GZ'V^{-1}(y - X\beta)$$

Matriz  $G = A\sigma_a^2$ :

A matriz  $G = A\sigma_a^2$  refere-se à matriz de covariâncias entre os efeitos aleatórios de genótipos, em que:

$\sigma_a^2$ : variância dos efeitos aleatórios dos tratamentos genéticos;

$A$ : matriz de relacionamento entre os tratamentos genéticos.

Incorporação da informação de parentesco entre os genótipos em avaliação:

$a_{ij}$  = Parentesco de Wright entre os genótipos  $i$  e  $j$ ;

$a_{ij} = 2 \times$  Parentesco de Malecot;

$$\text{Parentesco Aditivo} = g_{ij} = a_{ij} \sigma_A^2;$$

Na ausência de Parentesco  $A = I$ .

Observe que assumir um fator como de efeito aleatório envolve pressuposições extras, mas permite inferências mais amplas.

Modelo Linear Misto e Estimação de Componentes de Variância:

Método da Máxima Verossimilhança Restrita (Restricted Maximum Likelihood- REML) e Algoritmo EM (Expectation-Maximization):

São fornecidos valores arbitrários iniciais para os componentes de variância  $\sigma_a^2$  e  $\sigma_e^2$ :

$$\hat{\sigma}_e^2 = \frac{y'y - \hat{\beta}'X'y - \hat{g}'Z'y}{n - r(X)}$$

↔

$$\begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

$$\hat{\sigma}_a^2 = \frac{\hat{g}'A^{-1}\hat{g} + \hat{\sigma}_e^2 \operatorname{tr}(A^{-1}C^{22})}{q}$$

↔

$$C^{-1} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

Matriz G e significado do BLUP e  $\sigma_a^2$ :

Predição dos Efeitos Aleatórios:

$$BLUP(g) = GZ'V^{-1}(y - X\beta)$$

Matriz  $G = A\sigma_a^2$ :

Genótipos	Matriz A	$\sigma_a^2$	$BLUP(g)$
Relacionados	Parentesco Wright	$\sigma_A^2$ : Var. Aditiva	"Breeding Value"
Não Relacionados	$A = I$	$\sigma_g^2$ : Var. Genotípica	Valor Genotípico

O significado do BLUP depende dos tratamentos genéticos em avaliação:

1. Avaliação de linhagens ou de progênies: os BLUP se referem aos "breeding values" ou valores genéticos para melhoramento;
2. Avaliação de híbridos simples: os BLUP são para as CGC e CEC.

Exemplo 2 – Estimação e Predição com base no Modelo Linear Misto:

Experimento em DBC com três genótipos (Aleatórios) e duas repetições (Fixos):

$A = I$ : Genótipos Não Relacionados



BLUP dos genótipos são preditos somente com base em seus próprios desempenhos

$$\begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} XX & XZ \\ ZX & ZZ + A^{-1} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$



$$\begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{..} \\ y_{.1} \\ y_{.2} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix} \frac{\sigma_e^2}{\sigma_g^2}$$

Modelo Misto vs Modelo Fixo – Médias BLUP e BLUE:

Método REML/BLUP:

$$BLUP \begin{bmatrix} \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \end{bmatrix} = \begin{bmatrix} -0,3319 \\ 0,04485 \\ 0,2871 \end{bmatrix}$$

↓

$$\text{MédiaBLUP} \begin{bmatrix} \mu + \bar{b} + \tilde{g}_1 \\ \mu + \bar{b} + \tilde{g}_2 \\ \mu + \bar{b} + \tilde{g}_3 \end{bmatrix} = \begin{bmatrix} 4,8248 \\ 5,2016 \\ 5,4438 \end{bmatrix}$$

("Shrinkage")

$$\hat{\sigma}_e^2 = 0,0761 \text{ e } \hat{\sigma}_g^2 = 0,1265$$

Efeito "Shrinkage" ou Efeito de Encolhimento significa que os valores das médias preditas ficam mais próximos uns dos outros.

Método OLS/ANOVA:

$$BLUE \begin{bmatrix} \mu + \bar{b} + g_1 \\ \mu + \bar{b} + g_2 \\ \mu + \bar{b} + g_3 \end{bmatrix} = \begin{bmatrix} y_{1\cdot}/r \\ y_{2\cdot}/r \\ y_{3\cdot}/r \end{bmatrix} = \begin{bmatrix} 4,725 \\ 5,215 \\ 5,530 \end{bmatrix}$$

$$\hat{\sigma}_e^2 = 0,0761 \text{ e } \hat{\sigma}_g^2 = 0,1265$$

Modelo Fixo vs Modelo Misto - Ganho com Seleção:

$$\text{MédiaBLUE} = \begin{bmatrix} 4,725 \\ 5,215 \\ 5,530 \end{bmatrix}$$

$$MédiaBLUP = \begin{bmatrix} 4,8248 \\ 5,2016 \\ 5,4438 \end{bmatrix}$$

$$GS = \hat{h}^2 \cdot ds = \hat{h}^2 (\bar{y}_{i \cdot} - \bar{y}_{..})$$

$$\hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_F^2} = \frac{0,1265}{0,16455} = 0,7688$$

Para G1:

$$GS_1 = \hat{h}^2 (\bar{y}_{1 \cdot} - \bar{y}_{..}) = 0,7688(4,725 - 5,1567) = -0,3319$$

Para G2:

$$GS_2 = \hat{h}^2 (\bar{y}_{2 \cdot} - \bar{y}_{..}) = 0,7688(5,215 - 5,1567) = 0,04485$$

Para G3:

$$GS_3 = \hat{h}^2 (\bar{y}_{3 \cdot} - \bar{y}_{..}) = 0,7688(5,530 - 5,1567) = 0,2871$$

Para G1:

$$\bar{y}_{..} + GS_1 = 5,1567 - 0,3319 = 4,8248$$

Para G2:

$$\bar{y}_{..} + GS_2 = 5,1567 + 0,04485 = 5,2016$$

Para G3:

$$\bar{y}_{..} + GS_3 = 5,1567 + 0,2871 = 5,4438$$

Observe que maior  $h^2$  implica em menor “shrinkage” e, portanto, valores BLUE mais próximos dos BLUP.

### Desbalanceamento

A avaliação de grande número de genótipos e/ou a reduzida quantidade de material experimental pode tornar necessário o uso de delineamentos desbalanceados (não ortogonais) tais como Blocos Incompletos e Blocos Aumentados. A perda de parcelas também torna os delineamentos não ortogonais.

Como exemplo de desbalanceamento considere um experimento delineado em látice triplo 4 x 4:

Repetição I			
(1) 10,2	(2) 10,7	(3) 10,8	(4) 12,7
(5) 9,3	(6) 6,4	(7) 10,5	(8) 10,6
(9) 9,2	(10) 5,2	(11) 3,6	(12) 10,2
(13) 8,3	(14) 9,8	(15) 6,2	(16) 4,9

Repetição II			
(15) 5,5	(11) 5,5	(3) 7,3	(7) 10,0
(5) 8,3	(1) 6,2	(13) 7,2	(9) 10,0
(12) 5,0	(4) 9,5	(8) 11,0	(16) 7,5
(14) 10,5	(6) 9,5	(10) 11,2	(2) 10,3

Repetição III			
(1) 12,9	(11) 7,7	(6) 7,9	(16) 7,7
(8) 10,8	(3) 10,4	(9) 10,6	(14) 12,6
(4) 7,7	(13) 6,0	(10) 5,4	(7) 7,6
(2) 7,6	(5) 12,1	(12) 6,5	(15) 8,2

Modelo para Delineamentos Experimentais Blocos Incompletos do tipo Látice:

$$y_{ijk} = \mu + r_k + b_{j(k)} + t_i + e_{ijk}$$

Os tipos de análises que podem ser realizadas são:

Análise 1 – Análise Intrablocos (Média ajustada)

Análise 2 – Análise com Recuperação da Informação Interblocos (Média ajustada)

Análise 3 – Análise com Recuperação da Informação Interblocos e Intergenotípica (Média BLUP)

Resultados:

Genótipo (.)	Média Ajustada (Intrablocos)	Genótipo (.)	Média Ajustada (Interblockos)	Genótipo (.)	Média BLUP
8	10,763	8	10,773	8	9,862
7	10,742	5	10,364	<b>14</b>	<b>9,748</b>
5	10,546	<b>14</b>	<b>10,362</b>	7	9,673
9	10,400	7	10,355	5	9,645
4	10,358	9	10,269	4	9,602
<b>14</b>	<b>10,125</b>	4	10,248	9	9,596
13	8,967	1	9,162	1	9,012
1	8,925	3	8,952	3	8,775
3	8,738	13	8,460	2	8,536
2	7,967	2	8,408	13	8,421
<b>10</b>	<b>7,658</b>	<b>10</b>	<b>7,548</b>	<b>10</b>	<b>7,972</b>
<b>12</b>	<b>7,542</b>	<b>12</b>	<b>7,455</b>	<b>12</b>	<b>7,942</b>
<b>15</b>	<b>7,121</b>	<b>15</b>	<b>6,984</b>	<b>15</b>	<b>7,658</b>
6	6,588	6	6,966	6	7,648
16	6,425	16	6,502	16	7,395
<b>11</b>	<b>5,438</b>	<b>11</b>	<b>5,493</b>	<b>11</b>	<b>6,816</b>

Observações:

No BLUP a média populacional é a mesma para todos os tratamentos genéticos, uma vez que, em termos de processo de estimação, trata-se de uma mesma população.

Quanto mais eficiente for o delineamento látice maior será o efeito da recuperação de informação interblockos no ordenamento das médias de genótipos.

O efeito “shrinkage” ocorre quando se considera o efeito de tratamento genético aleatório e não quando se considera o efeito de blocos aleatório.

#### Informação de Parentesco no Modelo Misto via Genealogia

$$A = \begin{bmatrix} 2 & 2f_{(G1,G2)} & 2f_{(G1,G3)} \\ 2f_{(G2,G1)} & 2 & 2f_{(G2,G3)} \\ 2f_{(G3,G1)} & 2f_{(G3,G2)} & 2 \end{bmatrix}$$

$f_{(x,y)}$ : coeficiente parentesco de Malecot

$$A = \begin{bmatrix} 2 & 2 \times 1/2 & 2 \times 7/16 \\ 2 \times 1/2 & 2 & 2 \times 37/32 \\ 2 \times 7/16 & 2 \times 37/32 & 2 \end{bmatrix}$$

Elementos não nulos fora da diagonal de  $A$  refletem o uso da informação de relativos na predição dos efeitos dos genótipos.

Neste caso, para o Exemplo 1(DBC anterior) tem-se:

$$\begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

↓

$$\begin{bmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} n & t & t \\ t & t & 0 \\ t & 0 & t \end{bmatrix} & \begin{bmatrix} r & r & r \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ \begin{bmatrix} r & 1 & 1 \\ r & 1 & 1 \\ r & 1 & 1 \end{bmatrix} & \begin{bmatrix} r & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & r \end{bmatrix} \end{bmatrix} + \begin{bmatrix} 2 & 2f_{12} & 2f_{13} \\ 2f_{21} & 2 & 2f_{23} \\ 2f_{31} & 2f_{32} & 2 \end{bmatrix} \begin{bmatrix} \frac{\sigma_e^2}{\sigma_g^2} \\ \frac{\sigma_e^2}{\sigma_g^2} \\ \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} y_{..} \\ y_{.1} \\ y_{.2} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix}$$

$$BLUP \begin{bmatrix} \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \end{bmatrix} = \begin{bmatrix} -0,3103 \\ 0,1435 \\ 0,3325 \end{bmatrix}$$

$$Média BLUP \begin{bmatrix} \mu + \bar{b} + \tilde{g}_1 \\ \mu + \bar{b} + \tilde{g}_2 \\ \mu + \bar{b} + \tilde{g}_3 \end{bmatrix} = \begin{bmatrix} 4,7911 \\ 5,2449 \\ 5,4339 \end{bmatrix}$$

Efeito “Shrinkage”:

$$BLUE = \begin{bmatrix} 4,725 \\ 5,215 \\ 5,530 \end{bmatrix} \quad BLUP = \begin{bmatrix} 4,8248 \\ 5,2016 \\ 5,4438 \end{bmatrix}$$

$$BLUP_i = \begin{bmatrix} 4,7911 \\ 5,2449 \\ 5,4339 \end{bmatrix}$$

### BLUP de Combinações Híbridas Não Testadas

Nos programas de produção de híbridos comerciais apenas 10 a 15% dos híbridos simples (HS) são testados.

Grupo Heterótico1      x      Grupo Heterótico2

10	x	20	= 200 HS
100	x	150	= 15.000 HS

Interesse de melhoristas de alógamas – Predição do desempenho de híbridos simples não testados.

Dificuldade:

Heterose = f (divergência alélica; dominância)

$$HS_{ij} = m + CGC_i + CGC_j + CEC_{ij} + e_{ij}$$

CGC – Capacidade Geral de Combinação

CEC – Capacidade Específica de Combinação

Critérios utilizados para predição do desempenho de híbridos:

1. Desempenho das linhagens parentais – correlações baixas

$$HS_{ij} = m + CGC_i + CGC_j \text{ (Modelo Aditivo)}$$

2. Divergência entre as linhagens parentais – genealogia, marcadores, correlações baixas

$$HS_{ij} = m + CGC_i + CGC_j + pD_{ij} \text{ (Modelo de Distância)}$$

3. BLUP HS não testados – explora a informação de parentesco (genealogia ou marcadores) entre HS testados (T) e não testados (NT)

Correlações entre os desempenhos preditos e observados de  $HS_{NT}$  oriundos de 16 padrões heteróticos (Bernardo, 1996):

Caracteres (Milho)	Correlações $(BLUP_{NT}, Y)$
Produtividade de grãos	0,43 – 0,76
Umidade de grãos	0,75 – 0,93

$$r(BLUP_{NT,Y}) \rightarrow \sqrt{h^2}$$

Considere o modelo linear misto geral:

$$y = X\beta + Z_1g_1 + Z_2g_2 + Z_3s + e$$

$y$ : vetor de observações (desempenho de híbridos), de dimensões  $n \times 1$ ;

$X$ : matriz de incidência, de dimensões  $n \times p$ , que relaciona  $y$  a  $\beta$ ;

$\beta$ : vetor de efeitos fixos, de dimensões  $p \times 1$ ;

$Z_1$ ,  $Z_2$  e  $Z_3$ : matrizes de incidência que relacionam  $y$  a  $g_1$ ,  $y$  a  $g_2$  e  $y$  a  $s$ , respectivamente;

$g_1$ ,  $g_2$  e  $s$ : vetores de efeitos aleatórios das  $CGC1$ ,  $CGC2$  e  $CEC$ , respectivamente,

sendo  $g_1 \sim N(0, G_1 \sigma_{CGC(1)}^2)$ ,  $g_2 \sim N(0, G_2 \sigma_{CGC(2)}^2)$  e

$s \sim N(0, S \sigma_{CEC}^2)$ ;

$e$ : vetor de erros aleatórios,  $e \sim N(0, R)$  e  $R = I \sigma_e^2$ .

Considere que  $x$  e  $x'$  são linhagens do grupo heterótico 1 ( $GH1$ ) e que  $k$  e  $k'$  são linhagens do grupo heterótico 2 ( $GH2$ ).

A covariância entre híbridos simples  $HSxk$  e  $HSx'k'$ , considerando ausência de epistasia, é expressa por:

$$Cov(HS_{xk}, HS_{x'k'}) = f_{xx'} \sigma_{CGC(1)}^2 + f_{kk'} \sigma_{CGC(2)}^2 + f_{xx'} f_{kk'} \sigma_{CEC}^2$$

$f_{xx'}$ : coeficiente de coancestria entre as linhagens  $x$  e  $x'$ ;

$f_{kk'}$ : coeficiente de coancestria entre as linhagens  $k$  e  $k'$ .

Solução baseada no Sistema de Equações de Henderson (1950):

$$\begin{bmatrix} \hat{\beta} \\ \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{s} \end{bmatrix} = \begin{bmatrix} X'X & X'Z_1 & X'Z_2 & X'Z_3 \\ Z_1'X & Z_1'Z_1 + G_1^{-1} \frac{\sigma_e^2}{\sigma_{CGC(1)}^2} & Z_1'Z_2 & Z_1'Z_3 \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + G_2^{-1} \frac{\sigma_e^2}{\sigma_{CGC(2)}^2} & Z_2'Z_3 \\ Z_3'X & Z_3'Z_1 & Z_3'Z_2 & Z_3'Z_3 + S^{-1} \frac{\sigma_e^2}{\sigma_{CEC}^2} \end{bmatrix} \begin{bmatrix} X'y \\ Z_1'y \\ Z_2'y \\ Z_3'y \end{bmatrix}$$

$$\hat{\sigma}_e^2 = \frac{y'y - \hat{\beta}'X'y - \tilde{g}_1'Z_1'y - \tilde{g}_2'Z_2'y - \tilde{s}'Z_3'y}{n - r(X)}$$

$$\hat{\sigma}_{CGC(1)}^2 = \frac{\tilde{g}_1'G_1^{-1}\tilde{g}_1 + \hat{\sigma}_e^2 \text{tr}(G_1^{-1}C^{22})}{\text{No.LinhagensGH1}}$$

$$\hat{\sigma}_{CGC(2)}^2 = \frac{\tilde{g}_2'G_2^{-1}\tilde{g}_2 + \hat{\sigma}_e^2 \text{tr}(G_2^{-1}C^{33})}{\text{No.LinhagensGH2}}$$

$$\hat{\sigma}_{CEC}^2 = \frac{\tilde{s}'S^{-1}\tilde{s} + \hat{\sigma}_e^2 \text{tr}(S^{-1}C^{44})}{\text{No.HS}}$$

$$\begin{bmatrix} \hat{\beta} \\ \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{s} \end{bmatrix} = C^{-1} = \begin{bmatrix} C^{11} & C^{12} & C^{13} & C^{14} \\ C^{21} & C^{22} & C^{23} & C^{24} \\ C^{31} & C^{32} & C^{33} & C^{34} \\ C^{41} & C^{42} & C^{43} & C^{44} \end{bmatrix}$$

Passos para a predição dos  $HS_{NT}$  a partir da análise dos  $HS_T$  (Bernardo, 2002; Silva Filho, 2004):

1. Ajustar os desempenhos dos  $HS_T$  para os efeitos fixos do modelo;

$$\hat{y}_T = \left( Z_3' R^{-1} Z_3 \right)^{-1} Z_3' R^{-1} (y - X \hat{\beta})$$

2. Determinar as covariâncias entre os  $HS$ ;

$$Cov(HS_{xk}, HS_{x'k'}) = f_{xx'} \sigma_{CGC(1)}^2 + f_{kk'} \sigma_{CGC(2)}^2 + f_{xx'} f_{kk'} \sigma_{CEC}^2$$

Obter as matrizes de covariâncias  $C_{NTT}$  e  $C_{TT}$ :

$C_{NTT}$ : matriz de covariâncias genéticas entre os  $HS_{NT}$  e os  $HS_T$ ;

$C_{TT}$ : matriz de covariâncias fenotípicas entre os  $HS_T$ .

3. Predizer os desempenhos dos  $HS_{NT}$  a partir dos desempenhos dos  $HS_T$ .

$$\hat{y}_{NT} = C_{NTT} C_{TT}^{-1} \hat{y}_T$$

Exemplo: Considere os dados de produtividade de grãos de híbridos de milho (Adaptado de Bernardo, 2002)

Linhagens B73, B84 e H123 – Iowa Stiff Stalk Synthetic (GH1)  
 Linhagens Mo17 e N197 – Lancaster Sure Crop (GH2)

Locais	HS	Pedigree	Produtividade (t.ha <sup>-1</sup> )
1	HS1	B73 x Mo17	7,85
1	HS2	H123 x Mo17	7,36
1	HS3	B84 x N197	5,61
2	HS2	H123 x Mo17	7,47
2	HS3	B84 x N197	5,96

$$y = X\beta + Z_1 g_1 + Z_2 g_2 + Z_3 s + e$$

$$\begin{bmatrix} 7,85 \\ 7,36 \\ 5,61 \\ 7,47 \\ 5,96 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} g_{B73} \\ g_{B84} \\ g_{H123} \end{bmatrix} +$$

$$y = X\beta + Z_1 g_1 + Z_2 g_2 + Z_3 g_3 + e$$

$$+ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} g_{Mo17} \\ g_{N197} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} + e$$

$$+ Z_2 g_2 + Z_3 g_3 + e$$

$$g_1 \sim N(0, G_1 \sigma_{g1}^2); g_2 \sim N(0, G_2 \sigma_{g2}^2)$$

$$\begin{array}{c} \downarrow \\ f_{xx'} \end{array} \qquad \qquad \qquad \begin{array}{c} \downarrow \\ f_{kk'} \end{array}$$

$$s \sim N(0, S \sigma_s^2); e \sim N(0, R) - R = I \sigma_e^2$$

$$\begin{array}{c} \downarrow \\ f_{xx'} \cdot f_{kk'} \end{array}$$

$$Cov(HS_{xk}, HS_{x'k'}) = f_{xx'} \sigma_{CGC(1)}^2 + f_{kk'} \sigma_{CGC(2)}^2 + f_{xx'} f_{kk'} \sigma_{CEC}^2$$

$$G_1 = \begin{bmatrix} \text{B73} & \text{B84} & \text{H123} \\ 1 & 0,265 & 0,75 \\ 0,275 & 1 & 0,19875 \\ 0,75 & 0,19875 & 1 \end{bmatrix};$$

$$G_2 = \begin{bmatrix} \text{Mo17} & \text{N197} \\ 1 & 0,75 \\ 0,75 & 1 \end{bmatrix};$$

Os híbridos são: B73xMo17, H123xMo17 e B84xN197

$$S = \begin{bmatrix} 1 & 0,75 & 0,19875 \\ 0,75 & 1 & 0,14906 \\ 0,19675 & 0,14906 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \tilde{g}_{B73} \\ \tilde{g}_{B84} \\ \tilde{g}_{H123} \\ \tilde{g}_{Mol7} \\ \tilde{g}_{N197} \\ \tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \end{bmatrix} = \begin{bmatrix} X'X & X'Z_1 & X'Z_2 & X'Z_3 \\ Z'_1X & Z'_1Z_1 + G_1^{-1} \frac{\sigma_e^2}{\sigma_{CGC(1)}^2} & Z'_1Z_2 & Z'_1Z_3 \\ Z'_2X & Z'_2Z_1 & Z'_2Z_2 + G_2^{-1} \frac{\sigma_e^2}{\sigma_{CGC(2)}^2} & Z'_2Z_3 \\ Z'_3X & Z'_3Z_1 & Z'_3Z_2 & Z'_3Z_3 + S^{-1} \frac{\sigma_e^2}{\sigma_{CEC}^2} \end{bmatrix} \begin{bmatrix} X'y \\ Z'_1y \\ Z'_2y \\ Z'_3y \end{bmatrix}$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \tilde{g}_{B73} \\ \tilde{g}_{B84} \\ \tilde{g}_{H123} \\ \tilde{g}_{Mol7} \\ \tilde{g}_{N197} \\ \tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \end{bmatrix} = \begin{bmatrix} 6,77 \\ 6,77 \\ 0,40 \\ -0,45 \\ 0,37 \\ 0,07 \\ -0,07 \\ 0,15 \\ 0,13 \\ -0,16 \end{bmatrix}$$

$$\sigma_{CGC(1)}^2 = 0,30 \quad \sigma_{CGC(2)}^2 = 0,15$$

$$\sigma_{CEC}^2 = 0,10 \quad \sigma_e^2 = 0,30$$

Predição de híbridos simples (HS) não testados:

HS <sub>T</sub>	HS <sub>NT</sub>
B73 x Mo17	B73 x N197
H123 x Mo17	B84 x Mo17
B84 x N197	H123 x N197

- Ajustar os desempenhos dos  $HS_T$  para os efeitos fixos do modelo:

$$\hat{y}_T = \left( Z'_3 R^{-1} Z_3 \right)^{-1} Z'_3 R^{-1} (y - X\beta)$$

$$Z_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \begin{array}{l} HS1: B73 \times Mo17 \\ HS2: H123 \times Mo17 \\ HS3: B84 \times N197 \end{array}$$

$$\hat{y}_T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1,08 \\ 0,59 \\ -1,16 \\ 0,70 \\ -0,81 \end{bmatrix} = \begin{bmatrix} 1,08 \\ 0,645 \\ -0,985 \end{bmatrix}$$

- Determinar as covariâncias entre os  $HS$ :

Os híbridos são: B73xMo17, H123xMo17 e B84xN197

$$B73 \times N197 \quad C_{NTT} = \begin{bmatrix} 0,4875 & 0,39375 & 0,256 \\ 0,256 & 0,2295 & 0,4875 \\ 0,39375 & 0,4875 & 0,2295 \end{bmatrix}$$

$$B84 \times Mo17 \quad C_{TT} = \begin{bmatrix} 0,85 & 0,45 & 0,21188 \\ 0,45 & 0,70 & 0,18703 \\ 0,21188 & 0,18703 & 0,70 \end{bmatrix}$$

$$H123 \times N197$$

$$Cov(HS_{xk}, HS_{x'k'}) = f_{xx'}\sigma_{CGC(1)}^2 + f_{kk'}\sigma_{CGC(2)}^2 + f_{xx'}f_{kk'}\sigma_{CEC}^2$$

$$Cov(B73 \times N197, B73 \times Mo17) = 1(0,30) + 0,75(0,15) + 0,75(0,10) = 0,4875$$

$$B73 \times Mo17 \quad H123 \times Mo17 \quad B84 \times N197$$

$$B73 \times Mo17 \quad C_{TT} = \begin{bmatrix} 0,85 & 0,45 & 0,21188 \\ 0,45 & 0,70 & 0,18703 \\ 0,21188 & 0,18703 & 0,70 \end{bmatrix}$$

$$Cov(HS_{xk}, HS_{x'k'}) = f_{xx'}\sigma_{CGC(1)}^2 + f_{kk'}\sigma_{CGC(2)}^2 +$$

$$+ f_{xx'}f_{kk'}\sigma_{CEC}^2 + V_R / j$$

$$V_R / j \Rightarrow Z_3' R^{-1} Z_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$Cov(B73xMo17, B73xMo17) = 1(0,30) + 1(0,15) + 1(0,10) + 0,3/1 = 0,85$$

$$Cov(B73xMo17, H123xMo17) = 0,75(0,30) + 1(0,15) + 0,75(0,10) + 0 = 0,45$$

3. Predizer os desempenhos dos  $HS_{NT}$  a partir dos desempenhos dos  $HS_T$ :

$$\hat{y}_{NT} = C_{NTT} C_{TT}^{-1} \hat{y}_T$$

$$\hat{y}_{NT} = \begin{bmatrix} 0,4875 & 0,39375 & 0,256 \\ 0,256 & 0,2295 & 0,4875 \\ 0,39375 & 0,4875 & 0,2295 \end{bmatrix} \begin{bmatrix} 0,85 & 0,45 & 0,21188 \\ 0,45 & 0,70 & 0,18703 \\ 0,21188 & 0,18703 & 0,70 \end{bmatrix}^{-1} \begin{bmatrix} 1,08 \\ 0,645 \\ -0,985 \end{bmatrix}$$

$$\hat{y}_{NT} = \begin{bmatrix} \hat{y}_{B73xN197} \\ \hat{y}_{B84xMo17} \\ \hat{y}_{H123xN197} \end{bmatrix} = \begin{bmatrix} 0,42 \\ -0,47 \\ 0,37 \end{bmatrix}$$

$$\begin{bmatrix} g_i & g_j & s_{ij} \\ \hat{y}_{B73xMo17} \\ \hat{y}_{H123xMo17} \\ \hat{y}_{B84xN197} \end{bmatrix} = \begin{bmatrix} 0,40+0,07+0,15=0,62 \\ 0,37+0,07+0,13=0,57 \\ -0,45-0,07-0,16=-0,68 \end{bmatrix}$$

### Predição de Valores Genéticos via Metodologia de Modelos Mistos

(Extraído de Lynch e Walsh, 1998)

Considere um vetor coluna  $\mathbf{y}$  contendo os valores fenotípicos de uma característica medida em  $n$  indivíduos.

É assumido que estas observações são descritas adequadamente por um modelo linear com um vetor  $p \times 1$  de efeitos fixos ( $\beta$ ) e um vetor  $q \times 1$  de efeitos aleatórios ( $u$ ). O primeiro elemento do vetor  $\beta$  é tipicamente a média da população e os outros elementos (fatores) incluídos pode ser gênero, local, ano, tratamento experimental e outros. Os elementos do vetor  $u$  de efeitos aleatórios são geralmente efeitos genéticos tal como valores genéticos aditivos.

Em termos matriciais, o modelo é expresso como:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}u + \mathbf{e} \quad (1)$$

, em que:

$X$  : matriz de incidência  $n \times p$  de fator de efeito fixo;

$Z$  : matriz de incidência  $n \times q$  de fator de efeito aleatório;

$e$ : vetor coluna  $n \times 1$  de desvios residuais, assumido ser independentemente distribuído dos efeitos genéticos aleatórios.

Usualmente todos os elementos das matrizes de incidências são iguais a 1 ou 0, dependendo se o efeito considerado contribui ou não para o fenótipo do indivíduo. Devido ao modelo considerar juntamente efeitos fixos e aleatórios é referido como um modelo misto.

Exemplo: Suponha que três genitores são escolhidos ao acaso de uma população e cada um foi cruzado aleatoriamente. São avaliadas duas progênieis de cada cruzamento, algumas no ambiente 1 e outras no ambiente 2.

Seja  $y_{ijk}$ , que representa o valor fenotípico da  $k$ -ésima progênie do genitor  $i$  no ambiente  $j$ .

Então o modelo linear é expresso como:

$$y_{ijk} = \beta_j + u_i + e_{ijk}$$

Este modelo tem três efeitos aleatórios ( $u_1, u_2, u_3$ ), que medem a contribuição de cada genitor, e dois efeitos fixos ( $\beta_1, \beta_2$ ) que descrevem a influência dos dois ambientes. O modelo desconsidera o efeito da interação genitor x ambiente.

Foi avaliado um total de seis progênieis. Uma progênie do genitor 1 foi alocada no ambiente 1 e teve valor fenotípico  $y_{111}=9$ , enquanto que a segunda progênie foi alocada no ambiente 2 e teve valor fenotípico  $y_{121}=12$ . As duas progênieis do genitor 2 foram ambas alocadas no ambiente 1 e tiveram valores fenotípicos  $y_{211}=11$  e  $y_{212}=6$ . Uma progênie do genitor 3 foi alocada no ambiente 1 e teve valor fenotípico  $y_{311}=7$ , enquanto que a segunda progênie foi alocada no ambiente 2 e teve valor fenotípico  $y_{321}=14$ .

O vetor de observações (ou vetor de respostas) resultante pode ser escrito como:

$$y = \begin{bmatrix} y_{111} \\ y_{121} \\ y_{211} \\ y_{212} \\ y_{311} \\ y_{321} \end{bmatrix} = \begin{bmatrix} 9 \\ 12 \\ 11 \\ 6 \\ 7 \\ 14 \end{bmatrix}$$

, resultando num modelo linear misto:

$$y = X\beta + Zu + e$$

, onde as matrizes de incidências para os efeitos fixos e aleatórios e os vetores destes efeitos são respectivamente:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}; Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix};$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; \quad u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Considere agora as médias e as variâncias dos vetores componentes do modelo misto. Uma vez que  $E(\mathbf{u})=E(\mathbf{e})=\mathbf{0}$ , por definição,  $E(\mathbf{y})=X\beta$ . Denote a matriz de covariância ( $n \times n$ ) para o vetor  $\mathbf{e}$  de erros ou resíduos por  $\mathbf{R}$  e a matriz de covariância ( $q \times q$ ) para o vetor  $\mathbf{u}$  de efeitos genéticos aleatórios por  $\mathbf{G}$ . Excluindo a diferença entre indivíduos devido a efeitos fixos, a partir da equação  $\sigma(y, y) = \mathbf{A}\mathbf{V}\mathbf{A}'$  (fornecendo a matriz de covariância para  $\mathbf{y} = \mathbf{A}\mathbf{x}$ ) e da suposição que  $\mathbf{u}$  e  $\mathbf{e}$  são não correlacionados, a matriz de covariância para o vetor de observações  $\mathbf{y}$  é expressa como:

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \quad (2)$$

O primeiro termo considera a contribuição dos efeitos genéticos aleatórios, enquanto que o segundo considera a variância devida aos efeitos de erros ou resíduos. Geralmente é assumido que os erros ou resíduos têm variância constante e são não correlacionados, de forma que  $\mathbf{R}$  é uma matriz diagonal com  $\mathbf{R} = \sigma_E^2 \mathbf{I}$ .

Agora o modelo linear misto pode ser contrastado com o modelo linear geral:

O modelo linear geral é descrito por,  $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}^*$ , em que  $\mathbf{e}^* \perp \!\!\! \perp (\mathbf{0}, \mathbf{V})$ , o que implica em  $\mathbf{y} \perp \!\!\! \perp (\mathbf{X}\beta, \mathbf{V})$ .

Por outro lado, o modelo linear misto partitiona o vetor de efeitos resíduais em dois componentes com,  $\mathbf{e}^* = \mathbf{Z}\mathbf{u} + \mathbf{e}$ , resultando em  $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$ ,

em que  $\mathbf{u} \perp \!\!\! \perp (\mathbf{0}, \mathbf{G})$  e  $\mathbf{e} \perp \!\!\! \perp (\mathbf{0}, \mathbf{R})$ , o que implica em:

$$\mathbf{y} \perp \!\!\! \perp (\mathbf{X}\beta, \mathbf{V}) \text{ ou } \mathbf{y} \perp \!\!\! \perp (\mathbf{X}\beta, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}).$$

Quando a análise é apropriada, ambas as formulações produzem as mesmas estimativas para o vetor de efeitos fixos  $\beta$ , enquanto que a formulação por modelo linear misto leva em consideração ainda as estimativas do vetor de efeitos aleatórios  $\mathbf{u}$ .

Para o modelo linear misto, são observados  $\mathbf{y}$ ,  $\mathbf{X}$  e  $\mathbf{Z}$ , enquanto que  $\beta$ ,  $\mathbf{u}$ ,  $\mathbf{R}$  e  $\mathbf{G}$  são geralmente desconhecidos. Desta forma, a análise por modelos lineares mistos envolve dois aspectos complementares da estimação: (1) estimação dos vetores de efeito fixo e aleatório,  $\beta$  e  $\mathbf{u}$ , e (2) estimação das matrizes de covariância  $\mathbf{G}$  e  $\mathbf{R}$ .

As matrizes de covariância geralmente são assumidas com sendo funções de alguns componentes de variância desconhecidos. Primeiro serão considerados os estimadores de  $\beta$  e  $u$  sob a suposição de que  $y$ ,  $X$ ,  $Z$ ,  $G$  e  $R$  são conhecidos. Depois será realizada a estimação dos componentes de variância ( $G$  e  $R$ ) a partir de  $y$ ,  $X$  e  $Z$ .

As inferências sobre efeitos fixos são chamadas de estimativas, enquanto que aquelas relacionadas com efeitos aleatórios são conhecidas como previsões. Os métodos mais amplamente utilizados para obtenção destes estimadores e preditores são o BLUE e o BLUP, respectivamente.

Estes estimadores e preditores são os melhores no sentido de que eles minimizam a variância amostral, lineares no sentido de que eles são funções lineares dos fenótipos observados  $y$  e não viesados no sentido de que  $E[BLUE(\beta)] = \beta$  e  $E[BLUP(u)] = u$ .

Para o modelo linear misto expresso pela equação 1 ( $y = X\beta + Zu + e$ ), o BLUE de  $\beta$  é:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y \quad (3)$$

, com  $V$  expresso pela equação 2 ( $V = ZGZ' + R$ ).

O BLUP de  $u$  é expresso por:

$$\hat{u} = GZ'V^{-1}(y - X\hat{\beta}) \quad (4)$$

Como pode ser notado em (3) e (4), que a aplicação prática destas expressões requer que os componentes de variância ( $G$  e  $R$ ) sejam conhecidos. Assim, antes de uma análise BLUP, os componentes de variância precisam ser estimados por ANOVA ou REML.

Exemplo 2: Quais são os valores BLUP para os efeitos de genitores  $(u_1, u_2, u_3)$  no Exemplo 1?

Para obter estes valores é necessário ter as matrizes de covariância para os efeitos de genitores e de erros.

Assumindo que as variâncias residuais dentro dos dois ambientes são as mesmas  $(\sigma_E^2)$ , então

$R = \sigma_E^2 I$ , em que  $I$  é a matriz identidade  $6 \times 6$ .

Assumindo que todos os três genitores são não relacionados e retirados da mesma população, então  $G = \sigma_S^2 I$ , em que  $I$  é a matriz identidade  $3 \times 3$  e  $\sigma_S^2$  é a variância dos efeitos de genitores.

Assumindo apenas variância genética aditiva e que os genitores foram amostrados aleatoriamente a partir de uma população base de cruzamento livre,  $\sigma_A^2 = \sigma_A^2 / 4$ , em que  $\sigma_A^2$  é a variância genética aditiva. Supondo que  $\sigma_A^2 = 8$  e  $\sigma_E^2 = 6$ , a matriz de covariância  $V$  para o vetor de observações  $y$  dada por  $V = ZGZ' + R$ , ou:

$$V = \frac{8}{4} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + 6 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\text{, ou } V = \begin{bmatrix} 8 & 2 & 0 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & 0 & 2 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{bmatrix}, \text{ resultando em:}$$

$$V^{-1} = \frac{1}{30} \begin{bmatrix} 4 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{bmatrix}$$

Usando este resultado e alguns cálculos matriciais obtém-se:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'V^{-1}X)^{-1} X'V^{-1}y = \frac{1}{18} \begin{pmatrix} 148 \\ 235 \end{pmatrix}$$

$$\hat{u} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{pmatrix} = GZ'V^{-1}(y - X\hat{\beta}) = \frac{1}{18} \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Observe que as soluções para as equações do BLUE de  $\beta$  (Eq. 3) e BLUP de  $u$  (Eq. 4) requerem a inversa da matriz de covariância  $V$ . No exemplo anterior,  $V^{-1}$  não é difícil de ser obtida. Entretanto, quando  $y$  contém milhares de observações, a computação de  $V^{-1}$  pode ser muito difícil. Como uma forma de contornar este problema, Henderson (1950, 1963, 1973, 1984) ofereceu um método mais compacto para obter conjuntamente  $\hat{\beta}$  e  $\hat{u}$  na forma de suas equações de modelos mistos (EMM):

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \quad (5)$$

Embora estas expressões pareçam mais complexas do que as equações (3) e (4),  $\mathbf{R}^{-1}$  e  $\mathbf{G}^{-1}$  são obtidas de forma simples se  $\mathbf{R}$  e  $\mathbf{G}$  são diagonal, e, portanto as submatrizes das EMM são mais fáceis de computar do que  $\mathbf{V}^{-1}$ . Outra vantagem das EMM está relacionada com a dimensionalidade da matriz na esquerda. Relembre que  $\mathbf{X}$  e  $\mathbf{Z}$  são  $n \times p$  e  $n \times q$  respectivamente,  $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$  é  $p \times p$ ,  $\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}$  é  $p \times q$  e  $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}$  é  $q \times q$ . Então, a matriz que precisa ser invertida para obter a solução para  $\hat{\beta}$  e  $\hat{u}$  é de ordem  $(p+q) \times (p+q)$ , que em geral é consideravelmente menor do que a dimensionalidade de  $\mathbf{V}$ , que é uma matriz  $n \times n$ .

Originalmente, Henderson (1950) obteve as EMM assumindo que as matrizes de covariância  $\mathbf{G}$  e  $\mathbf{R}$  são conhecidas e que as densidades dos vetores  $\mathbf{u}$  e  $\mathbf{e}$  são cada uma normal multivariada com nenhuma correlação entre eles. Então a equação (5) produz as estimativas de máxima verossimilhança dos efeitos fixo e aleatório. Posteriormente, Henderson (1963) mostrou que as EMM realmente não dependem da normalidade e que  $\hat{\beta}$  e  $\hat{u}$  são BLUE e BLUP, respectivamente, sob condições gerais dado que as variâncias são conhecidas.

Exemplo 3: Usando os valores dos Exemplos 1 e 2 obtém-se que,

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} = \frac{1}{6} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} = \left( \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} \right)' = \frac{1}{6} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} = \frac{5}{6} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6} \begin{bmatrix} 33 \\ 26 \end{bmatrix},$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6} \begin{bmatrix} 21 \\ 17 \\ 21 \end{bmatrix}.$$

Então, as EMM para estes dados tornam-se:

$$\begin{bmatrix} 4 & 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 1 & 5 & 0 & 0 \\ 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \begin{bmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{bmatrix}$$

Tomando a inversa obtém-se a solução:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \frac{1}{270} \begin{bmatrix} 100 & 25 & -25 & -40 & -25 \\ 25 & 175 & -40 & -10 & -40 \\ -25 & -40 & 67 & 10 & 13 \\ -40 & -10 & 10 & 70 & 10 \\ -25 & -40 & 13 & 10 & 67 \end{bmatrix} \begin{bmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{bmatrix},$$

ou:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix} = \frac{1}{18} \begin{bmatrix} 148 \\ 235 \\ -1 \\ 2 \\ -1 \end{bmatrix}$$

, que é idêntico ao resultado obtido no Exemplo 2.

Para que os BLUP sejam efetivamente os melhores preditores não viesados, as variâncias genéticas apropriadas devem ser conhecidas sem erro. Entretanto, as previsões BLUP permanecem não viesadas quando estimativas de variâncias genéticas são usadas no lugar dos valores reais (o que geralmente é o caso), embora não se tenha garantia de ser o melhor para todos os preditores não viesados.

A questão de estimabilidade dos efeitos fixos ocorre quando em algumas situações é impossível obter estimativas BLUE únicas para todos os fatores de efeitos fixos de um modelo.

Por exemplo, suponha que:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \text{ com } X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Neste caso, os fatores 1 e 2 estão completamente confundidos, uma vez que eles contribuem igualmente para todas as observações. Então, estimativas únicas de  $\beta_1$  e  $\beta_2$  não podem ser obtidas. Geralmente, quando duas ou mais colunas de  $X$  não são independentes, ainda é possível obter BLUE únicos para determinadas combinações lineares de  $\beta$  por meio do uso de inversas generalizadas.

Para a matriz delineamento  $X$  anterior, a solução é simples, ou seja, pela combinação dos dois fatores em um único novo fator,  $\beta_1 + \beta_2$ , o novo modelo torna-se:

$$\beta_* = \begin{bmatrix} \beta_1 + \beta_2 \\ \beta_3 \end{bmatrix} \text{ com } X_* = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Uma vez que as colunas de  $X_*$  são independentes, existe uma solução única para  $X'_* V^{-1} X_*$  e os dois BLUE dos efeitos fixos são dados por:

$$\hat{\beta}_* = (X'_* V^{-1} X_*)^{-1} X'_* V^{-1} y.$$

Situações nas quais são requeridas combinações lineares surgem comumente quando um número muito grande de fatores de efeitos fixos é incluído no modelo, tal como ocorre nos grandes programas de melhoramento que envolve locais, épocas e anos.

Uma extensão relativamente simples das equações de modelos mistos de Henderson fornece as estimativas para os erros padrões dos efeitos fixos e aleatórios. Considere a inversa da matriz principal da esquerda da Equação 5 como:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} C_{11} & C_{12} \\ C'_{12} & C_{22} \end{bmatrix} \quad (6)$$

, em que  $C_{11}$ ,  $C_{12}$  e  $C_{22}$  são, respectivamente, submatrizes  $p \times p$ ,  $p \times q$  e  $q \times q$ .

Usando esta notação, Henderson (1975) mostrou que a matriz de covariância amostral para o BLUE de  $\beta$  é fornecida por:

$$\sigma(\hat{\beta}) = C_{11} \quad (7a)$$

, que a matriz de covariância amostral dos erros de predição  $(\hat{u} - u)$  é dada por:

$$\sigma(\hat{u} - u) = C_{22} \quad (7b)$$

, e que a matriz de covariância amostral dos efeitos estimados e dos erros de predição é dada por:

$$\sigma(\hat{\beta}, \hat{u} - u) = C_{12} \quad (7c)$$

Os erros padrões dos efeitos fixos e aleatórios são obtidos como as raízes quadradas dos elementos diagonais de  $C_{11}$  e  $C_{22}$ , respectivamente.

**Exemplo 4:** Considere as equações de modelos mistos do Exemplo 3. Neste caso, para os efeitos fixos,  $\beta_1$  e  $\beta_2$ , e efeitos aleatórios  $u_1$ ,  $u_2$  e  $u_3$ , a inversa da matriz de coeficientes é:

$$\begin{bmatrix} 4 & 0 & \vdots & 1 & 2 & 1 \\ 0 & 2 & & 1 & 0 & 1 \\ \dots & \dots & & \dots & & \\ 1 & 1 & & 5 & 0 & 0 \\ 2 & 0 & \vdots & 0 & 5 & 0 \\ 1 & 1 & & 0 & 0 & 5 \end{bmatrix}^{-1} = \frac{1}{270} \begin{bmatrix} 100 & 25 & \vdots & -25 & -40 & -25 \\ 25 & 175 & & -40 & -10 & -40 \\ \dots & \dots & & \dots & \dots & \\ -25 & -40 & & 67 & 10 & 13 \\ -40 & -10 & \vdots & 10 & 70 & 10 \\ -25 & -40 & & 13 & 10 & 67 \end{bmatrix}$$

Portanto:

$$C_{11} = \frac{1}{270} \begin{bmatrix} 100 & 25 \\ 25 & 175 \end{bmatrix} \quad e \quad C_{22} = \frac{1}{270} \begin{bmatrix} 67 & 10 & 13 \\ 10 & 70 & 10 \\ 13 & 10 & 67 \end{bmatrix}$$

Então, tem-se:

$$\sigma^2(\hat{\beta}_1) = \frac{100}{270}, \quad \sigma^2(\hat{\beta}_2) = \frac{175}{270},$$

$$\sigma(\hat{\beta}_1, \hat{\beta}_2) = \frac{25}{270} \quad e \quad \sigma^2(\hat{u}_1 - u_1) = \frac{67}{270},$$

$$\sigma^2(\hat{u}_2 - u_2) = \frac{70}{270}, \quad \sigma^2(\hat{u}_3 - u_3) = \frac{67}{270},$$

$$\sigma(\hat{u}_1 - u_1, \hat{u}_3 - u_3) = \frac{13}{270} \quad e \text{ assim por diante.}$$

# Capítulo 16

## Seleção de Modelos

### Introdução

Uma vez que um modelo é especificado e seus melhores parâmetros ajustados são encontrados, o pesquisador encontra-se numa posição de avaliar a viabilidade do modelo. Os pesquisadores têm proposto vários critérios que consideram serem importantes para avaliação de modelos (Myung et al., 2003).

Estes incluem três critérios qualitativos (adequação explicativa, interpretabilidade, fidelidade, exatidão) e quatro critérios quantitativos (qualidade de ajuste, simplicidade/complexidade, generalizabilidade).

Um modelo satisfaz o critério de adequação explicativa se suas pressuposições são plausíveis e consistentes com os objetivos e a consideração teórica é razoável para o processo estudado. Ou seja, o modelo deve ser capaz de fazer mais do que redescrivêr os dados observados.

O modelo também deve ser interpretável no sentido de que o modelo faz sentido e é compreensível. Os componentes do modelo, especialmente seus parâmetros, devem estar ligados aos processos e conceitos agronômicos.

O modelo é dito ser fidedigno no sentido da capacidade do modelo de capturar o processo mental subjacente originado dos princípios teóricos incorporados no modelo, ao invés de escolhas feitas na sua especificação computacional.

A refutabilidade é uma condição necessária para testar um modelo ou uma teoria. Refere-se a se existem observações potenciais que um modelo não pode descrever. Se existirem, então o modelo é dito ser refutável. Um modelo não refutável é aquele que pode descrever infalivelmente todos os possíveis padrões de dados numa dada situação experimental. Obviamente que não tem sentido testar um modelo irrefutável.

Uma regra heurística para determinar a refutabilidade de um modelo já é familiar: o modelo é refutável se e somente se o número de seus parâmetros livres é menor que o número de dados observados. Uma regra mais formal é a seguinte: um modelo é refutável se o posto de sua matriz Jacobiana é menor que o número de dados observados para todos os valores dos parâmetros. A matriz Jacobiana é definida em termos de derivadas parciais como:

$$J_{ij}(w) = \partial E(y_j) / \partial w_i \quad (i=1, \dots, k; j=1, \dots, m).$$

Um modelo deve também fornecer uma boa descrição dos dados observados. Qualidade de Ajuste refere-se à capacidade do modelo para ajustar um particular conjunto de dados. São exemplos de medidas de qualidade de ajuste a soma de quadrados do erro minimizada (SQE), o quadrado médio do erro (QME), a raiz quadrada do QME

(RQME), a percentagem da variância explicada (PVAE) e a máxima verossimilhança (ML). Estas medidas estão relacionadas da seguinte forma:

$$QME = SQE(w_{EQM}^*)/m$$

$$RQME = \sqrt{SQE(w_{EQM}^*)/m}$$

$$PVAE = 100 \left( 1 - SQE(w_{EQM}^*)/SQT \right)$$

$$ML = f(y|w_{EML}^*)$$

Nas equações,  $w_{EQM}^*$  é o parâmetro que minimiza  $SQE(w)$ , ou seja, uma estimativa de quadrados mínimos (LSE), e  $SQT$  significa soma de quadrados total definida como

$SQT = \sum_i (y_i - \bar{y})^2$ . Máxima verossimilhança (ML) é a função de distribuição de probabilidade maximizada com relação aos parâmetros do modelo, estimada a  $w_{EML}^*$  que é obtida por meio de estimativa de máxima verossimilhança (MLE).

Um modelo deve descrever bem um conjunto de dados, mas da forma menos complexa possível. Complexidade tem relação com uma inerente flexibilidade de um modelo que o torna capaz de ajustar uma ampla faixa de padrões de dados.

Existem pelo menos duas dimensões da complexidade de um modelo, o número de parâmetros e a forma funcional do modelo (forma de combinação dos parâmetros na equação do modelo).

Generalizabilidade refere-se à capacidade do modelo possibilitar a obtenção de inferências para outros conjuntos de dados, além dos dados ajustados. A qualidade de ajuste de um modelo reflete não apenas a sua habilidade para capturar o processo subjacente, mas também sua capacidade de ajustar ruídos aleatórios. Isto é conceitualmente colocado na seguinte equação:

Qualidade de Ajuste = Ajuste para Regularidade (Generalizabilidade) + Ajuste para Ruído (Sobreajuste). A generalizabilidade pode ser considerada o “padrão ouro” da avaliação de modelos porque ela fornece uma medida mais acurada da aproximação do modelo do processo subjacente. Quantitativamente, isto se refere ao modelo com a menor generalização do erro, que é definida como a predição média do erro que o modelo produz sobre todos os possíveis dados oriundos da mesma fonte.

### Métodos de seleção de modelos

Uma vez que a generalizabilidade de um modelo não é diretamente observável, ela deve ser estimada usando os dados observados. A medida desenvolvida para este propósito tem como objetivo selecionar o modelo que seja complexo o suficiente para capturar a regularidade nos dados, mas que não ultrapasse a complexidade para capturar a variação aleatória sempre presente.

Visto desta forma, a generalizabilidade formaliza o princípio da navalha de Occam.

Os principais critérios de generalizabilidade são: Critério de Informação de Akaike (AIC, Akaike, 1973) e Critério de Informação Bayesiano (BIC, Schwarz, 1978) e Validação Cruzada (VC, Stone, 1974; Browne, 2000). Em todos estes métodos, a função log-verossimilhança maximizada é usada como uma medida de qualidade de ajuste, mas eles diferem em como a complexidade do modelo é conceitualizada e medida.

### Critérios AIC e BIC

AIC e BIC para um dado modelo são definidos como a seguir:

$$AIC = -2 \ln f(y|w^*) + 2k$$

$$BIC = -2 \ln f(y|w^*) + k \ln n$$

em que,  $w^*$  é uma estimativa de máxima verossimilhança (MLE),  $\ln$  é o logaritmo natural de base  $e$ ,  $k$  é o número de parâmetros e  $n$  é o tamanho da amostra.

Em cada critério, o primeiro termo representa uma medida de falta de ajuste, o segundo termo representa uma medida de complexidade, e juntos eles representam uma medida de falta de generalizabilidade. Um menor valor do critério significa melhor generalizabilidade.

Correspondentemente, o modelo que minimiza um dado critério deve ser escolhido.

A complexidade em AIC e BIC é uma função apenas do número de parâmetros. A forma funcional não é considerada. Por isto, estes métodos não são recomendados para comparar modelos com o mesmo número de parâmetros, mas com diferente forma funcional. O método VC é sensível tanto ao número de parâmetros quanto à forma funcional.

### Validação cruzada (VC)

Na VC a generalizabilidade de um modelo é definida sem a definição explícita de uma medida de complexidade. Ao invés disto, modelos com complexidade maior que a necessária para capturar a regularidade nos dados são penalizados por meio de um procedimento de reamostragem, que é realizado da seguinte forma: a amostra de dados observada é dividida em duas subamostras, calibração e validação.

A amostra de calibração é então usada para obter os melhores valores ajustados dos parâmetros de um modelo por MLE ou LSE. Estes valores, denotados por  $w_{cal}^*$ , são então fixados e ajustados,

sem nenhuma afinação posterior dos parâmetros, à amostra de validação, denotada por  $y_{val}$ .

O resultado da divisão do ajuste para  $y_{val}$  pelo  $w_{cal}^*$  é chamado de índice VC do modelo e é tomado como estimativa da generalizabilidade do modelo.

Se desejado, o índice VC pode ser substituído pelo índice VC médio, calculado a partir de múltiplas divisões de amostras de calibração e de validação. O índice VC médio é uma estimativa mais acurada da generalizabilidade do modelo, embora seja também mais computacionalmente intensiva.

Diferente de AIC e BIC, a VC é sensível à dimensão da complexidade da forma funcional do modelo. Por isto, o método pode ser usado em todas as situações de modelagem, incluindo os casos de comparação de modelos que diferem em forma funcional, mas que têm o mesmo número de parâmetros. Entretanto, a acurácia de sua estimativa de generalizabilidade frequentemente é pior do que AIC e BIC.

### Comparando AIC, BIC e VC

Os critérios AIC e BIC são originalmente derivados da teoria da informação. Na VC, os parâmetros do modelo são estimados usando  $N-1$  observações e usados para prever os valores das observações remanescentes. Se este procedimento é repetido para todos os  $N$  possíveis subconjuntos de observações, a soma das raízes quadradas dos erros de predição pode ser usada como uma medida do desempenho do modelo (Weakliem, 2004).

Stone (1977) mostrou que escolhendo o modelo com o menor AIC é assintoticamente equivalente a escolher o modelo com o melhor desempenho na validação cruzada. Entretanto, não é tão simples dizer que escolhendo o AIC dará o melhor desempenho preditivo. Trabalhos mais recentes têm mostrado que o BIC também está relacionado com a forma mais generalizada de validação cruzada na qual são deixados  $d$  observações.

Shao (1997) mostrou que a validação cruzada generalizada é assintoticamente equivalente a diferentes critérios de seleção de modelos, dependendo do número de dados omitidos. Somente se  $d/N$  aproxima-se de zero quando  $N$  aproxima-se do infinito a VC é assintoticamente equivalente ao AIC. Se  $d/N$  aproxima-se de 1, a VC é equivalente a usar uma penalidade que aumenta com o tamanho da amostra. De fato, se for usada uma proporção crescente da amostra para fornecer previsões sobre a parte remanescente, o AIC desempenhará melhor; se for usada uma proporção decrescente, o BIC desempenhará melhor.

### 6. Comparando AIC, BIC e testes de hipóteses clássicos

Os testes de hipóteses clássicos envolvem modelos aninhados, ou seja, casos nos quais um modelo é um caso especial do outro, mas pode ser estendido para cobrir modelos não aninhados.

Usualmente a hipótese nula é que  $H_0: \theta = k$ , em que,  $\theta$  é um parâmetro e  $k$  é uma constante especificada previamente. Geralmente a hipótese testada é  $H_0: \theta = 0$ , mas outros valores de

$k$  são considerados. O modelo subtendido pela  $H_0$  é mais simples, no sentido de ter poucos parâmetros, do que o subtendido pela hipótese alternativa.

Uma das recentes e mais proeminentes objeções aos testes de hipóteses clássicos é que os resultados são fortemente influenciados pelo tamanho da amostra. Experiências têm mostrado que em amostras grandes, quase todas as hipóteses nulas são rejeitadas, de forma que o uso de testes de hipóteses para seleção de modelos leva a modelos muito complexos.

Diferente dos testes de hipóteses clássicos, os critérios AIC e BIC fornecem um ordenamento de todos os modelos. Ambos fornecem uma medida do grau no qual um modelo é superior ao outro.

Ambos os critérios podem mostrar que um modelo menor é melhor que um modelo maior, ao invés de dizer simplesmente que ele não é rejeitado contra o modelo maior. Com AIC, isto é uma simples evidência que o modelo menor fornecerá melhores previsões quando usar as estimativas de parâmetros da amostra. Com o BIC, em contraste, pode ser interpretado como evidência da veracidade do modelo menor.

O AIC e o BIC não recorrem a uma escolha arbitrária de níveis de significância. A tendência de rejeitar todas as hipóteses falsas em grandes amostras aplica-se tanto aos testes clássicos quanto ao BIC. Se uma hipótese é falsa, o valor esperado de  $-2\ln f(y|w^*)$  aumentará em proporção

ao valor de  $N$ . Embora o BIC inclua uma penalidade que aumenta com o tamanho da amostra, o aumento é a uma taxa muito lenta de  $\ln(N)$ . Uma vez que a penalidade no AIC não aumenta com o tamanho da amostra, este critério também favorece modelos maiores à medida que o tamanho da amostra aumenta.

Os critérios AIC e BIC têm significante atração em comparação com os testes de hipóteses clássicos. Eles são convenientes por serem capazes de ordenar modelos com diferentes graus de liberdade. Entretanto, como nos testes clássicos, eles irão favorecer modelos crescentemente mais complexos para tamanhos de amostras suficientemente grandes. Mais precisamente, eles irão favorecer crescentemente modelos complexos se o verdadeiro modelo é complexo. Entretanto, isto não constitui problema para os pesquisadores uma vez que é difícil entender e explicar um modelo contendo um grande número de parâmetros.

A escolha entre o AIC e o BIC depende, em certo grau, de se acreditar na inerente plausibilidade de hipóteses da forma  $\theta=0$ . Se existe uma chance real de que uma hipótese deste tipo seja verdadeira, provavelmente o BIC seja preferível. Entretanto, se existe uma pequena chance de que a hipótese nula seja verdadeira, parece melhor usar o menos conservador AIC.

Este critério surte efeito apenas para aqueles parâmetros que são tão pobremente estimados que prejudiquem o poder preditivo do modelo.

Estes argumentos sugerem que a melhor estratégia seria uma combinação de seleção de modelo pelo AIC e o uso de testes de hipóteses bilaterais para avaliar as previsões teóricas. Esta abordagem teria significantes implicações para a prática da pesquisa.

Os testes de hipóteses clássicos tornaram-se os métodos de seleção de modelos dominantes, não porque eles foram particularmente bem-sucedidos para a tarefa. Existe forte argumento para voltar aos critérios de verossimilhança penalizada tais como os AIC e BIC. Qual critério seria preferido é uma questão mais difícil. A escolha não é puramente uma questão estatística, mas envolve julgamentos sobre os propósitos da seleção de modelo e a natureza da realidade agronômica.

### Inferências para multimodelos: entendendo os valores de AIC e BIC selecionando modelos

**Valores  $\Delta_i$ :** Os valores individuais de AIC e BIC, obtidos com base nas expressões

$AIC = -2\log(L) + 2k$  e  $BIC = -2\log(L) + k\log(n)$ , respectivamente, não são interpretáveis uma vez que eles contêm constantes arbitrárias e são muito afetados pelo tamanho da amostra. Nas expressões anteriores  $L$  se refere a uma função de verossimilhança,  $k$  ao número de parâmetros e  $n$  ao tamanho da amostra utilizados na especificação de determinado modelo. Quando  $k$  é grande em relação ao tamanho da amostra  $n$ , existe a versão AIC corrigido,

cuja expressão é  $AICC = -2\log(L) + 2k + \frac{2k(k+1)}{n-k-1}$ . Podem ser observados valores de

AIC que variam desde 600 a 34000. Desta forma, é imperativo reescalonar os valores AIC para:

$\Delta_i = AIC_i - AIC_{\min}$ , em que  $AIC_i$  é o AIC do modelo de ordem  $i$  e  $AIC_{\min}$  é o menor dos  $r$  diferentes AIC, ambos tomados em valor absoluto. Esta transformação (reescalonamento) procura forçar que o melhor modelo tenha  $\Delta=0$ , enquanto que os restantes dos modelos tenham valores positivos. Veja exemplo hipotético a seguir:

$Modelo_i$	$AIC_i$	$\Delta_i$
M1	30	5
M2	25	0
M3	45	20

Os valores de  $\Delta_i$  são fáceis de interpretar e fornecem uma rápida e forte evidência para a comparação e o ordenamento de modelos candidatos. Quanto maior o valor de  $\Delta_i$ , menos plausível está o modelo  $i$  ajustado de ser o melhor modelo no conjunto dos modelos candidatos. Então, a expressão “o menor é o melhor” para os valores de AIC é sempre válida, ou seja, indiferentemente do sinal do AIC. Efetivamente, o melhor dentre os modelos candidatos é o que apresenta valor de AIC mais próximo de zero.

Geralmente, é importante saber qual modelo é o segundo melhor (o ordenamento dos modelos), bem como ter alguma medida desta posição em relação ao melhor modelo. Algumas regras práticas são frequentemente úteis para avaliar os méritos relativos de modelos em um conjunto. Uma delas é a seguinte: modelos tendo  $\Delta_i \leq 2$  têm suporte (evidência) substancial, aqueles em

que  $4 \leq \Delta_i \leq 7$  têm consideravelmente menos suporte, e, modelos que têm  $\Delta_i > 10$  apresentam essencialmente nenhum suporte. Todas estas inferências são válidas também para os valores de AICC e BIC.

**Ponderações de Akaike:** As ponderações de Akaike ( $w_i$ ) fornecem outra medida da força de evidência para cada modelo, e representam a razão de valores delta AIC ( $\Delta_i$ ) para cada modelo em relação a todo o conjunto de  $r$  modelos candidatos. A expressão da ponderação de Akaike é:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{i=1}^r \exp(-\Delta_i/2)}.$$

De fato, simplesmente é feita uma mudança da escala dos  $\Delta_i$ 's para compará-los numa escala de 1,0, ou seja, de forma que a soma dos  $w_i$  se iguala a 1,0. A interpretação do valor  $w_i$  é muito fácil: ele indica a probabilidade de que o modelo é o melhor dentre todo o conjunto de modelos candidatos. Por exemplo, uma ponderação Akaike de 0,75 para um modelo, indica que para os dados considerados, ele tem uma chance de 75% de ser o melhor dentre aqueles considerados no conjunto de modelos candidatos. Para o exemplo hipotético anterior tem-se:

Modelo $i$	$\Delta_i$	$w_i$
M1	5	0,075854
M2	0	0,924103
M3	20	0,000042

Desta forma, o modelo M2 teria 92,41% de probabilidade de ser o melhor dentre os três modelos candidatos.

**Razões de evidência:** Pode-se comparar a ponderação Akaike do “melhor” modelo com a dos modelos que estão competindo para determinar em que extensão um é melhor do que o outro.

Estas são denominadas de razões de evidência (RE) e são calculadas como:  $RE = \frac{w_j}{w_i}$ , onde o

modelo  $j$  é comparado com o modelo  $i$ . Por exemplo, uma razão de evidência de

$\frac{w_j}{w_i} = \frac{0,55}{0,40} = 1,375$  indicaria que o modelo  $j$  é apenas 1,375 mais evidente do que o modelo

$i$  para ser o melhor, dado o conjunto de  $r$  modelos candidatos e os dados.

Assim, o melhor modelo é o que apresenta o menor valor de AIC e de  $\Delta_i$ , a maior ponderação de Akaike ( $w_i$ ) e a maior razão de evidência (RE). Para o exemplo hipotético anterior têm-se as razões de evidência seguintes:

$$RE_1 = \frac{w_1}{w_2} = \frac{0,0758549}{0,9241031} = 0,0821$$

$$RE_2 = \frac{w_1}{w_3} = \frac{0,0758549}{0,00004195} = 1808,223$$

$$RE_3 = \frac{w_2}{w_1} = \frac{0,9241031}{0,0758549} = 12,183$$

$$RE_4 = \frac{w_2}{w_3} = \frac{0,9241031}{0,0000419} = 22054,967$$

$$RE_5 = \frac{w_3}{w_1} = \frac{0,00004195}{0,0758549} = 0,000553$$

$$RE_6 = \frac{w_3}{w_2} = \frac{0,0000419}{0,9241031} = 0,000045$$

Desta forma o modelo M2 seria 12 vezes melhor que o M1 e 22054 vezes melhor que o M3. O modelo M1 seria 1808 vezes melhor que o M3 e não seria melhor que o M2 ( $RE_1 = 0,082$ ).

### Teste razão de verossimilhança

Duas formas de usar as funções de verossimilhança para selecionar modelos ou para verificar/validar pressuposições são as seguintes:

1. Calcular a máxima verossimilhança dos dados amostrais com base num modelo de distribuição assumido (o máximo ocorre quando os parâmetros desconhecidos são substituídos pelas suas estimativas de máxima verossimilhança). Repetir estes cálculos para outros modelos de distribuição candidatos que parecem ajustar os dados. Se todos os modelos tiverem o mesmo número de parâmetros, e, não existe nenhuma razão convincente para escolher um determinado modelo em detrimento de outro, então escolha o modelo com o maior valor de verossimilhança.
2. Muitas pressuposições de modelos podem ser vistas como colocação de restrições sobre os parâmetros numa expressão de verossimilhança dos dados, que efetivamente reduza o número

de parâmetros desconhecidos. Nestes casos, existe uma forma simples e muito útil de testar se a suposição é consistente com os dados, que é o teste de razão de verossimilhança.

O teste de razão de verossimilhança é uma generalização dos teste-z, teste F, teste t e teste qui-quadrado. Neste teste é avaliada, por meio da estatística razão de verossimilhança (LR ou  $\lambda$ ), a aproximação de um modelo restrito (reduzido) ao modelo completo na determinação dos parâmetros. O Teste Razão de Verossimilhança é muito útil e amplamente aplicado, estando implementado nos principais aplicativos computacionais de estatística.

Procedimento do Teste Razão de Verossimilhança (TRV):

Seja  $L_1$  o valor máximo da verossimilhança dos dados sem a suposição adicional. Em outras palavras,  $L_1$  é a verossimilhança dos dados com todos os parâmetros não restritos e substituídos por estimativas de máxima verossimilhança.

Seja  $L_0$  o valor máximo da verossimilhança quando os parâmetros são restritos (e reduzidos em número) com base na suposição. Considere que  $k$  parâmetros foram perdidos (ou seja,  $L_0$  tem  $k$  parâmetros menos que  $L_1$ ).

Construa a razão  $\lambda = L_0 / L_1$ . Esta razão sempre está entre 0 e 1, e, quanto menos verdadeira (verossímil) for a suposição menor será o valor de  $\lambda$ . Isto pode ser quantificado para um dado nível de confiança como a seguir:

1. Calcule  $\chi^2 = -2 \ln \lambda$ . Quanto menor for  $\lambda$ , maior será o valor de  $\chi^2$ .
2. Pode-se saber quando o  $\chi^2$  é significativamente grande comparando-o com o  $100(1-\alpha)$  percentil superior de uma distribuição qui-quadrado com  $k$  graus de liberdade. O  $\chi^2$  tem uma distribuição qui-quadrado aproximada com  $k$  graus de liberdade e a aproximação é usualmente boa, mesmo para amostras de tamanho pequeno.
3. No Teste Razão de Verossimilhança (TRV) computa-se o  $\chi^2$  e rejeita-se a suposição se  $\chi^2$  é maior que o percentil qui-quadrado com  $k$  graus de liberdade, onde o percentil corresponde ao nível de confiança escolhido pelo analista.

Exemplo: Testar  $H_0: \lambda = \lambda_0$  versus  $H_1: \lambda \neq \lambda_0$  com base em  $n$  observações com distribuição de Poisson ( $\lambda$ ). Então, tem-se:

$$\begin{aligned}
-2 \ln \lambda(x) &= -2 \ln \left( \frac{e^{-n\lambda_0} \lambda_0^{\sum_{i=1}^n x_i}}{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}} \right) e \\
-2 \ln \lambda(x) &= 2n \left[ (\lambda_0 - \lambda) - \lambda \ln \left( \frac{\lambda_0}{\lambda} \right) \right], \text{ com } \lambda = \frac{\sum_{i=1}^n x_i}{n}.
\end{aligned}$$

Rejeita-se  $H_0$  se  $-2 \ln \lambda(x) > \chi^2_{1,\alpha}$

# CAPITULO 17

## Delineamentos Genéticos

### Introdução

Delineamento genético é um sistema planejado de cruzamento, realizado de forma que se conheça a relação de parentesco entre os indivíduos ou grupos de indivíduos envolvidos no estudo.

O delineamento genético possibilita a estimativa de componentes da variância genotípica que são utilizados para estimar parâmetros genéticos indispensáveis para a avaliação da potencialidade da população, orientação de esquemas apropriados de seleção e predição de ganhos genéticos.

Os principais delineamentos genéticos são:

**1) Teste de progênie:** avaliação de um único tipo de progênie, sendo um dos mais usados nos programas de melhoramento. Por exemplo, progênies de meios-irmãos, de irmãos completos e parcialmente endogâmicas.

Nos experimentos de avaliação de progênies têm-se, por exemplo, as seguintes análises estatístico-genéticas:

Considere o modelo linear  $Y_{ij} = m + g_i + b_j + \varepsilon_{ij}$ , em que:

$m$ : média geral;

$g_i$ : efeito da  $i$ -ésima progênie, com  $i=1,2,\dots,g$  e  $g_i \sim NID(0, \sigma_g^2)$ ;

$b_j$ : efeito do  $j$ -ésimo bloco, com  $j=1,2,\dots,r$  e  $b_j \sim NID(0, \sigma_b^2)$ ;

$\varepsilon_{ij}$ : erro associado a observação  $Y_{ij}$ , com  $\varepsilon_{ij} \sim NID(0, \sigma_e^2)$ .

Considera-se também que os efeitos aleatórios são independentes entre si.

O esquema de análise de variância associado a este modelo é o seguinte:

FV	GL	QM	E (QM)
Blocos	$r-1$	$QMB$	$\sigma^2 + g\sigma_b^2$

Progêneries	$g-1$	$QMG$	$\sigma^2 + r\sigma_g^2$
Resíduos	$(r-1)(g-1)$	$QMR$	$\sigma^2$

Com base nesta análise é possível realizar as seguintes estimativas de componentes de variância e parâmetros genéticos:

- a) Variância ambiental entre parcelas – refere-se à variância ambiental entre as  $gr$  parcelas avaliadas:

$$\hat{\sigma}^2 = QMR$$

- b) Variância ambiental entre médias de parcelas – refere-se à variância ambiental que existe entre as  $g$  médias das progêneries avaliadas:

$$\hat{\sigma}_e^2 = \frac{QMR}{r} = \frac{\hat{\sigma}^2}{r}$$

- c) Variância genética entre médias de progêneries – refere-se à variância genética entre as médias das progêneries avaliadas.

Esta variância é obtida como a covariância genética entre os indivíduos dentro da progénie, ou seja:

$$Cov(Y_{ij}, Y_{ij'}) = E(g_i^2) = \sigma_g^2.$$

A covariância genética entre progêneries de meios-irmãos é igual a 1/4 da variância genética aditiva. Então, tem-se:

$$\hat{\sigma}_g^2 = \frac{QMG - QMR}{r} = \frac{1}{4} \hat{\sigma}_A^2.$$

Com base neste modelo a estratégia de seleção mais apropriada é aquela em que as unidades seletivas são as médias das  $g$  progêneries. Neste caso, utiliza-se apenas a variabilidade entre e

desconsidera-se a variabilidade dentro de progêneries, que representa  $\frac{3}{4} \hat{\sigma}_A^2$ , além da variância de

desvios de dominância.

- d) Variância fenotípica entre médias de progêneries – ocorre que a seleção envolve a avaliação do contraste  $\bar{Y}_{i\cdot} - \bar{Y}_{i'}$  entre as médias  $\bar{Y}_{i\cdot}$  e  $\bar{Y}_{i'}$  de duas progêneries quaisquer, sendo o contraste expresso por:

$\bar{Y}_{i\cdot} - \bar{Y}_{i'} = (g_i - g_{i'}) + (\bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{i'})$ . Desta forma, a variância fenotípica entre unidades de seleção é estimada por meio de:

$$\hat{\sigma}_f^2 = \frac{QMG}{r} = \hat{\sigma}_g^2 + \frac{1}{r} \hat{\sigma}^2.$$

e) Herdabilidade – estimada com base nos componentes de variância, sendo expressa como:

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_f^2} = \frac{\hat{\sigma}_g^2}{QMG/r}.$$

#### Exemplo de Aplicação

A seguir são apresentados os resultados da análise de variância para o caráter produtividade de grãos avaliada em dez progênies de meios-irmãos, num experimento em blocos casualizados com três repetições (Extraído de Cruz, 2005):

FV	GL	QM	F
Blocos	2	356,6080	-
Progênies	9	709,4248	4,82 (p<0,01)
Resíduos	18	147,0680	

Então, podem ser estimados os seguintes componentes de variância:

Variância ambiental

$$\hat{\sigma}^2 = QMR = 147,0680$$

Variância ambiental entre médias de parcelas

$$\hat{\sigma}_e^2 = \frac{\hat{\sigma}^2}{r} = \frac{QMR}{r} = \frac{147,0680}{3} = 49,0227$$

Variância genética entre médias de progênies

$$\hat{\sigma}_g^2 = \frac{QMG - QMR}{r} = \frac{709,4248 - 147,0680}{3} = 187,4523$$

Variância fenotípica entre médias de progênies

$$\hat{\sigma}_f^2 = \frac{QMG}{r} = \frac{709,4248}{3} = 236,4749$$

Herdabilidade no sentido amplo entre médias de progênies

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_f^2} = \frac{187,4523}{236,4749} = 0,7927$$

Teste de progénie com avaliação de plantas dentro de parcelas – neste caso são obtidos os dados de plantas individuais dentro de parcelas, o que permite a seleção entre progênies e dentro de progênies, a seleção de plantas individuais (seleção massal) dentro do experimento ou a seleção de plantas individuais dentro de cada bloco (seleção massal estratificada).

O modelo estatístico-genético associado a este delineamento genético é o seguinte:

$$Y_{ijk} = m + g_i + b_j + \varepsilon_{ij} + \delta_{ijk}, \text{ em que:}$$

$m$ : média geral;

$g_i$ : efeito da  $i$ -ésima progénie, com  $i=1,2,\dots,g$  e  $g_i \sim NID(0, \sigma_g^2)$ ;

$b_j$ : efeito do  $j$ -ésimo bloco, com  $j=1,2,\dots,r$  e  $b_j \sim NID(0, \sigma_b^2)$

$\varepsilon_{ij}$ : efeito aleatório ambiental existente entre parcelas, com  $\varepsilon_{ij} \sim NID(0, \sigma_e^2)$ ;

$\delta_{ijk}$ : efeito aleatório existente entre plantas dentro de parcela, com  $k=1,2,\dots,n$  e

$\delta_{ijk} \sim NID(0, \sigma_d^2)$ .

O esquema de análise de variância de um experimento em blocos casualizados com avaliação de plantas dentro das parcelas é o seguinte:

FV	GL	QM	E (QM)
Blocos	$r-1$	QMB	$\sigma_d^2 + n\sigma_e^2 + ng\sigma_b^2$
Progénies	$g-1$	QMG	$\sigma_d^2 + n\sigma_e^2 + nr\sigma_g^2$
Entre Parcelas	$(g-1)(r-1)$	QME	$\sigma_d^2 + n\sigma_e^2$
Dentro de Parcada		QMD	
	$(n-1)gr$		$\sigma_d^2$

Observe que estes componentes de variância, que são de natureza genética ou ambiental, estão associados a plantas individuais, parcelas ou média de progénies.

Então se tem os seguintes estimadores:

a) Estimador do componente de variância genética proporcionada pelas diferenças entre as médias de  $g$  progénies avaliadas  $(\hat{\sigma}_g^2)$  - os componentes genéticos aditivos e de dominância

desta variância são obtidos por meio da avaliação da covariância entre as plantas individuais parentadas que estão sendo avaliadas.

Por exemplo, se as progénies são de meios-irmãos tem-se que  $\hat{\sigma}_g^2 = \frac{1}{4}\hat{\sigma}_A^2$  e se são de irmãos

completos tem-se que  $\hat{\sigma}_g^2 = \frac{1}{2}\hat{\sigma}_A^2 + \frac{1}{4}\hat{\sigma}_D^2$ .

- b) Estimador do componente de variância ambiental proporcionada pelas diferenças entre as médias de  $r$  blocos  $(\hat{\sigma}_b^2)$ .
- c) Estimador do componente de variância fenotípica entre plantas dentro de parcela  $(\hat{\sigma}_d^2)$ , proporcionada pelas diferenças entre  $ngr$  observações de plantas individuais.

Este estimador é expresso por:  $\hat{\sigma}_d^2 = \hat{\sigma}_{gd}^2 + \hat{\sigma}_{ed}^2$ , em que:

$\hat{\sigma}_{gd}^2$ : variância genética dentro de progênies ou entre plantas individuais dentro de parcela; para progênies de meios-irmãos tem-se que  $\hat{\sigma}_{gd}^2 = 3\hat{\sigma}_g^2 = \frac{3}{4}\hat{\sigma}_A^2$ ;

$\hat{\sigma}_{ed}^2$ : variância ambiental dentro de progênies, proporcionada pelas diferenças entre plantas individuais.

- d) Estimador do componente de variância ambiental entre parcelas  $(\hat{\sigma}_e^2)$ , proporcionada pelas diferenças entre as  $gr$  médias.

Observe que o componente de variância fenotípica entre médias de progênies, expressa como  $\sigma_g^2 + (1/r)\sigma_e^2 + (1/nr)\sigma_d^2$ , tem um coeficiente apropriado para cada componente porque  $\sigma_g^2$ ,  $\sigma_e^2$  e  $\sigma_d^2$  estão associados a  $g$ ,  $gr$  e  $ngr$  observações, respectivamente.

Os estimadores dos componentes de variância são descritos por meio das expressões seguintes:

$$\hat{\sigma}_b^2 = \frac{QMB - QME}{ng}$$

$$\hat{\sigma}_g^2 = \frac{QMG - QME}{nr}$$

$$\hat{\sigma}_e^2 = \frac{QME - QMD}{n}$$

$$\hat{\sigma}_d^2 = QMD$$

O coeficiente de herdabilidade é um atributo do caráter e da população sob seleção e sua estimação deve estar de acordo com o método de seleção a ser utilizado. A herdabilidade é importante na predição de ganho com a seleção em várias estratégias de melhoramento. Neste contexto o coeficiente de herdabilidade deve ser expresso como:

$$h^2 = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_F^2}, \text{ em que:}$$

$\hat{\sigma}_A^2$ : Variância genética aditiva entre unidades de seleção;

$\hat{\sigma}_F^2$ : Variância fenotípica entre unidades de seleção.

As estimativas de coeficientes de herdabilidades podem ser feitas para os seguintes métodos de seleção (esquemas seletivos):

- a) **Seleção entre médias de progênies** – refere-se à seleção das melhores progênies usando como critério o desempenho médio  $(\bar{Y}_{i..})$  das  $g$  progênies avaliadas. Neste caso tem-se:

Unidade de seleção – médias de progênies  $(\bar{Y}_{i..})$ .

Variância genética – estimador

$$\hat{\sigma}_g^2 = \frac{QMG - QME}{nr}.$$

Variância fenotípica – estimador

$$\hat{\sigma}_f^2 = \hat{\sigma}_g^2 + (1/r)\hat{\sigma}_e^2 + (1/nr)\hat{\sigma}_d^2 = \frac{QMG}{nr}.$$

Herdabilidade – estimador

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_f^2} = \frac{\hat{\sigma}_g^2}{QMG/nr}.$$

- b) **Seleção entre plantas de progênies selecionadas** – refere-se à seleção de plantas individuais dentro de cada parcela de progênies superiores, tendo como critério os valores  $Y_{ijk}$  para cada progénie  $i$  selecionada, dentro de cada repetição  $j$ . Neste caso tem-se:

Unidade de seleção – valores  $Y_{ijk}$  de  $n$  plantas individuais dentro de cada progénie selecionada em cada bloco.

Variância genética dentro de progénie – estimador  $\hat{\sigma}_{gd}^2 = 3\hat{\sigma}_g^2 = \frac{3}{4}\hat{\sigma}_A^2$ .

Variância fenotípica dentro de progénie – estimador  $\hat{\sigma}_d^2 = QMD$ .

Herdabilidade – estimador  $h^2 = \frac{3\hat{\sigma}_g^2}{\hat{\sigma}_d^2}$ .

c) **Seleção massal estratificada** – refere-se à seleção de plantas individuais dentro de cada bloco, tendo como critério os valores  $Y_{ijk}$ , para  $j=1,2,\dots,r$ . Neste caso tem-se:

Unidade de seleção – valores  $Y_{ijk}$  de  $ng$  plantas dentro de cada bloco.

Variância genética entre plantas, que corresponde à variância genética total – estimador

$$\hat{\sigma}_A^2 = \hat{\sigma}_{ge}^2 + \hat{\sigma}_{gd}^2 = 4\hat{\sigma}_g^2 = \frac{1}{4}\hat{\sigma}_A^2 + \frac{3}{4}\hat{\sigma}_A^2.$$

Variância fenotípica entre plantas – estimador

$$\begin{aligned}\hat{\sigma}_f^2 &= \hat{\sigma}_{ge}^2 + (\hat{\sigma}_{gd}^2 + \hat{\sigma}_{ed}^2) + \hat{\sigma}_{ee}^2 \\ \hat{\sigma}_f^2 &= \hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2\end{aligned}$$

Herdabilidade – estimador

$$h^2 = \frac{4\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2}$$

d) **Seleção massal** – refere-se à seleção de plantas individuais no experimento, ignorando os blocos. Neste caso tem-se:

Unidade de seleção – valores  $Y_{ijk}$  de  $ngr$  plantas dentro do experimento.

Variância genética entre plantas individuais, que corresponde à variância genética total – estimador

$$\hat{\sigma}_A^2 = 4\hat{\sigma}_g^2.$$

Variância fenotípica entre plantas individuais, que neste caso inclui a variância ambiental entre parcelas proporcionada pela variação de blocos – estimador

$$\hat{\sigma}_f^2 = \hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2 + \hat{\sigma}_b^2.$$

Herdabilidade – estimador

$$h^2 = \frac{4\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2 + \hat{\sigma}_b^2}.$$

### Exemplo de Aplicação

Considere os resultados da análise de variância da avaliação de oito progênies de meios-irmãos em blocos casualizados com seis repetições e com observações de seis plantas dentro de cada parcela (Extraído de Cruz, 2005):

FV	GL	QM	E (QM)
Blocos	5	358039,787	$\sigma_d^2 + n\sigma_e^2 + ng\sigma_b^2$
Progênies	7	632466,158	
Entre Parcelas	35	199981,387	$\sigma_d^2 + n\sigma_e^2 + nr\sigma_g^2$
Dentro de Parcelas	240	44044,257	$\sigma_d^2$

Média = 415,879; n = 6; r = 6; g = 8;

Então, as estimativas dos componentes de variância são:

$$\hat{\sigma}_b^2 = \frac{QMB - QME}{ng} = 3292,883$$

$$\hat{\sigma}_g^2 = \frac{QMG - QME}{nr} = 12013,466$$

$$\hat{\sigma}_e^2 = \frac{QME - QMD}{n} = 25989,522$$

$$\hat{\sigma}_d^2 = QMD = 44044,257$$

As estimativas de herdabilidades para diferentes esquemas seletivos são as seguintes:

Seleção entre médias de progênies –

$$h^2 = \frac{\hat{\sigma}_g^2}{QMG/nr} = 0,6838.$$

Seleção dentro da parcela, entre plantas individuais de progênies selecionadas –

$$h^2 = \frac{3\hat{\sigma}_g^2}{\hat{\sigma}_d^2} = 0,8183.$$

Seleção massal estratificada –

$$h^2 = \frac{4\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2} = 0,5857.$$

Seleção massal –

$$h^2 = \frac{4\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2 + \hat{\sigma}_b^2} = 0,5631.$$

1) Delineamento I de Comstock e Robinson – consiste em cruzamentos que envolvem um número  $m$  de genitores masculinos, sendo que cada um é cruzado com um número variável de diferentes genitores femininos. Ilustração:

M1	F1	D111	D112	D113	D114	D115
	F2	D121	D122	D123	D124	D125
	F3	D131	D132	D133	D134	D135
	F4	D141	D142	D143	D144	D145
	F5	D151	D152	D153	D154	D155
M2	F6	D261	D262	D263	D264	D265
	F7	D271	D272	D273	D274	D275
	F8	D281	D282	D283	D284	D281
	F9	D291	D292	D293	D294	D295
	F10	D2101	D2102	D2103	D2104	D2105
M3	F11	D3111	D3112	D3113	D3114	D3115
	F12	D3121	D3122	D3123	D3124	D3125
	F13	D3131	D3132	D3133	D3134	D3135
	F14	D3141	D3142	D3143	D3144	D3145
	F15	D3151	D3152	D3153	D3154	D3155

M: machos; F: fêmeas; D: descendentes.

Observe que dentro de cada macho (M) nas linhas estão os descendentes que são irmãos completos (IC) e nas colunas os descendentes que são meios-irmãos paternos (MI).

Este delineamento é caracterizado pela classificação hierárquica, na qual são avaliados machos (M) e fêmeas dentro de machos (F/M), conforme descrito pelo seguinte modelo estatístico-genético:

$$Y_{ijk} = m + M_i + F/M_{ij} + \varepsilon_{ijk}, \text{ em que:}$$

$Y_{ijk}$ : observação no  $k$ -ésimo descendente resultante do cruzamento do  $i$ -ésimo macho (M) com

a  $j$ -ésima fêmea (F);

$m$ : média geral;

$M_i$ : efeito do  $i$ -ésimo macho, com  $i=1,2,\dots,m$  e  $M_i \sim NID(0, \sigma_m^2)$ ;

$F/M_{ij}$ : efeito da  $j$ -ésima fêmea dentro do  $i$ -ésimo macho e considerando que cada macho é

cruzado com um grupo fixo de fêmeas, tem-se:  $j=1,2,\dots,f$  e  $F/M_{ij} \sim NID(0, \sigma_{f/m}^2)$ ;

$\varepsilon_{ijk}$ : efeito do  $k$ -ésimo descendente dos  $i$ -ésimo macho e  $j$ -ésima fêmea, com  $k=1,2,\dots,n$

e  $\varepsilon_{ijk} \sim NID(0, \sigma_d^2)$ .

O esquema de análise de variância do delineamento I de Comstock e Robinson é o seguinte:

FV	GL	QM	E (QM)
Tratamentos	$mf-1$	QMT	$\sigma_d^2 + n\sigma_{f/m}^2 + n\frac{mf-f}{mf-1}\sigma_m^2$
Machos (M)	$m-1$	QMM	$\sigma_d^2 + n\sigma_{f/m}^2 + nf\sigma_m^2$
Fêmeas/Machos (F/M)	$m(f-1)$	QMF	$\sigma_d^2 + n\sigma_{f/m}^2$
Descendentes/(F/M)	$(n-1)mf$	QMD	$\sigma_d^2$

Os estimadores dos componentes de variância são os seguintes:

$$\hat{\sigma}_d^2 = QMD$$

$$\hat{\sigma}_{f/m}^2 = \frac{QMF - QMD}{n}$$

$$\hat{\sigma}_m^2 = \frac{QMM - QMF}{nf}$$

O conteúdo genético destes componentes é estabelecido por meio do conhecimento da covariância entre os indivíduos aparentados. Neste caso são avaliados indivíduos que são meios-irmãos e irmãos completos cujas covariâncias são as seguintes:

a) Covariância entre irmãos completos – expressa por

$$Cov(IC) = Cov(Y_{ijk}, Y_{ijk'})$$
. Para o modelo estatístico aleatório tem-se:

$$Cov(IC) = E(M_i^2) + E(F/M_{ij}^2) = \sigma_m^2 + \sigma_{f/m}^2$$

A covariância genética entre irmãos completos é expressa como:  $Cov(IC) = \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$ .

b) Covariância entre meios-irmãos – expressa por:

$$Cov(MI) = Cov\left(Y_{ijk}, Y_{ij'k'}\right). \quad \text{Para o modelo estatístico aleatório tem-se:}$$

$$Cov(MI) = E\left(M_i^2\right) = \sigma_m^2.$$

A covariância genética entre meios-irmãos é expressa como:

$$Cov(MI) = \frac{1}{4}\sigma_A^2.$$

Então, têm-se os seguintes estimadores:

$$\hat{Cov}(IC) = \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2 = \frac{1}{2}\hat{\sigma}_A^2 + \frac{1}{4}\hat{\sigma}_D^2 \text{ e}$$

$$\hat{Cov}(MI) = \hat{\sigma}_m^2 = \frac{1}{4}\hat{\sigma}_A^2.$$

$$\hat{\sigma}_D^2 = 4\left[\left(\hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2\right) - \frac{1}{2}\left(4\hat{\sigma}_m^2\right)\right], \text{ ou de forma}$$

$$\hat{\sigma}_D^2 = 4\left(\hat{\sigma}_{f/m}^2 - \hat{\sigma}_m^2\right)$$

Deduz-se então que:  $\hat{\sigma}_A^2 = 4\hat{\sigma}_m^2$  e

equivalente tem-se que:

$$\hat{\sigma}_m^2 = \frac{1}{4}\hat{\sigma}_A^2 \text{ e } \hat{\sigma}_{f/m}^2 = \frac{1}{4}\hat{\sigma}_A^2 + \frac{1}{4}\hat{\sigma}_D^2.$$

Considerando o modelo estatístico aleatório:

$$Y_{ijk} = m + M_i + F/M_{ij} + \varepsilon_{ijk}, \text{ tem-se que:}$$

$$\hat{\sigma}_y^2 = \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2 + \hat{\sigma}_d^2 = \hat{\sigma}_g^2 + \hat{\sigma}_e^2, \text{ em que:}$$

$\hat{\sigma}_g^2$ : variância genética total, sendo  $\hat{\sigma}_g^2 = \hat{\sigma}_A^2 + \hat{\sigma}_D^2$ ,

$\hat{\sigma}_e^2$ : variância ambiental.

Então se deduz que:

$$\hat{\sigma}_d^2 = \frac{1}{2}\hat{\sigma}_A^2 + \frac{3}{4}\hat{\sigma}_D^2 + \hat{\sigma}_e^2.$$

Para este delineamento genético, dependendo do método de seleção adotado, podem ser estimados os seguintes coeficientes de herdabilidade:

1) **Seleção com base na média de machos** – a seleção é realizada entre os valores médios obtidos para cada macho.

São comparados valores de

$$\bar{Y}_{i..} = \mu + M_i + \frac{1}{f} \sum_{j=1}^f F/M_{ij} + \frac{1}{nf} \sum_{j=1}^f \sum_{k=1}^n \varepsilon_{ijk} \text{ com}$$

$$\bar{Y}_{i'..} = \mu + M_{i'} + \frac{1}{f} \sum_{j=1}^f F/M_{i'j} + \frac{1}{nf} \sum_{j=1}^f \sum_{k=1}^n \varepsilon_{i'jk}$$

Tem-se que:

$$V(\bar{Y}_{i..}) = \hat{\sigma}_{Fm}^2 = \hat{\sigma}_m^2 + \frac{\hat{\sigma}_{f/m}^2}{f} + \frac{\hat{\sigma}_d^2}{nf} = \frac{QMM}{nf}, \text{ sendo } \hat{\sigma}_{Fm}^2 \text{ o estimador da variância fenotípica entre as médias de machos.}$$

A variância genotípica total entre as médias de machos submetidos à seleção é expressa por

$$\hat{\sigma}_m^2 + \left( \hat{\sigma}_{f/m}^2 / f \right).$$

$$\text{Então, a herdabilidade no sentido amplo é calculada como } h^2 = \frac{\hat{\sigma}_m^2 + \left( \hat{\sigma}_{f/m}^2 / f \right)}{\hat{\sigma}_{Fm}^2}.$$

Pode ser verificado que:

$$\hat{\sigma}_m^2 + \frac{\hat{\sigma}_{f/m}^2}{f} = \frac{1}{4f} \left[ (f+1)\hat{\sigma}_A^2 + \hat{\sigma}_D^2 \right].$$

Desta forma, a variância genotípica pode ser estimada com base na covariância que ocorre dentro da unidade de seleção:

Se cada macho foi cruzado com uma única fêmea ( $f=1$ ), então a variância genética entre unidades de seleção seria aquela própria de progêniess de irmãos completos, ou seja:

$$\hat{\sigma}_m^2 + \frac{\hat{\sigma}_{f/m}^2}{f} = \frac{1}{2}\hat{\sigma}_A^2 + \frac{1}{4}\hat{\sigma}_D^2.$$

Se cada macho foi acasalado com um número infinito de fêmeas ( $f=\infty$ ), então a variância genética seria semelhante à de progênieis de meios-irmãos, ou seja:

$$\hat{\sigma}_m^2 + \frac{\hat{\sigma}_{f/m}^2}{f} = \frac{1}{4} \hat{\sigma}_A^2.$$

Considerando que neste delineamento genético são comparados todos os machos que foram acasalados com um número finito de fêmeas ( $f$ ), dois indivíduos dentro de um mesmo macho podem ser irmãos completos, com probabilidade  $1/f$  ou meios-irmãos, com probabilidade  $(f-1)/f$ . Então, espera-se que:

$$Cov(M) = \frac{1}{f} Cov(IC) + \frac{f-1}{f} Cov(MI), \text{ em que:}$$

**Cov(M)**: covariância dentro de machos, que quantifica a variância genética entre as médias dos machos selecionados;

**Cov(IC)**: covariância genética entre irmãos completos, sendo expressa por

$$Cov(IC) = \frac{1}{2} \sigma_A^2 + \frac{1}{4} \sigma_D^2;$$

**Cov(MI)**: covariância genética entre meios-irmãos, sendo expressa por  $Cov(MI) = \frac{1}{4} \sigma_A^2$ .

Então, deduz-se que:

$$Cov(M) = \frac{1}{f} Cov(IC) + \frac{f-1}{f} Cov(MI)$$

$$Cov(M) = \frac{1}{4f} \left[ (f+1)\sigma_A^2 + \sigma_D^2 \right]$$

2) **Seleção com base na média de fêmeas/macho** – progênieis de cada grupo de fêmeas são selecionadas dentro de cada macho.

Isto caracteriza a seleção estratificada de fêmeas. Os valores médios para cada fêmea, dentro de cada macho são comparados, ou seja, comparam-se os seguintes valores:

$$\bar{Y}_{ij.} = \mu + M_i + F/M_{ij} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ijk} \text{ e}$$

$$\bar{Y}_{ij'.} = \mu + M_i + F/M_{ij'} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ij'k}$$

Tem-se que:

$$V(\bar{Y}_{ij.}) = \hat{\sigma}_{Ff}^2 = \hat{\sigma}_{f/m}^2 + \frac{\hat{\sigma}_d^2}{n} = \frac{QMF}{n}, \text{ em que:}$$

$\hat{\sigma}_{Ff}^2$ : variância fenotípica entre médias de fêmeas, para determinado macho.

A variância genotípica que ocorre entre as médias de fêmeas para um dado macho é expressa por

$$\hat{\sigma}_{f/m}^2. \text{ Então, a herdabilidade no sentido amplo pode ser calculada como } h^2 = \frac{\hat{\sigma}_{f/m}^2}{\hat{\sigma}_{Ff}^2}.$$

Neste caso, a variância genética  $\hat{\sigma}_{f/m}^2$  não pode ser estimada a partir da covariância que se manifesta na unidade de seleção  $(f/m)$ . Ocorre que a seleção dentro de fêmeas, para um determinado macho, envolve a covariância entre progênies de irmãos completos que é expressa por  $Cov(IC) = (1/2)\sigma_A^2 + (1/4)\sigma_D^2$ . Entretanto, a variância genética estimada com base nesta covariância está superestimada, porque não considera que as progênies são provenientes de um único macho. Então, neste caso tem-se que:

$$\hat{\sigma}_{f/m}^2 = Cov(IC) - \hat{\sigma}_m^2 \text{ e então tem-se:}$$

$$\hat{\sigma}_{f/m}^2 = \left( \frac{1}{2}\hat{\sigma}_A^2 + \frac{1}{4}\hat{\sigma}_D^2 \right) - \frac{1}{4}\hat{\sigma}_A^2 = \frac{1}{4}(\hat{\sigma}_A^2 + \hat{\sigma}_D^2)$$

3) **Seleção com base na média de fêmeas** – a seleção é feita entre fêmeas, independente do macho com que elas foram acasaladas. Neste caso são feitas comparações de valores médios de fêmeas acasaladas com um mesmo macho e de fêmeas acasaladas com diferentes machos. Os valores médios fenotípicos para as comparações entre fêmeas acasaladas com um mesmo macho são expressos como:

$$\bar{Y}_{ij.} = \mu + M_i + F/M_{ij} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ijk} \text{ e}$$

$$\bar{Y}_{ij'.} = \mu + M_i + F/M_{ij'} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ij'k}, \text{ sendo a variância destes valores expressa por:}$$

$$V(\bar{Y}_{ij.}) = \hat{\sigma}_{Ff}^2 = \hat{\sigma}_{f/m}^2 + \frac{\hat{\sigma}_d^2}{n} = \frac{QMF}{n}.$$

Os valores médios fenotípicos para as comparações que envolvem fêmeas acasaladas com diferentes machos são expressos como:

$$\bar{Y}_{ij\cdot} = \mu + M_i + F/M_{ij} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{ijk} \text{ e}$$

$$\bar{Y}_{i'j'} = \mu + M_{i'} + F/M_{i'j'} + \frac{1}{n} \sum_{k=1}^n \varepsilon_{i'j'k}$$

Observe que neste caso, fêmeas acasaladas com diferentes machos poderão ser comparadas no processo de seleção. Então, na comparação dos valores médios de duas fêmeas quaisquer os efeitos de fêmeas, de macho e do erro são variáveis. Assim, a variância geral dos valores fenotípicos médios seria expressa como:

$$V(\bar{Y}_{ij\cdot}) = \hat{\sigma}_{Fgf}^2 = \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2 + \frac{\hat{\sigma}_d^2}{n}.$$

Ocorre que neste delineamento genético, duas fêmeas quaisquer podem estar associadas a um mesmo macho, com probabilidade igual a  $(f-1)/(mf-1)$ , ou a machos diferentes com probabilidade igual a  $(mf-f)/(mf-1)$ . Desta forma, a variância fenotípica entre médias de fêmeas é expressa por:

$$V(Fenotípica) = \hat{\sigma}_{Ffm}^2$$

$$\hat{\sigma}_{Ffm}^2 = \frac{mf-f}{mf-1} \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2 + \frac{1}{n} \hat{\sigma}_d^2.$$

A variância genética total que ocorre entre as médias de fêmeas submetidas à seleção é expressa por:

$$V(Genotípica) = \frac{mf-f}{mf-1} \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2, \text{ sendo } \theta = \frac{2mf-f-1}{mf-1}.$$

$$V(Genotípica) = \theta \frac{1}{4} \hat{\sigma}_A^2 + \frac{1}{4} \hat{\sigma}_D^2$$

A herdabilidade no sentido amplo pode ser obtida com a seguinte expressão:

$$h^2 = \frac{\frac{mf-f}{mf-1} \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2}{\hat{\sigma}_{Ffm}^2}, \text{ em que:}$$

$\hat{\sigma}_{Ffm}^2$ : variância fenotípica entre médias de fêmeas, que é obtida como,

$$\hat{\sigma}_{Ffm}^2 = \frac{(m-1)QMM + m(f-1)QMF}{n(mf-1)}.$$

### Exemplo de Aplicação

Considere os resultados da análise de variância da avaliação de quatro plantas individuais obtidas com os cruzamentos de seis genitores masculinos com grupos de oito diferentes genitores femininos de acordo com o delineamento I de Comstock e Robinson, apresentados a seguir (Extraído de Cruz, 2005):

FV	GL	SQ	QM	F
Tratamentos	47	1203,80	25,6	3,01*
Machos (M)	5	263,00	52,6	2,35*
Fêmeas/Machos (F/M)	42	940,80	22,4	2,63**
Descendentes/F/M	144	1224,00	8,5	

Pede-se: a) Estimar a variância genética aditiva e de dominância; b) Estimar a herdabilidade no sentido amplo para diferentes esquemas seletivos.

Resolução:

a) Com base nos resultados da análise de variância são obtidos os seguintes componentes de variância:

$$\hat{\sigma}_d^2 = QMD = 8,500$$

$$\hat{\sigma}_{f/m}^2 = \frac{QMF - QMD}{n} = \frac{22,4 - 8,5}{4} = 3,475$$

$$\hat{\sigma}_m^2 = \frac{QMM - QMF}{nf} = \frac{52,6 - 22,4}{(4)(8)} = 0,944$$

Então, a variância genética aditiva e a variância devido à dominância são estimadas como:

$$\hat{\sigma}_A^2 = 4\hat{\sigma}_m^2 = 4(0,944) = 3,775$$

$$\hat{\sigma}_D^2 = 4(\hat{\sigma}_{f/m}^2 - \hat{\sigma}_m^2) = 10,124$$

Para a seleção com base na média de machos, a herdabilidade no sentido amplo é calculada como:

$$h^2 = \frac{\hat{\sigma}_m^2 + (\hat{\sigma}_{f/m}^2/f)}{QMM/nf} = \frac{0,944 + (3,475/8)}{52,6/32} = 0,8385$$

Para a seleção com base na média de fêmeas/macho, a herdabilidade no sentido amplo é calculada como:

$$h^2 = \frac{\hat{\sigma}_{f/m}^2}{QMF/n} = \frac{3,475}{22,4/4} = 0,6205$$

Para a seleção com base na média de fêmeas, a herdabilidade no sentido amplo é calculada como:

$$\hat{\sigma}_{Ffm}^2 = \frac{(m-1)QMM + m(f-1)QMF}{n(mf-1)}$$

$$\hat{\sigma}_{Ffm}^2 = \frac{5(52,6) + 42(22,4)}{4(47)} = 6,4011$$

$$h^2 = \frac{\frac{mf-f}{mf-1} \hat{\sigma}_m^2 + \hat{\sigma}_{f/m}^2}{\hat{\sigma}_{Ffm}^2}$$

$$h^2 = \frac{\frac{40}{47}(0,944) + 3,475}{6,4011} = 0,6684$$

4) **Delineamento II de Comstock e Robinson** – neste delineamento genético um grupo de  $m$  genitores masculinos é acasalado com as mesmas  $f$  fêmeas, de forma que cada fêmea é acasalada com todos os machos e cada macho é acasalado com todas as fêmeas, produzindo todas as possíveis  $mf$  progênies. Ilustração:

M1	F1	D111	D112	D113	D114	D115
	F2	D121	D122	D123	D124	D125
	F3	D131	D132	D133	D134	D135
	F4	D141	D142	D143	D144	D145
	F5	D151	D152	D153	D154	D155
M2	F1	D211	D212	D213	D214	D215
	F2	D221	D222	D223	D224	D225
	F3	D231	D232	D233	D234	D231
	F4	D241	D242	D243	D244	D245
	F5	D251	D252	D253	D254	D255
M3	F1	D311	D312	D313	D314	D315
	F2	D321	D322	D323	D324	D325
	F3	D331	D332	D333	D334	D335
	F4	D341	D342	D343	D344	D345
	F5	D351	D352	D353	D354	D355

M: machos; F: fêmeas; D: descendentes.

Este delineamento se caracteriza como factorial porque são avaliados os machos, as fêmeas e a interação entre macho e fêmea, como expresso no seguinte modelo linear:

$$Y_{ijk} = m + M_i + F_j + MF_{ij} + \varepsilon_{ijk}, \text{ em que:}$$

$Y_{ijk}$ : observação do  $k$ -ésimo descendente do cruzamento do  $i$ -ésimo macho com a  $j$ -ésima fêmea;

$m$ : média geral;

$M_i$ : efeito do  $i$ -ésimo macho, com  $i=1,2,\dots,m$  e  $M_i \sim NID(0, \sigma_m^2)$ ;

$F_j$ : efeito da  $j$ -ésima fêmea, com  $j=1,2,\dots,f$  e  $F_j \sim NID(0, \sigma_f^2)$ ;

$MF_{ij}$ : efeito da interação entre o  $i$ -ésimo macho e a  $j$ -ésima fêmea, com

$MF_{ij} \sim NID(0, \sigma_{mf}^2)$ ;

$\varepsilon_{ijk}$ : efeito do  $k$ -ésimo descendente do cruzamento entre o  $i$ -ésimo macho e a  $j$ -ésima fêmea, sendo

$\varepsilon_{ijk} \sim NID(0, \sigma^2)$ .

O esquema para a análise de variância do delineamento II de Comstock e Robinson é o seguinte:

FV	GL	QM	E (QM)
Machos (M)	$m-1$	QMM	$\sigma^2 + n\sigma_{mf}^2 + nf\sigma_m^2$
Fêmeas (F)	$f-1$	QMF	$\sigma^2 + n\sigma_{mf}^2 + nm\sigma_f^2$
$M \times F$	$(m-1)(f-1)$	QMI	$\sigma^2 + n\sigma_{mf}^2$
Descendentes/ $M \times F$	$(n-1)mf$	QMD	$\sigma^2$

Com base na análise de variância são obtidos os seguintes estimadores de componentes de variância:

$$\hat{\sigma}^2 = QMD$$

$$\hat{\sigma}_{mf}^2 = \frac{QMI - QMD}{n}$$

$$\hat{\sigma}_f^2 = \frac{QMF - QMI}{nm}$$

$$\hat{\sigma}_m^2 = \frac{QMM - QMI}{nf}$$

Neste delineamento genético as covariâncias entre os indivíduos aparentados são as seguintes:

a) Covariância entre irmãos completos -  $Cov(IC) = Cov(Y_{ijk}, Y_{ijk'})$ :

Na análise estatística tem-se que:

$$Cov(IC) = \sigma_m^2 + \sigma_f^2 + \sigma_{mf}^2.$$

Na análise genética tem-se que:

$$Cov(IC) = \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2.$$

- b) Covariância entre meios-irmãos maternos -  $Cov(MIM) = Cov(Y_{ijk}, Y_{i'jk'})$ :

Pela análise estatística tem-se que:

$$Cov(MIM) = \sigma_f^2.$$

Na análise genética tem-se que:

$$Cov(MIM) = \frac{1}{4}\sigma_A^2.$$

- c) Covariância entre meios-irmãos paternos -  $Cov(MIP) = Cov(Y_{ijk}, Y_{ij'k'})$ :

Pela análise estatística tem-se que:

$$Cov(MIP) = \sigma_m^2.$$

Na análise genética tem-se que:

$$Cov(MIP) = \frac{1}{4}\sigma_A^2.$$

Com base nas expressões anteriores têm-se os seguintes estimadores:

$$\hat{\sigma}_m^2 = \frac{1}{4}\hat{\sigma}_A^2$$

$$\hat{\sigma}_f^2 = \frac{1}{4}\hat{\sigma}_A^2$$

$$\hat{\sigma}_{mf}^2 = \frac{1}{4}\hat{\sigma}_D^2$$

$$\hat{\sigma}^2 = \frac{1}{2}\hat{\sigma}_A^2 + \frac{3}{4}\hat{\sigma}_D^2 + \hat{\sigma}_e^2$$

$$\hat{\sigma}_A^2 = 2(\hat{\sigma}_m^2 + \hat{\sigma}_f^2)$$

$$\hat{\sigma}_D^2 = 4\hat{\sigma}_{mf}^2.$$

Neste delineamento genético, os estimadores de herdabilidade para diferentes esquemas seletivos são:

- 1) Seleção entre genitores masculinos com base na média de desempenho da progênie:

As médias a serem comparadas são:

$$\bar{Y}_{i..} = \mu + M_i + \frac{1}{f} \sum_j F_j + \frac{1}{f} \sum_j MF_{ij} + \frac{1}{nf} \sum_j \sum_k \varepsilon_{ijk} \text{ e}$$

$$\bar{Y}_{i'..} = \mu + M_{i'} + \frac{1}{f} \sum_j F_j + \frac{1}{f} \sum_j MF_{i'j} + \frac{1}{nf} \sum_j \sum_k \varepsilon_{i'jk}$$

Os contrastes entre as médias são expressos como:

$$\begin{aligned} \bar{Y}_{i..} - \bar{Y}_{i'..} &= (M_i - M_{i'}) + \frac{1}{f} \left( \sum_j MF_{ij} - \sum_j MF_{i'j} \right) \\ &+ \frac{1}{nf} \left( \sum_j \sum_k \varepsilon_{ijk} - \sum_j \sum_k \varepsilon_{i'jk} \right) \end{aligned}$$

, com variância

$$\hat{\sigma}_{F\text{Entre}}^2 = \hat{\sigma}_m^2 + \frac{1}{f} \hat{\sigma}_{mf}^2 + \frac{1}{nf} \hat{\sigma}^2 = \frac{QMM}{nf}.$$

Então, a estimativa de herdabilidade é calculada como:

$$h_m^2 = \frac{\hat{\sigma}_m^2}{\hat{\sigma}_{F\text{Entre}}^2} = \frac{\hat{\sigma}_m^2}{(QMM/nf)}.$$

## 2) Seleção entre fêmeas com base na média de desempenho da progênie:

As médias a serem comparadas são:

$$\bar{Y}_{.j.} = \mu + \frac{1}{m} \sum_i M_i + F_j + \frac{1}{m} \sum_i MF_{ij} + \frac{1}{nm} \sum_i \sum_k \varepsilon_{ijk} \text{ e}$$

$$\bar{Y}_{.j'.} = \mu + \frac{1}{m} \sum_i M_i + F_{j'} + \frac{1}{m} \sum_i MF_{ij'} + \frac{1}{nm} \sum_i \sum_k \varepsilon_{ij'k}$$

Os contrastes entre as médias são expressos como:

$$\begin{aligned} \bar{Y}_{.j.} - \bar{Y}_{.j'.} &= (F_j - F_{j'}) + \frac{1}{m} \left( \sum_i MF_{ij} - \sum_i MF_{ij'} \right) \\ &+ \frac{1}{nm} \left( \sum_i \sum_k \varepsilon_{ijk} - \sum_i \sum_k \varepsilon_{ij'k} \right) \end{aligned}$$

, com variância

$$\hat{\sigma}_{F\text{Entre}}^2 = \hat{\sigma}_f^2 + \frac{1}{m} \hat{\sigma}_{mf}^2 + \frac{1}{nm} \hat{\sigma}^2 = \frac{QMF}{nm}.$$

Então, a estimativa de herdabilidade pode ser obtida por meio da seguinte expressão:

$$h_m^2 = \frac{\hat{\sigma}_f^2}{\hat{\sigma}_{FEntre}^2} = \frac{\hat{\sigma}_f^2}{(QMF/nm)}.$$

3) Seleção individual dos descendentes:

Os valores individuais a serem comparados são:

$$Y_{ijk} = \mu + M_i + F_j + MF_{ij} + \varepsilon_{ijk} \text{ e}$$

$$Y_{i'j'k'} = \mu + M_{i'} + F_{j'} + MF_{i'j'} + \varepsilon_{i'j'k'}$$

Os contrastes entre os valores individuais são expressos como:

$$Y_{ijk} - Y_{i'j'k'} = (M_i - M_{i'}) + (F_j - F_{j'})$$

$$+ (MF_{ij} - MF_{i'j'}) + (\varepsilon_{ijk} - \varepsilon_{i'j'k'})$$

, com variância

$$\hat{\sigma}_{FEntre}^2 = \hat{\sigma}_m^2 + \hat{\sigma}_f^2 + \hat{\sigma}_{mf}^2 + \hat{\sigma}^2$$

Para a obtenção das estimativas de herdabilidade pode ser usada uma das seguintes expressões:

$$h_{i1}^2 = \frac{4\hat{\sigma}_m^2}{\hat{\sigma}_{FEntre}^2}$$

$$h_{i2}^2 = \frac{4\hat{\sigma}_f^2}{\hat{\sigma}_{FEntre}^2}$$

$$h_{i3}^2 = \frac{2(\hat{\sigma}_m^2 + \hat{\sigma}_f^2)}{\hat{\sigma}_{FEntre}^2}.$$

#### Exemplo de Aplicação

Considere os resultados da análise de variância da avaliação de quatro descendentes obtidos do cruzamento envolvendo dez genitores masculinos acasalados com dose genitores femininos (Extraído de Cruz, 2005):

FV	GL	SQ	QM	F
Machos (M)	9	2523,60	280,4	3,75*
Fêmeas (F)	11	3415,50	310,5	4,16*
M x F	99	7385,40	74,6	6,66**
Descendentes/(MxF)	360	4320,00	12,0	-

Com base nos resultados da análise de variância são obtidas as seguintes estimativas:

$$\hat{\sigma}^2 = QMD = 12,0$$

$$\hat{\sigma}_{mf}^2 = \frac{QMI - QMD}{n} = \frac{74,6 - 12,0}{4} = 15,65$$

$$\hat{\sigma}_f^2 = \frac{QMF - QMI}{nm} = \frac{310,5 - 74,6}{(4)(10)} = 5,8975$$

$$\hat{\sigma}_m^2 = \frac{QMM - QMI}{nf} = \frac{280,4 - 74,6}{(4)(12)} = 4,2875$$

Então, as estimativas da variância aditiva e de dominância são:

$$\hat{\sigma}_A^2 = 2\left(\hat{\sigma}_m^2 + \hat{\sigma}_f^2\right) = 2(4,2875 + 5,8975) = 20,37$$

$$\hat{\sigma}_D^2 = 4\hat{\sigma}_{mf}^2 = 4(15,65) = 62,60$$

Para a seleção de genitores masculinos tem-se a seguinte herdabilidade:

$$h_m^2 = \frac{\hat{\sigma}_m^2}{(QMM/nf)} = \frac{4,2875}{(280,4/48)} = 0,7340$$

Para a seleção de genitores femininos tem-se a seguinte herdabilidade:

$$h_f^2 = \frac{\hat{\sigma}_f^2}{(QMF/nm)} = \frac{5,8975}{(310,5/40)} = 0,7597$$

Para a seleção de plantas individuais têm-se as seguintes herdabilidades:

$$\hat{\sigma}_y^2 = \hat{\sigma}_f^2 + \hat{\sigma}_m^2 + \hat{\sigma}_{mf}^2 + \hat{\sigma}^2$$

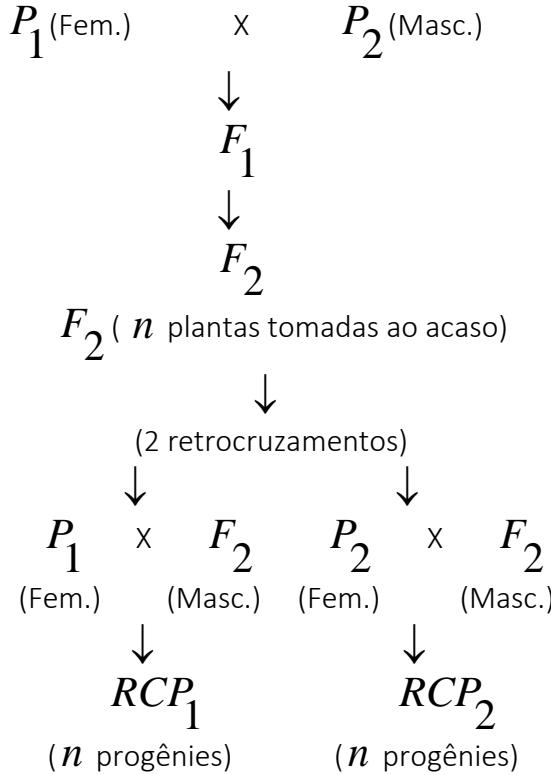
$$\hat{\sigma}_y^2 = 5,8975 + 4,2875 + 15,65 + 12,0 = 37,835$$

$$h_{i1}^2 = \frac{4\hat{\sigma}_m^2}{\hat{\sigma}_y^2} = \frac{4(4,2875)}{37,835} = 0,4533$$

$$h_{i2}^2 = \frac{4\hat{\sigma}_f^2}{\hat{\sigma}_y^2} = \frac{4(5,8975)}{37,835} = 0,6235$$

$$h_{i3}^2 = \frac{2\left(\hat{\sigma}_m^2 + \hat{\sigma}_f^2\right)}{\hat{\sigma}_y^2} = \frac{2(4,2875 + 5,8975)}{37,835} = 0,5384$$

5) Delineamento III de Comstock e Robinson – neste delineamento genético a população de genótipos experimentais é composta por  $n$  pares de progênies obtidas por meio do cruzamento de  $n$  plantas  $F_2$ , tomadas ao acaso, com as linhagens genitoras homozigotas, contrastantes para o caráter de interesse. Considere a ilustração seguinte:



Observe que os componentes de cada par de progênies possuem o mesmo genitor masculino, que é a planta  $F_2$ , mas possuem diferentes genitores femininos, que são as duas linhagens endogâmicas contrastantes cruzadas inicialmente. Então, são produzidas  $2n$  progênies que são avaliadas e as análises estatísticas dos dados realizadas com base no seguinte modelo linear:

$$Y_{ijk} = \mu + P_i + L_j + PL_{ij} + \varepsilon_{ijk}, \text{ em que:}$$

$\mu$ : média geral;

$P_i$ : efeito aleatório da  $i$ -ésima progénie  $F_2$ , com  $i=1,2,\dots,n$ ;

$L_j$ : efeito fixo da  $j$ -ésima linhagem genitora, com  $j=1,2$ ;

$PL_{ij}$ : efeito aleatório da interação entre progênies  $F_2$  e linhagem genitora;

$\varepsilon_{ijk}$ : efeito aleatório do  $k$ -ésimo descendente (progénie) do retrocruzamento entre a  $i$ -ésima progénie  $F_2$  e a  $j$ -ésima linhagem genitora, com  $k=1,2,\dots,r$ .

O esquema para a análise de variância de caracteres avaliados em progênies obtidas em retrocruzamentos realizados conforme o delineamento III de Comstock e Robinson é o seguinte:

FV	GL	QM	E (QM)
Linhagens (L)	1	-	-

---

Progêneres F <sub>2</sub> (P)	n - 1	QMP	$\sigma^2 + 2r\sigma_P^2$
Interação P x L	n - 1	QMPL	$\sigma^2 + r\sigma_{PL}^2$
Descendentes	(r-1)2n	QMD	$\sigma^2$

---

Com base nesta análise de variância são obtidos os estimadores dos componentes de variância seguintes:

$$\text{Variância de progêneres } F_2 : \hat{\sigma}_P^2 = \frac{QMP - QMD}{2r}$$

$$\text{Variância da interação entre progêneres } F_2 \text{ e linhagens genitoras: } \hat{\sigma}_{PL}^2 = \frac{QMPL - QMD}{r}$$

$$\text{Variância de descendentes: } \hat{\sigma}^2 = QMD$$

Na interpretação genética dos componentes de variância considera-se que a variância associada à  $\sigma_P^2$  é dada pela variância dos valores médios das progêneres F<sub>2</sub>, sendo expressa como:

$$\sigma_P^2 = \frac{1}{4} \sigma_A^2$$

$$\text{A variância associada à } \sigma_{PL}^2 \text{ é expressa como: } \sigma_{PL}^2 = \sigma_D^2$$

Então, os estimadores da variância genética aditiva e devido à dominância são expressos por:

$$\sigma_A^2 = 4\sigma_P^2 \text{ e } \sigma_D^2 = \sigma_{PL}^2$$

**6) Dialelos** – neste delineamento genético  $p$  genitores são cruzados entre produzindo combinações híbridas. Os principais métodos de análise de dialelos são os de Hayman (1954), Griffing (1956) e Gardner e Eberhart (1966).

Por exemplo, os métodos de Griffing (1956) descrevem a capacidade geral e específica de combinação dos genitores, considerando a inclusão dos seguintes genótipos nos experimentos:

- 1) Dialelo com os  $p$  genitores,  $p(p-1)/2$  F<sub>1</sub>s e  $p(p-1)/2$  recíprocos dos F<sub>1</sub>s, totalizando  $p^2$  genótipos;
- 2) Dialelo com os  $p$  genitores e os  $p(p-1)/2$  F<sub>1</sub>s, totalizando  $p(p+1)/2$  genótipos;
- 3) Dialelo com os  $p(p-1)/2$  híbridos F<sub>1</sub>s e os  $p(p-1)/2$  recíprocos dos F<sub>1</sub>s, totalizando  $p(p-1)$  genótipos;
- 4) Dialelo com os  $p(p-1)/2$  híbridos F<sub>1</sub>s.

Considere como exemplo ilustrativo um dialelo que inclui apenas os híbridos  $F_1$ , cujo esquema é o seguinte:

Genitor	G1	G2	G3	G4	G5
G1	-	$H_{12}(Y_{12})$	$H_{13}(Y_{13})$	$H_{14}(Y_{14})$	$H_{15}(Y_{15})$
G2	-	-	$H_{23}(Y_{23})$	$H_{24}(Y_{24})$	$H_{25}(Y_{25})$
G3	-	-	-	$H_{34}(Y_{34})$	$H_{35}(Y_{25})$
G4	-	-	-	-	$H_{45}(Y_{45})$
G5	-	-	-	-	-

O modelo genético-estatístico que descreve cada observação  $(Y_{ij})$  no  $k$ -ésimo descendente do cruzamento entre os genitores  $i$  e  $j$  é expresso como:

$$Y_{ijk} = \mu + g_i + g_j + s_{ij} + \varepsilon_{ijk}, \text{ em que:}$$

$Y_{ijk}$ : observação no  $k$ -ésimo descendente do cruzamento entre os genitores  $i$  e  $j$ , com

$i < j = 1, 2, \dots, p$  e  $k = 1, 2, \dots, n$ ;

$\mu$ : média geral;

$g_i$  e  $g_j$ : efeitos da capacidade geral de combinação (CGC) dos genitores  $i$  e  $j$ ,

respectivamente, com  $g_i$  e  $g_j \sim NID(0, \sigma_g^2)$ ;

$s_{ij}$ : efeito da capacidade específica de combinação (CEC) entre os genitores  $i$  e  $j$ , com

$s_{ij} \sim NID(0, \sigma_s^2)$ ;

$\varepsilon_{ijk}$ : erro aleatório, com  $\varepsilon_{ijk} \sim NID(0, \sigma^2)$ .

O esquema de análise de variância para este delineamento genético é o seguinte:

FV	GL	QM	E(QM)
Genótipos	$[p(p-1)/2]-1$	QMG	-
CGC	$p-1$	QMCGC	$\sigma^2 + n\sigma_s^2 + n(p-2)\sigma_g^2$
CEC	$p(p-3)/2$	QMCEC	$\sigma^2 + n\sigma_s^2$
Descendentes	$p(p-1)(n-1)/2$	QMD	$\sigma^2$

Com base nesta análise são obtidas as seguintes estimativas de componentes de variância:

$$\hat{\sigma}^2 = QMD$$

$$\hat{\sigma}_s^2 = \frac{QMCEC - QMD}{n}$$

$$\hat{\sigma}_g^2 = \frac{QMCGC - QMCEC}{n(p-2)}$$

Por meio deste delineamento genético são obtidos e avaliados irmãos completos ( $Y_{ijk}$  e  $Y_{ijk'}$ ), meios-irmãos paternos ( $Y_{ijk}$  e  $Y_{ij'k'}$ ) e meios-irmãos maternos ( $Y_{ijk}$  e  $Y_{i'jk'}$ ).

Por exemplo, os descendentes H<sub>12</sub>, H<sub>13</sub>, H<sub>14</sub> e H<sub>15</sub> são meios-irmãos maternos, enquanto que os descendentes H<sub>13</sub>, H<sub>23</sub>, H<sub>43</sub> e H<sub>53</sub> são meios-irmãos paternos. O conjunto de todos os descendentes, H<sub>12</sub>, H<sub>13</sub>, ..., H<sub>45</sub> é de irmãos completos.

# CAPITULO 18

## Índices de Seleção

Geralmente, a obtenção de ganhos genéticos em relação a uma característica é realizada por meio da seleção de genótipos com base nesta característica. Entretanto, a seleção baseada em apenas uma característica pode provocar mudanças, desejáveis ou indesejáveis em outras que sejam correlacionadas com a mesma.

Então, para a seleção de genótipos superiores, que reúnam várias características agronômicas desejáveis, pode-se utilizar índices de seleção. Segundo Neves et al. (2011), índice de seleção é uma técnica multivariada que associa as observações de várias características de interesse com as informações genéticas da população sob avaliação.

O índice de seleção é um valor numérico, resultante da combinação de características escolhidas, que funciona como uma característica adicional para a prática de seleção simultânea. Por este método, o ganho genético sobre uma determinada característica é reduzido, mas essa redução é compensada por uma melhor distribuição dos ganhos no conjunto das características desejáveis (Cruz et al., 2014).

### Indices de Seleção Clássicos

A metodologia de índice de seleção foi idealizada por Smith (1936) para a seleção de múltiplas características. Os índices de seleção clássicos são classificados como paramétricos e não paramétricos. Os paramétricos ainda são classificados como restritos e não restritos.

Para a obtenção de índices de seleção paramétricos são necessárias a estimação das matrizes de covariâncias genotípicas e fenotípicas e a determinação dos pesos econômicos relativos às várias características avaliadas. Neste caso, os ganhos genéticos são obtidos a partir da maximização da correlação entre o valor genotípico e o índice, visando a máxima eficiência na seleção e a melhoria gradativa na frequência dos alelos favoráveis para o conjunto de características desejáveis. Os pesos econômicos podem ser estimados a partir de estatísticas dos dados, como por exemplo o coeficiente de variação genotípico e a herdabilidade, ou ainda podem ser atribuídos de forma aleatória pelo melhorista.

Os índices de seleção paramétricos não restritos são construídos com base nas matrizes de covariâncias genotípicas e fenotípicas de cada característica em avaliação e os pesos econômicos atribuídos. Para os índices de seleção paramétricos restritos são estabelecidas restrições para as características.

### Indices de seleção paramétricos não restritos

#### Indice de seleção de Smith (1936) e Hazel (1943)

O índice de seleção de Smith (1936) e Hazel (1943) é obtido por meio de uma função/combinação linear dos valores fenotípicos observados de várias características de importância econômica, cujos

coeficientes de ponderação são estimados de forma que a correlação entre o índice de seleção e o agregado genotípico seja maximizada.

O agregado genotípico é constituído por outra combinação linear envolvendo os valores genotípicos desconhecidos, estimada por meio da expressão:

$$H = a_1 g_1 + a_2 g_2 + \dots + a_n g_n = \sum_{i=1}^n a_i g_i = a'g, \text{ em que:}$$

**H**: agregado genotípico;

**a**: vetor  $n \times 1$  de pesos econômicos previamente estabelecidos;

**g**: matriz  $n \times p$  de valores genotípicos desconhecidos das **n** características consideradas no índice;

**n**: número de características no índice;

**p**: número de progênies avaliadas.

O índice de seleção é estimado usando a seguinte expressão:

$$I = b_1 y_1 + b_2 y_2 + \dots + b_n y_n = \sum_{i=1}^n b_i y_i = b'y, \text{ em que:}$$

**I**: índice de seleção;

**b**: vetor  $n \times 1$  de coeficientes de ponderação do índice;

**y**: matriz  $n \times p$  de médias fenotípicas conhecidas das **n** características consideradas no índice;

**n**: número de características no índice.

**p**: número de progênies avaliadas.

Para o processo de estimativação do índice de seleção de cada progénie, que consiste em maximizar a correlação entre o agregado genotípico e o índice, deve-se considerar:

**P**: matriz  $n \times n$  de variâncias e covariâncias fenotípicas e

**G**: matriz  $n \times n$  de variâncias e covariâncias genotípicas.

Desta forma, o vetor **b**, estimador dos coeficientes de ponderação do índice, é obtido por meio da seguinte expressão:

$$b = P^{-1}G_a$$

O ganho esperado na característica **j**, quando a seleção for praticada sobre o índice, é expresso como:

$$\Delta g_{j/I} = DS_{j/I} h_j^2, \text{ em que:}$$

$\Delta g_{j/I}$ : ganho esperado para a característica **j**, quando a seleção for baseada no índice **I**;

$DS_{j/I}$ : diferencial de seleção da característica **j**, para a seleção feita com base no índice **I**;

$h_j^2$ : herdabilidade da característica **j**.

### Índice de Pesek e Baker (1969)

O índice de seleção de Pesek e Baker (1969) é calculado utilizando os ganhos genéticos desejados de cada característica no lugar dos pesos econômicos. Neste caso, é necessário estimar as médias dos

genótipos e as matrizes de covariâncias fenotípicas e genotípicas. Este método resulta em um ganho máximo para cada característica, de acordo com sua importância relativa, assumida pelo melhorista, na especificação do ganho desejado.

A construção do índice é realizada com base na expressão do ganho esperado para as características consideradas no índice, da seguinte forma:

$$\Delta g = \frac{\hat{G}\hat{b}_I}{\hat{\sigma}_I}, \text{ em que:}$$

$\Delta g$  : ganho genético estimado pelo índice;

$G$  : matriz  $n \times n$  de variâncias e covariâncias genotípicas entre as características;

$\hat{b}$  : vetor  $1 \times n$  dos coeficientes de ponderação estimados do índice de seleção;

$i$  : diferencial de seleção do índice  $I$  em unidades de desvio padrão;

$\hat{\sigma}_I$  : desvio padrão do índice  $I$ .

Substituindo  $\Delta g$  pelo vetor dos ganhos desejados,  $\Delta gd$ , estima-se o vetor  $\hat{b}$  por meio da seguinte expressão:

$\hat{b} = G^{-1} \Delta gd \frac{\hat{\sigma}_I}{i}$ , mas  $\frac{\hat{\sigma}_I}{i}$  não afeta a proporcionalidade dos  $\hat{b}_i$ , então  $\hat{b}$  pode ser estimado como:

$$\hat{b} = G^{-1} \Delta gd$$

### Índice base de Willians (1962)

O índice de Willians (1962) pondera os valores fenotípicos pelos seus respectivos pesos econômicos, evitando as imprecisões das matrizes de covariâncias. Este índice foi obtido por meio da estimação de uma combinação linear dos valores fenotípicos médios das características, ponderados pelos respectivos pesos econômicos, utilizando a seguinte expressão:

$$I = a_1 y_1 + a_2 y_2 + \dots + a_n y_n = \sum_{i=1}^n a_i y_i = a'y, \text{ em que:}$$

$a$  : vetor  $n \times 1$  de pesos econômicos atribuídos a cada uma das características;

$y$  : vetor  $n \times 1$  de médias das características.

### Índices de seleção paramétricos restritos

#### Índice de seleção de Kempthorne e Nordskog (1959)

O índice de seleção de Kempthorne e Nordskog (1959) é obtido por meio de funções lineares obtidas restringindo-se um conjunto de características para maximizar o ganho simultâneo em um outro conjunto de características de maior interesse. Este método utiliza as matrizes de covariâncias fenotípicas e genotípicas. A restrição é imposta fazendo com que a covariância entre o índice e o valor genotípico de uma ou mais características seja, igual a zero, ou seja,  $Cov(I, g_i) = 0$ . Os coeficientes de ponderação do índice são estimados usando a seguinte expressão:

$$b = \left[ I - P^{-1} G C \left( C' G P^{-1} G C \right)^{-1} \right] P^{-1} G a, \text{ em que:}$$

$b$ : vetor de coeficientes estimados do índice;

$I$ : valor do índice de seleção;

$P$ : matriz de variâncias e covariâncias fenotípicas;

$G$ : matriz de variâncias e covariâncias genotípicas;

$a$ : vetor de pesos econômicos das características no índice;

$C$ : matriz  $n \times r$  de restrições, sendo  $r$  o número de restrições feitas no índice, de forma que  $b' G C = 0$ .

Os elementos da matriz  $C$  são  $C_{ij}$ , para  $i = 1, 2, \dots, n$  e  $j = 1, 2, \dots, r$ . Verifica-se então que

$C_{ij} = 1$ , para  $i = j$  e  $C_{ij} = 0$ , para  $i \neq j$ .

### Indice de seleção de Tallis (1962)

Tallis (1962) modificou o índice de seleção de Kempthorne e Nordskog, propondo que para tornar a metodologia de restrição mais eficaz a dedução das expressões do índice seja feita considerando algum valor para a restrição, ou seja,  $\text{Cov}(I, g_i) = k_j$ , ( $j = 1, 2, \dots, r$ ), sendo  $r$  o número de restrições.

### Indice de seleção de James (1968)

No índice de James (1968) os coeficientes de ponderação são estimados de forma a maximizar a correlação entre o índice e um agregado genotípico, mas sujeito à restrição de que, para certas características, o ganho possa ser estabelecido a partir de covariância não nula entre a característica e o índice, ou seja,  $\text{Cov}(I, g_i) \neq 0$ . Neste caso, são impostas restrição aos ganhos genéticos e aos coeficientes de ponderação das características nos índices.

### Indices de seleção não paramétricos

Estes índices têm como objetivo a classificação dos genótipos e os mais conhecidos são: índice de soma de postos (Mulamba e Mock, 1978), índice livre de pesos e parâmetros (Elston, 1963), índice de distância genótipo – ideótipo (Schwarzbach, 1972)) e índice multiplicativo (Subandi et al., 1973).

### Indice de seleção de soma de postos de Mulamba e Mock (1978)

O índice de soma de postos de Mulamba e Mock (1978) consiste em ordenar os genótipos a partir da média fenotípica de cada uma das características, de acordo com o objetivo da seleção. Após realizar o ordenamento, são somadas as ordens de cada genótipo e esta soma é tomada como índice de seleção. Este índice não necessita de estimativas de covariâncias genotípicas e fenotípicas, mas pode ser utilizado para obtenção de ganhos genéticos.

Então, este método consiste em primeiro ordenar os genótipos em relação a cada uma das características, em ordem desejável para o melhoramento. Em seguida é obtido o índice, multiplicando o peso econômico de cada característica pelo posto de cada genótipo na característica, conforme a seguinte expressão:

$$I_i = \sum_j p_j r_{ij}, \text{ em que:}$$

$I_i$ : índice do  $i$ -ésimo genótipo;

$p_j$ : peso econômico atribuído à  $j$ -ésima característica;

$r_{ij}$ : posto do  $i$ -ésimo genótipo na  $j$ -ésima característica.

#### Índice de seleção livre de pesos e parâmetros de Elston (1963)

O índice livre de pesos e parâmetros de Elston (1963) é baseado apenas nos valores fenotípicos das características, sendo que todas as características têm a mesma importância e não é necessária a estimativa de parâmetros genéticos. Este método é útil quando existe pouco conhecimento sobre as características, exceto que devem ser selecionados valores altos ou baixos de cada uma das características.

Neste método é utilizado as médias das características como ponto de corte para a seleção dos genótipos. O índice é obtido por meio da seguinte expressão:

$$I_{Ei} = \log \prod_{j=1}^n (y_{ij} - k_j) = \log [(y_{i1} - k_1)(y_{i2} - k_2) \cdots (y_{in} - k_n)], \text{ em que:}$$

$I_{Ei}$ : índice de Elston no  $i$ -ésimo genótipo;

$y_{ij}$ : média do  $i$ -ésimo genótipo para a  $j$ -ésima característica;

$k_j$ : menor valor selecionável, expresso como  $k_j = \frac{n(\min y_{ij}) - \max y_{ij}}{n-1}$ ;

$n$ : número de genótipos.

#### Índice de distância genótipo-ideótipo de Schwarzbach (1972)

Este índice foi proposto por Schwarzbach (1972), citado por Wricke e Weber (1986). Segundo Cruz (2006) o índice é obtido da seguinte forma: primeiro é estabelecido um valor ideal para cada característica, formando um ideótipo que pode ou não existir na população. Em seguida obtém-se a diferença entre a média de cada característica e o valor atribuído ao ideótipo e calcula-se, para cada genótipo, a sua distância em relação ao ideótipo, sendo o índice o valor desta distância.

O índice distância genótipo-ideótipo é descrito por Cruz (2006) da seguinte forma: é definido, para cada característica, os valores ótimos e os intervalos de valores considerados desejáveis para o melhoramento. É considerado o valor fenotípico médio do  $i$ -ésimo genótipo em relação à  $j$ -ésima característica,  $X_{ij}$ , o valor fenotípico médio transformado,  $Y_{ij}$  e uma constante referente à

depreciação da média do genótipo por estar fora dos padrões desejáveis para o melhoramento,  $C_j$ . Define-se também:

$LI_j$ : limite inferior da característica  $j$  que deve ser apresentado pelo genótipo, conforme desejado para o melhoramento;

$LS_j$ : limite superior da característica  $j$  que deve ser apresentado pelo genótipo, conforme desejado para o melhoramento;

$VO_j$ : valor ótimo da característica  $j$  que deve ser apresentado pelo genótipo, conforme desejado para o melhoramento;

Se  $LI_j \leq X_{ij} \leq LS_j$ , então  $Y_{ij} = X_{ij}$ ;

Se  $X_{ij} < LI_j$ , então  $Y_{ij} = X_{ij} + VO_j - LI_j - C_j$ ;

Se  $X_{ij} > LS_j$ , então  $Y_{ij} = X_{ij} + VO_j - LS_j + C_j$ .

É considerado que  $C_j = LS_j - LI_j$ . O valor  $C_j$  garante que qualquer valor  $X_{ij}$  dentro do intervalo de variação em torno do ótimo resulta num valor  $Y_{ij}$  com valor próximo do valor ótimo,

$VO_j$ . Desta forma, a transformação de  $X_{ij}$  é realizada para garantir a depreciação dos valores fenotípicos fora do intervalo. Os valores  $Y_{ij}$ , obtidos por transformação, são padronizados e ponderados pelos pesos atribuídos a cada característica, obtendo então os valores  $y_{ij}$ , conforme expressão dada a seguir:

$$y_{ij} = \sqrt{a_j} \frac{Y_{ij}}{s(Y_j)}, \text{ em que:}$$

$s(Y_j)$ : desvio padrão da característica  $j$ ;

$a_j$ : peso econômico da característica  $j$ .

Os valores  $VO_j$  também são padronizados e ponderados de acordo com a seguinte expressão:

$$vo_j = \sqrt{a_j} \frac{VO_j}{s(Y_j)}$$

Com base nos valores obtidos anteriormente, são calculados os valores do índice DGI expressos como as distâncias entre os genótipos e o ideótipo, conforme expressão a seguir:

$$I_{DGI} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_{ij} - v_{o_j})^2}$$

### Índice multiplicativo de Subandi et al. (1973)

Este índice é obtido por meio da multiplicação dos valores padronizados de cada característica para cada genótipo, da seguinte forma:

$$I = y_1^{k_1} y_2^{k_2} \dots y_n^{k_n}, \text{ em que:}$$

$I$ : índice multiplicativo;

$y_j$ : média da característica  $j$ ;

$k_j = 1$ , para o caso da relação direta do índice com a característica;

$k_j = -1$ , para o caso da relação inversa do índice com a característica.

Na utilização de índices de seleção, algumas características podem ser selecionadas no sentido de acréscimo e outras no sentido de decréscimo de seus valores. Os pesos econômicos podem ser estabelecidos com base nas estimativas do coeficiente de variação genética ( $CV_g$ ), do desvio padrão genético ( $DP_g$ ), de herdabilidade ( $h^2$ ) e também em pesos aleatórios (PA). Em geral, são utilizados pesos aleatórios maiores para as características principais e menores para as características secundárias.

### Índices de Seleção Novos

#### Índice FAI-BLUP-Índice multicaracterísticas baseado na análise de fatores e delineamento-ideótipo

Selecionar genótipos de alto desempenho para multicaracterísticas simultaneamente pode ser uma tarefa difícil. O primeiro índice para seleção simultânea foi proposto por Smith (1936) para melhoramento vegetal e Hazel (1943) para melhoramento animal. Este índice é baseado na seleção de valores genéticos desconhecidos. Desta forma, o uso de valores fenotípicos e covariâncias genéticas é necessário para determinar como um vetor de pesos deve ser escolhido de forma a maximizar a correlação entre valores genéticos desconhecidos e valores fenotípicos (Hazel et al., 1994). Uma das dificuldades de usar este índice é a falta de um procedimento para atribuir pesos econômicos para as características.

No processo de seleção, geralmente o melhorista manuseia multicaracterísticas. Entretanto, problemas de multicolinearidade certamente irá aparecer e se tornar um obstáculo para o índice clássico Smith-Hazel. Multicolinearidade entre características tem sido uma questão sistemática na análise multivariada e evidentemente que causa sérias dificuldades para a interpretação apropriada dos resultados, com o risco de conclusões errôneas e mal direcionamento das pesquisas. Além disso, o índice Smith-Hazel não utiliza como vantagem a correlação entre características.

A análise de fatores pode produzir eixos não correlacionados ou ortogonais entre os escores finais de fatores, e, portanto, eles são livres de multicolinearidade. Este método concentra nas poucas primeiras variáveis latentes, e geralmente as variáveis latentes menos importantes são descartadas, levando a uma redução dimensional, o que consequentemente simplifica a análise.

A fundamentação teórica da modelagem de equações estruturais (SEM) surge pela junção das técnicas tradicionais de análise de fatores (Análise de Fatores Exploratória) com o delineamento-

ideótipo (Análise de Fatores Confirmatória). A SEM possibilitará o uso das correlações (covariâncias) entre as características. Desta forma, a influência que uma característica exerce sobre outra será computada, fornecendo informação a respeito de sua magnitude e sentido.

Melhoristas de plantas experientes têm em mente uma planta ideótipo, que os leva a selecionar plantas de alto desempenho. Neste contexto, a ideia por trás do ideótipo é que ele fornece para o melhorista um alvo definitivo para seleção.

O método do índice FAI-BLUP é descrito por Rocha et al, 2018 da seguinte forma:

Análise de fatores exploratória: A análise de fatores, baseada no método de componentes principais, é utilizada para a extração das cargas fatoriais da matriz de correlação genética, obtida por meio de valores genéticos. O critério varimax é usado para a rotação analítica. Para o cálculo dos escores dos fatores é usado o método dos quadrados mínimos ponderados.

Delineamento-ideótipo: O número de ideótipos é definido com base na combinação de fatores desejados e não desejados para o objetivo da seleção. O número de ideótipos é obtido por meio do seguinte algoritmo:

$$NI = 2^n, \text{ onde } NI = \text{número de ideótipos e } n = \text{número de fatores.}$$

O número de fatores (**n**) usado para delinear os ideótipos deve ser igual ao número de autovalores (variâncias dos componentes principais) maiores que 1,0 (Kaiser, 1958). Este número indica também o número de coordenadas que devem ser calculadas (ou seja, **n** escores fatoriais) para cada ideótipo. Então, cada ideótipo e suas descrições baseadas na combinação de fatores desejados e não desejados são descritos como a seguir:

Por exemplo, se três autovalores são maiores que 1,0 (**n=3**), serão delineados oito ideótipos

$$(NI=2^3=8) \text{ com três coordenadas cada um.}$$

O escore do fator é uma combinação linear de valores genéticos padrão (médias BLUP) ponderados pelas cargas do fator obtidas pela análise de fatores. Portanto, um fator desejável deve ter os valores genéticos desejáveis para todas as características sob seleção. Os valores genéticos desejáveis pode ser o valor máximo, mínimo, médio ou um valor genético específico. Um fator indesejável deve ter valores genéticos indesejáveis (máximo, mínimo, médio ou um valor genético específico) para todos as características sob seleção. Então, cada ideótipo e suas coordenadas (escores fatoriais) pode ser delineado.

Índice multicaracterísticas baseado na análise de fatores e na distância genótipo-ideótipo (índice FAI-BLUP):

Depois dos ideótipos serem determinados, as distâncias de cada genótipo em relação a cada ideótipo são estimadas e convertidas em probabilidades espaciais, possibilitando o ordenamento dos genótipos. Para isto é utilizada a seguinte expressão:

$$P_{ij} = \frac{1}{\sum_{i=1; j=1}^n \frac{1}{d_{ij}}}, \text{ no qual: } P_{ij} = \text{Probabilidade do } i\text{-ésimo genótipo } (i=1,2,\dots,n) \text{ ser similar ao } j\text{-ésimo ideótipo } (j=1,2,\dots,m); d_{ij} = \text{Distância genótipo-ideótipo do } i\text{-ésimo genótipo ao } j\text{-ésimo ideótipo baseada na distância Euclidiana média padronizada.}$$

Na análise de fatores exploratória (AFE), o pesquisador tem um grande conjunto de características e hipotetiza que as características observadas podem estar ligadas em virtude das correlações entre as características (estrutura subjacente desconhecida) e o objetivo de uma EFA é revelar esta estrutura e levar a uma redução dimensional. Numa análise de fatores confirmatória (AFC), o pesquisador tem uma ideia sobre o número de fatores, as relações entre os fatores e o relacionamento entre os fatores e as características medidas. O objetivo da AFC é hipotetizar um fator estruturado *a priori* e verificar empiricamente (ou por testes), ao invés de deriva-lo a partir dos dados.

A fundamentação teórica da modelagem de equações estruturais (SEM) surge pela junção das técnicas tradicionais de análise de fatores (Análise de Fatores Exploratória) com o delineamento-ideótipo (Análise de Fatores Confirmatória). A SEM possibilitará o uso das correlações (covariâncias) entre as características. Desta forma, a influência que uma característica exerce sobre outra será computada, fornecendo informação a respeito de sua magnitude e sentido. Entretanto, para utilizar esta fundamentação teórica no processo de seleção em programas de melhoramento, o delineamento do ideótipo deve ser usado, uma vez que ele considera todos os relacionamentos desejáveis entre as características e os valores para as características que se deseja encontrar.

Além de lidar com problemas de multicolinearidade e falta de atribuição de pesos (em ambos os casos devido a análise de fatores), o índice FAI-BLUP leva em conta a estrutura de correlação obtida dos dados e direciona para a seleção de genótipos próximos aos hipotetizados pelo melhorista (ideótipo). A junção da análise de fatores exploratória (EFA, ou seja, amostragem dos dados) e o delineamento do ideótipo (semelhante a análise de fatores confirmatória, CFA) estabelece a base teórica da modelagem de equações estruturais (SEM).

Uma das principais vantagens da SEM é que ela pode ser usada para estudar o relacionamento entre estruturas latentes (fatores) que são indicados por múltiplas medições. É aplicável também a dados experimentais e não-experimentais bem como a dados cruzados e longitudinais. A modelagem de equações estruturais tem sido aplicada no melhoramento multicaracterísticas (Valente et al., 2010; Rosa et al., 2011, Lamb et al., 2011 e Valente et al., 2013).

Comparação do índice FAI-BLUP com o índice clássico Smith-Hazel (SH):

O índice de Smith-Hazel (SH) é usado para validar o potencial do índice FAI-BLUP por meio da comparação dos dois índices. O índice SH visa determinar como um vetor de pesos deve ser escolhido de forma a maximizar a correlação entre valores genéticos desconhecidos e valores fenotípicos. Isto pode ser obtido resolvendo a seguinte equação:

$b = P^{-1}G a$ , em que:  $b$  = vetor de pesos do índice a ser estimado;  $a$  : vetor de pesos econômicos relativos conhecidos. Neste estudo, o coeficiente de variação genético pode ser atribuído como peso econômico relativo, considerando o sentido da característica sob seleção – sinal positivo ou negativo;  $P$  = matriz variância-covariância fenotípica;  $G$  = matriz variância-covariância genotípica.

O diagnóstico de multicolinearidade é conduzido na matriz de correlação fenotípica de acordo com recomendações de Montgomery & Peck (1992), e algumas variáveis podem ser descartadas para resolver os problemas de multicolinearidade e, portanto, aplicar o índice clássico Smith-Hazel (SH). Comparações entre o índice SH e o índice FAI-BLUP são conduzidas por meio de ganhos genéticos preditos. Para fazer uma comparação mais válida, os ganhos genéticos preditos são calculados usando os genótipos indicados pelo índice SH, com base em valores genéticos (SH-BLUP) e usando os genótipos indicados pelo índice FAI-BLUP. Deve-se adotar uma intensidade de seleção de acordo com os objetivos e estratégias de melhoramento.

### **Índice MGIDI-Índice distância genótipo-ideótipo multi-características**

O índice de seleção fenotípico mais amplamente utilizado é o índice de Smith-Hazel (SH). Para computar o índice Smith-Hazel (SH), são utilizadas as matrizes fenotípicas e de variância-covariância genotípica bem como um vetor de pesos econômicos para determinar como um vetor de coeficientes do índice deve ser escolhido de forma a maximizar a correlação entre valores genéticos desconhecidos e valores fenotípicos. Devido ao índice SH requerer a inversão de uma matriz de covariância fenotípica (Smith, 1936), a presença de multicolinearidade pode resultar em matrizes pouco condicionadas e coeficientes do índice viesados, afetando assim as estimativas de ganho genético.

Além do problema de multicolinearidade, os melhoristas frequentemente enfrentam dificuldades na escolha realística dos valores econômicos das características. Uma vez que a combinação de técnicas multivariadas é eficiente para levar em conta o problema de multicolinearidade em índices multi-características, parece ser de valor uma investigação para desenvolver um índice que cubra a fraqueza do índice SH e no qual todas as características sejam selecionadas favoravelmente e com ganhos satisfatórios para aplicação em programas de melhoramento.

O índice distância genótipo-ideótipo multi-características (MGIDI) é descrito por Olivoto e Nardino, 2021:

#### **A) Reescalonando as características**

O primeiro passo para computar o índice MGIDI é reescalonar a matriz  $X$  de forma que todos os valores tenham uma faixa de 0-100. O valor reescalonado para a  $j$ -ésima característica do  $i$ -ésimo genótipo  $(rX_{ij})$  é obtido como descrito na equação a seguir:

$$rX_{ij} = \frac{\eta_{nj} - \varphi_{nj}}{\eta_{oj} - \varphi_{oj}} \times (\theta_{ij} - \eta_{oj}) + \eta_{nj} \quad (1)$$

Onde,  $\eta_{nj}$  e  $\varphi_{nj}$  são os novos valores máximo e mínimo, respectivamente, para a característica  $j$  depois de reescalonar;  $\eta_{oj}$  e  $\varphi_{oj}$  são os valores originais máximo e mínimo, respectivamente, para a característica  $j$  e  $\theta_{ij}$  é o valor original para a  $j$ -ésima característica do  $i$ -ésimo genótipo.

Os valores para  $\eta_{nj}$  e  $\varphi_{nj}$  são escolhidos como a seguir:

Para as características nas quais valores mais baixos são desejados são utilizados  $\eta_{nj}=0$  e  $\varphi_{nj}=100$ . Para as características nas quais valores mais altos são desejados são utilizados  $\eta_{nj}=100$  e  $\varphi_{nj}=0$ .

Depois do reescalonamento, é obtida uma tabela de dupla entrada de valores reescalados ( $rX$ ).

Cada coluna de  $rX$  tem uma faixa 0-100 que considera o sentido de seleção desejado (acrúscimo ou decréscimo) e mantém a estrutura de correlação do conjunto original de variáveis/características.

### B) Análise de Fatores

O segundo passo consiste em computar uma análise de fatores exploratória para agrupar as características correlacionadas em fatores não correlacionados e então estimar os escores fatoriais para cada genótipo. A análise de fatores é baseada na seguinte equação:

$$X = \mu + Lf + \varepsilon \quad (2)$$

Onde,  $X$  é um vetor  $p \times 1$  de observações;  $\mu$  é um vetor  $p \times 1$  de médias padronizadas;  $L$  é uma matriz  $p \times f$  de cargas fatoriais;  $f$  é um vetor  $p \times 1$  de fatores comuns;  $\varepsilon$  é um vetor  $p \times 1$  de resíduos/fatores específicos, sendo  $p$  o número de características e  $f$  o número de fatores comuns retidos.

Os autovalores e autovetores são obtidos a partir da matriz de correlação da tabela de dupla entrada  $rX$ . As cargas fatoriais iniciais são obtidas considerando apenas fatores com autovalores maiores do que 1,0. O critério de rotação varimax (Kaiser, 1958) é usado para a rotação analítica e estimação das cargas fatoriais finais. Os escores fatoriais para genótipos são obtidos de acordo com a seguinte equação:

$$F = Z \left( A^T R^{-1} \right)^T \quad (3)$$

Onde,  $F$  é uma matriz  $g \times f$  com os escores fatoriais;  $Z$  é uma matriz  $g \times p$  com as médias padronizadas;  $A$  é uma matriz  $p \times f$  de cargas fatoriais;  $R$  é uma matriz de correlação  $p \times p$  entre as características;  $g$  representa o número de genótipos;  $f$  representa o número de fatores retidos;  $p$  representa o número de características analisadas.

### C) Delineando o Ideótipo

O terceiro passo no cálculo do MGIDI consiste em delinejar o ideótipo. Por definição (Equação 1), o ideótipo tem o mais alto valor reescalonado (100) para todas as características analisadas. Desta forma, o ideótipo foi definido por um vetor  $I$   $p \times 1$ , tal que  $I = [100, 100, \dots, 100]$ . Os escores do ideótipo também foram estimados de acordo com a Equação 3.

4-Estimativa do Índice Distância Genótipo-Ideótipo Multi-características (MGIDI)

O quarto passo consiste na estimativa do índice distância genótipo-ideótipo multi-características (MGIDI) de acordo com a seguinte equação:

$$MGIDI_i = \left[ \sum_{j=1}^f (\gamma_{ij} - \gamma_j)^2 \right]^{0,5} \quad (4)$$

Onde,  $MGIDI_i$  é o índice distância genótipo-ideótipo multi-características para o  $i$ -ésimo genótipo;  $\gamma_{ij}$  é o escore do  $i$ -ésimo genótipo para o  $j$ -ésimo fator ( $i=1,2,\dots,g; j=1,2,\dots,f$ ), sendo  $g$  o número de genótipos e  $f$  o número de fatores;  $\gamma_j$  é o escore do ideótipo no  $j$ -ésimo fator.

Então, o genótipo com o mais baixo MGIDI está mais próximo do ideótipo e, portanto, apresenta valores desejados para todas as características analisadas. A proporção do índice MGIDI do  $i$ -ésimo genótipo explicada pelo  $j$ -ésimo fator ( $\omega_{ij}$ ) é calculada por meio da seguinte equação:

$$\omega_{ij} = \frac{\sqrt{D_{ij}^2}}{\sum_{j=1}^f \sqrt{D_{ij}^2}} \quad (5)$$

Onde,  $D_{ij}$  é a distância entre o  $i$ -ésimo genótipo e o ideótipo para o  $j$ -ésimo fator.

Fatores com baixas contribuições, para um dado genótipos, sugere que este genótipo está próximo do ideótipo para as características dentro deste fator.

O ganho com a seleção em percentagem,  $SG(%)$ , é calculado para cada característica, considerando uma intensidade de seleção ( $IS$ ) desejada, com base na seguinte equação:

$$SG(%) = \frac{(\bar{X}_s - \bar{X}_o) \times h^2}{\bar{X}_o} \times 100 \quad (6)$$

Onde,  $\bar{X}_s$  é a média dos genótipos selecionados,  $\bar{X}_o$  é a média da população original e  $h^2$  é a herdabilidade sentido amplo.

#### E) Comparação do índice MGIDI com outros índices multi-características

O índice clássico Smith-Hazel (SH) e o índice FAI-BLUP (Rocha et al., 2018) podem ser usados para validar o potencial do índice MGIDI em termos de ganhos genéticos.

Para o índice FAI-BLUP, o ideótipo desejado (ID) é definido por um vetor com “min” para características nas quais valores mais baixos são desejados e “max” para características nas quais valores mais altos são desejados. Então, as distâncias de cada um dos genótipos aos ideótipos são estimadas e convertidas em probabilidades espaciais, como base na seguinte equação:

$$P_{ij} = \frac{1}{\sum_{i=1; j=1}^n \frac{1}{d_{ij}}} \quad (7)$$

Onde,  $P_{ij}$  é a probabilidade do  $i$ -ésimo ( $i=1,2,\dots,n$ ) ser similar ao  $j$ -ésimo ideótipo ( $j=1,2,\dots,m$ );  $d_{ij}$  é a distância genótipo-ideótipo, do  $i$ -ésimo genótipo ao  $j$ -ésimo ideótipo baseada na distância Euclidiana média padronizada.

O índice SH é computado de acordo com a seguinte equação:

$$\mathbf{b} = \mathbf{P}^{-1} \mathbf{Gw} \quad (8)$$

Onde,  $\mathbf{P}$  e  $\mathbf{G}$  são as matrizes variância-covariância fenotípica e genotípica, respectivamente;  $\mathbf{b}$  e  $\mathbf{w}$  são os vetores de coeficientes dos índices e de pesos econômicos, respectivamente. Todas as características podem ser consideradas como tendo o mesmo peso econômico ( $w=1$ ) com sinal negativo (-1) para características nas quais valores mais baixos são desejados e positivo (1) para características nas quais valores mais altos são desejados, possibilitando assim que todas as características estejam na direção desejada para a seleção.

Então, o valor/mérito genético ( $I$ ) do genótipo individual baseado nas características avaliadas é computado por meio da seguinte equação:

$$I = \mathbf{Xb} \quad (9)$$

Onde,  $\mathbf{X}$  é uma matriz  $g \times p$  de valores BLUP para os  $g$  genótipos e  $p$  características;  $\mathbf{b}$  é um vetor  $1 \times p$  de coeficientes de índice para as  $p$  características.

Devem ser consideradas duas possibilidades de cálculo do índice: na primeira (SH-1), o índice SH é computado com todas as características; na segunda (SH-2), as características que geram multicolinearidade são excluídas e o índice é computado com as características remanescentes. As características a serem excluídas são escolhidas com a função de forma a não haver nenhum fator de inflação da variância (FIV) maior que 10 (Olivoto et al., 2017).

Para quantificar a coincidência de seleção de genótipos, pode ser computado o índice de coincidência (IC) entre cada par de índices, como a seguir (Hamblin e Zimmermann, 1986):

$$IC = \frac{A-C}{M-C} \times 100 \quad (10)$$

Onde,  $A$  é o número de genótipos selecionados comum aos diferentes índices;  $M$  é o número de genótipos selecionados de acordo com a intensidade de seleção ( $IS$ ) adotada;  $C$  é o número de genótipos esperados serem selecionados por chance ( $M \times IS$ ).

# CAPITULO 19

## Estabilidade e adaptabilidade

Existem diversos métodos de análise de estabilidade e adaptabilidade destinadas à avaliação de um grupo de genótipos em diversos ambientes. A diferença entre elas se dá, basicamente, pelos parâmetros adotados em sua avaliação, nos procedimentos biométricos empregados para medi-las e na informação ou detalhamento de sua análise, uma vez que todas são fundamentadas na existência de interações de genótipos com ambientes (Vencovsky e Barriga, 1992; Cruz et al., 2012).

### Métodos clássicos de avaliação de estabilidade e adaptabilidade

#### Método de Wricke (1964)

Este método, também conhecido como ecovalência é um dos que avalia a estabilidade fenotípica de genótipos utilizando apenas a análise de variância. O parâmetro  $W_i$ , estimado com base na decomposição da soma de quadrados da interação G x E, é obtido por meio da seguinte expressão:

$$W_i = r \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2, \text{ em que:}$$

$r$ : número de repetições;

$Y_{ij}$ : média do genótipo  $i$  no ambiente  $j$ ;

$\bar{Y}_{i\cdot}$ : média do genótipo  $i$ ;

$\bar{Y}_{\cdot j}$ : média do ambiente  $j$ ;

$\bar{Y}_{\cdot\cdot}$ : média geral.

O estimador da estabilidade,  $W_i$ , é uma medida da contribuição de cada genótipo para a interação G x E, em que os genótipos que menos contribuem para a interação são considerados os mais estáveis. O termo ecovalência se refere à capacidade do genótipo responder às variações ambientais. Então, uma alta ecovalência implica em baixo  $W_i$  e significa que o genótipo é estável (Wricke, 1964; Wricke e Weber, 1986)

#### Método de Shukla (1972)

Neste método, o componente de variância de cada genótipo ao longo dos ambientes é proposto como uma medida de estabilidade fenotípica. Ele mede a estabilidade ao invés do desempenho do genótipo. Portanto, na obtenção da estabilidade de variância  $(\sigma_i^2)$  a soma de quadrados da interação G x E é partitionada em componentes, um para cada genótipo e estimada por meio da seguinte expressão:

$$\sigma_i^2 = \frac{1}{(g-1)(e-1)} \times \left[ g(g-1) \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2 - \sum_i \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2 \right], \text{ em que:}$$

$g$ : número de genótipos;  $e$ : número de ambientes;

$Y_{ij}$ : média de produtividade do genótipo  $i$  no ambiente  $j$ ;

$\bar{Y}_{i\cdot}$ : média de produtividade do genótipo  $i$  em todos os ambientes;

$\bar{Y}_{\cdot j}$ : média de produtividade de todos os genótipos no ambiente  $j$ :

$\bar{Y}_{\cdot\cdot}$ : média geral.

Outra forma de expressar o estimador da estabilidade de variância é a seguinte:

$$\sigma_i^2 = g / [(g-1)(e-1)] W_i - QMGE / g-2, \text{ onde } QMGE \text{ é o quadrado médio da interação genótipos x ambientes e } W_i \text{ obtido como } W_i = \sum_{j=1}^e \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2.$$

A interpretação de  $\sigma_i^2$  é realizada da seguinte forma: se a estabilidade de variância de um genótipo for igual a variância ambiental, o que implica em  $\sigma_i^2 = 0$ , então o genótipo é identificado como

estável; um valor relativamente grande de  $\sigma_i^2$  indica maior instabilidade para o genótipo  $i$ ; valor significativo de  $\sigma_i^2$  significa que o desempenho do genótipo ao longo dos ambientes foi instável; genótipos com valor de  $\sigma_i^2$  não significativo ou negativo são considerados estáveis ao longo dos ambientes.

Observe que, como  $\sigma_i^2$  é obtido por meio da diferença entre duas somas de quadrados pode ocorrer valor negativo, que pode ser considerado igual a zero. Observe também que  $\sigma_i^2$  não pode ser estimado para dados desbalanceados.

### Método de Annicchiarico (1992)

Os métodos que avaliam a estabilidade com base na análise de variância das são os mais antigos e consistem na análise de grupos de experimentos sendo a variação de ambientes, dentro de cada genótipo, usada como estimador do parâmetro de estabilidade, de forma que o genótipo que apresentar menor quadrado médio, ou seja, menor variância, será o mais estável (CRUZ et al., 2012).

Um dos métodos mais utilizados é o de Annicchiarico (1992). Este método vem sendo muito utilizado pelos melhoristas na análise da estabilidade fenotípica, pois o mesmo apresenta uma relativa facilidade de aplicação e de interpretação dos resultados gerados. Esta é baseado em análise de variância e considera a estimativa de um índice de confiança ( $W_i$ ) que representa a chance de um genótipo  $i$  apresentar desempenho fenotípico superior à média geral do conjunto de genótipos que está sendo avaliado.

Neste método, os valores absolutos da variável analisada são convertidos para valores em porcentagem relativa à média de cada ambiente e depois são calculados os desvios relativos de cada genótipo nos diversos ambientes. Posteriormente, a média e os desvios relativos são então utilizados no cálculo do índice de confiança/recomendação. Assim quanto maior o índice de confiança maior será a estabilidade e adaptabilidade da cultivar, traduzida na confiança da indicação do genótipo.

De acordo com Annicchiarico (1992), quanto maior a estimativa de  $W_i$ , mais estável é considerado o genótipo, sendo preferidos os genótipos que apresentem estimativa superior a 100%. Por essa proposta, as cultivares que apresentarem valor de  $W_i$  superior a 100% não deverão apresentar médias fenotípicas inferiores à média geral.

Este método estima o índice de confiança ( $W_i$ ) de um determinado genótipo apresentando desempenho abaixo da média do ambiente, de acordo com o seguinte modelo estatístico:

$$W_i = Y_i - Z_{(1-\alpha)} S_i, \text{ em que:}$$

$W_i$ : índice de confiança, em percentagem;

$Y_i$ : media do genótipo  $i$ , em porcentagem;

$Z$ : percentil  $(1-\alpha)$  da função de distribuição normal acumulada;

$\alpha$ : probabilidade de erro Tipo I;

$S_i$ : desvio padrão dos valores, em percentagem.

### Método de Eberhart e Russell (1966)

O método proposto por Eberhart e Russell (1966), dentre os que se baseiam na regressão linear, tem sido um dos mais utilizadas na recomendação de genótipos em função da simplicidade dos cálculos, facilidade de interpretação e informações fornecidas. Neste método o comportamento de cada genótipo, diante das variações ambientais, é estimado por meio de uma análise de regressão linear simples da variável dependente sobre um índice ambiental, definido como a diferença entre a média de cada ambiente e a média de todos os ambientes. É estimada uma equação de regressão para cada genótipo sob avaliação.

No método de Eberhart e Russel (1966), considera-se o coeficiente de interseção ( $\beta_{0i}$ ), o coeficiente de regressão ( $\beta_{1i}$ ), os desvios da regressão ( $\delta_{ij}$ ), o índice ambiental ( $I_j$ ), ou

qualidade do ambiente, que pode ser obtido por meio da média do ambiente menos a média geral, e o erro experimental médio  $(\varepsilon_{ij})$ . Além disso, são estimados os coeficientes de determinação  $(R^2)$  para as equações obtidas para cada um dos genótipos, para verificar o ajuste da equação, para o genótipo em questão.

O modelo de regressão linear simples utilizado é o seguinte:

$$Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \delta_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\beta_{0i}$ : média geral do  $i$ -ésimo genótipo;

$\beta_{1i}$ : coeficiente de regressão linear, que mede a resposta do  $i$ -ésimo genótipo à variação do ambiente;

$I_j$ : índice ambiental;

$\delta_{ij}$ : desvio da regressão do  $i$ -ésimo genótipo sobre o  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : erro experimental médio associado a observação  $Y_{ij}$ .

A estimativa do parâmetro de estabilidade  $(S_{d_i}^2)$  é obtida de acordo com a seguinte expressão:

$S_{d_i}^2 = QMD_i - QMR/r$ , onde:  $QMD_i$  é o quadrado médio dos desvios de regressão do  $i$ -ésimo genótipo,  $QMR$  é o quadrado médio do resíduo e  $r$  é o número de repetições.

A estimativa do parâmetro de adaptabilidade  $(\beta_i)$  foi obtida por meio da seguinte expressão:

$$\beta_i = \frac{n}{j=1} Y_{ij} I_j / \sum_{j=1}^n I_j^2, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$I_j$ : índice ambiental, sendo  $I_j = (Y_j/p) - (Y_\infty/pn)$ , onde  $Y_j$  é a média de todos os genótipos no  $j$ -ésimo ambiente,  $Y_\infty$  é a média geral,  $n$  é o número de genótipos e  $p$  o número de ambientes.

O coeficiente de determinação  $(R^2)$  é obtido de acordo com a seguinte expressão:

$R_i^2 = [(SQRL_{i\bar{i}})/(SQ(E/G_i))] \times 100$ , onde  $SQRL_{i\bar{i}}$  é a soma de quadrados da regressão linear do  $i$ -ésimo genótipo e  $SQ(E/G_i)$  é a soma de quadrados de ambientes

dentro do *i*-ésimo genótipo. As estimativas para  $\beta_i$  são testadas segundo a hipótese  $H_0: \beta_i = 1$  e hipótese alternativa  $H_1: \beta_i \neq 1$ , avaliada por meio da estatística  $t$ .

Segundo Eberhart e Russell (1966), os genótipos podem ser classificados quanto à adaptabilidade em três grupos:

- a) adaptabilidade geral, com  $\beta_i = 1$ , que apresenta média acima da média geral, sendo o tipo desejável em ambientes com muitas variações imprevisíveis;
- b) adaptabilidade específica a ambientes favoráveis, com  $\beta_i > 1$ , que agrupa os genótipos com alto desempenho em ambientes favoráveis;
- c) adaptabilidade específica a ambientes desfavoráveis, com  $\beta_i < 1$ , que agrupa os genótipos que se destacam em ambientes desfavoráveis.

Os genótipos podem ser classificados quanto à estabilidade em genótipos de alta estabilidade ( $S_{d_i}^2 = 0$ ) e genótipos de baixa estabilidade ( $S_{d_i}^2 \neq 0$ ).

### Método de Cruz, Torres e Vencovsky (1989)

Por ser baseado em apenas uma regressão linear para cada genótipo, o método de Eberhart e Russel (1966) é considerado conservador, uma vez que favorece apenas genótipos com desempenho médio em relação ao conjunto analisado. Diante da hipótese de se identificar genótipos com performance desejável nos ambientes considerados desfavoráveis e favoráveis, considerou-se a alternativa de modificar este método empregando-se dois segmentos de reta, ou seja, regressão linear bissegmentada (Silva e Barreto, 1986; Silva, 1995a; Silva, 1995b).

O método de Cruz et al. (1989) baseia-se na análise de regressão bissegmentada, possuindo três parâmetros de adaptabilidade: a média ( $\beta_{0i}$ ), a resposta linear aos ambientes desfavoráveis ( $\beta_{1i}$ ) e a resposta linear aos ambientes favoráveis ( $\beta_{2i}$ ). A estabilidade é avaliada pelo desvio da regressão ( $\delta_{ij}$ ) de cada cultivar em função das variações ambientais.

Por esse método as estimativas  $\beta_{1i}$  e  $\beta_{1i} + \beta_{2i}$  são não correlacionadas entre si, o que leva a independência dos dois segmentos de reta. O genótipo ideal é aquele que apresenta alta média de produtividade ( $\beta_{0i}$ ), baixo  $\beta_{1i} < 1$  (adaptabilidade a ambientes desfavoráveis),  $\beta_{1i} + \beta_{2i} > 1$  (responsividade à melhoria ambiental), e  $\delta_{ij} = 0$  (estabilidade fenotípica).

Portanto, nesse método, a média ( $\beta_0$ ), a resposta linear a ambientes desfavoráveis ( $\beta_1$ ) e favoráveis ( $\beta_1 + \beta_2$ ) são os parâmetros que estimam a adaptabilidade dos genótipos, e os

desvios da regressão  $(\delta^2)$  de cada genótipo e o coeficiente de determinação  $(R^2)$ , constituem os parâmetros que estimam a estabilidade. O modelo estatístico deste método é o seguinte:

$$Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \beta_{2i} T(I_j) + \delta_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\beta_{0i}$ : média geral do  $i$ -ésimo genótipo;

$\beta_{1i}$ : coeficiente de regressão linear, que mede a resposta do  $i$ -ésimo genótipo nos ambientes desfavoráveis;

$I_j$ : índice ambiental;

$\beta_{1i} + \beta_{2i}$ : mede a resposta do  $i$ -ésimo genótipo nos ambientes desfavoráveis;

$T(I_j)$ : variável do eixo das abcissas;

$\delta_{ij}$ : desvio da regressão do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : erro experimental médio associado a observação  $Y_{ij}$ .

### Método de Lin e Binns (1988)

O método proposto por Lin e Binns (1988), baseado em métodos não paramétricos, é simples e fácil de interpretar, possibilitando identificar um ou mais genótipos com desempenho próximo ao máximo nos vários ambientes testados. Neste método, os genótipos superiores são identificados por meio de um simples parâmetro  $(P_i)$ , associado à estabilidade e à produtividade, e define um genótipo superior como aquele que apresenta performance próxima do máximo nos vários ambientes testados. A estimativa  $P_i$  é o quadrado médio da distância em relação à resposta máxima em cada ambiente. O genótipo ideal é aquele com menor valor de  $P_i$ .

A medida de estabilidade é definida como o quadrado médio da distância entre a média do genótipo e a resposta média máxima para todos os ambientes. Essa definição é detalhada por Rocha (2002), mostrando que o parâmetro  $P_i$  representa de fato o quadrado médio da distância entre a resposta de um determinado genótipo em relação à resposta do genótipo que apresenta produtividade máxima, dentre todos os genótipos, em um determinado ambiente. Então, quanto menor a distância entre a resposta do genótipo e a resposta do genótipo de produtividade máxima nos diversos ambientes, menor será o valor de  $P_i$  e mais estável o genótipo. O valor de  $P_i$  é estimado por meio da seguinte expressão:

$$P_i = \sum_{j=1}^n \left( Y_{ij} - M_j \right)^2 / 2n, \text{ em que:}$$

$P_i$ : índice de superioridade do  $i$ -ésimo genótipo;

$Y_{ij}$ : produtividade do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$M_j$ : resposta máxima obtida dentre todos os genótipos no  $j$ -ésimo ambiente;

$n$ : número de ambientes.

A expressão anterior pode ser desdobrada da seguinte forma:

$$P_i = \left[ n(\bar{Y}_i - \bar{M})^2 + \sum_{j=1}^n (Y_{ij} - \bar{Y}_i - M_j + \bar{M})^2 \right] / 2n, \text{ onde } \bar{Y}_i \text{ é a média de}$$

produtividade do  $i$ -ésimo genótipo nos  $n$  ambientes, sendo expressa como:

$$\bar{Y}_i = \frac{\sum_{j=1}^n Y_{ij}}{n} \text{ e } \bar{M} \text{ é a média das produtividades máximas de todos os genótipos em todos os}$$

ambientes, sendo expressa como:

$$\bar{M} = \frac{\sum_{j=1}^n M_j}{n}. \text{ Estas estimativas são utilizadas para testar a hipótese de nulidade da}$$

estimativa  $P_i$ , para cada genótipo por meio do teste F. A significância do teste indica que o genótipo difere estatisticamente do máximo ao longo dos ambientes.

#### Método de Lin e Binns modificado por Carneiro (1998)

As modificações no método de Lin e Binns (1988) propostas por Carneiro (1998) têm como objetivo a recomendação de genótipos dividindo o conjunto de ambientes em favoráveis e desfavoráveis, de forma a refletir ambiente onde há emprego de alta e baixa tecnologia, respectivamente, além de introduzir um referencial mais apropriado do que a distância do genótipo a pontos máximos, como preconizado na metodologia de Lin e Binns (1988). Com isso, o método torna-se de aplicação mais ampla tanto em relação aos caracteres avaliados como aos genótipos disponíveis. A classificação dos ambientes segundo essa metodologia é baseada nos índices ambientais que nada mais são do que a diferença da média dos genótipos em cada ambiente e a média geral.

Neste método, a decomposição da estimativa  $P_i$ , para ambientes favoráveis e desfavoráveis é estimada da seguinte forma:

$$P_{if} = \frac{\sum_{j=1}^f (Y_{ij} - M_j)^2}{2f} \text{ e } P_{id} = \frac{\sum_{j=1}^{d-f} (Y_{ij} - M_j)^2}{2(d-f)}, \text{ em que:}$$

$P_{if}$ : índice de superioridade do  $i$ -ésimo genótipo nos  $j$  ambientes favoráveis;

$P_{id}$ : índice de superioridade do  $i$ -ésimo genótipo nos  $j$  ambientes desfavoráveis;

$Y_{ij}$ : produtividade do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$M_j$ : resposta máxima obtida dentre todos os genótipos no  $j$ -ésimo ambiente favorável ou desfavorável;

$f$  : número de ambientes favoráveis;

$d$  : número de ambientes desfavoráveis;

## Métodos novos de avaliação de estabilidade e adaptabilidade

### Método AMMI-Biplot

O método denominado efeitos aditivos e interação multiplicativa (AMMI-Biplot) é uma combinação entre a análise de variância e a análise de componentes principais (ACP). Neste caso, os componentes aditivos são utilizados para estudar os efeitos principais (genótipos e ambientes) e os componentes multiplicativos, para estudar a interação genótipos x ambientes. Na ACP a variação contida nos componentes principais significativos é denominada padrão, e a contida nos componentes não significativos, denomina-se de ruído (Zobel et al., 1988). O modelo estatístico associado a este método é o seguinte:

$$\bar{Y}_{ij} = \mu + g_i + a_j + \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} + r_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$\bar{Y}_{ij}$ : média de desempenho do genótipo  $i$  ( $i=1,2,\dots,g$ ) no ambiente  $j$  ( $j=1,2,\dots,a$ );

$\mu$ : média geral dos experimentos;

$g_i$ : efeito fixo do genótipo  $i$ ;

$a_j$ : efeito fixo do ambiente  $j$ ;

$\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$ : efeito fixo da interação genótipos x ambientes multiplicativa, onde  $\lambda_k$  é o valor singular,  $\gamma_{ik}$  e  $\alpha_{jk}$  são os escores do eixo  $k$  da ACP, para genótipo e ambiente, respectivamente, e  $n$  é o número de eixos ou de componentes principais retidos para descrever o padrão da interação GxA na análise AMMI;

$r_{ij}$ : efeito residual do modelo AMMI (ruído);

$\varepsilon_{ij}$ : erro experimental considerado de efeito aleatório.

Para a definição do número de eixos a serem retidos na análise para explicar o padrão relacionado a interação GxA, pode ser adotado o critério proposto por Gauch e Zobel (1988), que considera a proporção da soma de quadrados da interação genótipos x ambientes acumulada nos primeiros eixos. A significância do teste F de Gollob pode ser utilizado como determinação do modelo para cada família AMMI (AMMI1, AMMI2, ..., AMMIn) e os resíduos dos eixos testados pelo teste F de Cornelius.

Também pode ser calculado o valor de estabilidade AMMI ("AMMI stability value-ASV"). O ASV é obtido com base no cálculo da distância euclidiana entre a origem do plano cartesiano e a coordenada do ponto do genótipo ou ambiente (PURCHASE et al., 2000). Como o valor do IPCA1 contribui mais para a interação GxA, é necessário o uso de um valor ponderado. Este valor é calculado para cada genótipo e ambiente, de acordo com a contribuição relativa do IPCA1 e o IPCA2 para a interação GxA, por meio da seguinte expressão:

$$ASV = \sqrt{\left[ \frac{SQ_{IPCA1}}{SQ_{IPCA2}} (Escore_{IPCA1})^2 \right] + (Escore_{IPCA2})}, \text{ em que:}$$

$SQ_{IPCA1}$ : soma de quadrados do IPCA1;

$SQ_{IPCA2}$ : soma de quadrados do IPCA2.

Quanto menor o valor de ASV, maior será a estabilidade do genótipo.

### Método GGE-Biplot

O modelo associado ao método GGE, proposto por Yan et al. (2007), é expresso como:

$$Y_{ij} - \mu - E_j = y_1 e_{i1} \rho_{j1} + y_2 e_{i2} \rho_{j2} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : desempenho do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\mu$ : média geral das observações;

$E_j$ : efeito principal do  $j$ -ésimo ambiente;

$y_1$  e  $y_2$ : valores singulares associados ao IPCA1 e IPCA2, respectivamente;

$e_{i1}$  e  $e_{i2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $i$ -ésimo genótipo;

$\rho_{j1}$  e  $\rho_{j2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : efeito residual não explicado por nenhum dos fatores (“ruído”).

Neste método, após obtenção dos escores associados aos ambientes e genótipos, são construídos os gráficos biplot e feitas as análises de “which-won-where”, média x estabilidade, discriminativo x representativo e genótipo ideal (YAN; TINKER, 2006).

### Método MHPRVG via REML/BLUP

A análise por meio da metodologia de modelos lineares mistos utiliza a máxima verossimilhança restrita (REML) para estimar os componentes de variância do modelo utilizado e o melhor preditor linear não viesado (BLUP) para estimar o valor genotípico dos genótipos. Esse método é conhecido como REML/BLUP (RESENDE, 2007).

No método BLUP, o efeito de genótipos pode ser considerado como aleatório, tendo como vantagem a facilidade de implementação nas análises quando comparado a outros estimadores do tipo “shrinkage” e o uso em experimentos desbalanceados. O efeito aleatório da interação genótipos x ambientes também permite inferências para uma população de ambientes.

Com base no método REML/BLUP, a medida simultânea de adaptabilidade e estabilidade para cada genótipo é obtida por meio da média harmônica do desempenho relativo dos valores genotípicos (MHPRVG). Este método tem como princípio que quanto menor o valor do desvio-padrão do comportamento genotípico entre os ambientes, maior será a média harmônica dos valores genotípicos. A seleção pelos maiores valores da média harmônica dos valores genotípicos (MHVG)

determina, simultaneamente a produtividade e a estabilidade. O desempenho relativo associado aos valores genotípicos (PRVG) nos diferentes ambientes exprime a adaptabilidade (RESENDE, 2002).

Por exemplo, para o método da máxima verossimilhança restrita/melhor predição linear não viesada (REML/BLUP) pode ser utilizado o modelo 52 – Delineamento em blocos incompletos e vários locais e uma observação por parcela (RESENDE, 2007), associado ao seguinte modelo estatístico:

$$y = X_r + Z_g + W_b + T_i + e, \text{ em que:}$$

$y$ : vetor de observações da característica;

$r$ : vetor de efeitos fixos de repetição somados com a média geral;

$g$ : vetor de efeitos genotípicos aleatórios;

$b$ : vetor de efeitos aleatórios de blocos;

$i$ : vetor de efeitos aleatórios da interação genótipo x ambiente ( $ga$ );

$e$ : vetor aleatório de erros ou resíduos.

As matrizes  $X$ ,  $Z$ ,  $W$  e  $T$  representam as incidências dos efeitos dos fatores  $r$ ,  $g$ ,  $b$  e  $i$ , respectivamente. A média e as variâncias deste modelo são estruturadas e distribuídas da seguinte forma:

$$\begin{aligned} y/r, V &\square N(Xr, V); g/\sigma_g^2 \square N(0, I\sigma_g^2); b/\sigma_b^2 \square N(0, \sigma_b^2); \\ i/\sigma_i^2 &\square N(0, I\sigma_i^2); e/\sigma_e^2 \square N(0, I\sigma_e^2) \end{aligned}$$

Os valores genotípicos são obtidos por meio da solução das equações de modelos mistos a seguir:

$$\left[ \begin{array}{cccc} XX & XZ & XW & XT \\ Z'X & Z'Z + I \frac{\sigma_e^2}{\sigma_g^2} & Z'W & Z'T \\ W'X & W'Z & W'W + I \frac{\sigma_e^2}{\sigma_b^2} & W'T \\ T'X & T'Z & T'W & T'T + I \frac{\sigma_e^2}{\sigma_{ga}^2} \end{array} \right] \begin{bmatrix} \hat{r} \\ \hat{g} \\ \hat{b} \\ \hat{i} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ T'y \end{bmatrix}$$

A partir dos valores genotípicos preditos ( $\hat{g}$ ), as percentagens relativas dos valores genotípicos (**PRVG**) são estimados para cada genótipo nos diferentes ambientes. A medida simultânea de adaptabilidade e estabilidade para cada genótipo é obtida por meio da Média Harmônica da Performance Relativa dos Valores Genotípicos (**MHPRVG**) com base na seguinte expressão:

$$MHPRVG_{ij} = n \sqrt[k]{\sum_{j=1}^k \frac{1}{PRVG_{ij}}}, \text{ em que:}$$

$n$ : número de ambientes;

$$PRVG_{ij} = VG_{ij} / VG_j$$
, sendo:

$VG_{ij}$ : valor genotípico do genótipo  $i$  no ambiente  $j$  e

$VG_j$ : média genotípica no ambiente  $j$ .

Para melhorar a interpretação dos resultados, os valores de MHPRVG são multiplicados pela média geral (**MG**) de todos os ambientes (**MHPRVG×MG**), representando os resultados na mesma magnitude da característica estudada. Depois os ambientes são agrupados em favoráveis e desfavoráveis de acordo com a média geral da característica e são conduzidas análises separadas para cada grupo. Os ambientes com média acima da média geral são considerados favoráveis e os ambientes com média abaixo da média geral são considerados desfavoráveis (Mendes et al., 2012). Então, para analisar em conjunto a adaptabilidade e a estabilidade, os resultados dos valores **MHPRVG×MG** de cada grupo são comparados em gráfico de dispersão, que contrasta ambientes favoráveis (eixo das abscissas) e ambientes desfavoráveis (eixo das ordenadas). O plano cartesiano do gráfico é dividido em quatro quadrantes, da seguinte forma: I(Inférieur Esquerdo)-genótipos com baixo desempenho em ambos grupos de ambientes ; II(Inférieur Direito)- genótipos com adaptabilidade específica a ambientes favoráveis; III(Superior Direito)- genótipos com desempenho superior em ambos grupos de ambientes; IV(Superior Esquerdo)-genótipos com adaptabilidade específica a ambientes desfavoráveis (Mendes et al., 2012; Yamamoto et al., 2021).

# CAPITULO 20

## Análise de Trilha

### Conceitos

Os estudos de correlação entre características não permitem concluir sobre relações de causa e efeito. Desta forma, a existência de uma correlação entre as características  $X$  e  $Y$  não implica que  $Y$  é causado por  $X$ , ou vice-versa. A correlação mede apenas o grau de associação entre características.

A análise de trilha é um método que consiste no estudo dos efeitos diretos e indiretos de várias variáveis/características sobre uma variável/característica básica, com base em relações de causa e efeito previamente estabelecidas. As estimativas dos efeitos são obtidas por meio de equações de regressão, em que as variáveis originais são previamente padronizadas.

Este método fornece quantidades, chamadas de coeficientes de trilha que medem a influência direta de uma variável sobre outra, independentemente das demais, no contexto de relações de causa e efeito. Permite também desdobrar coeficientes de correlação simples em seus efeitos diretos e indiretos. As relações de causa e efeito são especificadas por meio do diagrama de trilha, como mostrado na figura a seguir:

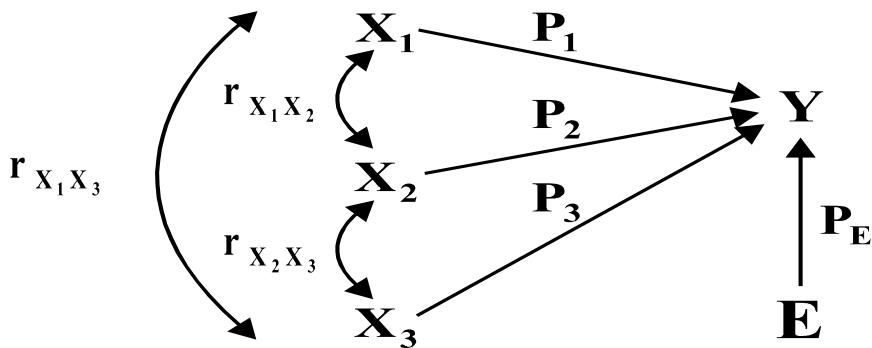


Figura 1. Diagrama causal de efeitos diretos – coeficientes de trilha  $P_1$ ,  $P_2$ ,  $P_3$ ,  $P_E$  e indiretos –  $r_{12}P_2$ ,  $r_{13}P_3$ ,  $r_{21}P_1$ ,  $r_{23}P_3$ ,  $r_{31}P_1$ ,  $r_{32}P_2$  das variáveis  $X_1$ ,  $X_2$ ,  $X_3$  e  $E$  sobre a variável básica  $Y$ .

Diferentemente das análises de correlação múltipla e parcial, esta análise requer uma relação de causa e efeito entre as variáveis. A causa e o efeito são estabelecidos “*a priori*” ou experimentalmente, de tal forma que uma variável é a causa das variações de outras, e, que, determinados pares de variações são correlacionados como o efeito de determinada causa em comum.

Coeficientes de trilha são coeficientes de regressão parciais padronizados. Como no caso de coeficientes de regressão parciais, eles medem o efeito direto de uma variável independente ( $X$ )

sobre uma variável dependente ( $Y$ ), após a “remoção” da influência de todas as outras variáveis independentes incluídas na análise.

### Padronizações de Variáveis e Estimação de Coeficientes de Trilha

Numa regressão linear múltipla, a soma de quadrados da variável dependente pode ser decomposta por meio de formas padronizadas das variáveis incluídas no modelo. A transformação de uma variável  $Y \sim N(\mu_Y, \sigma_Y^2)$  para sua forma padronizada  $y \sim N(0, 1)$  é feita da seguinte

$$\text{forma: } y = \frac{Y - \mu_Y}{\sigma_Y}, \quad \mu_Y = \bar{Y}.$$

Então, o modelo de regressão linear múltipla,

$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ , por exemplo, pode ser reescrito usando-se os desvios das variáveis em relação às suas respectivas médias, ou seja:

$$(Y - \mu_Y) = \beta_0(X_0 - \mu_0) + \beta_1(X_1 - \mu_1) + \beta_2(X_2 - \mu_2) + \beta_3(X_3 - \mu_3)$$

Aqui, a variável  $X_0$  é denominada de variável “dummy” e seu valor é sempre 1, o que torna o termo  $(X_0 - \mu_0)$  igual à zero. Os outros termos do modelo podem ser expressos como uma operação do desvio padrão de  $Y(\sigma_Y)$ , o que resulta em:

$$\frac{(Y - \mu_Y)}{\sigma_Y} = \beta_1 \frac{1}{\sigma_Y} (X_1 - \mu_1) + \beta_2 \frac{1}{\sigma_Y} (X_2 - \mu_2) + \beta_3 \frac{1}{\sigma_Y} (X_3 - \mu_3)$$

Os termos do lado direito da equação anterior podem ser multiplicados por  $\sigma_1/\sigma_1$ ,  $\sigma_2/\sigma_2$  e  $\sigma_3/\sigma_3$ , sem alterar a igualdade, o que resulta em:

$$\left( \frac{Y - \mu_Y}{\sigma_Y} \right) = \left( \beta_1 \frac{\sigma_1}{\sigma_Y} \right) \left( \frac{X_1 - \mu_1}{\sigma_1} \right) + \left( \beta_2 \frac{\sigma_2}{\sigma_Y} \right) \left( \frac{X_2 - \mu_2}{\sigma_2} \right) + \left( \beta_3 \frac{\sigma_3}{\sigma_Y} \right) \left( \frac{X_3 - \mu_3}{\sigma_3} \right)$$

Desta forma, o modelo com variáveis padronizadas pode ser escrito como:

$$y = P_1 x_1 + P_2 x_2 + P_3 x_3, \text{ onde:}$$

$$y = \left( \frac{Y - \mu_Y}{\sigma_Y} \right), \text{ forma padronizada de } Y$$

$$x_1 = \left( \frac{X_1 - \mu_1}{\sigma_1} \right), \text{ forma padronizada de } X_1$$

$$x_2 = \left( \frac{X_2 - \mu_2}{\sigma_2} \right), \text{ forma padronizada de } X_2$$

$$x_3 = \left( \frac{X_3 - \mu_3}{\sigma_3} \right), \text{ forma padronizada de } X_3$$

$$P_1 = \left( \beta_1 \frac{\sigma_1}{\sigma_Y} \right), \text{ coeficiente de regressão parcial padronizado de } y \text{ em função de } x_1$$

$$P_2 = \left( \beta_2 \frac{\sigma_2}{\sigma_Y} \right), \text{ coeficiente de regressão parcial padronizado de } y \text{ em função de } x_2$$

$$P_3 = \left( \beta_3 \frac{\sigma_3}{\sigma_Y} \right), \text{ coeficiente de regressão parcial padronizado de } y \text{ em função de } x_3$$

O coeficiente de regressão parcial padronizado reflete a proporção do desvio padrão em  $Y$ , que é atribuída a variação em  $X$ . A interpretação mais simples surge quando se transforma todas as variáveis para um sistema em que os desvios padrão são iguais a 1. Os coeficientes passam a ser definidos como a mudança em  $Y$ , medida em unidades de desvio padrão, para a mudança em desvio padrão em  $X$ , mantendo as outras variáveis independentes constantes.

A decomposição da variância da variável dependente pode ser ilustrada para o modelo padronizado da seguinte forma:

Considere, por exemplo, a variância de  $y$  quando  $y = P_1 x_1 + P_2 x_2$  que é expressa por:

$$\sigma_y^2 = P_1^2 \sigma_1^2 + P_2^2 \sigma_2^2 + 2P_1 P_2 \rho_{12}. \text{ Uma vez que todas as variáveis são padronizadas, tem-se}$$

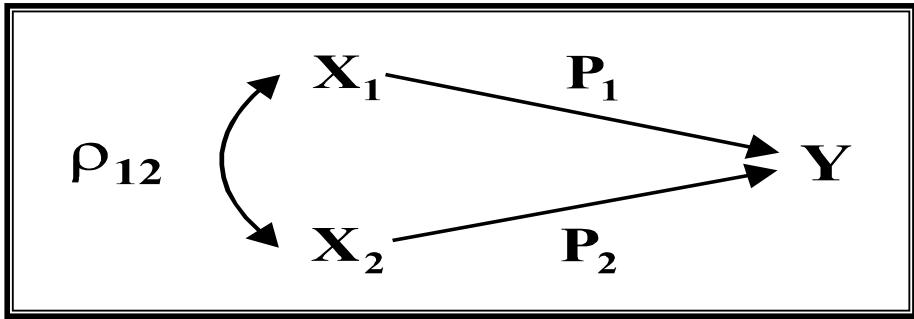
que  $\sigma_y^2 = \sigma_1^2 = \sigma_2^2 = 1$  e então a variância de  $y$  pode ser escrita como:

$$1 = P_1^2 + P_2^2 + 2P_1 P_2 \rho_{12}.$$

A variância total em  $y$ , uma variável padronizada, é a unidade. Esta variância total pode ser decomposta em uma porção  $P_1^2$ , pelo efeito direto de  $x_1$  e uma porção  $P_2^2$ , pelo efeito direto de  $x_2$ . A porção remanescente,  $2P_1 P_2 \rho_{12}$  pode ser atribuída aos efeitos conjuntos de  $x_1$  e  $x_2$ , que se não são correlacionados, ou seja  $\rho_{12} = 0$ , então, a variância em  $y$  é simplesmente a soma

dos quadrados dos dois efeitos diretos,  $P_1^2 + P_2^2$ , e os efeitos diretos são simplesmente as correlações entre  $x_1$  e  $y$ , e entre  $x_2$  e  $y$ .

O modelo anterior pode ser representado num *diagrama causal* (ou *diagrama de trilha*) da seguinte forma:

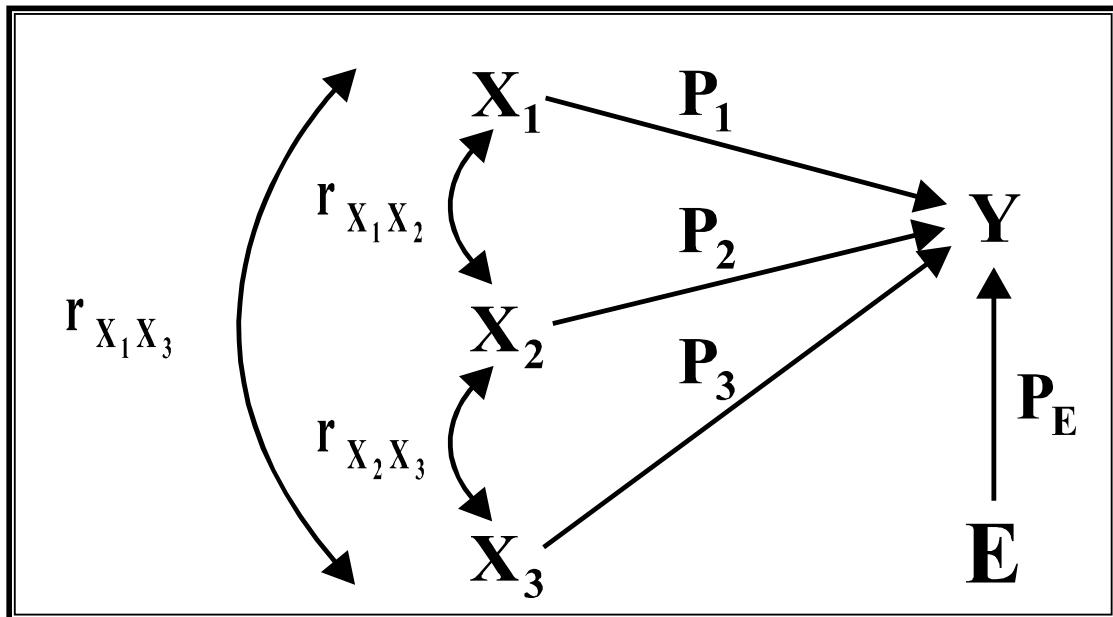


Este diagrama informa que existe “*a priori*” um conhecimento dos fatores que influenciam  $Y$ , ou que alguma hipótese a respeito de possíveis relacionamentos de causa e efeito está sob investigação. A influência de  $X_1$  sobre  $Y$ , quando  $\rho_{12}$  é zero, é simplesmente o quadrado da trilha direta de  $X_1$  para  $Y$  ou seja  $P_1^2$ . A influência de  $X_1$  sobre  $Y$ , quando  $\rho_{12} \neq 0$ , tem dois componentes, a trilha direta  $P_1^2$  e a trilha indireta via  $X_2$ , dada por  $P_2\rho_{12}$ .

Estes diagramas, usados com variáveis padronizadas, têm demonstrado ser um instrumento valioso, em muitas áreas de pesquisa. Na prática ocorrem situações em que se trabalha com um grande número de variáveis independentes, com relacionamento entre elas, e também componentes aleatórios não especificados, influenciando a variância da variável dependente.

#### Exemplo de Aplicação (Extraído de Vencovsky e Barriga, 1992)

Considere, por exemplo, a produtividade de grãos ( $Y$ ) de uma planta como variável básica (efeito). Este efeito é função de vários componentes (fatores causais). Em algumas espécies,  $Y$  é função do número de vagens por planta ( $X_1$ ), do número de grãos por vagem ( $X_2$ ) e do peso dos grãos ( $X_3$ ). Geralmente, não é suficiente considerar apenas a variável principal  $Y$ , sendo necessário investigar suas componentes. O exemplo em consideração pode ser representado num diagrama de trilha como a seguir:



Este é um diagrama das relações de causa e efeito (diagrama causal ou diagrama de trilha) da produtividade de grãos e seus componentes. No diagrama pode-se observar que a produtividade ( $Y$ ) é o efeito resultante dos fatores  $X_1$ ,  $X_2$ ,  $X_3$  e de outro fator, formado por um conjunto de fatores de erros, denominado de  $E$ . As variáveis  $X_1$ ,  $X_2$  e  $X_3$  são correlacionadas entre si;  $P_1$ ,  $P_2$ ,  $P_3$  e  $P_E$  são os coeficientes de trilha determinantes de  $Y$ . Subentende-se que  $Y$  é o efeito, e  $X_1$ ,  $X_2$ ,  $X_3$  e  $E$  as causas.

A base estatística da análise de trilha é uma regressão linear múltipla que pode ser representada por:

$$Y = A + B_1 X_1 + B_2 X_2 + B_3 X_3 + E.$$

Tomam-se variáveis padronizadas na regressão considerando:

$$y = \frac{Y - \bar{Y}}{\sigma_Y} \text{ e } x_i = \frac{X_i - \bar{X}_i}{\sigma_{X_i}}.$$

Nestas condições a equação de regressão passa a ser:

$y = P_1 x_1 + P_2 x_2 + P_3 x_3 + P_E e$ , sendo  $P_1$ ,  $P_2$ ,  $P_3$  e  $P_E$  os coeficientes de trilha;  $e$  também é uma variável padronizada.

A relação entre os coeficientes de regressão com as variáveis originais ( $B$ ) e os coeficientes de trilha ( $P$ ) é expressa por:

$$P_1 = \frac{B_1 \sigma_{X_1}}{\sigma_Y}, P_2 = \frac{B_2 \sigma_{X_2}}{\sigma_Y} \text{ e } P_3 = \frac{B_3 \sigma_{X_3}}{\sigma_Y}$$

Observa-se, portanto, que os coeficientes  $P$  são coeficientes de regressão parcial padronizados, como já mencionado.

As correlações simples entre os caracteres são:

$$\begin{aligned} r(X_1 X_2) &= r_{12}; r(X_1 X_3) = r_{13}; r(X_2 X_3) = r_{23}; \\ r(X_1 Y) &= r_{1Y}; r(X_2 Y) = r_{2Y}; r(X_3 Y) = r_{3Y}; \end{aligned}$$

Então, na análise de trilha se tem que:

$$r_{1Y} = P_1 + r_{12} P_2 + r_{13} P_3$$

$$r_{2Y} = r_{12} P_1 + P_2 + r_{23} P_3$$

$$r_{3Y} = r_{13} P_1 + r_{23} P_2 + P_3$$

Nestas equações nota-se que a correlação entre  $X_1$  e  $Y$  pode ser desdobrada em três partes: 1) efeito direto de  $X_1$  sobre  $Y$ , quantificado por  $P_1$ ; 2) efeito indireto de  $X_1$  sobre  $Y$ , via  $X_2$ , quantificado por  $r_{12} P_2$ ; e 3) efeito indireto de  $X_1$  sobre  $Y$ , via  $X_3$ , quantificado por  $r_{13} P_3$ . O mesmo raciocínio vale para as outras correlações:  $r_{2Y}$  e  $r_{3Y}$ .

Entende-se que os componentes de cada correlação, sendo coeficientes de regressão ou função destes, podem atingir valores maiores do que 1,0 ou menores do que -1,0. Percebe-se que as avaliações diretas apenas das correlações  $r_{1Y}$ ,  $r_{2Y}$  e  $r_{3Y}$  não são totalmente informativas. O desdobramento destas correlações esclarece grandemente a maneira como os caracteres causais influem na expressão, ou variação, de  $Y$ .

As estimativas dos coeficientes de trilha ( $P$ ) são obtidas resolvendo o sistema de equações seguinte:

$$\begin{bmatrix} r_{1Y} \\ r_{2Y} \\ r_{3Y} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$$

Tem-se que  $r_{12} = r_{21}$ ;  $r_{13} = r_{31}$ ;  $r_{23} = r_{32}$ .

De posse das estimativas de  $P_1$ ,  $P_2$  e  $P_3$  calculam-se os valores dos efeitos indiretos. Por exemplo, o efeito indireto do número de vagens por planta ( $X_1$ ), via número de grãos por vagem ( $X_2$ ) será  $r_{12}P_2$ .

O coeficiente de trilha devido aos efeitos de erros ou causas não controladas ( $E$ ) obtém-se por:

$$P_E = \sqrt{1 - (P_1 r_{1Y}) - (P_2 r_{2Y}) - (P_3 r_{3Y})}$$

Esta expressão é válida para qualquer número de variáveis  $X$ , bastando estendê-la com termos do tipo  $P_i r_{iY}$  ( $i=1, 2, \dots, n$  variáveis  $X$ ).

Os dados apresentados a seguir referem-se às médias de produtividade de grãos ( $Y$ ) de 28 genótipos de feijoeiro, que foram avaliados também para número de vagens por planta ( $X_1$ ), número de grãos por vagem ( $X_2$ ) e peso de 100 grãos ( $X_3$ ):

Genótipos	$Y$	$X_1$	$X_2$	$X_3$
1	5,67	3,78	2,44	61,67
2	2,23	2,20	2,36	43,89
3	5,82	4,81	3,23	37,34
4	3,91	5,35	4,28	19,15
5	4,21	2,55	3,58	47,84
6	4,83	3,84	3,28	38,57
7	4,05	4,24	4,04	23,76
8	3,36	2,79	2,30	52,43
9	6,15	4,38	2,94	47,71
10	4,41	5,22	2,67	31,40
11	4,07	2,74	2,59	57,07
12	4,87	3,33	2,77	52,86
13	3,83	3,59	3,16	34,25
14	5,35	4,42	2,84	43,04
15	4,20	5,36	2,57	31,44
16	3,73	2,90	2,89	44,56
17	3,48	2,87	2,69	45,22
18	4,83	5,58	3,46	28,56
19	5,79	7,46	3,28	23,61
20	5,80	3,97	3,45	42,36
21	5,32	4,18	3,06	41,99
22	4,97	5,59	3,29	27,06
24	5,31	6,84	2,85	27,85
25	4,30	5,96	3,75	19,10
26	5,23	3,53	3,27	45,07
27	4,92	4,77	3,20	33,99
28	5,53	5,34	3,35	31,40

As estimativas das correlações simples são:

$$r_{1Y} = 0,482; r_{2Y} = 0,203; r_{3Y} = 0,09$$

$$r_{12} = 0,340; r_{13} = 0,767; r_{23} = 0,613$$

Substituindo estes valores no sistema de equações tem-se:

$$\begin{bmatrix} 0,482 \\ 0,203 \\ 0,009 \end{bmatrix} = \begin{bmatrix} 1,00 & 0,340 & -0,767 \\ 0,340 & 1,000 & -0,613 \\ -0,767 & -0,613 & 1,000 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$$

Pela resolução deste sistema obtém-se:

$$P_1 = 1,3844; P_2 = 0,6227; P_3 = 1,4525 \text{ e } P_E = 0,4395$$

A análise detalhada dos efeitos diretos e indiretos das características sobre a produtividade de grãos do feijoeiro é a seguinte:

A) Produtividade de grãos ( $Y$ ) e número de vagens por planta ( $X_1$ ):

Efeitos	Valores
Direto	$P_1 = 1,3844$
Indireto via nº grãos/vagem ( $X_2$ )	$P_{212} = 0,2117$
Indireto via peso 100 grãos ( $X_3$ )	$P_{313} = -1,1141$
Total (direto + indireto)	$r_{1Y} = 0,4820$

B) Produtividade de grãos ( $Y$ ) e número de grãos por vagem ( $X_2$ )

Efeitos	Valores
Direto	$P_2 = 0,6227$
Indireto via nº vagens/planta ( $X_1$ )	$P_{112} = 0,4707$
Indireto via peso 100 grãos ( $X_3$ )	$P_{323} = -0,8904$
Total (direto + indireto)	$r_{2Y} = 0,2030$

C) Produtividade de grãos ( $Y$ ) e peso de 100 grãos ( $X_3$ )

Efeitos	Valores
Direto	$P_3 = 1,4525$
Indireto via nº vagens/planta ( $X_1$ )	$P_{113} = -1,0618$
Indireto via nº grãos/vagem ( $X_2$ )	$P_{223} = -0,3817$
Total (direto + indireto)	$r_{3Y} = 0,0090$

D) Efeito de Erro  $P_E = 0,4395$

Uma forma alternativa de apresentação dos resultados é por meio de um quadro resumido dos efeitos diretos (na diagonal) e indiretos (fora da diagonal) dos componentes sobre a produtividade de grãos do feijoeiro, como a seguir:

Caracteres Componentes	Vagens/Planta	Grãos/Vagem	Peso 100 Grãos	Correlação com Produtividade (*)
Vagens/Planta	1,3844	0,2117	-1,1141	0,4820
Grãos/Vagem	0,4707	0,6227	-0,8904	0,2030
Peso 100 Grãos	-1,0618	-0,3817	1,4525	0,0090

\* Correlações  $r_{1Y}$ ,  $r_{2Y}$  e  $r_{3Y}$

Pode ser estimado também o coeficiente de determinação,  $R^2_{Y.123} = P_1 r_{1Y} + P_2 r_{2Y} + P_3 r_{3Y}$ , que corresponde à determinação tanto do modelo causal (diagrama de trilha) quanto do modelo de regressão linear múltipla pressuposto para  $Y$  em função de  $X_1$ ,  $X_2$  e  $X_3$ . Neste exemplo, a estimativa de  $R^2_{Y.123}$  é obtida como:

$$R^2_{Y.123} = 1,3844 \times 0,4820 + 0,6227 \times 0,2030 +$$

$$+ 1,4525 \times 0,0090$$

$$R^2_{Y.123} = 0,8068$$

As principais propriedades da análise de trilha são (Cruz e Regazzi, 1993):

1) a trilha ("path") é direcional e seus coeficientes, que expressam os efeitos diretos de caracteres, podem assumir valores maiores que a unidade, negativos ou positivos;

2) pode ser usada para comparar efeitos de caracteres mensuráveis em escalas diferentes, pois é um coeficiente de regressão padronizado;

3) duas variáveis podem não ser correlacionadas, mas o coeficiente de trilha pode assumir valor diferente de zero;

4) duas variáveis podem ser completamente determinadas pela mesma causa em comum e, mesmo assim, não apresentarem correlação.

Com o desdobramento de coeficientes de correlação em seus efeitos diretos e indiretos, surgem algumas situações interessantes (Vencovsky e Barriga, 1992):

1) a correlação semelhante em sinal e magnitude com o efeito direto evidencia que a variável independente é determinante das variações na variável básica;

2) se a correlação é positiva, mas seu efeito direto é negativo ou pequeno, devem-se considerar os efeitos indiretos;

3) a correlação negativa associada com o efeito direto positivo e alto indica que a variável independente não deve ser descartada.

Geralmente, as análises de trilha não consideram os efeitos adversos da multicolinearidade sobre os estimadores de quadrados mínimos, adotados para a resolução dos sistemas de equações. A multicolinearidade ocorre quando as observações da amostra das variáveis independentes, ou suas combinações lineares, são altamente correlacionadas.

# CAPITULO 21

## Análise de Componentes Principais

### Caracterização do método

O método de componentes principais foi originalmente descrito por Karl Pearson, em 1901, e posteriormente consolidado por Hotelling em 1933 e 1936, com o propósito de analisar estruturas de correlações. Os objetivos deste método são:

- a) avaliar as correlações entre caracteres estudados;
- b) reduzir um grande conjunto de caracteres para evidenciar um sentido biológico;
- c) promover a eliminação de caracteres que contribuem pouco com a variação de um grupo de indivíduos (ou objetos) avaliados;
- d) possibilitar o agrupamento de indivíduos similares, por meio de exames visuais em dispersões gráficas no espaço bi ou tridimensional.

Neste método as variáveis originais são substituídas por novas variáveis abstratas, os componentes principais (CP), esperando que os primeiros componentes guardem em si quase toda a variação que possuem as variáveis originais. Cada componente principal é uma combinação linear das variáveis originais analisadas. Os componentes principais são independentes entre si e estimados com o objetivo de reter, em ordem de estimação, o máximo de informação em relação à variação total contida nos dados originais.

O coeficiente de cada variável original no componente principal representa a contribuição dessa variável para o componente principal; o valor desse coeficiente depende das outras variáveis que são incluídas na análise. Para diferentes conjuntos de variáveis, os coeficientes estimados serão diferentes. Os coeficientes dos componentes principais são também denominados de elementos de autovetores e a variância associada a cada componente principal é denominada de autovalor.

Cada componente principal (CP) retém uma percentagem da variância total original, e as variâncias retidas são decrescentes do primeiro ao último PC, ou seja, o primeiro componente possui a maior variância, o segundo possui a segunda maior e assim sucessivamente. A importância de um componente é avaliada por meio da percentagem de variância que o mesmo retém. Assim, o primeiro componente é o mais importante, pois retém a maior parte da variação total encontrada nos dados originais.

A independência ou ortogonalidade (ausência de correlação) entre os componentes é uma propriedade útil porque significa que os componentes estão avaliando diferentes dimensões dos dados. A consequência disso é que a soma das variâncias dos componentes principais (autovalores) corresponde à soma das variâncias das variáveis originais, não havendo, portanto, perda de informação na transformação das variáveis originais em componentes principais.

Na maioria das situações, o número de componentes principais necessários para extrair completamente a informação contida em um grupo de variáveis, é igual ao número de variáveis do grupo. Entretanto, os primeiros componentes poderão conter a maior parte da variação das variáveis originais analisadas. Assim, uma questão que surge é referente a quantos componentes

reter na análise. Um dos métodos para usados para determinar quantos componentes reter, consiste em reter componentes que expliquem uma percentagem relativamente alta da variação total, geralmente acima de 80%.

Outro método, baseado na ideia que se os caracteres originais são não correlacionados então o conjunto de componentes principais é o mesmo dos caracteres originais. No caso da matriz de correlação, todos os caracteres têm variância um. Portanto, qualquer componente principal cuja variância é menor que um não é selecionado, uma vez que é assumido que ele contém muito menos informação que qualquer um dos caracteres originais (Khattree e Naik, 2000). Jolliffe (1972) sugere que sejam considerados apenas os componentes associados a autovalores acima de 0,7.

Existe ainda um terceiro método que é gráfico e usa o chamado diagrama *scree*, que, como o primeiro método, é aplicável tanto para matrizes de correlação como de covariância. Um diagrama *scree* é um gráfico de autovalores  $\lambda_j$  contra  $j$ ,  $j=1,2,\dots,p$ . Com base neste gráfico, o número

( $j$ ) de componentes principais a serem retidos na análise é determinado de tal forma que a inclinação do gráfico seja um declive à esquerda de  $j$ , mas ao mesmo tempo não seja um declive à direita. A ideia deste gráfico é que o número de componentes principais a serem retidos é tal que as diferenças entre autovalores consecutivos vão se tornando cada vez menores (Khattree e Naik, 2000).

### Descrição estatística do método

Seja  $X_{ij}$  a média original e seja  $Z_{ij}$  a média padronizada da j-ésima variável ( $j= 1, 2, \dots, p$ ) avaliada

no i-ésimo indivíduo ( $i= 1, 2, \dots, n$ ), dada por  $Z_{ij} = \frac{X_{ij}}{\sigma_j}$ , em que:  $X_{ij}$  são os valores observados de

cada indivíduo  $i$  para uma dada variável (característica)  $j$ ;  $\sigma_j$  é o desvio padrão da j-ésima variável (característica). Tem-se que:

$$\hat{Cov}(X_j, X_{j'}) = \frac{\left[ \sum_{i=1}^n (X_{ij} X_{ij'}) - \left( \sum_{i=1}^n X_{ij} \sum_{i=1}^n X_{ij'} \right) / n \right]}{n-1} \text{ ou}$$

$$\hat{Cov}(X_j, X_{j'}) = \left[ \sigma^2(X_j + X_{j'}) - \sigma^2(X_j) - \sigma^2(X_{j'}) \right] / 2.$$

$$\text{Tem-se ainda que: } r(X_j, X_{j'}) = \hat{Cov}(Z_j, Z_{j'}) = \frac{\hat{Cov}(X_j, X_{j'})}{\sqrt{\sigma^2(X_j) \sigma^2(X_{j'})}}$$

ou de forma equivalente:  $r_{jj'} = \frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}_{j'})}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^n (X_{ij'} - \bar{X}_{j'})^2}}$ , em que:

$\hat{Cov}(X_j, X_{j'})$  e  $\hat{Cov}(Z_j, Z_{j'})$  são estimadores da covariância entre duas variáveis originais e duas variáveis padronizadas, respectivamente;  $r(X_j, X_{j'}) = r_{jj'}$  é a correlação entre duas variáveis originais (coeficiente de correlação de Pearson), sendo  $\bar{X}_j$  a média para a  $j$ -ésima variável (característica).

Em virtude de as variáveis originais ( $X_j$ ) possuírem diferentes unidades de medidas, utiliza-se as variáveis padronizadas ( $Z_j$ ) na estimação dos componentes principais. A análise de componentes principais consiste na transformação do conjunto de  $p$  variáveis  $Z_1, Z_2, \dots, Z_p$  em um novo conjunto (os componentes principais)  $Y_1, Y_2, \dots, Y_p$ . Em relação aos componentes principais, as seguintes propriedades são observadas:

a) Seja  $Y_j$  um componente principal, então,  $Y_j$  é uma combinação linear de  $Z_j$ , dada por

$$Y_j = a_1 Z_1 + a_2 Z_2 + \dots + a_p Z_p;$$

b) Seja  $Y_{j'}$  outro componente principal, dado por  $Y_{j'} = b_1 Z_1 + b_2 Z_2 + \dots + b_p Z_p$ . Os componentes principais são independentes entre si (não correlacionados ou ortogonais), ou seja,

$$\sum_{j=1}^p a_j b_j = 0 \text{ e } \sum_{j=1}^p a_j^2 = \sum_{j=1}^p b_j^2 = 1;$$

c) Entre todos os componentes estimados,  $Y_1$  apresenta a maior variância,  $Y_2$  a segunda maior variância e assim sucessivamente, ou seja,  $\sigma^2(Y_1) > \sigma^2(Y_2) > \dots > \sigma^2(Y_p)$ ;

d) Tem-se que  $\sum_{j=1}^p \sigma^2(Y_j) = \text{Traço}(R) = p$ , em que, traço ( $R$ ) é a soma dos elementos da

diagonal ou a soma dos autovalores de  $R$ , sendo  $R$  a matriz de covariâncias entre as variáveis padronizadas ou a matriz de correlação entre as variáveis originais.

Com base nas propriedades anteriores dos componentes, avalia-se a dispersão dos indivíduos estudados em sistemas de eixos cartesianos, de forma que o aproveitamento da variabilidade disponível seja maximizado. O problema estatístico consiste, fundamentalmente, em estimar os coeficientes de ponderação das variáveis em cada componente, de modo que a variância a eles associada seja maximizada, observando-se as restrições descritas no item b das propriedades.

O primeiro componente principal pode ser definido como uma combinação linear das variáveis analisadas cujos coeficientes  $a_j$  ( $j = 1, 2, \dots, p$ ) são os elementos do autovetor (vetor característico)

associado ( $\alpha_j$ ) com o maior autovalor (raiz característica)  $\lambda_1$  da matriz de covariância (ou correlação) das variáveis. Se  $\alpha'_1\alpha_1=1$ , então  $\lambda_1$  é interpretado como a variância amostral do primeiro componente principal. Assim, a variância associada a cada componente é estimada pelos autovalores da matriz R e os coeficientes de ponderação das variáveis originais, pelos elementos dos autovetores correspondentes.

As estimativas dos autovalores,  $\lambda_j$  (variância associada a cada componente principal) e dos autovetores,  $\alpha_j$  (conjunto de coeficientes de cada componente principal) são obtidas pela solução dos seguintes sistemas, respectivamente:  $\det(R-\lambda I)=0$  e  $(R-\lambda_j I)\alpha_j=\Phi$ , em que:

$\lambda_j$ : autovalores (ou raízes características) da matriz de correlações entre as variáveis originais (ou de covariância entre as variáveis padronizadas); existem p autovalores correspondentes às variâncias de cada um dos p componentes principais;

$\alpha_j$ : autovetor (ou vetor característico), de dimensão ( $p \times 1$ ), que representa o conjunto de transformações ortogonais pelas quais as variáveis padronizadas devem ser multiplicadas para produzir os componentes principais;

I: matriz identidade, de dimensão ( $p \times p$ );

R: matriz de correlações entre pares de variáveis originais de dimensão ( $p \times p$ );

$\Phi$ : vetor nulo, de dimensão ( $p \times 1$ ).

A solução do sistema  $(R-\lambda_j I)\alpha_j=\Phi$  deve ser tal que  $\alpha_j \neq \Phi$ . Para tanto, impõe-se a condição  $|R-\lambda I|=0$ , de modo que a solução possa ser escolhida entre aquelas que satisfazem a restrição  $\alpha'_1\alpha_1=1$ , que é necessária para que exista uma solução única.

A importância relativa de um componente principal é avaliada pela percentagem da variância total que ele explica. A soma dos primeiros k autovalores dividida pela soma de todos os autovalores, ou seja,  $(\lambda_1+\lambda_2+\dots+\lambda_k)/(\lambda_1+\lambda_2+\dots+\lambda_p)$ , que representa a proporção da variância total explicada pelos primeiros k componentes principais, ou seja, a proporção da informação retida na redução de p para k dimensões.

### Exemplo de aplicação

Serão utilizados dados obtidos por Almeida (1980) em um estudo de resistência horizontal do cafeiro ‘Catimor’ à ferrugem. Foram avaliados quatro componentes de resistência: período de incubação (PI), período latente médio (PLM), severidade (SEV) e número de lesões com esporos por folha (NPF). Foram estudadas 12 progénies de cafeiro (Quadro 8.1).

Quadro 8.1. Componentes de resistência à ferrugem em progénies de cafeiro ‘Catimor’ (Almeida, 1980)

Progêneres	PI (dias)	PLM (dias)	NPF	SEV (%)
UFV 2144	23,15	32,00	9,10	11,96
UFV 1340	25,50	33,50	4,09	4,98
UFV 2861	29,37	38,04	2,34	2,22
UFV 2862	30,37	37,46	2,14	2,32
UFV 2863	30,12	41,76	2,48	2,18
UFV 1307	28,62	36,87	1,50	1,74
UFV 3684	26,62	37,12	3,66	3,16
UFV 3686	27,37	37,52	1,83	1,85
UFV 3687	30,75	38,92	2,33	1,67
UFV3658	35,12	47,30	1,75	2,00
UFV 4303	35,20	50,22	1,25	1,25
UFV 4305	36,40	48,42	2,12	2,25

PI: período de incubação; PLM: período latente médio; NPF: número de lesões com esporos por folha; SEV: severidade.

O procedimento para estimativa (extração) dos componentes principais envolve os passos seguintes:

$$1) \text{ Padronização das variáveis por meio de } Z_{ij} = \frac{X_{ij}}{\sigma_j}$$

$$Z = \begin{bmatrix} 5,682 & 5,491 & 4,286 & 4,112 \\ 6,259 & 5,748 & 1,926 & 1,712 \\ 7,209 & 6,527 & 1,102 & 0,763 \\ 7,454 & 6,428 & 1,008 & 0,798 \\ 7,393 & 7,165 & 1,168 & 0,749 \\ 7,025 & 6,326 & 0,707 & 0,598 \\ 6,534 & 6,369 & 1,724 & 1,086 \\ 6,718 & 6,438 & 0,862 & 0,636 \\ 7,548 & 6,678 & 1,097 & 0,918 \\ 8,620 & 8,116 & 0,824 & 0,688 \\ 8,640 & 8,617 & 0,589 & 0,430 \\ 8,935 & 8,308 & 0,999 & 0,774 \end{bmatrix}$$

2) Estimação da matriz de correlação (R)

$$R = \begin{bmatrix} 1,0000 & 0,9519 & -0,6888 & -0,6512 \\ & 1,0000 & -0,5983 & -0,5758 \\ & & 1,0000 & 0,9890 \\ & & & 1,0000 \end{bmatrix}$$

3) Estimação dos autovalores de R, por meio de  $\det(R - \lambda I) = 0$ , obtendo-se as estimativas  $\lambda_1 = 3,2289$ ,  $\lambda_2 = 0,7190$ ,  $\lambda_3 = 0,0443$  e  $\lambda_4 = 0,0079$ ;

4) Estimação dos autovetores, por meio de  $(R - \lambda_j I)\alpha_j = \Phi$ . Para a obtenção do autovetor do primeiro componente principal utiliza-se a seguinte equação:  $(R - \lambda_1 I)\alpha_1 = \Phi$  ou  $(R - 3,2289 I)\alpha_1 = \Phi$ , sendo  $\alpha'_1 = [a_1 \ a_2 \ a_3 \ a_4]$ . Assim, obtém-se o seguinte sistema de equações:

$$\begin{aligned} -2,2289a_1 + 0,9519a_2 - 0,6888a_3 - 0,6512a_4 &= 0 \\ 0,9519a_1 - 2,2289a_2 - 0,5983a_3 - 0,5758a_4 &= 0 \\ -0,6888a_1 - 0,5983a_2 - 2,2289a_3 + 0,9890a_4 &= 0 \\ -0,6512a_1 - 0,5758a_2 + 0,9890a_3 - 2,2289a_4 &= 0 \end{aligned}$$

Este sistema é indeterminado. Para a obtenção de uma solução pode-se tomar, por exemplo,  $a_4 = 1$  e considerar somente as três primeiras equações. Assim, o vetor solução deste sistema é :

$\alpha'_1 = [-1,0173 \ -0,9635 \ 1,0167 \ 1,0000]$ . Deve-se considerar a restrição imposta  $\alpha'_1 \alpha_1 = 1$ . Então, o vetor  $\alpha'_1$  normalizado corresponde aos coeficientes do primeiro componente principal, ou seja:

Norma de  $\alpha'_1 = \|\alpha'_1\| = +(\alpha'_1 \alpha_1)^{1/2} = 1,9992$ ; então  $\alpha'_1 norm = (1/1,9992)\alpha'_1$  e

$\alpha'_1 norm = [-0,5092 \ -0,4829 \ 0,5084 \ 0,4991]$ . Assim, fica estimado o primeiro componente principal:

$$Y_1 = -0,5092Z_1 - 0,4829Z_2 + 0,5084Z_3 + 0,4991Z_4.$$

Se, na solução do sistema de equação anterior, fosse adotado  $a_1 = 1$  (ao invés de  $a_4 = 1$ ), o primeiro componente principal seria

$$Y_1 = 0,5092Z_1 + 0,4829Z_2 - 0,5084Z_3 - 0,4991Z_4.$$

Isto não alteraria os resultados da análise. Apenas haveria, no aspecto geométrico, uma mudança de sentido no eixo representado pelo primeiro componente principal, ou seja, uma rotação na configuração espacial dos indivíduos em relação aos componentes principais, sem alteração da mesma. Os autovetores dos demais componentes são obtidos de modo análogo.

Quadro 8.2. Estimativas das variâncias (autovalores) associadas aos componentes principais e respectivos coeficientes (autovetores) de quatro componentes principais referentes à resistência de progênies de ‘Catimor’ à ferrugem

Autovalor (Variância)	Variância Acumulada (%)	Z <sub>1</sub>	Z <sub>2</sub>	Z <sub>3</sub>	Z <sub>4</sub>
3,2289	80,72	-0,5092	-0,4829	0,5084	0,4991
0,7190	98,70	0,4416	0,5633	0,4726	0,5141
0,0443	99,80	0,7099	-0,6530	-0,1331	0,2281
0,0079	100,00	-0,2045	0,1521	-0,7074	0,6592

Z<sub>1</sub>, Z<sub>2</sub>, Z<sub>3</sub> e Z<sub>4</sub> correspondem às variáveis PI, PLM, NPF e SEV, respectivamente, todas padronizadas com variância igual a unidade.

A análise de componentes principais permite identificar as variáveis de resistência que contribuem pouco para a divergência entre as progênies por serem redundantes, devido a serem altamente correlacionadas com outras. As variáveis de resistência NPF e PI poderiam ser descartadas, pois possuem os maiores coeficientes de ponderação absolutos associados aos componentes principais com variância inferior a 0,7 (Quadro 2).

5) Escolha dos componentes principais a serem considerados na dispersão gráfica. Serão utilizados os dois primeiros componentes principais, uma vez que os mesmos retêm 98,7% da variância total original (Quadro 2), que é um valor acima de 80%, como preconizado na literatura sobre análise de componentes principais.

Então, tomando-se como exemplo o indivíduo 1 (progénie ‘Catimor’ UFV 2144), obtém-se o escore deste indivíduo em relação ao primeiro componente principal da seguinte forma:

$Y_{11} = -0,5092 (5,682) - 0,4829 (5,491) + 0,5084 (4,286) + 0,4991 (4,112) = -1,31$ . O escore da progénie UFV 2144 em relação ao segundo componente principal é dado por:

$Y_{12} = 0,4416 (5,682) + 0,5633 (5,491) + 0,4726 (4,286) + 0,5141 (4,112) = 9,74$ . Os demais escores são apresentados no Quadro 8.3.

Quadro 8.3. Escores das progênies de ‘Catimor’ em relação aos dois primeiros componentes principais ( $Y_1$  e  $Y_2$ )

Progênies	Y <sub>1</sub>	Y <sub>2</sub>
UFV 2144	-1,31	9,74
UFV 1340	-4,13	7,79
UFV 2861	-5,88	7,77
UFV 2862	-5,99	7,80
UFV 2863	-6,26	8,24
UFV 1307	-5,97	7,31
UFV 3684	-4,98	7,85
UFV 3686	-5,77	7,33
UFV 3687	-6,05	8,09
UFV 3658	-7,54	9,12
UFV 4303	-8,05	9,17
UFV 4305	-7,67	9,50

6) Construção da dispersão gráfica das progênies em relação aos dois componentes principais.

Quanto maior a proximidade entre duas progênies maior é a similaridade entre elas. A composição dos grupos de progênies é feita visualmente.

Quadro 8.4. Valores médios das variáveis de resistência dos grupos de progênies de cafeeiro, obtidos com base na análise de componentes principais

Grupos de Progênies	PI	PLM	NPF	SEV
A UFV2144	23,2	32,0	9,1	12,0
B UFV3658 UFV4303 UFV4305	35,6	48,6	1,7	1,8
C UFV1340 UFV3684	26,1	35,3	3,9	4,1
D UFV1307 UFV2861 UFV2862 UFV2863 UFV3686 UFV3687	29,4	38,4	2,1	2,2

PI e PLM: período de incubação e período latente médio (dias); NPF: número de pústulas por folha; SEV: severidade (%).

### Orientações Gerais para Análise de Componentes Principais (ACP)

Metas da ACP:

1-Extrair as informações mais importantes de um conjunto de dados multivariados, ou seja, dados de  $n$  observações referentes a  $p$  variáveis correlacionadas.

2-Reduzir o tamanho do conjunto de dados multivariados deixando apenas as informações mais importantes.

3-Simplificar a descrição do conjunto de dados multivariados.

4-Analisar a estrutura das observações (indivíduos ou objetos) e das variáveis.

Para atingir estas metas, são computadas novas variáveis não observadas chamadas de componentes principais (CP) que são obtidas como combinações lineares das variáveis originais observadas.

Princípios da ACP:

A ACP procura obter uma projeção linear de dados de alta dimensionalidade em um subespaço de dimensionalidade mais baixa, tal que:

1-A variância retida é maximizada.

2-O erro de reconstrução espacial por quadrados mínimos é minimizado.

Passos para a ACP:

A ACP consiste em transformar linearmente uma matriz  $X_{n \times p}$  em uma matriz  $Y_{n \times r}$ , com  $r < p$ .

1 – Centralizar os dados (subtrair a média);

2 – Calcular a matriz de covariâncias  $S = \frac{1}{n-1} X'X$ , sendo:

$$S_{jj'} = \frac{1}{n-1} \sum_{i=1}^n X_{ij} X_{ij'}$$

$S_{jj}$  (diagonal): variância da variável  $X_j$ , com  $j=1,2,\dots,p$ ;

$S_{jj'}$  (fora da diagonal): covariância entre as variáveis  $X_j$  e  $X_{j'}$ ;

3 – Calcular os autovalores e autovetores (ortonormais) da matriz de covariâncias  $S$ ;

4 – Selecionar  $r$  autovetores que correspondam aos  $r$  maiores autovalores para serem a nova base (subespaço);

Autovetores:

Se  $A$  é uma matriz quadrada, então um vetor não nulo  $v$  é um autovetor de  $A$  se existe um escalar  $\lambda$  (autovalor) tal que  $Av = \lambda v$ . Por exemplo:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Se a matriz quadrada  $A$  for considerada como uma matriz de transformação linear, então a sua multiplicação pelo autovetor não muda sua direção.

Exemplo de ACP:

Considere uma matriz de dados de variáveis originais,  $X_{n \times p}$ , com  $n=11$  observações (indivíduos ou objetos) e  $p=2$  dimensões (variáveis originais):

$$1) \text{ Centralizar os dados } (x_{ij} = X_{ij} - \bar{X}_j)$$

A matriz de covariância de variáveis originais centradas é:  $S = \begin{bmatrix} 0,716 & 0,615 \\ 0,615 & 0,616 \end{bmatrix}$

2) Calcular os autovalores e autovetores da matriz de covariância:

$$\lambda_1 = 1,28; v_1 = \begin{bmatrix} -0,677 & -0,735 \end{bmatrix}'$$

$$\lambda_2 = 0,49; v_2 = \begin{bmatrix} -0,735 & 0,677 \end{bmatrix}'$$

Observe que  $v_1$  e  $v_2$  são ortonormais, ou seja:

$$\|v_1\| = \|v_2\| = 1 \text{ e } v_1 \cdot v_2 = 0;$$

3) Projetar os dados na nova base (subespaço):

Seja  $V = [v_1 \ v_2 \ \cdots \ v_r]$  uma matriz  $p \times r$  onde as colunas  $v_j$  são os autovetores que correspondem aos  $r$  maiores autovalores.

A matriz de dados projetados,  $Y = XV$ , é uma matriz  $n \times r$ . Se  $r=p=posto(X)$

toda a variação das variáveis originais  $X$  é utilizada na análise, o que não é recomendado.

Tem-se que:

$$\lambda_1 = 1,28; v_1 = \begin{bmatrix} -0,677 & -0,735 \end{bmatrix}'$$

$$\lambda_2 = 0,49; v_2 = \begin{bmatrix} -0,735 & 0,677 \end{bmatrix}'$$

Então tem-se:

$$V = \begin{bmatrix} -0,677 & -0,735 \\ -0,735 & 0,677 \end{bmatrix}$$

Os elementos do autovetor associado ao mais alto autovalor são os coeficientes do primeiro componente principal ( $Y_1$ ) e os elementos do autovetor associado ao segundo maior autovalor são os coeficientes do segundo componente principal ( $Y_2$ ).

Se for desejado utilizar apenas uma dimensão, o  $Y_1 = -0,677X_1 - 0,735X_2$  é a melhor direção, ou seja, retém a máxima variância.

Propriedades Importantes da ACP:

1 – A matriz de covariância é sempre simétrica:

$$S' = \left( \frac{1}{n-1} X' X \right)' = \frac{1}{n-1} X' (X')' = S$$

2 – Os componentes principais de  $X$  são ortonormais:

$$\nu'_j \nu_{j'} = \begin{cases} 1, \text{ se } j=j' \\ 0, \text{ se } j \neq j' \end{cases}$$

$$V = \begin{bmatrix} \nu_1 & \nu_2 & \cdots & \nu_r \end{bmatrix}, \text{ então } V' = V^{-1} \text{ e}$$

$$V'V = I$$

3 – Se a matriz quadrada  $S_{p \times p}$  é uma matriz real e simétrica ( $S = S'$ ), então:

$S = V \Lambda V'$ , onde  $V = \begin{bmatrix} \nu_1 & \nu_2 & \cdots & \nu_p \end{bmatrix}$  são os autovetores e  $\Lambda = \text{diagonal}(\lambda_1 \ \lambda_2 \ \cdots \ \lambda_p)$  são os autovalores de  $S$ .

4 – Seja os dados projetados  $Y = XV$ . A matriz de covariância de  $Y$  é expressa como:

$$S_Y = \frac{1}{n-1} Y Y' = \frac{1}{n-1} V X' X V = V' S_X V$$

Uma vez que a matriz de covariância  $S_X$  é simétrica tem-se que:

$$S_Y = V' V \Lambda V' V$$

Como a matriz  $V$  é ortonormal obtém-se:

$$S_Y = V' V \Lambda V' V, \text{ então } S_Y = \Lambda$$

Observe que após a transformação/combinação linear a matriz de covariância torna-se diagonal. Assume-se que a melhor transformação é aquela que maximiza a variância dos dados projetados.

Decomposição em Valores Singulares (DVS):

Qualquer matriz  $\mathbf{X}$ ,  $n \times p$ , pode ser expressa singularmente como:

$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}'$ , em que:

$\mathbf{U}$ : matriz coluna-ortonormal,  $n \times r$ , sendo  $r$  o posto da matriz  $\mathbf{X}$ ;

$\Sigma$ : matriz diagonal,  $r \times r$ , onde os valores singulares  $\sigma_j$  são arranjados em ordem decrescente;

$\mathbf{V}$ : matriz coluna-ortonormal,  $p \times r$ .

Teorema da Relação entre ACP e DVS:

Seja  $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}'$  a expressão da DVS de uma matriz  $\mathbf{X}$ ,  $n \times p$ , e  $\mathbf{S} = \frac{1}{n-1}\mathbf{X}\mathbf{X}'$  a expressão da matriz de covariância. Os autovetores de  $\mathbf{S}$  são os mesmos que os vetores singulares direitos de  $\mathbf{X}$ .

Prova:

$\mathbf{X}\mathbf{X}' = \mathbf{V}\Sigma\mathbf{U}'\mathbf{U}\Sigma\mathbf{V}' = \mathbf{V}\Sigma\Sigma\mathbf{V}' = \mathbf{V}\Sigma^2\mathbf{V}'$ , então  $\mathbf{S} = \mathbf{V}\frac{\Sigma^2}{n-1}\mathbf{V}'$ . Mas,  $\mathbf{S}$  é simétrica, logo

$\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}'$ . Portanto, os autovetores da matriz de covariância  $\mathbf{S}$  são os mesmos que os vetores singulares direitos da matriz  $\mathbf{V}$  da DVS e os autovalores de  $\mathbf{S}$  podem ser calculados a partir dos

valores singulares, como  $\lambda_j = \frac{\sigma_j^2}{n-1}$ .

Observe então que a decomposição em valores singulares e a decomposição em autovalores da ACP estão estreitamente relacionadas, da seguinte forma:

- Os vetores singulares direitos de  $\mathbf{X}$  são autovetores de  $\mathbf{X}\mathbf{X}'$ ;
- Os vetores singulares esquerdos de  $\mathbf{X}$  são autovetores de  $\mathbf{X}'\mathbf{X}$ ;
- Os valores singulares não nulos de  $\mathbf{X}$ , encontrados sobre a diagonal de  $\Sigma$ , são as raízes quadradas dos autovalores não nulos tanto de  $\mathbf{X}\mathbf{X}'$  quanto de  $\mathbf{X}'\mathbf{X}$ .

# CAPITULO 22

## Análise de Agrupamento

Análise de agrupamento é um conjunto de procedimentos numéricos simples que classifica indivíduos, objetos, processos, métodos ou características em grupos distintos. Classifica observações (linhas) ou variáveis (colunas) de uma matriz de dados, utilizando algoritmos que se baseiam em distâncias (dissimilaridades) ou parecença (similaridade) entre os casos (indivíduos, objetos, métodos) ou entre as variáveis/características.

### Principais medidas de dissimilaridade e de similaridade

1) **Distância Euclidiana** – é a medida de dissimilaridade mais utilizada para dados quantitativos, sendo expressa por:

$$d_{ii'} = \left[ \sum_{j=1}^p (Y_{ij} - Y_{i'j})^2 \right]^{1/2}, \text{ em que:}$$

$Y_{ij}$ : valor da variável (característica) de ordem  $j$  na unidade avaliativa  $i$ ;

$Y_{i'j}$ : valor da variável (característica) de ordem  $j$  na unidade avaliativa  $i'$ ;

$p$ : número de variáveis (características);

2) **Distância Euclidiana Média** – utilizada para contornar a influência do número de variáveis (características) na análise; é expressa por:

$$d_{ii'} = \left[ \frac{1}{p} \sum_{j=1}^p (Y_{ij} - Y_{i'j})^2 \right]^{1/2}, \text{ em que:}$$

$Y_{ij}$ : valor da variável de ordem  $j$  na unidade avaliativa  $i$ ;

$Y_{i'j}$ : valor da variável de ordem  $j$  na unidade avaliativa  $i'$ ;

$p$ : número de variáveis;

3) **Quadrado da Distância Euclidiana Média**

$$d_{ii'}^2 = \frac{1}{p} \sum_{j=1}^p (Y_{ij} - Y_{i'j})^2$$

$Y_{ij}$ : valor da variável de ordem  $j$  na unidade avaliativa  $i$ ;

$Y_{i'j}$ : valor da variável de ordem  $j$  na unidade avaliativa  $i'$ ;

$p$ : número de variáveis;

4) **Distância Euclidiana Média Padronizada** – utilizada para contornar o problema do número de caracteres e de escala de medição das características; é expressa por:

$$d_{ii'} = \left[ \frac{1}{p} \sum_{j=1}^p (y_{ij} - y_{i'j})^2 \right]^{1/2}, \text{ em que:}$$

$$y_{ij} = \frac{Y_{ij}}{s(Y_j)}, \text{ sendo } s(Y_j) \text{ o desvio padrão da variável/ característica de ordem } j;$$

5) **Distância Generalizada** – a distância generalizada ou ponderada entre dois elementos (indivíduos ou objetos)  $Y_i$  e  $Y_{i'}$  é expressa por:

$$d_{ii'} = d(Y_i, Y_{i'}) = \left[ (Y_i - Y_{i'})' A (Y_i - Y_{i'}) \right]^{1/2}$$

, em que  $A_{p \times p}$  é uma matriz de ponderação positiva definida.

Quando a matriz  $A_{p \times p}$  é a matriz identidade, a distância generalizada é a distância Euclidiana. Se

$A_{p \times p}$  é igual a  $S_{p \times p}^{-1}$ , tem-se a distância generalizada de Mahalanobis. Quando

$A_{p \times p} = diag\left(\frac{1}{p}\right)$  tem-se a distância Euclidiana média.

Uma matriz do tipo  $A_{p \times p} = diag\left[\left(S_j^2\right)^{-1}\right]$ , onde  $S_j^2$  é a variância da variável aleatória  $j$ ,

$j=1,2,\dots,p$ , considera na ponderação das distâncias apenas as diferenças de variâncias entre as

variáveis. Já uma matriz do tipo  $A_{p \times p} = S_{p \times p}^{-1}$  considera na ponderação as diferenças de variâncias e as relações lineares entre as variáveis, medidas pelas covariâncias.

**6) Distância Generalizada de Mahalanobis** – usada apenas para dados quantitativos, sendo expressa por:

$$D^2 = d' S^{-1} d , \text{ em que:}$$

$d$  : vetor das diferenças entre as médias das unidades de observação, para todas as características (variáveis), ou seja,

$$d' = [d_1 \ d_2 \ \cdots \ d_p] \text{ e } d_j = \bar{Y}_{ij} - \bar{Y}_{i'j}, \text{ sendo } \bar{Y}_{ij} = \frac{Y_{ij}}{r};$$

$S^{-1}$ : inversa da matriz de variâncias – covariâncias residual  $p \times p$  entre variáveis (características);  $r$ : número de repetições.

**7) Distância de Minkowsky** – a distância entre dois elementos (indivíduos ou objetos)  $Y_i$  e  $Y_{i'}$  é definida como:

$$d_{ii'} = d(Y_i, Y_{i'}) = \left[ \sum_{j=1}^p w_j |Y_{ij} - Y_{i'j}|^\lambda \right]^{1/\lambda}, \text{ em que os } w_j \text{ são os pesos de ponderação}$$

para as variáveis/características.

Para  $\lambda=1$  esta distância é chamada de *city-block* ou *Manhattan* e para  $\lambda=2$  tem-se a distância Euclidiana. A distância de Minkowsky é menos afetada pela presença de valores discrepantes na amostra do que a distância Euclidiana.

As distâncias entre os elementos (indivíduos, objetos, métodos) são armazenadas numa matriz de dimensão  $n \times n$ , denominada de matriz de distâncias ou matriz de dissimilaridade, como mostrado a seguir, para uma matriz 4x4:

$$D_{4 \times 4} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{21} & 0 & d_{23} & d_{24} \\ d_{31} & d_{32} & 0 & d_{34} \\ d_{41} & d_{42} & d_{43} & 0 \end{bmatrix}$$

**8) Coeficiente de Jaccard** – é utilizada para variáveis qualitativas, sendo expresso por:

$$S = \frac{a}{a+b+c}, \text{ em que:}$$

*a, b* e *c*: valores assumidos pelas variáveis qualitativas.

9) **Coeficiente de Similaridade** – o mais utilizado é o coeficiente de correlação de Pearson, que é expresso por:

$$r_{ii'} = \frac{\sum_{j=1}^p (Y_{ij} - \bar{Y}_j)(Y_{i'j} - \bar{Y}_j)}{\left[ \sum_{j=1}^p (Y_{ij} - \bar{Y}_j)^2 \right] \left[ \sum_{i=1}^p (Y_{i'j} - \bar{Y}_j)^2 \right]}.$$

#### Exemplos de cálculos de medidas de distância

Considere os dados apresentados a seguir, que representam os valores médios de dois caracteres (SEM-número de sementes por vagem e MCS-massa de cem sementes) avaliados em sete genótipos de feijoeiro (A, B, C, D, E, F).

Genótipos	SEM	MCS
A	4,9	23,5
B	5,7	21,3
C	5,5	19,8
D	4,9	23,5
E	4,9	21,2
F	4,2	24,7
G	5,6	18,3
Média $(\bar{Y}_j)$	5,1	21,8
Desvio Padrão $(s_j)$	0,4928	2,1057

A distância Euclidiana entre os genótipos A e B é igual a:

$$d_{AB} = d(Y_A, Y_B) = \left[ \sum_{j=1}^2 (Y_{Aj} - Y_{Bj})^2 \right]^{1/2}$$

$$d_{AB} = \left[ (4,9 - 5,7)^2 + (23,5 - 21,3)^2 \right]^{1/2} = 2,34$$

A distância Euclidiana média entre os genótipos A e B é:

$$d_{AB} = d(Y_A, Y_B) = \left[ \frac{1}{2} \sum_{j=1}^2 (Y_{Aj} - Y_{Bj})^2 \right]^{1/2}$$

$$d_{AB} = \left\{ \frac{1}{2} \left[ (4,9 - 5,7)^2 + (23,5 - 21,3)^2 \right] \right\}^{1/2} = 1,65$$

A distância generalizada de Mahalanobis entre os genótipos A e B, considerando a matriz de ponderação  $A_{2 \times 2}$  igual à matriz inversa da matriz de variâncias e covariâncias amostrais, é:

$$d_{AB} = D^2 = d'S^{-1}d, \text{ sendo } d' = [d_1 \quad d_2] \text{ e } d_j = Y_{ij} - Y_{i'j}, j=1,2.$$

Então,  $d_{AB} = \left[ (-0,8 \quad 2,2) S^{-1} \begin{pmatrix} -0,8 \\ 2,2 \end{pmatrix} \right]^{1/2}$ , onde

$$S^{-1} = \begin{bmatrix} 0,2428 & -1,7200 \\ -1,7200 & 4,4339 \end{bmatrix}^{-1} = \begin{bmatrix} -2,3561 & -0,9139 \\ -0,9139 & -0,1290 \end{bmatrix}$$

Logo,  $d_{AB} = \left[ (-0,8 \quad 2,2) \begin{bmatrix} -2,3561 & -0,9139 \\ -0,9139 & -0,1290 \end{bmatrix} \begin{pmatrix} -0,8 \\ 2,2 \end{pmatrix} \right]^{1/2}$

e  $d_{AB} = D^2 = 1,0414$ .

Agora utilizando a matriz  $A$  como

$$A_{2 \times 2} = \begin{bmatrix} 0,2428 & 0 \\ 0 & 4,4339 \end{bmatrix}^{-1} = \begin{bmatrix} 4,1186 & 0 \\ 0 & 0,2253 \end{bmatrix} \text{ tem-se:}$$

$$d_{AB} = \left[ (-0,8^2)(4,1186) + (2,2^2)(0,2253) \right]^{1/2} = 1,9304$$

Para os mesmos dois genótipos, a distância de Minkowsky é expressa por:

$$d_{AB} = \left[ \sum_{j=1}^2 w_j |Y_{Aj} - Y_{Bj}|^\lambda \right]^{1/\lambda}$$

Para  $w_1 = w_2 = 1$  e  $\lambda = 1$  tem-se:

$$d_{AB} = |4,9 - 5,7| + |23,5 - 21,3| = 3,0$$

Os outros pares de distâncias envolvendo todos os outros genótipos, para todos os tipos de distâncias, são obtidos de forma análoga à descrita anteriormente.

Observações:

- a) Para todas as medidas é exigido que as variáveis tenham distribuição normal multivariada e uma relação linear entre elas;
- b) Para o uso de  $D^2$  pressupõe-se a existência de distribuição multinormal p-dimensional e a homogeneidade da matriz de covariância residual das unidades amostrais;
- c) Uma vez obtida a matriz de similaridade ou de dissimilaridade, os grupos distintos serão formados mediante o uso de um algoritmo de agrupamento e representados graficamente.

### Principais métodos de agrupamento

**1) Método da Ligação Simples ou do Vizinho Mais Próximo ou da Distância Mínima** – os novos indivíduos que se agregarão aos grupos já existentes o farão a um nível de dissimilaridade igual à distância entre estes e o indivíduo do grupo mais próximo deles; neste caso ocorre uma contração do gráfico.

Desta forma, a distância entre um indivíduo  $k$  e um grupo, formado pelos indivíduos  $i$  e  $j$ , é expressa por:

$d_{(ij)k} = \min\{d_{ik}; d_{jk}\}$ , ou seja,  $d_{(ij)k}$  é dada pelo menor elemento do conjunto das distâncias entre dois pares de indivíduos ( $i$  e  $k$ ) e ( $j$  e  $k$ ).

A distância entre dois grupos é expressa por:

$d_{(ij)(kl)} = \min\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$ , ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos ( $i$  e  $j$ ) e ( $k$  e  $l$ ) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos ( $i$  e  $k$ ), ( $i$  e  $l$ ), ( $j$  e  $k$ ) e ( $j$  e  $l$ ).

**2) Método da Ligação Completa ou do Vizinho Mais Distante ou da Distância Máxima** – os novos indivíduos se agregam aos grupos já formados a um nível de dissimilaridade igual à distância entre eles e o indivíduo do grupo mais afastado deles; neste caso ocorre uma dilatação do gráfico.

Assim, a distância entre um indivíduo  $k$  e um grupo, formado pelos indivíduos  $i$  e  $j$ , é expressa por:

$d_{(ij)k} = \max\{d_{ik}; d_{jk}\}$ , ou seja,  $d_{(ij)k}$  é dada pelo maior elemento do conjunto das distâncias dos pares de indivíduos ( $i$  e  $k$ ) e ( $j$  e  $k$ ).

A distância entre dois grupos é expressa por:

$d_{(ij)(kl)} = \max\{d_{ik}; d_{il}; d_{jk}; d_{jl}\}$ , ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos ( $i$  e  $j$ ) e ( $k$  e  $l$ ) é dada pelo maior elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos ( $i$  e  $k$ ), ( $i$  e  $l$ ), ( $j$  e  $k$ ) e ( $j$  e  $l$ ).

**3) Método da Ligação Média Entre Grupos ou UPGMA** – inicialmente são identificados os pares de indivíduos com menores distâncias; em seguida são calculados valores médios destes pares originando um novo elemento singular.

Neste caso, para os cálculos dos valores médios atribui-se sempre o mesmo peso aos dois elementos que estão sendo integrados. Desta forma, cada um dos novos indivíduos (ou objetos) que se incorporarão aos grupos já existentes o fará a um nível de dissimilaridade igual à média das distâncias dos mesmos em relação a todos os membros do grupo.

Então, a distância entre um indivíduo  $k$  e um grupo, formado pelos indivíduos  $i$  e  $j$ , é expressa por:

$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}$ , ou seja,  $d_{(ij)k}$  é dada pela média do conjunto das distâncias entre os pares de indivíduos ( $i$  e  $k$ ) e ( $j$  e  $k$ ).

A distância entre dois grupos é expressa por:

$d_{(ij)(kl)} = \frac{d_{ik} + d_{il} + d_{jk} + d_{jl}}{4}$ , ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos ( $i$  e  $j$ ) e ( $k$  e  $l$ ) é dada pela média do conjunto cujos elementos são as distâncias entre os pares de indivíduos ( $i$  e  $k$ ), ( $i$  e  $l$ ), ( $j$  e  $k$ ) e ( $j$  e  $l$ ).

**4) Método de Ward** – para a formação inicial do grupo consideram-se os indivíduos que proporcionam a menor soma de quadrados dos desvios dentro do grupo. A soma de quadrados dos desvios dentro é calculada considerando apenas os indivíduos (ou objetos) dentro do grupo em

formação; a soma de quadrados dos desvios total é calculada considerando todos os indivíduos disponíveis para a análise de agrupamento. A perda de informação devida ao agrupamento pode ser quantificada pela razão entre a soma de quadrados dos desvios dentro do grupo e a soma de quadrados dos desvios total.

O agrupamento pode ser feito a partir das somas de quadrados dos desvios entre indivíduos  $(SQD_{ii'})$  ou a partir do quadrado da distância euclidiana  $(d_{ii'}^2)$ , sendo:

$$SQD_{ii'} = \sum_{j=1}^p SQD_{j(ii')} \text{ e } d_{ii'}^2 = \sum_{j=1}^p (Y_{ij} - Y_{i'j})^2, \text{ em que:}$$

$SQD_{j(ii')}$ : soma de quadrados dos desvios entre indivíduos para a variável/característica de ordem  $j$ ;

$Y_{ij}$ : valor da característica  $j$  para o indivíduo (objeto ou método)  $i$ ;

$p$ : número de variáveis/características avaliadas.

A relação  $SQD_{ii'} = \frac{1}{2} d_{ii'}^2$  é verificada.

A soma de quadrados dos desvios total é dada por:  $SQDTotal = \frac{1}{n} \sum_{i < i'}^n d_{ii'}^2$ , sendo  $n$  o número de indivíduos.

Na aplicação deste método, identifica-se na matriz  $D$ , que tem como elementos os valores de  $d_{ii'}^2$ , ou na matriz  $S$ , que tem como elementos os valores de  $SQD_{ii'}$ , o par de indivíduos que proporciona a menor soma de quadrados dos desvios. Com estes indivíduos agrupados, uma nova matriz de dissimilaridade, de menor dimensão, é recalculada, considerando que:

$SQD_{(ijk)} = \frac{1}{k} d_{(ijk)}^2$ , sendo  $k$  o número de indivíduos no grupo a ser formado, neste caso igual a 3, e  $d_{(ijk)}^2 = d_{(ij)}^2 + d_{(ik)}^2 + d_{(jk)}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2$ . Em seguida considera que:

$SQD_{(ijkm)} = \frac{1}{k} d_{(ijkm)}^2$ , sendo  $k$  o número de indivíduos no grupo a ser formado, que neste caso é igual a 4, e  $d_{(ijkm)}^2 = d_{ij}^2 + d_{ik}^2 + d_{jk}^2 + d_{im}^2 + d_{jm}^2 + d_{km}^2$ , e o processo continua

até os  $n-1$  passos de agrupamento, que agrupa todos os indivíduos e forma o diagrama em árvore (ou dendrograma).

**5) Método de Tocher** – são identificados os dois indivíduos mais próximos, ou seja, aqueles que apresentam a menor estimativa de distância, para formar o primeiro grupo; a seguir avalia-se a

possibilidade de inclusão de novos indivíduos no grupo, seguindo-se o critério de que a média das distâncias intragrupo seja menor que as distâncias intergrupo.

Como a entrada de um indivíduo em um grupo sempre aumentará o valor médio da distância intragrupo, toma-se a decisão de incluí-lo neste grupo comparando-se o acréscimo no valor da distância média do grupo e um nível máximo permitido ( $\theta$ ), o qual pode ser estabelecido adotando-se a maior distância do conjunto de menores distâncias envolvendo cada indivíduo.

Quando a adição de um indivíduo acarreta aumento na média da distância intragrupo maior que a permitida, este indivíduo não é adicionado ao grupo. Similarmente, um segundo grupo é formado e o processo continua até que todos os indivíduos sejam incluídos em um ou outro grupo, que são independentes entre si. Desta forma, a decisão de inclusão ou não do indivíduo  $k$  num grupo é feita com base no critério seguinte:

Se  $\frac{d_{(grupo)k}}{n} \leq \theta$ , inclui-se o indivíduo  $k$  no grupo;

Se  $\frac{d_{(grupo)k}}{n} > \theta$ , não se inclui o indivíduo  $k$  no grupo, sendo  $n$  o número de indivíduos do grupo original. Neste caso, a distância entre o indivíduo  $k$  e o grupo formado pelos indivíduos  $i$  e  $j$  é dada por:  $d_{(ij)k} = d_{ik} + d_{jk}$ .

#### Observações:

1) A escolha do método de agrupamento depende do espectro de variação da matriz de distâncias – se for estreito recomenda-se usar a distância máxima; se for amplo pode-se utilizar a distância mínima ou média;

2) Os métodos ligação simples, ligação completa, UPGMA e de Ward são denominados de métodos hierárquicos porque os indivíduos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o diagrama de árvore ou dendrograma. Neste caso, o interesse maior está nas ramificações da árvore e não no número de grupos. As delimitações dos grupos podem ser estabelecidas por um exame visual do dendrograma, procurando detectar pontos de alta mudança de nível;

3) O método ligação simples tem como desvantagem a incapacidade de não discernir grupos pobramente separados. Entretanto, apresenta a vantagem de delinear grupos não-elipsóides, ou seja, evita-se estabelecer grupos únicos quando os indivíduos se dispõem numa estrutura de filamentos conhecida como encadeamento, uma vez que indivíduos em extremidades opostas da cadeia podem ser completamente dissimilares;

4) O método UPGMA é utilizado com maior frequência em ecologia, sistemática e taxonomia numérica. Uma vez que utiliza médias aritméticas das medidas de dissimilaridade, evita caracterizar as dissimilaridades por valores extremos (máximo ou mínimo) entre os indivíduos considerados;

5) O método de Tocher é um método de agrupamento por otimização; neste método os grupos são formados pela adequação de algum critério de agrupamento e o objetivo é alcançar uma partição dos indivíduos que otimize (maximize ou minimize) alguma medida predefinida.

### Exemplos de aplicações de métodos de agrupamento

Considere os resultados básicos das análises de variâncias das avaliações de oito parentais em relação a quatro caracteres ( $X_1, X_2, X_3, X_4$ ), num experimento em blocos casualizados (Extraído de Cruz et al., 2004):

Quadro 1. Resumo das análises de variâncias de quatro caracteres e seus Produtos Médios (PM)

Caractere	QM ou PM Blocos (3) <sup>1</sup>	QM ou PM Genótipos (7)	QM ou PM Resíduo (21)	F
X1	46,6992	163,9392	18,7459	8,74**
X2	0,2018	1,1408	0,1879	6,07**
X3	0,0045	0,3448	0,0096	35,92**
X4	123,4010	892,8013	72,1533	12,37**
X1 e X2	-0,6209	0,4143	0,0151	-
X1 e X3	0,1429	-0,1370	-0,0146	-
X1 e X4	74,6738	291,7570	30,8056	-
X2 e X3	0,0151	-0,2929	-0,0036	-
X2 e X4	-1,5408	0,7343	0,8517	-
X3 e X4	0,2529	-2,6835	0,0871	-

\*\*Significativo a 1% de probabilidade pelo teste F.

1 – Valores entre parênteses correspondem aos graus de liberdade.

Quadro 2. Médias de oito parentais em relação a quatro caracteres

Parentais	$X_1$	$X_2$	$X_3$	$X_4$
1	41,900	20,300	3,900	85,675
2	43,800	19,750	3,650	98,250
3	37,300	18,725	4,600	74,575
4	40,150	20,300	4,300	91,625
5	32,500	20,250	4,100	54,125
6	52,750	19,725	4,375	100,375
7	43,900	20,225	4,275	91,000
8	49,250	20,025	4,150	82,175

Quadro 3. Médias e desvios padrões de quatro caracteres avaliados em oito parentais

Caractere	Média $(X_j)$	Desvio Padrão $[s(X_j)]$
X1	42,6937	6,4020
X2	19,9125	0,5340
X3	4,1687	0,2936
X4	84,7250	14,9361

Com base nos resultados do Quadro 2 obtem-se as médias padronizadas  $(x_1, x_2, x_3, x_4)$  obtidas por meio de  $x_j = X_j / s(X_j)$  para os oito parentais.

Quadro 4. Médias padronizadas de oito parentais em relação a quatro caracteres

Parentais	$x_1$	$x_2$	$x_3$	$x_4$
1	6,5448	38,0150	13,2834	5,7361
2	6,8416	36,9850	12,4319	6,5780
3	5,8263	35,0655	15,6675	4,9929
4	6,2715	38,0150	14,6458	6,1344
5	5,0765	37,9213	13,9646	3,6338
6	8,2396	36,9382	14,9012	6,7203
7	6,8572	37,8745	14,5606	6,0926
8	7,6929	37,5000	14,1349	5,5016

#### Distância Euclidiana

A distância Euclidiana média padronizada entre os parentais 1 e 2 tem a seguinte estimativa:

$$d_{12} = \left\{ \left[ (6,5448 - 6,8416)^2 + \dots + (5,7361 - 6,5780)^2 \right] / 4 \right\}^{1/2}$$

$$d_{12} = 0,8035$$

Esta e as demais distâncias entre pares de parentais são armazenadas na matriz de distâncias ( $D$ ) apresentada como no quadro a seguir.

Quadro 5. Distâncias Euclidianas médias padronizadas entre pares de oito parentais

Parentais	2	3	4	5	6	7	8
1	0,804	1,965	0,723	1,331	1,380	0,685	0,769
2		2,103	1,273	1,941	1,421	1,179	1,123
3			1,677	1,837	1,796	1,687	1,733
4				1,432	1,166	0,305	0,858
5					2,315	1,551	1,626
6						0,908	0,819
7							0,585

$d_{ii'}$ , máximo: 2,315 (entre os parentais 5 e 6);

$d_{ii'}$ , mínimo: 0,305 (entre os parentais 4 e 7);

Total do quadrado das distâncias  $\left( \sum_{i < i'} \sum d_{ii'}^2 \right)$ : 56,00

*Distância Generalizada de Mahalanobis*  $(D^2)$

Estimação de  $D^2$  a partir de médias originais:

$$D_{ii'}^2 = \delta' \psi^{-1} \delta, \text{ em que } \delta' = [d_1 \quad d_2 \quad \cdots \quad d_p]$$

Para o cálculo da distância de Mahalanobis entre os parentais 1 e 2 tem-se:

$$\psi = \begin{bmatrix} 18,7459 & 0,0151 & -0,0146 & 30,8056 \\ & 0,1879 & -0,0036 & 0,8517 \\ & & 0,0096 & 0,0871 \\ & & & 72,1533 \end{bmatrix}$$

$$\psi^{-1} = \begin{bmatrix} 0,2217 & 0,4724 & 1,4508 & -0,1020 \\ & 6,7022 & 5,8672 & -0,2879 \\ & & 116,3802 & -0,8304 \\ & & & 0,0618 \end{bmatrix}$$

$$\delta = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} = \begin{bmatrix} X_{11} - X_{21} \\ X_{12} - X_{22} \\ X_{13} - X_{23} \\ X_{14} - X_{24} \end{bmatrix} = \begin{bmatrix} -1,900 \\ 0,550 \\ 0,250 \\ -12,575 \end{bmatrix}$$

Então,  $D_{12}^2 = \delta' \psi^{-1} \delta = 23,455$ . De modo análogo obtém-se as demais estimativas de distâncias que são armazenadas numa matriz de distâncias ( $D$ ) como a seguir:

Quadro 6. Distância generalizada de Mahalanobis entre pares de parentais

País	2	3	4	5	6	7	8
1	23,46	62,94	17,63	29,74	34,52	15,49	29,61
2		125,03	56,59	100,43	91,21	63,59	97,48
3			26,20	47,69	22,00	23,68	44,03
4				37,44	17,51	3,16	35,26
5					40,02	25,97	15,66
6						7,66	13,74
7							17,66

$D^2$  máximo: 125,03 (parentais 2 e 3);

$D^2$  mínimo: 3,16 (parentais 4 e 7);

Total de  $D^2 \left( \sum_i \sum_{i' < i} D_{ii'}^2 \right) : 1125,39$

#### Método de Agrupamento Vizinho Mais Próximo

A distância entre um parental  $k$  e um grupo formado pelos parentais  $i$  e  $j$  é dada por:

$$d_{(ij)k} = \min \{ d_{ik}; d_{jk} \}$$

A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \min \{ d_{ik}; d_{il}; d_{jk}; d_{jl} \}$$

O método do vizinho mais próximo será aplicado para o agrupamento dos oito parentais com base nas medidas de distâncias  $D^2$  (matriz de dissimilaridade). São considerados os seguintes passos:

Passo 1: Entidades mais similares: parentais 4 e 7

Distância entre entidades: 3,16

Nova matriz de dissimilaridade:

	(2)	(3)	(4,7)	(5)	(6)	(8)
(1)	23,46	62,94	15,49	29,74	34,5	29,61
(2)		125,03	56,59	100,43	91,21	97,48
(3)			23,68	47,69	22,00	44,03
(4,7)				25,97	7,66	17,66
(5)					40,02	15,66
(6)						13,74

Passo 2: Entidades mais similares: parental 6 e o grupo (4,7)

Distância entre entidades: 7,66

Nova matriz de dissimilaridade:

	(2)	(3)	(4,6,7)	(5)	(8)
(1)	23,46	62,94	15,49	29,74	29,61
(2)		125,03	56,59	100,43	97,48
(3)			22,00	47,69	44,02
(4,6,7)				25,97	13,74
(5)					15,66

Passo 3: Entidades mais similares: parental 8 e o grupo (4,6,7)

Distância entre entidades: 13,74

Nova matriz de dissimilaridade:

	(2)	(3)	(4,6,7,8)	(5)
(1)	23,46	62,94	15,49	29,74
(2)		125,03	56,59	100,43
(3)			22,00	47,69
(4,6,7,8)				15,66

Passo 4: Entidades mais similares: parental 1 e o grupo (4,6,7,8)

Distância entre entidades: 15,49

Nova matriz de dissimilaridade:

$$(1,4,6,7,8) \begin{bmatrix} (2) & (3) & (5) \\ 23,46 & 22,00 & 15,66 \\ (2) & & 125,03 & 100,43 \\ (3) & & & 47,69 \end{bmatrix}$$

Passo 5: Entidades mais similares: parental 5 e o grupo (1,4,6,7,8)

Distância entre entidades: 15,66

Nova matriz de dissimilaridade:

$$(1,4,5,6,7,8) \begin{bmatrix} (2) & (3) \\ 23,46 & 22,00 \\ (2) & 125,03 \end{bmatrix}$$

Passo 6: Entidades mais similares: parental 3 e o grupo (1,4,5,6,7,8)

Distância entre entidades: 22,00

Nova matriz de dissimilaridade:

$$(1,3,4,5,6,7,8) \begin{bmatrix} (2) \\ 23,46 \end{bmatrix}$$

Passo 7: Formação do grupo final: 1,2,3,4,5,6,7,8

Distância entre entidades: 23,46

No dendrograma as distâncias entre entidades podem ser convertidas em percentagem, tomando-se como base o valor obtido na formação do grupo final igual a 100%. O estabelecimento dos grupos, com base no dendrograma pode ser feito tendo por base as mudanças acentuadas de níveis, associadas ao conhecimento do pesquisador sobre o material estudado. Com base nestes princípios pode-se pressupor, para este caso, a existência dos grupos: (I) P4 e P7; (II) P6; (III) P8, P1 e P5; (IV) P3 e P2.

Veja a Figura a seguir:

Grupo I: P4 e P7 (distância = 3,16)

Grupo II: P6 (distância = 7,66)

Grupo III: P8, P1 e P5 (distâncias = 13,74; 15,49 e 15,66)

Grupo IV: P3 e P2 (distâncias = 22,00 e 23,46)

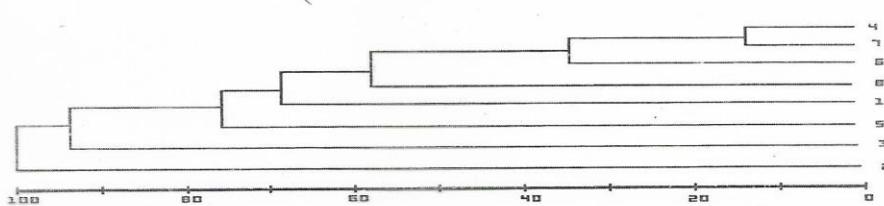


Figura – Dendrograma de similaridade entre oito progenitores, obtido pelo método da ligação simples com base nas distâncias generalizadas de Mahalanobis.

### Métodos não hierárquicos de agrupamento

São métodos desenvolvidos para agrupar itens (indivíduos, objetos, métodos) em um conjunto de  $k$  grupos, que pode ser definido antecipadamente ou determinado durante a execução do procedimento. Os passos básicos destes métodos são:

1. Selecionar  $k$  grupos ou sementes iniciais, onde  $k$  é o número de grupos desejado;
  2. Designar cada observação ao grupo mais próximo;
  3. Realocar cada observação a um dos  $k$  grupos de acordo com uma regra de parada pré-determinada;
  4. Parar o processo se não existe mais nenhuma realocação de pontos ou se a realocação satisfizer o critério estipulado na regra de parada. Caso contrário, voltar para o passo número 2.
- Os algoritmos não hierárquicos diferem com respeito ao método usado para obter os pontos iniciais ou os *pontos sementes* ou ainda quanto à regra usada para realocar os itens.

### Método de $k$ -Médias ( $k$ -means method):

O termo  $k$ -means é usado para descrever um algoritmo que aloca cada item ao grupo que tem o *centroide* (média) mais próximo. O procedimento é composto de três passos básicos:

1. Dividir os itens em  $k$  grupos iniciais;
2. Prosseguir, alocando cada item ao grupo cujo centroide está mais próximo (usualmente são calculadas as distâncias euclidianas). Recalcular o centroide do grupo que recebeu um novo item e do grupo que perdeu um item;
3. Repetir o passo 2 até que não seja mais necessário fazer uma realocação.

Ao invés de iniciar com uma partição de todos os itens em  $k$  grupos preliminares no passo 1, pode-se especificar  $k$  centroides iniciais (pontos sementes) e então seguir com o passo 2. Desta forma, a alocação final dos itens nos grupos será dependente da partição inicial ou da seleção inicial dos *pontos sementes*.

Exemplo de Aplicação: Suponha que foram medidas as variáveis  $X_1$  e  $X_2$  nos itens  $A$ ,  $B$ ,  $C$  e  $D$ :

Item	$X_1$	$X_2$
A	5	3
B	-1	1
C	1	-2
D	-3	-2

O objetivo é dividir esses itens em  $k=2$  grupos tais que os itens dentro de cada grupo estejam mais próximos um do outro do que em relação aos itens de um outro grupo. Para aplicar o método  $k=2$ -means, divide-se os itens arbitrariamente em dois grupos: (AB) e (CD) e calculam-se as coordenadas  $(\bar{X}_1, \bar{X}_2)$  do centroide de cada grupo.

Assim, no passo 1 tem-se:

Grupo	Coordenadas do Centroide	
	$\bar{X}_1$	$\bar{X}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

No passo 2 calcula-se a distância euclidiana de cada item aos centroides e realoca-se cada item ao grupo mais próximo. Se algum item for movido da configuração inicial, os centroides dos grupos devem ser atualizados antes de prosseguir. Então, tem-se:

$$d^2[A, (AB)] = (5-2)^2 + (3-2)^2 = 10$$

$$d^2[A, (CD)] = [5 - (-1)]^2 + [3 - (-2)]^2 = 61$$

Uma vez que o item A está mais próximo do grupo (AB), ele não será realocado.

Continuando o passo 2:

$$d^2[B, (AB)] = [-1 - (2)]^2 + (1 - 2)^2 = 10$$

$$d^2[B, (CD)] = [-1 - (-1)]^2 + [1 - (-2)]^2 = 9$$

Uma vez que o item B está mais próximo do grupo (CD) ele será realocado no grupo (CD) formando um novo grupo (BCD). Então, as coordenadas dos novos centroides deverão ser atualizadas:

Grupo	Coordenadas do Centroide	
	$\bar{X}_1$	$\bar{X}_2$
A		3
(BCD)	-1	-1

Novamente, deve-se verificar a necessidade de realocar cada um dos itens. Calculando o quadrado das distâncias euclidianas, tem-se:

Grupo	Distâncias aos Centroides			
	Item			
	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

Pode-se observar que os itens estão bem alocados nos grupos com centroides mais próximos. Então, os  $k = 2$  grupos finais são (A) e (BCD).

### Observações

- Para conferir a estabilidade do agrupamento, é desejável reinicializar o algoritmo com uma nova partição inicial.

2. Uma vez que os grupos foram determinados, as interpretações podem ser auxiliadas com o rearranjo da lista de itens de modo que aqueles do primeiro grupo apareçam primeiro, aqueles do segundo grupo apareçam depois e assim por diante.
3. Um quadro dos centroides dos grupos e as variâncias dentro dos grupos também auxiliam a delinear as diferenças entre os grupos.

### **Escolha dos Grupos Iniciais ou Sementes**

A escolha dos grupos/sementes iniciais de agrupamento influencia no agrupamento final. Algumas sugestões para a escolha das sementes são as seguintes:

1. Uso de métodos hierárquicos: inicialmente utilizam-se algum dos métodos hierárquicos de agrupamento para se obter os  $k$  grupos iniciais. Em seguida calcula-se o vetor de médias de cada grupo formado, sendo estes vetores de médias as sementes iniciais usadas no método das  $k$ -médias.
2. Escolha aleatória: as  $k$  sementes iniciais são escolhidas por meio de amostragem aleatória simples sem reposição dentro do conjunto de dados. Uma forma de melhorar a eficiência deste processo é selecionar  $m$  amostras aleatórias constituídas de  $k$  sementes e calcular o vetor de médias das  $m$  sementes selecionadas para cada grupo. Estes vetores constituem os centroides iniciais do método das  $k$ -médias.
3. Escolha por meio de uma variável aleatória: escolhe-se a variável aleatória de maior variância dentre as  $p$  componentes do vetor aleatório  $X$  considerado. Divide-se o domínio da variável escolhida em  $k$  intervalos. A semente inicial será o centroide de cada intervalo.
4. Valores discrepantes do conjunto de dados: por meio de análise estatística busca-se  $k$  elementos discrepantes no conjunto de dados em relação às  $p$  variáveis. Cada um desses elementos constituirá a semente de um grupo inicial.
5. Os  $k$  primeiros valores do conjunto de dados: a maioria dos aplicativos computacionais estatísticos usa como padrão para a escolha de sementes iniciais, as  $k$  primeiras observações do conjunto de dados, a menos que se especifiquem quais sementes devem ser usadas para iniciar o algoritmo. Esse processo traz bons resultados quando os  $k$  primeiros valores são discrepantes entre si.
6. Escolha prefixada: as sementes são escolhidas arbitrariamente. É um método muito subjetivo. Entretanto, pode ser usado quando o pesquisador tem um bom conhecimento do problema estudado.

### **Comentários**

1. Se dois ou mais pontos sementes inadvertidamente situam-se em um único grupo, os grupos resultantes da aplicação do método serão pouco diferenciados.
2. A existência de um *outlier* pode produzir no mínimo um grupo com itens muito dispersos.
3. Mesmo se soubermos que a população é formada por  $k$  grupos, pode ser que os dados de um grupo muito raro não apareçam na amostra. Neste caso, forçar a existência de  $k$  grupos pode gerar grupos absurdos.

4. Com base nos argumentos anteriores não é recomendável se fixar antecipadamente o número de grupos.
5. Após realizar o agrupamento é recomendável dar nome aos grupos ou traçar um perfil de cada grupo usando os centroides, o que pode ajudar na interpretação dos resultados.
6. Os métodos hierárquicos são também chamados de métodos de ligação e os métodos não hierárquicos de métodos de partição.

### Avaliação da qualidade dos agrupamentos e determinação do número de grupos

Algumas estatísticas que podem auxiliar são:

**RMSSTD** (*root-mean-square standard deviation*) de um grupo: é o desvio padrão ponderado de todos os elementos que formam o grupo. Uma vez que o objetivo da análise de agrupamento é formar grupos homogêneos, o RMSSTD de um grupo deverá ser o menor possível. Um valor alto de RMSSTD indica que o novo grupo não será homogêneo.

**RS** (*R-squared*): é calculado dividindo-se a soma de quadrados entre grupos  $(SQ_e)$  pela soma de quadrados dentro do grupo  $(SQ_d)$ . Como  $SQ_t = SQ_e + SQ_d$ , a um maior valor de  $SQ_e$  corresponde um menor valor de  $SQ_d$ . Para um determinado conjunto de dados, maior diferença entre grupos implica em grupos mais homogêneos e vice-versa. Então,  $0 \leq RS \leq 1$  mede o quanto cada grupo é diferente de cada outro. Valores de **RS** próximos de zero indicam pouca diferença entre grupos e valores próximos de um indicam diferenças máximas entre grupos.

**SPR** (*semipartial R-squared*): na análise de agrupamento, um novo grupo formado em certo passo, é obtido juntando-se dois grupos formados em passos anteriores. A diferença entre a  $SQ_d$  ponderada do novo grupo e a soma ponderada das  $SQ_d$  dos grupos juntados para formar o novo grupo é chamada de perda de homogeneidade.

Se a perda de homogeneidade é nula, então o novo grupo foi obtido juntando-se dois grupos perfeitamente homogêneos. Se a perda de homogeneidade é grande, então o novo grupo foi obtido juntando-se grupos muito heterogêneos. Usualmente, SPR é a razão entre o valor dessa diferença e a  $SQ_t$ . Desta forma, SPR representa a perda de homogeneidade devida à junção de dois grupos para formar um novo grupo. Um valor pequeno de SPR significa que está se juntando dois grupos homogêneos e vice-versa.

### Orientações Gerais para Análise de Agrupamento

1) A escolha do método de agrupamento depende do espectro de variação da matriz de distâncias – se for estreito recomenda-se usar a distância máxima; se for amplo pode-se utilizar a distância mínima ou média;

2) Os métodos LS, LC, UPGMA e de Ward são denominados de **métodos hierárquicos** porque os indivíduos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o diagrama de árvore ou dendrograma. Neste caso, o interesse maior está nas ramificações da árvore e não no número de grupos. As delimitações dos grupos podem ser estabelecidas por um exame visual do dendrograma, procurando detectar pontos de alta mudança de nível;

3) O método LS tem como desvantagem a incapacidade de não discernir grupos pobramente separados. Entretanto, apresenta a vantagem de delinejar grupos não-elipsoides, ou seja, evita-se estabelecer grupos únicos quando dos indivíduos se dispõem numa estrutura de filamentos conhecida como encadeamento, uma vez que indivíduos em extremidades opostas da cadeia podem ser completamente dissimilares;

4) O método UPGMA é utilizado com maior frequência em ecologia, sistemática e taxonomia numérica. Uma vez que utiliza médias aritméticas das medidas de dissimilaridade, evita caracterizar as dissimilaridades por valores extremos (máximo ou mínimo) entre os indivíduos considerados;

5) O método de Tocher é um dos **métodos de agrupamento de otimização**; nestes métodos os grupos são formados pela adequação de algum critério de agrupamento e objetivo é alcançar uma partição dos indivíduos que otimize (maximize ou minimize) alguma medida predefinida.

**Procedimentos Básicos para Determinar a Formação dos Grupos** - Uma vez obtido o dendrograma, a formação dos grupos pode ser feita com base em:

a) **Regra de decisão fixa** – neste caso se estabelece um nível (ou umbral) fixo de partição do gráfico, que geralmente coincide com a metade do espectro total de variação das distâncias;

b) **Regra de decisão variável** – neste caso se estabelece umbrais a níveis variáveis, dependendo do ajustamento dos grupos formados com a realidade sob estudo.

**Consistência e Adequação do Agrupamento** – Os principais métodos de avaliação da consistência e adequação do padrão de agrupamento são:

a) **Método da Análise Discriminante de Anderson (1948)** – neste caso obtém-se a matriz de distâncias dos dados originais que é utilizada para promover o agrupamento pela análise de otimização; retorna-se aos dados originais, classificando-os conforme resultado obtido pela análise de otimização efetuada.

Em seguida, realiza-se então a estimativa de funções discriminantes, que permitem classificar novos indivíduos não incluídos na análise e ratificar a alocação daqueles em estudo nos grupos considerados. Neste último caso, obtém-se a taxa de erro aparente, que quantifica o número de indivíduos classificados de forma diferente da prevista pela análise de otimização, medindo a adequação do agrupamento anteriormente realizado;

b) **Método do Coeficiente de Correlação Cofenético** – neste caso, depois de obtido o dendrograma tem-se uma nova leitura da similaridade (ou dissimilaridade) entre os indivíduos avaliados. Os novos coeficientes de semelhança indicados no dendrograma são estabelecidos de

acordo com o método de agrupamento utilizado e podem dar origem a uma nova matriz de dissimilaridade, denominada de matriz de coeficientes de semelhança cofenéticos.

Com a formação do dendrograma pode ocorrer considerável simplificação das informações originais e podem ser geradas distorções no padrão de dissimilaridade entre os indivíduos estudados, daí a necessidade de avaliar a adequação dos resultados.

A adequação dos métodos hierárquicos é feita por meio do **coeficiente de correlação cofenético (CCC)**, que é um coeficiente de correlação entre os elementos da matriz de dissimilaridade e os elementos da matriz cofenética. Portanto, é uma medida de concordância entre os valores originais de dissimilaridade e aqueles representados no dendrograma, usando-se somente os valores encontrados acima da diagonal das referidas matrizes. Quanto maior o valor de CCC, menor será a distorção provocada ao agrupar os indivíduos.

# CAPITULO 23

## Análise de AMMI/GGE – Biplot

### Introdução

Os experimentos multiambientes são muito importantes em melhoramento de plantas para testar a adaptação geral e específica de cultivares. Um cultivar pode apresentar flutuações no desempenho de produtividade quando cultivados em diferentes ambientes, sendo essas flutuações referidas como interação genótipo x ambiente (IGE).

Na análise de experimentos multiambientes é estruturada uma matriz  $Y$  típica, na qual as linhas são as médias de genótipos e as colunas são as médias de ambientes onde os genótipos foram avaliados. A presença de IGE impossibilita o uso de modelos interpretativos simples, que têm somente efeitos principais aditivos de genótipos e de ambientes.

O modelo linear típico para a resposta média,  $\bar{y}_{ij}$ , do i-ésimo genótipo no j-ésimo ambiente com  $n$  repetições em cada uma das  $i \times j$  caselas é expresso como:

$$\bar{y}_{ij} = \mu + g_i + e_j + (ge)_{ij} + \bar{\varepsilon}_{ij}, \text{ em que:}$$

$\mu$ : média geral;

$g_i$ : efeito aditivo do i-ésimo genótipo;

$e_j$ : efeito aditivo do j-ésimo ambiente;

$(ge)_{ij}$ : componente de interação GE para o i-ésimo genótipo no j-ésimo ambiente;

$\bar{\varepsilon}_{ij}$ : erro, assumido como  $NID(0, \sigma^2/n)$ , sendo  $\sigma^2$  a variância residual dentro de ambiente, considerada constante.

Esse modelo não é parcimonioso, uma vez que cada casela da IGE tem seu próprio parâmetro de interação, e pouco informativo, porque os parâmetros de interação independentes são de difícil interpretação. Então, Yates e Cochran (1938) propuseram tratar os termos da IGE como sendo linearmente relacionados ao efeito de ambiente, ou seja, colocar  $(ge)_{ij} = \xi_i e_j + d_{ij}$ , em que:

$\xi_i$ : coeficiente de regressão linear do i-ésimo genótipo na média ambiental;

$d_{ij}$ : desvio da regressão;

Este procedimento foi posteriormente usado por Finlay e Wilkinson (1963) e ligeiramente modificado por Eberhart e Russell (1966). Tukey (1949) propôs um teste para a IGE usando

$(ge)_{ij} = K g_i e_j$ , sendo  $K$  uma constante. Mandel (1961) generalizou o modelo de Tukey colocando  $(ge)_{ij} = \lambda \alpha_i e_j$  para genótipos e  $(ge)_{ij} = \lambda g_i \gamma_j$  para ambientes, obtendo assim um feixe de linhas retas que podem ser testadas para concorrência (isto é, se os  $\alpha_i$ , ou os  $\gamma_j$  são todos os mesmos) ou não concorrência.

Gollob (1968) e Mandel (1969, 1971) propuseram um termo bilinear para a IGE, da seguinte forma:

$(ge)_{ij} = \sum_{k=1}^s \lambda_k \alpha_{ik} \gamma_{jk}$ , no qual  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$  satisfazem a restrição de ortonormalidade, ou seja,  $\sum_i \alpha_{ik} \alpha_{ik'} = \sum_j \gamma_{jk} \gamma_{jk'} = 0$  para  $k \neq k'$  e  $\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1$ .

Isto conduz ao seguinte modelo linear-bilinear:

$$\bar{Y}_{ij} = \mu + g_i + e_j + \sum_{k=1}^s \lambda_k \alpha_{ik} \gamma_{jk} + \bar{\varepsilon}_{ij}$$

, o qual é uma generalização do modelo da regressão, com mais flexibilidade para descrever a IGE porque mais do que uma dimensão genótipo x ambiente é considerada. Então, Zobel et al. (1988) e Gauch (1988) chamaram esse modelo de “Additive Main Effects and Multiplicative Interaction Model” (AMMI), traduzido como Efeitos Principais Aditivos e Interação Multiplicativa.

### Descrição do Modelo AMMI

Considere um conjunto de  $g$  genótipos avaliados em  $e$  ambientes. A média de cada combinação de genótipo e ambiente, obtida com  $n$  repetições de um experimento, pode ser representada pela matriz seguinte:

$$Y_{(g \times e)} = \begin{bmatrix} \bar{Y}_{11} & \bar{Y}_{12} & \dots & \bar{Y}_{1e} \\ \bar{Y}_{21} & \bar{Y}_{22} & \dots & \bar{Y}_{2e} \\ \bar{Y}_{31} & \bar{Y}_{32} & \dots & \bar{Y}_{3e} \\ \dots & \dots & \ddots & \dots \\ \bar{Y}_{g1} & \bar{Y}_{g2} & \dots & \bar{Y}_{ge} \end{bmatrix} \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \\ \vdots \\ \bar{Y}_g \end{bmatrix}, \text{ em } \begin{bmatrix} \bar{Y}_{.1} & \bar{Y}_{.2} & \bar{Y}_{.3} & \dots & \bar{Y}_{.e} \end{bmatrix} \begin{bmatrix} \bar{Y}_{..} \end{bmatrix}$$

que os vetores marginais são os vetores de médias de linhas e colunas dos elementos da matriz  $Y$  e  $\begin{bmatrix} \bar{Y}_{..} \end{bmatrix}$  é a média geral dos dados.

O modelo AMMI postula componentes aditivos para os efeitos principais de genótipos ( $g_i$ ) e ambientes ( $e_j$ ) e componentes multiplicativos para o efeito da interação ( $ge$ )<sub>*ij*</sub>. Assim, a resposta média do genótipo  $i$  em um ambiente  $j$  é modelada por:

$$\bar{Y}_{ij} = \mu + g_i + e_j + \sum_{k=1}^m \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij} + \varepsilon_{ij}, \text{ no qual } (ge)_{ij} \text{ é representado por:}$$

$$\sum_{k=1}^m \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij}, \text{ sob as restrições:}$$

$$\sum_i g_i = \sum_j e_j = \sum_i (ge)_{ij} = \sum_j (ge)_{ij} = 0.$$

As estimativas da média geral ( $\mu$ ) e dos efeitos principais ( $g_i$  e  $e_j$ ) são obtidas de uma simples análise de variância com dois fatores a partir da matriz de médias  $Y_{(g \times e)} = [\bar{Y}_{ij}]$ . Os resíduos obtidos a partir dessa matriz constituem a matriz de interações  $GE_{(g \times e)} = [(\hat{g}e)_{ij}]$ , onde  $GE = \bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$ , e os termos da interação multiplicativa são estimados por meio da decomposição em valor singular (DVS) dessa matriz. Dessa forma,  $\lambda_k$  é estimado pelo k-ésimo valor singular de GE,  $\alpha'_{ik}$  é estimado pelo i-ésimo elemento do vetor singular esquerdo  $\alpha'_{k(g \times 1)}$  e  $\gamma'_{jk}$  é estimado pelo j-ésimo elemento do vetor singular direito  $\gamma'_{k(1 \times e)}$  associado com  $\lambda_k$  (Piepho, 1995).

Correspondências entre DVS e análise de componentes principais (ACP) são estabelecidas como a seguir:  $\lambda_k$  é o k-ésimo valor singular ou a raiz quadrada do k-ésimo maior autovalor da matriz  $(GE)(GE)'$  e  $(GE)'(GE)$ , as quais têm iguais autovalores não nulos;  $\alpha'_{ik}$  é o i-ésimo elemento do autovetor de  $(GE)(GE)'$  associado com  $\lambda_k^2$ ;  $\gamma'_{jk}$  é o j-ésimo elemento do autovetor de  $(GE)'(GE)$  associado com  $\lambda_k^2$ .

Portanto, a IGE neste modelo é expressa como uma soma de componentes, cada um deles multiplicado por  $\lambda_k$ , para um efeito genotípico ( $\alpha_{ik}$ ) e um efeito ambiental ( $\gamma_{jk}$ ).

O termo  $\lambda_k$  fornece a proporção da variância devida à IGE no k-ésimo componente. Os efeitos  $\alpha_{ik}$  e  $\gamma_{jk}$  representam pesos para o genótipo  $i$  e o ambiente  $j$  naquele componente da interação  $(\lambda_k^2)$ .

O posto da matriz GE é  $s = \min[(g-1), (e-1)]$ , de tal forma que o índice  $k$  na soma de componentes multiplicativos pode variar de 1 até  $s$ . O uso de todos os  $s$  componentes recupera toda a variação,  $SQ(IGE) = \sum_{k=1}^s \lambda_k^2$ , e o modelo é saturado de tal forma que ele produz um ajuste exato dos dados, sem termo residual para testar os efeitos. Quando se tem  $m < s$  componentes, o modelo é dito ser truncado.

Para a análise AMMI não se procura recuperar toda a  $SQ(IGE)$  mas somente aqueles componentes mais fortemente determinados por genótipos e ambientes. Consequentemente, o índice  $k$  é colocado a variar até  $m < s$ , de tal forma que as estimativas são obtidas para os primeiros  $m$  termos da DVS da matriz GE. Essa é uma análise de quadrados mínimos que conduz a um resíduo denotado por  $\rho_{ij}$ . Assim, a interação do genótipo  $i$  com o ambiente  $j$  é descrita pelo

padrão  $\sum_{k=1}^m \lambda_k \alpha_{ik} \gamma_{jk}$ , descartando o ruído dado por  $\sum_{k=m+1}^s \lambda_k \alpha_{ik} \gamma_{jk}$ .

Aqui como na ACP, o componente “explica”, sucessivamente, proporções decrescentes da variação presente na matriz GE, ou seja,  $\lambda_1^2 \geq \lambda_2^2 \geq \dots \lambda_k^2$ . Dessa forma, o método AMMI é visto como um procedimento capaz de separar padrão e ruído na análise da IGE.

### Determinação do número ótimo de termos multiplicativos no modelo AMMI

O principal objetivo da análise AMMI é a predição da verdadeira característica da resposta na casela em tabelas de dupla entrada de genótipos e ambientes. Para alcançar esse objetivo, um modelo AMMI truncado deve ser usado e então critérios para determinar o número de componentes necessários para explicar o padrão nos termos da IGE têm sido objeto de algumas pesquisas (Gollob, 1968;

Mandel, 1971; Gauch e Zobel, 1988; Piepho, 1994 e 1995; Cornelius, 1993; Cornelius et al., 1996).

Um dos procedimentos envolvidos na determinação do número ótimo de termos multiplicativos a serem retidos nos componentes da IGE usa o método de validação cruzada, no qual os dados são aleatoriamente divididos, sendo uma parte para modelagem e a outra para validação do modelo. O modelo AMMI é ajustado para modelar os dados e o quadrado médio do erro de predição, expresso como a raiz quadrada da diferença preditiva média (RMSPD), é determinado nos dados de validação.

Outro procedimento para determinar o melhor modelo truncado preditivo é usar testes de hipóteses sobre o k-ésimo componente,  $H_0: \lambda_k = 0$ , usando o conjunto de dados completos. Esses testes são baseados em somas de quadrados seqüenciais explicadas por termos multiplicativos.

### Testes de significância de termos multiplicativos

A soma de quadrados sequencial dos modelos AMMI para o k-ésimo componente,  $S_k'$ , é dada por

$n\lambda_k^2$ , para  $k=1,2,\dots, \text{posto}(GE)$ . Como em ACP, todos os critérios dos testes envolvem, pelo menos indiretamente, a razão da soma de quadrados acumulada para os primeiros  $m$  componentes da  $SQ(IGE)$ , ou seja,

$$\sum_{k=1}^m \lambda_k^2 / SQ(IGE).$$

Desde que se tenha uma partição ortogonal da soma de quadrados da interação, a razão entre o quadrado médio de qualquer componente da interação e o quadrado médio do resíduo é então assumida seguir uma distribuição F com os correspondentes graus de liberdade.

Uma vez que  $\lambda_k^2$ , não são variáveis aleatórias independentes que seguem uma distribuição qui-quadrado, um teste F não seria válido. Entretanto, a seleção do modelo ótimo é frequentemente baseada em teste F aproximado para os termos sucessivos da interação, e o número de termos incluídos corresponde ao número de componentes significativos.

O teste F aproximado de Gollob (1968) assume que  $n\hat{\lambda}_k / \sigma^2$  é distribuído como qui-quadrado e então obviamente não seria válido. Simulações feitas por Cornelius (1993) mostraram que o teste de Gollob a uma probabilidade 0,05 é muito liberal, com a taxa de erro Tipo I de 66% para testar  $H_{01}: \lambda_1 \neq 0$ .

Os testes F aproximados  $F_{GH1}$ ,  $F_{GH2}$  (Cornelius et al., 1993), efetivamente controla a taxa de erro Tipo I, e são mais parcimoniosos que o teste de Gollob. Entretanto, esses testes são conservativos para testar termos multiplicativos cujos valores são pequenos. O resíduo AMMI, obtido nos últimos termos da  $SQ(IGE)$ , pode também ser testado para confirmar sua não significância.

Gauch e Zobel (1996) apresentam alguns métodos para atribuir graus de liberdade para componentes do modelo AMMI, e aqueles devido a Gollob (1968) e Mandel (1971) são particularmente populares. Entretanto, existem discordâncias entre esses métodos.

O procedimento de Gollob (1968) é muito fácil de aplicar, uma vez que o número de graus de liberdade para o m-ésimo componente da interação é simplesmente definido como

$GL(IPCA_m) = g + e - 1 - 2m$ , enquanto outros procedimentos requerem simulações extensivas.

Gauch (1992) recomenda o uso do procedimento de Gollob e sugere que, em casos onde parece existir uma clara divisão entre grandes componentes determinando a *parte sistemática (padrão)* e pequenos componentes determinando a *parte de ruído*, associar igual número de graus de

liberdade,  $GL(IPCA_k) = [(g-1)(e-1)]/e$ , é especialmente útil para os primeiros componentes porque normalmente existirá pouco interesse na partição dos componentes devidos a ruídos. Desta forma, pelo procedimento de Gollob, a análise de variância conjunta completa, computada a partir das médias, tem a estrutura seguinte:

Quadro 1. Análise de variância conjunta completa, computada a partir das médias, usando o procedimento de Gollob

Fonte de Variação	Grau de Liberdade	SQ Gollob
Genótipo (G)	$g-1$	$SQ(G)$
Ambiente (E)	$e-1$	$SQ(E)$
Interação (IGE)	$(g-1)(e-1)$	$SQ(IGE)$
IPCA <sub>1</sub>	$g+e-1-(2 \times 1)$	$\lambda_1^2$
IPCA <sub>2</sub>	$g+e-1-(2 \times 2)$	$\lambda_2^2$
IPCA <sub>3</sub>	$g+e-1-(2 \times 3)$	$\lambda_3^2$
...	...	...
IPCA <sub>s</sub>	$g+e-1-(2 \times s)$	$\lambda_s^2$
Erro Médio/n	$ge(n-1)$	$SQ(ErroMedio)$
Total	$gen-1$	$SQ(Total)$

Pelo procedimento de Cornelius, a análise de variância conjunta completa, computada a partir das médias, tem a estrutura seguinte:

Quadro 2. Análise de variância conjunta completa, computada a partir das médias, usando o procedimento de Gollob

Fonte de Variação	GL Cornelius	SQ Cornelius
Genótipo (G)	-	-
Ambiente (E)	-	-
Interação (IGE)	-	-
IPCA <sub>1</sub>	$(g-1-1)(e-1-1)$	

---

IPCA <sub>2</sub>	$(g-1-2)(e-1-2)$	$\sum_{k=2}^s \lambda_k^2$
IPCA <sub>3</sub>	$(g-1-3)(e-1-3)$	$\sum_{k=3}^s \lambda_k^2$
...	...	...
IPCA <sub>s</sub>	...	$\sum_{k=4}^s \lambda_k^2$
Erro Médio/n	-	...
	-	$\sum_{k=1}^s \lambda_k^2$
Total	-	-

Piepho (1995) relata que o teste F aplicado de acordo com Gollob é liberal na seleção de componentes de um modelo AMMI, no sentido que seleciona muitos termos multiplicativos. Dos testes investigados pelo autor, o proposto por Cornelius et al. (1992) é o mais robusto para as suposições de homogeneidade e normalidade dos erros, recomendando que a validade das suposições deva ser verificada para outros testes.

A estatística-teste de Cornelius para  $m$  termos multiplicativos no modelo é a seguinte:

$$F_{R,m} = \left[ SQ(\text{IGE}) - \sum_{k=1}^m \lambda_k^2 \right] / \left[ f_2 QM(\text{ErroMedio}) \right]$$

, com  $f_2 = (g-1-m)(e-1-m)$ . Este é o teste  $F_R$  de Cornelius et al. (1992) que pode tornar-se liberal quando comparado com  $F_{GH1}$ ,  $F_{GH2}$ , ou testes iterativos de simulação. Sob a hipótese nula de que não mais que  $m$  termos determinam a interação, o numerador (ou seja, a SQ (GEI) residual para o modelo AMMI ajustado) é, aproximadamente, uma variável aleatória qui-quadrado (Piepho, 1995), de tal forma que o teste tem uma distribuição F com  $f_2$  gl e graus de liberdade do quadrado médio do resíduo.

Desta forma, um resultado significativo para o teste  $F_R$  indica que no mínimo um ou mais termos multiplicativos devem ser adicionados aos  $m$  já incluídos. Isso pode ser visto como um teste de significância para os primeiros  $m+1$  termos de interação. Quando  $m=0$ , ou seja, nenhum termo multiplicativo é incluído, o teste é equivalente ao teste F para a IGE na análise de variância, sendo então um teste exato. Nota-se também que o número de graus de liberdade do numerador de  $F_R$  é igual aos graus de liberdade para toda a interação menos os graus de liberdade atribuídos por

Gollob (1968) para os  $m$  primeiros termos. Desta forma, conclui-se que a aplicação de  $F_R$  é equivalente ao teste do resíduo AMMI para a GEI.

O modelo assim ajustado é denominado AMM10, AMM11, AMM12, ..., AMM1m, dependendo do número de termos (componentes) de interação retidos. Em AMM10 nenhum componente de interação é ajustado (modelo completamente aditivo), em AMM11 ajusta-se apenas o primeiro componente de interação, e assim por diante, até AMM1m (o modelo completo, com  $m = s$ ).

### Considerações adicionais sobre a análise AMMI

A análise AMMI combina num único modelo, termos aditivos para os efeitos principais, de genótipo ( $g_i$ ) e ambiente ( $e_j$ ), e termos multiplicativos para o efeito da interação ( $(ge)_{ij}$ ), conforme modelo a seguir:

$$Y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon, \text{ em que:}$$

$Y_{ij}$ : resposta média do genótipo  $i$  no ambiente  $j$ ;

$\mu$ : resposta média geral;

$g_i$ : efeito fixo do genótipo  $i$  ( $i=1,2,\dots,g$ );

$e_j$ : efeito fixo do ambiente  $j$  ( $j=1,2,\dots,e$ );

$(ge)_{ij}$ : efeito fixo da interação do genótipo  $i$  com o ambiente  $j$ ;

$\varepsilon$ : erro experimental médio, assumido como independente e  $\varepsilon \sim N(0, \sigma^2)$ .

A interação  $(ge)_{ij}$  é modelada como:

$$\sum_{k=1}^m \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij}.$$

No modelo AMMI,  $\alpha_{ik}$  e  $\gamma_{jk}$  estão sujeitos a restrições de ortonormalidade,

$$\sum_i \alpha_{ik}^2 = \sum_j \gamma_{jk}^2 = 1 \quad \text{e para } k \neq k', \quad \sum_i \alpha_{ik} \alpha_{ik'} = \sum_j \gamma_{jk} \gamma_{jk'} = 0; \quad \text{além disso,}$$

$$\sum_i \alpha_{ik} = \sum_j \gamma_{jk} = 0. \quad \text{O número de componentes multiplicativos é } m \leq s, \text{ em que } S \text{ é o}$$

número de componentes requeridos para saturar o modelo AMMI, se provido por mínimos quadrados, ou seja,  $\min[(g-1), (e-1)]$ .

No modelo, os termos que modelam a interação  $(ge)_{ij}$  resultam da decomposição em valores

singulares (DVS) da matriz de interação  $G \times E = (ge)_{ij}$ . Esta matriz é obtida como resíduo do

ajuste dos efeitos principais, por análise de variância (ANOVA), aplicada à matriz de médias  $(Y_{ij})$ .

Assim, por meio da DVS da matriz de interação  $\mathbf{G} \times \mathbf{E}$  obtém-se  $\sum_k \lambda_k \alpha_{ik} \gamma_{jk}$ , em que  $\lambda_k$  é o

k-ésimo autovalor de  $\mathbf{G} \times \mathbf{E}$  e ambos  $\alpha_{ik}$  e  $\gamma_{jk}$  são os respectivos autovetores relacionados ao

genótipo  $i$  e ao ambiente  $j$  que estão associados a  $\lambda_k$  (Piepho, 1995; Duarte e Vencovsky, 1999).

O índice  $k$  ( $k=1,2,\dots,s$ ), em que  $s = \min[(g-1),(e-1)]$ , o posto da matriz de interação

$\mathbf{G} \times \mathbf{E}$ , quando tomado até  $m$  no somatório ( $m \leq s$ ), determina uma aproximação de quadrados mínimos para a matriz de interação  $\mathbf{G} \times \mathbf{E}$  pelos  $m$  primeiros componentes da DVS (Duarte e Vencovsky, 1999; Duarte, 2003).

O número de componentes de interação retidos no modelo tem sido definido por meio do teste  $F$  ou  $F_R$  em função da proporção da  $SQ_{\mathbf{G} \times \mathbf{E}}$  acumulada até o m-ésimo componente, ou ainda

por validação cruzada, simulação e/ou iteração. Como numa análise de componentes principais, estes componentes de interação captam, sucessivamente, porções cada vez menores da variação

presente na matriz de interação  $\mathbf{G} \times \mathbf{E}$ , ou seja,  $(\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_s^2)$ . Assim, para  $m < s$ , a

análise AMMI é vista como um procedimento capaz de separar um *padrão de interação*

$\left( \sum_{k=1}^m \lambda_k \alpha_{ik} \gamma_{jk} \right)$  de seu *resíduo ou ruído*  $\left( \rho_{ij} = \sum_{k=m+1}^s \lambda_k \alpha_{ik} \gamma_{jk} \right)$ , no estudo da

$SQ_{\mathbf{G} \times \mathbf{E}}$ .

O *padrão de interação*  $\mathbf{G} \times \mathbf{E}$ , que é responsável pela variabilidade observada na matriz de médias  $(Y_{ij})$ , passa a ser explicado pelos  $m$  primeiros componentes de interação da análise AMMI. O

estudo destes componentes possibilita a identificação de fatores ambientais e genotípicos mais diretamente relacionados à interação. O resíduo presente na matriz de interação  $\mathbf{G} \times \mathbf{E}$ , que não é fortemente determinado por genótipos e ambientes, é descartado buscando melhorar a eficiência preditiva do modelo. Isto traz benefícios diretos à seleção de genótipos ou mesmo de ambientes para experimentos. Tais aspectos são tidos como vantagens da análise AMMI no estudo da interação  $\mathbf{G} \times \mathbf{E}$ , e que tem contribuído decisivamente para a popularidade dessa metodologia.

### Representação gráfica biplot

A análise AMMI apresenta ainda a vantagem de possibilitar a representação, num único gráfico, dos efeitos de interação de cada genótipo e cada ambiente. O gráfico, denominado *biplot* (Gabriel, 1971), baseia-se na aproximação da DVS de uma matriz, por outra de posto inferior. A aproximação da DVS da matriz  $\mathbf{G} \times \mathbf{E}$  pode ser escrita da seguinte forma:

$$\mathbf{G} \times \mathbf{E} = \sum_{k=1}^m \lambda_k \alpha_{ik} \gamma'_{jk} = \mathbf{U} \mathbf{S} \mathbf{V}' = \left( \mathbf{U} \mathbf{S}^{1/2} \right) \left( \mathbf{S}^{1/2} \mathbf{V}' \right) = (\mathbf{G})(\mathbf{H}')$$

, em que:

$U$  : tem em suas colunas os  $m$  autovetores  $\alpha$ ;

$V'$  : tem em suas linhas os  $m$  autovetores  $\gamma'$ ;

$S$  : tem na sua diagonal os  $m$  autovalores  $\lambda$ , com  $m \leq 3$ .

Então, a matriz  $G$  terá I autovetores denominados *marcadores de genótipos*, cada um com  $i$  linhas que correspondem às coordenadas dos genótipos para cada componente de interação retido no modelo AMMI. Da mesma forma, a matriz  $H'$  terá J autovetores denominados *marcadores de ambientes*, cada um com  $j$  linhas, que correspondem às coordenadas dos ambientes para os  $m$  componentes de interação retidos. Dada a ortogonalidade dos componentes, pode-se posicionar os I genótipos e os J ambientes num sistema cartesiano com  $m$  eixos (componentes) perpendiculares, para  $m \leq 3$  (Duarte, 2003).

Uma propriedade importante da representação gráfica *biplot* é que, multiplicando-se uma linha  $i$  de  $G$  por uma coluna  $j$  de  $H'$  obtém-se a estimativa *AMMIn* para a interação do genótipo  $i$  com o ambiente  $j$ . Desta forma, existe a possibilidade de se identificar combinações favoráveis de genótipos e ambientes, ou seja, aquelas com interações positivas de elevada magnitude, e assim capitalizá-las no processo de seleção. No gráfico *biplot*, isto corresponde a identificar combinações de genótipos e ambientes específicas, ou seja, com coordenadas de mesmo sinal e relativamente distantes da origem dos eixos.

Genótipos e ambientes com valores baixos de coordenadas para os eixos de interação são aqueles que relativamente menos contribuíram para a  $SQ_{GxE}$ , sendo, portanto fenotipicamente estáveis.

Estes genótipos podem ser recomendados amplamente, desde que apresentem também desempenhos médios elevados (estabilidade e adaptabilidade). Em ambientes estáveis espera-se que o ordenamento dos genótipos deva ser mais consistente. Esse tipo de informação permite aos melhoristas selecionar ambientes de teste conforme seus interesses (estratificação de ambientes). Portanto, a análise AMMI é útil tanto para a identificação de genótipos de alto desempenho e ampla adaptação, como para a realização de estratificação de ambientes, possibilitando identificar ambientes mais adequados para a condução de programas de melhoramento (Gauch e Zobel, 1996; Duarte e Vencovsky, 1999; Yan e Tinger, 2005).

### Valor Estabilidade AMMI (ASV)

Purchase (1997) apresentou a proposta de uma medida quantitativa de estabilidade baseada no modelo AMMI, com o objetivo de realizar o ordenamento de genótipos em termos de estabilidade de produtividade. O valor estabilidade AMMI (AMMI Stability Value-ASV), calculado para cada genótipo, é expresso como (Purchase et al., 2000):

$$ASV = \sqrt{\left[ \frac{SQ_{IPCA1}}{SQ_{IPCA2}} (Escore_{IPCA1}) \right]^2 + (Escore_{IPCA2})^2}, \quad \text{onde}$$

$SQ_{IPCA1}/SQ_{IPCA2}$  é o peso dado ao valor do escore *IPCA1*, uma vez que este escore é o que mais contribui para a soma de quadrados GxE.

Observe que quanto mais elevado for o valor do escore ***IPCA***, seja negativo ou positivo, mais o genótipo é adaptado especificamente a certos ambientes. Valores de escore ***ASV*** mais baixos indicam genótipos com maior estabilidade de produtividade ao longo dos ambientes.

Exemplo de Aplicação de Modelagem AMMI-Biplot na Análise de Interação Genótipos x Ambientes  
(Adaptado de Duarte e Vencovsky, 1999)

Considere a matriz  $\mathbf{Y}_{(g \times e)}$  de médias de produtividade de grãos de 10 genótipos de feijoeiro avaliados em 5 ambientes:

	E1	E2	E3	E4	E5	
G1	1589,9	560,9	1020,6	1449,7	1489,6	
G2	1386,7	691,7	1160,8	1591,2	1235,3	
G3	1572,9	770,4	1316,2	1835,9	1698,1	
G4	1139,7	734,5	862,7	1143,8	974,9	
G5	1348,9	458,2	663,8	1021,2	1116,3	
G6	1061,5	459,4	576,7	948,3	1013,7	
G7	1590,9	1104,1	1286,3	1250,9	1502,7	
G8	1223,9	546,9	674,3	1053,7	1053,4	
G8	1402,1	570,9	664,4	993,4	672,9	
G10	1313,6	580,1	596,3	1185,9	913,7	

O ajuste dos efeitos principais (G e E) por meio de análise de variância (ANOVA) dos dados da matriz  $\mathbf{Y}$  é o seguinte:

FV	GL	SQ	QM	F
Genótipos (G)	9	2230894,83	247877,20	10,81**
Ambientes (E)	4	3398566,23	849641,56	37,07**
Resíduo = GE	36	825226,64	22922,96	-
Total (Tratamentos)	49	6454687,71		

Neste caso,  $SQ_{GE} = 825226,64$ , que será utilizada para a DVS, representa 13% da  $SQ_{Total}(SQ_{Tratamentos} = G + E + GE)$ . As estimativas de médias de genótipos, de ambientes e geral, ajustadas com base na ANOVA (modelo sem interação) são as seguintes:

Genótipo (G)	Média $(\bar{Y}_{i.})$	Ambiente (E)	Média $(\bar{Y}_{.j})$
G1	1221,74	E1	1362,80
G2	1213,14	E2	647,70
G3	1438,68	E3	882,21
G4	971,12	E4	1247,40
G5	921,68	E5	1167,05
G6	811,92		
G7	1346,96		
G8	910,44		
G9	860,72		
G10	917,92		
		Média $(\bar{Y}_{..}) = 1061,43$	Geral

Aqui, o resíduo do ajuste dos efeitos principais com base na matriz de médias corresponde exatamente ao termo geral de interação  $(\hat{g}\hat{e})_{ij}$ , ou seja:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} \Leftrightarrow (\hat{g}\hat{e})_{ij} = Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$$

A aplicação da expressão anterior à matriz de médias  $Y_{(g \times e)}$ , resulta na matriz de interações  $GE_{(g \times e)}$  seguinte:

	E1	E2	E3	E4	E5	
G1	66,228	-247,472	-22,272	41,638	161,878	
G2	-127,972	-107,672	126,928	192,138	-83,422	
G3	-167,332	-254,532	56,768	211,278	153,818	
G4	-132,952	177,148	70,848	-13,242	-101,802	
G5	125,688	-49,712	-78,612	-86,402	89,038	
G6	-51,952	61,248	-55,952	-49,542	96,198	$= GE_{(10 \times 5)}$
G7	-57,612	170,888	118,588	-282,002	50,138	
G8	11,928	50,228	-56,872	-42,662	37,378	
G9	239,828	123,928	-17,072	-53,262	-293,422	
G10	94,148	75,948	-142,352	82,058	-109,802	

Observe que este resultado pode ser obtido operando-se diretamente com as matrizes correspondentes aos termos  $Y_{ij}$ ,  $\bar{Y}_{i.}$ ,  $\bar{Y}_{.j}$  e  $\bar{Y}_{..}$ . Denotando-se por  $L$ ,  $C$  e  $M$  as matrizes com

as médias de genótipos em suas linhas ( $L$ ), com as médias de ambientes em suas colunas ( $C$ ) e com a média geral em todas as posições ( $M$ ), respectivamente (todas de dimensão  $g \times e$ ) tem-se:

$$GE = Y - L - C + M$$

, em que:

$$L = \begin{bmatrix} \bar{Y}_{1.} & \bar{Y}_{1.} & \cdots & \bar{Y}_{1.} \\ \bar{Y}_{2.} & \bar{Y}_{2.} & \cdots & \bar{Y}_{2.} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{Y}_{g.} & \bar{Y}_{g.} & \cdots & \bar{Y}_{g.} \end{bmatrix}; \quad C = \begin{bmatrix} \bar{Y}_{.1} & \bar{Y}_{.2} & \cdots & \bar{Y}_{.e} \\ \bar{Y}_{.1} & \bar{Y}_{.2} & \cdots & \bar{Y}_{.e} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{Y}_{.1} & \bar{Y}_{.2} & \cdots & \bar{Y}_{.e} \end{bmatrix};$$

$$M = \begin{bmatrix} \bar{Y}_{..} & \bar{Y}_{..} & \cdots & \bar{Y}_{..} \\ \bar{Y}_{..} & \bar{Y}_{..} & \cdots & \bar{Y}_{..} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{Y}_{..} & \bar{Y}_{..} & \cdots & \bar{Y}_{..} \end{bmatrix}$$

O posto da matriz  $GE$  é  $p = \min(g-1, e-1) = \min(9, 4) = 4$ . Então, a  $SQ_{GE}$  pode ser decomposta em até 4 componentes ortogonais (ou somas de quadrados parciais). A operação básica para obtenção destes componentes é a aplicação da DVS na matriz  $GE$ , o que corresponde ao ajuste do termo de interação do modelo  $Y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \varepsilon_{ij}$  por meio da modelagem AMMI, em que  $(\hat{g}e)_{ij} = \sum_{k=1}^p \lambda_k \gamma_{ik} \alpha_{jk}$ . No SAS/IML esta operação é feita por meio do comando CALL SVD (U, S, V, GE).

Para esse exemplo, as matrizes resultantes que determinam a DVS da matriz de interações ( $GE = USV'$ ), são as seguintes:

$$U_{(10x4)} = \begin{bmatrix} -0,368471 & -0,061665 & 0,439858 & 0,287069 \\ -0,273094 & -0,202279 & -0,520215 & 0,269490 \\ -0,618207 & 0,005384 & -0,151055 & -0,006563 \\ 0,193414 & 0,166876 & -0,500910 & -0,177193 \\ 0,019655 & 0,018146 & 0,485182 & 0,046479 \\ 0,014484 & 0,247074 & 0,093268 & -0,415659 \\ 0,308700 & 0,657226 & -0,022498 & 0,357418 \\ 0,070399 & 0,084903 & 0,124821 & -0,255168 \\ 0,493786 & -0,519692 & 0,007667 & 0,416175 \\ 0,159333 & -0,395973 & 0,043883 & -0,522047 \end{bmatrix}$$

$$\begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \end{array}$$

$$S_{(4x4)} = \begin{bmatrix} 647,9814 \rightarrow \lambda_1 & 0,0000 & 0,0000 & 0,0000 \\ 0,0000 & 437,1808 \rightarrow \lambda_2 & 0,0000 & 0,0000 \\ 0,0000 & 0,0000 & 409,7126 \rightarrow \lambda_3 & 0,0000 \\ 0,0000 & 0,0000 & 0,0000 & 216,1715 \rightarrow \lambda_4 \end{bmatrix}$$

$$V_{(5x4)} = \begin{bmatrix} 0,318642 & -0,481743 & 0,616209 & 0,294410 \\ 0,681657 & 0,232309 & -0,280260 & -0,450367 \\ -0,075173 & 0,253732 & -0,437854 & 0,733658 \\ -0,473297 & -0,572047 & -0,364278 & -0,340667 \\ -0,451830 & 0,567749 & 0,466182 & -0,237034 \end{bmatrix}$$

$$\begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \end{array}$$

As informações do desdobramento da  $SQ_{GE}$  por DVS, relacionadas aos quadrados dos quatro valores singulares  $(\lambda_k^2)$  extraídos, ou, equivalentemente, aos quatro autovalores de  $(GE)(GE')$  ou de  $(GE')(GE)$  são as seguintes:

Componente (Comp)	Autovalor $(\lambda_k^2)$	Proporção $SQ_{GE}/Comp$	da Proporção Acumulada (%)
1	420101,0907	0,5091	50,9074
2	191042,6293	0,2315	74,0577
3	167424,3290	0,2029	94,3460
4	46658,5922	0,0565	100,0000
Total $\left(\sum_k \lambda_k^2\right)$	825226,6400	1,0000	-

Observa-se que o modelo AMMI2 explica 74% da  $SQ_{GE}$ , desdobrando-a em:

$$SQ_{GE(Padrão)} = 611143,72 \quad (74\% \text{ da variação da interação}) \quad \text{e}$$

$$SQ_{GE(Ruído)} = 214082,92 \quad (26\% \text{ da variação da interação}).$$

Um modelo AMMI saturado contemplaria todos os quatro componentes de interação, resgatando toda a  $SQ_{GE} = 825226,64 = \sum_k \lambda_k^2$ . Conforme o método de Gollob (1968) o número de graus de liberdade associados aos componentes da interação  $(IPCA_k)$  é:  
 $GL_{IPCA_n} = g + e - 1 - 2n$ . Então, o quadro de análise de variância completa (análise conjunta e decomposição de  $SQ_{GE}$  por modelagem AMMI) é o seguinte:

FV	GL	SQ	QM
Genótipos (G)	9	2230894,83	247877,20
Ambientes (E)	4	3398566,23	849641,56
Interação GxE	36	825226,64	22922,96
IPCA 1	12	420101,09	35008,42
IPCA 2	10	191042,63	19104,26
IPCA 3	8	167424,33	20928,04
IPCA 4	6	46658,59	7776,43
Erro Médio/r	90	871163,73	9679,5970

continua ...

continuação

FV	$F_{Gollob}$	$GL_{Res\ AMMI}$	$QM_{Res\ AMMI}$	$F_R/Cornelius$
Genótipos (G)	10,81**	-	-	-
Ambientes (E)	37,06**	-	-	-
Interação GxE	2,37**	36	22922,96	2,37**
IPCA 1	3,62**	24	16880,23	1,74*
IPCA 2	1,97*	14	15291,64	1,58 <sup>ns</sup>
IPCA 3	2,16*	6	7776,43	0,80 <sup>ns</sup>
IPCA 4	0,80 <sup>ns</sup>	0	0	-
Erro Médio/r	-	-	-	-

Podem ser observados também no quadro anterior os resultados do *resíduo AMMI* da interação, correspondente a cada membro da família de modelos. Desta forma, o resíduo de interação para AMM10 é toda a interação GE, com 36 GL, enquanto que para o modelo AMM11 é o restante da interação, depois de subtraídos os 12 GL e a soma de quadrados atribuída ao primeiro componente ( $IPCA_1$ ), ou seja, é a soma de quadrados dos efeitos atribuídos aos demais componentes ( $IPCA_2, IPCA_3, IPCA_4$ ), e assim por diante. Estes resultados são utilizados para avaliação dos modelos por meio do teste  $F_R$  de Cornelius (1992).

Diante da maior simplicidade de representação e da maior robustez do teste  $F_R$ , neste caso, o modelo AMM12 é recomendado como o que melhor descreve o padrão da resposta diferencial dos genótipos aos ambientes. Após a escolha do modelo deve-se fazer a estimativa de respostas para cada combinação de genótipo e ambiente. Considerando a seleção do modelo AMM12 esta operação deve ser realizada com base nas expressões seguintes:

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{..} + \left( \sum_{k=1}^{n=2} \lambda_k \gamma_{ik} \alpha_{jk} \right), \text{ ou}$$

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{..} + \lambda_1 \gamma_{i1} \alpha_{j1} + \lambda_2 \gamma_{i2} \alpha_{j2}$$

O último termo da expressão anterior corresponde a uma *aproximação DVS de posto dois* para o termo específico de interação  $(g\hat{e})_{ij}$  e reflete o *padrão da interação* GE presente no conjunto de dados. Este termo é função dos dois primeiros valores singulares  $(\lambda_1; \lambda_2)$  e dos elementos dos dois pares de vetores singulares a eles associados  $(\gamma_1, \gamma_2; \alpha_1, \alpha_2)$ . Desta forma, com base nas matrizes  $U$ ,  $S$  e  $V$  é possível obter a estimativa AMMI da resposta de um genótipo  $i$  num ambiente  $j$ .

Considere três exemplos de estimação de respostas com base no modelo AMMI2 ajustado:

1) Genótipo 5 no Ambiente 1

$$\begin{aligned}\hat{Y}_{51} &= \bar{Y}_{5.} + \bar{Y}_{.1} - \bar{Y}_{..} + \lambda_1 \gamma_{51} \alpha_{11} + \lambda_2 \gamma_{52} \alpha_{12} \\ \hat{Y}_{51} &= 921,68 + 1362,80 - 1061,43 + \\ &+ [648,1521(0,019971)(0,319555) + 437,0842(0,022439)(-0,475216)] = \\ &= 1223,05 + [-0,5244] = 1222,52 \text{ kg ha}^{-1} \text{ (resposta determinada basicamente pelos efeitos principais, o que indica alta estabilidade do genótipo 5)}\end{aligned}$$

2) Genótipo 9 no Ambiente 5

$$\begin{aligned}\hat{Y}_{95} &= \bar{Y}_{9.} + \bar{Y}_{.5} - \bar{Y}_{..} + \lambda_1 \gamma_{91} \alpha_{51} + \lambda_2 \gamma_{92} \alpha_{52} \\ \hat{Y}_{95} &= 860,72 + 1167,05 - 1061,43 + \\ &+ [648,1521(0,494042)(-0,452022) + 437,0842(-0,519068)(0,571761)] = \\ &= 966,34 + [-274,4630] = 691,88 \text{ kg ha}^{-1} \text{ (efeitos principais baixos e forte efeito negativo da interação, o que indica baixa adaptação do genótipo 9 no ambiente 5).}\end{aligned}$$

3) Genótipo 7 no Ambiente 2

$$\begin{aligned}\hat{Y}_{72} &= \bar{Y}_{7.} + \bar{Y}_{.2} - \bar{Y}_{..} + \lambda_1 \gamma_{71} \alpha_{21} + \lambda_2 \gamma_{72} \alpha_{22} \\ \hat{Y}_{72} &= 1346,96 + 647,70 - 1061,43 + \\ &+ [648,1521(0,308265)(0,681160) + 437,0842(0,657367)(0,229817)] = \\ &= 933,23 + [202,1296] = 1135,36 \text{ kg ha}^{-1} \text{ (resposta determinada pelo efeito da interação específica, o que indica boa adaptação do genótipo 7 ao ambiente 2).}\end{aligned}$$

Para obtenção dessas estimativas por meio de álgebra matricial, além das matrizes  $L$ ,  $C$  e  $M$  já definidas é necessária ainda uma aproximação de posto dois para a matriz  $GE$ , aqui denotada por  $\hat{GE}_2$ . Note que, este procedimento corresponde a obter uma estimativa da verdadeira matriz populacional de respostas (da qual  $Y$  representa uma amostra) com base no modelo escolhido. Considerando a seleção do modelo AMMI2, o estimador tem a notação  $\hat{Y}_2$  e a expressão matricial seguinte:

$$\begin{aligned}\hat{Y}_2 &= L + C - M + \hat{GE}_2 \\ \text{, em que:}\end{aligned}$$

$$GE_2 = \sum_{k=1}^{n=2} \lambda_k \gamma_k \alpha'_k$$

Observe que o termo  $[L+C-M]$  corresponde à matriz de respostas ajustadas pelo modelo AMMIO (sem interação), aqui denotada por  $\hat{Y}_0$ . Então, tem-se que:

$$\hat{Y}_2 = \hat{Y}_0 + GE_2, \text{ e os resultados são mostrados a seguir:}$$

$$\hat{Y}_{0(10 \times 5)} = \begin{bmatrix} 1523,11 & 808,01 & 1042,52 & 1407,71 & 1327,36 \\ 1514,51 & 799,41 & 1033,92 & 1399,11 & 1318,76 \\ 1740,05 & 1024,95 & 1259,46 & 1624,65 & 1544,30 \\ 1272,49 & 557,39 & 791,90 & 1157,09 & 1076,74 \\ 1223,05 & 507,95 & 742,46 & 1107,65 & 1027,30 \\ 1113,29 & 398,19 & 632,70 & 997,89 & 917,54 \\ 1648,33 & 933,23 & 1167,74 & 1532,93 & 1452,58 \\ 1211,81 & 496,71 & 731,22 & 1096,41 & 1016,06 \\ 1162,09 & 446,99 & 681,50 & 1046,69 & 966,34 \\ 1219,29 & 504,19 & 738,70 & 1103,89 & 1023,54 \end{bmatrix}$$

$$GE_{2(10 \times 5)} = \begin{bmatrix} -64,7892 & -168,5378 & 11,9477 & 127,2729 & 94,1065 \\ -13,4364 & -141,3848 & -9,2990 & 136,0014 & 28,1188 \\ -128,5884 & -272,5269 & 30,5702 & 188,7792 & 181,7659 \\ 6,2971 & 101,5679 & 8,2648 & -100,0980 & -16,0318 \\ -0,5244 & 11,0712 & 1,4727 & -11,7761 & -0,2434 \\ -48,4435 & 31,2593 & 26,3361 & -66,8043 & 57,6524 \\ -72,6934 & 202,1296 & 56,6950 & -260,0973 & 73,9662 \\ -3,2362 & 39,7401 & 5,9255 & -43,2376 & 0,8081 \\ 210,1413 & 165,9773 & -80,8730 & -20,7826 & -274,4630 \\ 115,2732 & 30,7040 & -51,0399 & 50,7424 & -145,6797 \end{bmatrix}$$

$$\hat{Y}_{2(10 \times 5)} = \begin{bmatrix} 1458,32 & 639,47 & 1054,47 & 1534,98 & 1421,46 \\ 1501,07 & 658,02 & 1024,62 & 1535,11 & 1346,88 \\ 1611,46 & 752,42 & 1290,03 & 1813,43 & 1726,06 \\ 1278,79 & 658,96 & 800,16 & 1056,99 & 1060,71 \\ 1222,52 & 519,02 & 743,93 & 1095,87 & 1027,05 \\ 1064,84 & 429,45 & 659,03 & 931,08 & 975,19 \\ 1575,63 & 1135,36 & 1224,43 & 1272,83 & 1526,54 \\ 1208,57 & 536,45 & 737,14 & 1053,17 & 1016,87 \\ 1372,23 & 612,97 & 600,63 & 1025,91 & 691,88 \\ 1334,56 & 534,89 & 687,66 & 1154,63 & 877,86 \end{bmatrix}$$

(Matrizes extraídas de Duarte e Vencovsky, 19990)

Para a representação gráfica dos genótipos e ambientes no **biplot** é necessária à determinação de suas coordenadas para os componentes de interação, obtidas por meio das matrizes  $G_{(g \times p)}$  e

$H'_{(p \times e)}$ , tal que:

$GE_{(g \times e)} = GH'$ , em que:

$G = US^{1/2}$  e  $H' = S^{1/2}V'$ . Os resultados são apresentados a seguir:

$$G_{(10 \times 4)} = \begin{bmatrix} -0,361 & -0,056 & 0,439 & 0,286 \\ -0,273 & -0,208 & -0,518 & 0,269 \\ -0,618 & 0,003 & -0,150 & -0,006 \\ 0,193 & 0,162 & -0,503 & -0,179 \\ 0,020 & 0,022 & 0,486 & 0,048 \\ 0,014 & 0,248 & 0,091 & -0,416 \\ 0,308 & 0,657 & -0,030 & 0,357 \\ 0,070 & 0,086 & 0,125 & -0,255 \\ 0,494 & -0,519 & 0,012 & 0,417 \\ 0,160 & -0,396 & 0,049 & -0,522 \end{bmatrix} \begin{bmatrix} (648,2)^{1/2} & 0,0 & 0,0 & 0,0 \\ 0,0 & (437,1)^{1/2} & 0,0 & 0,0 \\ 0,0 & 0,0 & (409,2)^{1/2} & 0,0 \\ 0,0 & 0,0 & 0,0 & (216,0)^{1/2} \end{bmatrix}$$

$$G_{(10 \times 4)} = \begin{bmatrix} -9,3942 & -1,1713 & | & 8,8708 & 4,2098 \\ -6,9507 & -4,3393 & | & -10,4850 & 3,9518 \\ -15,7320 & 0,0605 & | & -3,0418 & -0,0917 \\ 4,9170 & 3,3925 & | & -10,1757 & -2,6294 \\ 0,5084 & 0,4691 & | & 9,8282 & 0,7078 \\ 0,3681 & 5,1774 & | & 1,8498 & -6,1110 \\ 7,8481 & 13,7433 & | & -0,6002 & 5,2466 \\ 1,7943 & 1,7950 & | & 2,5192 & -3,7464 \\ 12,5777 & -10,8519 & | & 0,2504 & 6,1289 \\ 4,0633 & -8,2753 & | & 0,9842 & -7,6664 \end{bmatrix}$$

$$H'_{(4 \times 5)} = \begin{bmatrix} (648,2)^{1/2} & 0,0 & 0,0 & 0,0 & \\ 0,0 & (437,1)^{1/2} & 0,0 & 0,0 & \\ 0,0 & 0,0 & (409,2)^{1/2} & 0,0 & \\ 0,0 & 0,0 & 0,0 & (216,0)^{1/2} & \end{bmatrix} \begin{bmatrix} 0,320 & 0,681 & -0,076 & -0,473 & -0,452 \\ -0,475 & 0,230 & 0,250 & -0,576 & 0,572 \\ 0,620 & -0,282 & -0,441 & -0,358 & 0,461 \\ 0,296 & -0,451 & 0,733 & -0,341 & -0,237 \end{bmatrix}$$

$$H'_{(4 \times 5)} = \begin{bmatrix} 8,1355 & 17,3415 & -1,9231 & -12,0460 & -11,5079 \\ -9,9351 & 4,8047 & 5,2235 & -12,0466 & 11,9536 \\ - & - & - & - & - \\ 12,5457 & -5,7108 & -8,9289 & -7,2371 & 9,3310 \\ 4,3448 & -6,6300 & 10,7704 & -5,0079 & -3,4772 \end{bmatrix}$$

(Matrizes extraídas de Duarte e Vencovsky, 1999)

Para a construção do biplot em duas ou três dimensões basta que se tomem os dois ou os três primeiros componentes, respectivamente, para os marcadores de genótipos e de ambientes. Nas matrizes anteriores as duas primeiras dimensões estão separadas por linhas pontilhadas.

Considerando a escolha do modelo AMMI2, observe que se tomando apenas as duas primeiras colunas de  $\mathbf{G}$  e as duas primeiras linhas de  $\mathbf{H}'$ , se obtém as respectivas matrizes  $\tilde{\mathbf{G}}$  e  $\tilde{\mathbf{H}'}$ , por meio das quais se reproduz também a aproximação de posto dois para a matriz de interações  $(\mathbf{G}\hat{\mathbf{E}}_2 = \tilde{\mathbf{G}}\tilde{\mathbf{H}'})$ . Então, o biplot AMMI2 equivale à representação gráfica da matriz  $\mathbf{G}\hat{\mathbf{E}}_2$  que, na

modelagem AMMI, captura substancialmente o padrão da interação  $\mathbf{GE}$ .

As coordenadas dos eixos de abscissas (X) e de ordenadas (Y) para construção dos gráficos biplot AMMI1 (médias vs  $IPCA_1$ ) e biplot AMMI2 ( $IPCA_1$  vs  $IPCA_2$ ) são as apresentadas a seguir:

Genótipos e Ambientes	Biplot AMMI 1		Biplot AMMI 2		Ordem por Média
	X (Médias)	Y (IPCA1)	X (IPCA1)	Y (IPCA2)	
G1	1221,74	-9,3942	-9,3942	-1,1713	(3)
G2	1213,14	-6,9507	-6,9507	-4,3393	(4)
G3	1438,68	-15,7320	-15,7320	0,0605	(1)
G4	971,12	4,9170	4,9170	3,3925	(5)
G5	921,68	0,5084	0,5084	0,4691	(6)
G6	811,92	0,3681	0,3681	5,1774	(10)
G7	1346,96	7,8481	7,8481	13,7433	(2)
G8	910,44	1,7943	1,7943	1,7950	(8)
G9	860,72	12,5777	12,5777	-10,8519	(9)
G10	917,92	4,0633	4,0633	-8,2753	(7)
A1	1362,80	8,1355	8,1355	-9,9351	(1)
A2	647,70	17,3415	17,3415	4,8047	(5)
A3	882,21	-1,9231	-1,9231	5,2235	(4)
A4	1247,40	-12,0460	-12,0460	-12,0466	(2)
A5	1167,05	-11,5079	-11,5079	11,9536	(3)

(Quadro extraído de Duarte e Vencovsky, 1999)

Os dois gráficos biplot são apresentados a seguir. A partir destes gráficos são feitas interpretações, procurando identificar genótipos e ambientes que menos contribuíram para a interação **GE**, combinações de genótipos e ambientes desejáveis em termos de adaptabilidade e relações entre os componentes de interação e características genotípicas e ambientais conhecidas, dentre outras.

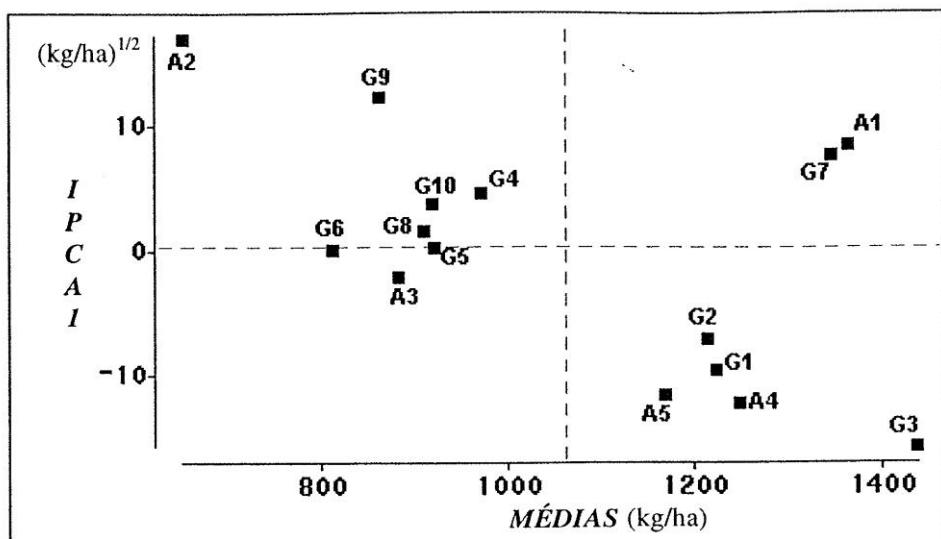


Figura 7 - Biplot AMMI para dados de produtividade de grãos (kg/ha), em feijoeiro, com dez genótipos (G) e cinco ambientes (A) (dados de Ramalho *et al.*, 1993).

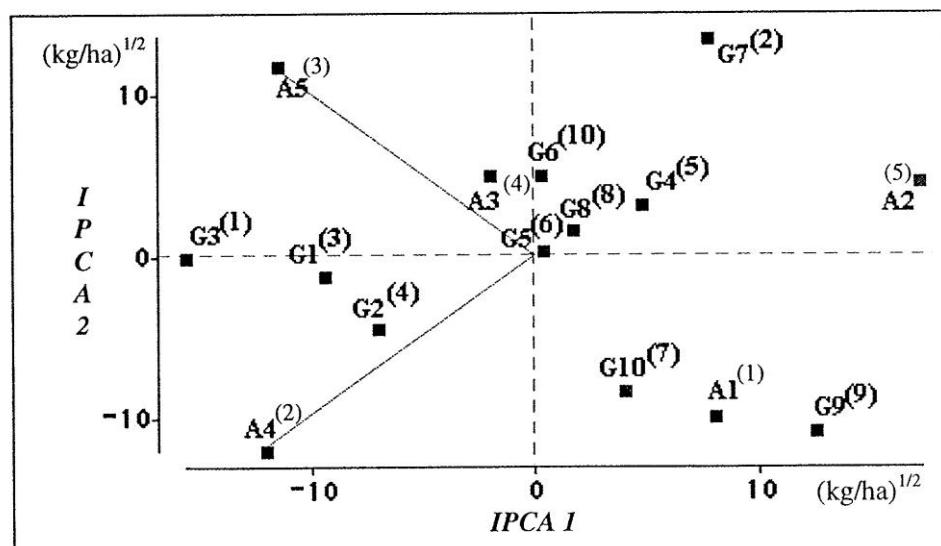


Figura 8 - Biplot AMMI2 para dados de produtividade de grãos (kg/ha), em feijoeiro. O expoente indica o posto do genótipo (G) ou do ambiente (A), nas listas de médias em ordem decrescente (dados de Ramalho *et al.*, 1993).

(Figuras 7 e 8, extraídas de Duarte e Vencovsky, 1999)

Os valores de estabilidade AMMI (ASV) para produtividade de grãos são apresentados a seguir:

Genótipos e Ambientes	Médias	IPCA1	IPCA2	ASV	Ordem por ASV
G1	1221,74	-9,3942	-1,1713	20,69	(7)
G2	1213,14	-6,9507	-4,3393	15,89	(6)
G3	1438,68	-15,7320	0,0605	34,59	(10)
G4	971,12	4,9170	3,3925	11,33	(4)
G5	921,68	0,5084	0,4691	1,21	(1)
G6	811,92	0,3681	5,1774	5,24	(3)
G7	1346,96	7,8481	13,7433	22,06	(8)
G8	910,44	1,7943	1,7950	4,33	(2)
G9	860,72	12,5777	-10,8519	29,71	(9)
G10	917,92	4,0633	-8,2753	12,18	(5)
A1	1362,80	8,1355	-9,9351	20,46	(2)
A2	647,70	17,3415	4,8047	38,43	(5)
A3	882,21	-1,9231	5,2235	6,72	(1)
A4	1247,40	-12,0460	-12,0466	29,10	(4)
A5	1167,05	-11,5079	11,9536	27,99	(3)

Observe que o genótipo 5 foi o mais estável e o genótipo 3, o que apresentou a maior média de produtividade ( $1438,68 \text{ kg ha}^{-1}$ ), foi o mais instável. O ambiente 3 foi o que apresentou maior estabilidade de produtividade de grãos.

### Aplicação da metodologia AMMI-Biplot na análise de dialelos

O método 2 da metodologia de análise de dialélica proposta por Griffing (1956), em que são incluídos  $\frac{p(p+1)}{2}$  tratamentos genéticos formados pelos  $p$  genitores e seus híbridos  $F_1$ , é um dos métodos que tem sido utilizado como referência na análise de dialelos. Neste método o modelo genético-estatístico é expresso como:

$$y_{ij} = m + g_i + g_j + s_{ij} + \bar{e}_{ij}, \text{ em que:}$$

$y_{ij}$ : valor médio da combinação híbrida ( $i \neq j$ ) ou do genitor ( $i = j$ );

$m$ : média geral;

$g_i, g_j$ : efeitos da capacidade geral de combinação (**CGC**) do  $i$ -ésimo ou  $j$ -ésimo genitor,

respectivamente ( $i, j = 1, 2, \dots, p$ );

$s_{ij}$ : efeito da capacidade específica de combinação (**CEC**) para os cruzamentos entre os

genitores  $i$  e  $j$ , considera-se  $s_{ij} = s_{ji}$ ;

$\bar{e}_{ij}$ : erro experimental médio associado à observação  $y_{ij}$ , estimado no experimento de avaliação dos tratamentos genéticos do dialelo.

Na forma matricial o modelo é expresso por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ em que:}$$

$\mathbf{y}_{n \times 1}$ : vetor dos valores observados;

$\mathbf{X}_{n \times (p+1)}$ : matriz de incidência dos efeitos genéticos (conhecida);

$\boldsymbol{\beta}_{(p+1) \times n}$ : vetor dos efeitos da média geral e das capacidades geral e específica de combinação (desconhecido);

$\boldsymbol{\varepsilon}_{n \times 1}$ : vetor do erro experimental associado aos elementos de  $\mathbf{y}$ .

Como neste caso a matriz  $\mathbf{X}$  não é de posto coluna completo, as soluções do sistema de equações normais associado ao modelo são obtidas por meio de  $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ , em que  $(\mathbf{X}'\mathbf{X})^{-1}$  é alguma inversa generalizada de  $\mathbf{X}'\mathbf{X}$ . Então, a estimativa dos efeitos e de suas somas de quadrados variam conforme a escolha da inversa generalizada.

Uma forma de contornar este problema, obtendo-se soluções únicas do sistema de equações, consiste em cortes no espaço de soluções pela imposição de restrições que sejam convenientes. Geralmente, adotam-se restrições do tipo soma zero, ou seja, expressando cada efeito como um desvio em relação à média. As restrições impostas simplificam as análises e proporcionam estimadores com significado biológico de interesse, além do fato de que as restrições nas soluções não interferem no modelo. Procedendo-se desta forma, as expressões das somas de quadrados para cada efeito são então expressas como:

$$SQ_m = R(m) = \hat{m}y_{..} = \frac{y_{..}^2}{p^2}$$

$$SQ_g = R(g|m) = \sum_i \hat{g}_i (y_{i.} + y_{.i}) = \frac{1}{2p} \sum_i (y_{i.} + y_{.i})^2 - \frac{2}{p^2} y_{..}^2$$

$$SQ_s = R(s|g,m) = \sum_{i < j} \hat{s}_{ij} (y_{ij} + y_{ji}) + \sum_i \hat{s}_{ii} y_{ii}, \text{ em que:}$$

$m$ : média geral;

$g$ : efeitos da capacidade geral de combinação (**CGC**);

$s$ : efeito da capacidade específica de combinação (**CEC**);

$R(m)$ : redução na soma de quadrados do resíduo corrigido para  $m$ ;

$R(g|m)$ : redução na soma de quadrados do resíduo após o ajuste do modelo para  $g$ , corrigido para  $m$ ;

$R(s|g,m)$ : redução na soma de quadrados do resíduo após o ajuste do modelo para  $s$ , corrigido para  $g$  e  $m$ .

Na metodologia AMMI Biplot, baseada na análise de componentes principais, dois componentes da matriz de dados obtidos a partir de múltiplos ambientes são utilizados para visualizar efeitos de genótipo e de interação genótipo x ambiente, usando um gráfico bidimensional. Como proposto por Yan e Hunt (2002) e descrito por Duarte e Pinto (2002), a metodologia AMMI Biplot pode ser utilizado também para a análise de capacidade de combinação em dialelos. Da mesma forma que

nos ensaios de múltiplos ambientes, os dois primeiros componentes principais podem ser usados para visualizar os efeitos da capacidade geral de combinação (**CGC**) e da capacidade específica de combinação (**CEC**) (BERTOIA et al., 2006).

Na descrição de Duarte e Pinto (2002), a análise dialélica é utilizada em conjunto com o método AMMI Biplot para identificar as melhores combinações híbridas. Utilizando a matriz da capacidade específica de combinação  $\mathbf{Y}_{(f \times m)}$  obtida a partir da análise dialélica foi possível à representação

gráfica, por meio da análise AMMI biplot dos dois primeiros componentes principais, melhorando a identificação de combinações hibridas favoráveis, permitindo a visualização do agrupamento dos genitores em grupos heteróticos. Então, a representação gráfica da capacidade específica de combinação com base na metodologia AMMI-Biplot é descrita da seguinte forma (Duarte e Pinto, 2002):

A decomposição em valor singular (DVS) é o método matricial utilizado como base para a representação gráfica biplot. Na análise AMMI a DVS é utilizada para obter a aproximação de uma matriz  $\mathbf{Y}$  de interação por outra de posto mais baixo com o objetivo de interpretar a interação de genótipos com ambientes. O método DVS consiste em decompor uma matriz  $\mathbf{Y}$  de posto  $p$  como a soma de  $p$  matrizes ortogonais de posto unitário, ou seja:

$$\mathbf{Y}_{(l \times c)} = \sum_{k=1}^p \lambda_k u_k v'_k, k=1,2,\dots,p \text{ e } p \leq \min(l,c), \text{ em que:}$$

$\lambda_k$ :  $k$ -ésimo valor singular da matriz  $\mathbf{Y}$ , que corresponde à raiz quadrada do  $k$ -ésimo autovalor não nulo de  $\mathbf{YY}'$  ou  $\mathbf{Y}'\mathbf{Y}$ ;  $u_k$  e  $v'_k$  são os vetores singulares coluna e linha, respectivamente, associados ao  $k$ -ésimo valor singular.

Na interpretação da capacidade específica de combinação (**CEC**) por meio da análise AMMI-Biplot é utilizada a matriz  $\mathbf{Y}_{(f \times m)}$ , onde  $f$  é o número de linhas (genitores femininos) e  $m$  é o número de colunas (genitores masculinos) na tabela do dialelo. No caso de cruzamentos apenas entre genótipos de diferentes grupos (dialelo parcial), como ocorre é o caso de testecrosses, a matriz  $\mathbf{Y}_{(f \times m)}$  é retangular e assimétrica, mas isto será resolvido com a aplicação da DVS.

Observe que a DVS determina a partição da soma de quadrados dos elementos da matriz original  $(SQ_Y)$  e considerando que  $\mathbf{Y}$  é a matriz das estimativas dos desvios da CEC  $(\hat{s}_{ij})$ , então esta soma de quadrados é diretamente relacionada com a  $SQ_{CEC}$  da análise de variância do dialelo.

Para a representação gráfica biplot da capacidade específica de combinação (CEC) será considerada uma aproximação DVS  $n$  para a matriz  $\mathbf{Y}_{(f \times m)}$  de posto  $p > n$ , da seguinte forma:

$$\hat{\mathbf{Y}}_{(f \times m)} = \sum_{k=1}^n \lambda_k u_k v'_k = \tilde{U} \tilde{S} \tilde{V}', \text{ em que } \tilde{U}_{(f \times n)} \text{ têm nas suas colunas apenas os } n \text{ primeiros } u_{(f \times 1)} \text{ vetores, } \tilde{V}'_{(n \times m)} \text{ tem nas suas linhas apenas os } n \text{ primeiros } v'_{(1 \times m)} \text{ vetores}$$

e  $\tilde{S}_{(n)}$  é a matriz diagonal com os primeiros valores singulares,  $\lambda_1, \lambda_2, \dots, \lambda_n$ . A expressão anterior pode ser escrita como:  $\hat{Y}_{(f \times m)} = (\tilde{U} \tilde{S}^{1/2}) (\tilde{S}^{1/2} \tilde{V}')$ , ou simplesmente como  $\hat{Y}_{(f \times m)} = (\tilde{G})(\tilde{H}')$ , sendo  $\tilde{G}_{(f \times n)} = (\tilde{U} \tilde{S}^{1/2})$  e  $\tilde{H}'_{(n \times m)} = (\tilde{S}^{1/2} \tilde{V}')$ . Portanto, a matriz  $\tilde{G}_{(f \times n)}$  terá  $f$  vetores, cada um com  $n$  elementos, chamados de marcadores de linhas, que correspondem às coordenadas dos parentais femininos (F), para cada eixo singular selecionado. De forma análoga, a matriz  $\tilde{H}'_{(n \times m)}$  terá  $m$  vetores, cada um com  $n$  elementos, chamados de marcadores de colunas, que correspondem às coordenadas dos parentais masculinos (M), para os mesmos eixos singulares selecionados. Uma vez que os eixos singulares são ortogonais, os  $f$  parentais femininos e os  $m$  parentais masculinos podem ser plotados num único sistema cartesiano de  $n$  eixos perpendiculares, numa representação gráfica em duas ou três dimensões, dependendo do número de eixos singulares selecionados na análise.

Exemplo de Aplicação (Extraído de Duarte e Pinto, 2002)

Considere os resultados de estimativas  $\hat{s}_{ij}$  para produtividade de grãos de milho, de um dialelo com híbridos  $F_1$ , sem genitores e híbridos recíprocos (método 4 de Griffing, 1956), que são expressas por meio de uma matriz simétrica  $B_{(9 \times 9)}$  (as colunas indicam os genitores masculinos,  $M1, M2, \dots, M9$  e as linhas os genitores femininos,  $F1, F2, \dots, F9$ ):

$$B = \begin{array}{ccccccccc|c} M1 & M2 & M3 & M4 & M5 & M6 & M7 & M8 & M9 & \\ \hline 0,00 & 4,94 & 14,46 & -2,76 & 2,80 & -12,31 & -16,93 & -8,67 & 18,47 & F1 \\ 4,94 & 0,00 & -6,54 & 14,14 & 8,90 & 5,68 & 13,07 & -2,47 & -37,73 & F2 \\ 14,46 & -6,54 & 0,00 & -29,94 & 19,01 & 15,80 & 4,48 & -6,96 & -10,31 & F3 \\ -2,76 & 14,14 & -29,94 & 0,00 & 11,50 & -8,21 & 3,77 & -19,47 & 30,97 & F4 \\ 2,80 & 8,90 & 19,01 & 11,50 & 0,00 & -36,16 & -0,77 & 4,99 & -10,27 & F5 \\ -12,31 & 5,68 & 15,80 & -8,21 & -36,16 & 0,00 & -10,78 & 32,77 & 13,22 & F6 \\ -16,93 & 13,07 & 4,48 & 3,77 & -0,77 & -10,78 & 0,00 & 5,66 & 1,50 & F7 \\ -8,67 & -2,47 & -6,96 & -19,47 & 4,99 & 32,77 & 5,66 & 0,00 & -5,85 & F8 \\ 18,47 & -37,73 & -10,31 & 30,97 & -10,27 & 13,22 & 1,50 & -5,85 & 0,00 & F9 \end{array}$$

Observe que devido à simetria, que advém da pressuposição de nenhum efeito recíproco, os valores  $\hat{s}_{ij}$  estão repetidos acima e abaixo da diagonal principal. Desta forma, a soma de quadrados ( $SQ$ )

dos elementos da matriz  $B$ ,  $\sum b_{ij}^2 = \sum \hat{s}_{ij}^2 = 18328,55$ , é equivalente a duas vezes a

capacidade específica de combinação,  $SQ_{CEC} = 9164,28$ , que corresponde ao quadrado médio de CEC igual a 339,44 com 27 graus de liberdade (Griffing, 1956). Então, a partição da  $SQ$  dos elementos da matriz  $B$  tem correspondência com a partição da  $SQ_{CEC}$ .

A decomposição em valores singulares (DVS) da matriz  $B$  é expressa como:  $B = USV'$ , sendo as matrizes  $U$ ,  $S$  e  $V'$  obtidas por meio de programa computacional de álgebra de matrizes, como por exemplo o Proc IML do SAS. Então, as matrizes que determinam a decomposição são:

1-Matriz  $U_{(9 \times 8)}$  de vetores singulares associados com as linhas (genitores femininos)

$$U = \begin{bmatrix} 0,2640 & 0,1759 & -0,1622 & 0,0515 & 0,5091 & 0,5569 & 0,4353 & -0,0210 \\ -0,3450 & -0,3612 & -0,0589 & -0,4124 & -0,3872 & 0,0198 & 0,5486 & -0,1207 \\ -0,3671 & 0,2619 & 0,2526 & -0,2661 & 0,5169 & -0,3951 & -0,1281 & -0,3335 \\ -0,0160 & 0,5778 & -0,5138 & 0,2488 & -0,3196 & -0,3324 & 0,1029 & 0,0760 \\ 0,3470 & -0,3116 & -0,2969 & -0,4184 & 0,1818 & -0,2327 & -0,2924 & 0,4853 \\ 0,5286 & -0,1209 & 0,5549 & 0,2996 & -0,1352 & -0,3438 & 0,2396 & -0,0566 \\ 0,2452 & 0,0337 & -0,0381 & -0,1533 & -0,3288 & 0,3567 & -0,5120 & -0,5524 \\ -0,3097 & 0,2535 & 0,4556 & 0,0130 & -0,1937 & 0,3505 & -0,1962 & 0,5676 \\ -0,3470 & -0,5090 & -0,1933 & 0,6373 & 0,1566 & 0,0201 & -0,1975 & -0,0446 \end{bmatrix}$$

2-Matriz  $S_{(8)}$  de valores singulares

$$S = \text{diag}[77,1865 \quad 63,3445 \quad 60,1728 \quad 54,7776 \quad 36,8391 \quad 17,6392 \quad 8,1048 \quad 1,7241]$$

3-Matriz  $V'_{(8 \times 9)}$  de vetores singulares associados com as colunas (genitores masculinos)

$$V' = \begin{bmatrix} -0,2640 & 0,3450 & 0,3671 & 0,0160 & -0,3470 & -0,5286 & -0,2452 & 0,3097 & 0,3470 \\ -0,1759 & 0,3612 & -0,2619 & -0,5778 & 0,3116 & 0,1209 & -0,0337 & -0,2535 & 0,5090 \\ -0,1622 & -0,0589 & 0,2526 & -0,5138 & -0,2969 & 0,5549 & -0,0381 & 0,4546 & -0,1933 \\ 0,0515 & -0,4124 & -0,2661 & 0,2488 & -0,4184 & 0,2996 & -0,1533 & 0,0130 & 0,6373 \\ 0,5091 & -0,3872 & 0,5169 & -0,3196 & 0,1818 & -0,1352 & -0,3288 & -0,1937 & 0,1566 \\ -0,5569 & -0,0198 & 0,3951 & 0,3324 & 0,2327 & 0,3438 & -0,3567 & -0,3505 & -0,0201 \\ 0,4353 & 0,5486 & -0,1281 & 0,1029 & -0,2924 & 0,2396 & -0,5120 & -0,1962 & -0,1975 \\ 0,0210 & 0,1207 & 0,3335 & -0,0760 & -0,4853 & 0,0566 & 0,5524 & -0,5676 & 0,0446 \end{bmatrix}$$

Observe que a matriz  $S$  tem dimensão  $8 \times 8$ , o que indica que o posto da matriz  $B$  é oito (9-1). Então,  $SQ_{CEC}$  pode ser particionada em até oito componentes ou eixos de componentes principais da interação (**IPCA**). Como neste exemplo o dialelo é completo, no particionamento cada componente corresponde à metade do quadrado de cada valor singular. Então, tem-se que:

$SQ_{CEC(IPCAk)} = \lambda_k^2 / 2, k = 1, 2, \dots, 8$ . O resultado deste particionamento está apresentado na tabela a seguir:

Tabela 1. Particionamento de soma de quadrados de capacidade específica de combinação ( $SQ_{CEC}$ ) por decomposição em valores singulares ( $DVS$ ), proporção retida em cada eixo de componente principal de interação ( $IPCAk$ ) e percentagem acumulada até o  $k$ -ésimo eixo, na análise AMMI de um dialelo

$IPCAk$	$SQ_{IPCAk}$	$SQ_{IPCAk}$	% Acumulada
		$SQ_{CEC}$	
1	2978,8753	0,3251	32,5053
2	2006,2649	0,2189	54,3975
3	1810,3817	0,1975	74,1523
4	1500,2918	0,1637	90,5234
5	678,5611	0,0740	97,9278
6	55,5711	0,0170	99,6254
7	32,8437	0,0036	99,9838
8	1,4862	0,0002	100,0000
Total	9164,2760	1,0000	-

Para construir o gráfico biplot é necessário obter as matrizes  $G = US^{1/2}$  e  $H' = S^{1/2}V'$ , de forma que  $B = GH'$ . Estas matrizes são:

1-Matriz  $G_{(9 \times 8)}$  de marcadores de linhas (genitores femininos  $F1, F2, \dots, F9$ )

$$G = \begin{bmatrix} 2,3198 & 1,4001 & -1,2579 & | & 0,3811 & 3,0903 & 2,3389 & 1,2392 & -0,0276 & F1 \\ -3,0311 & -2,8750 & -0,4565 & | & -3,0524 & -2,3502 & 0,0834 & 1,5617 & -0,1585 & F2 \\ -3,2249 & 2,0848 & 1,9591 & | & -1,9692 & 3,1374 & -1,6595 & -0,3646 & -0,4379 & F3 \\ -0,1409 & 4,5983 & -3,9856 & | & 1,8411 & -1,9398 & -1,3960 & 0,2928 & 0,0998 & F4 \\ 3,0482 & -2,4803 & -2,3030 & | & -3,0966 & 1,1035 & -0,9773 & -0,8325 & 0,6372 & F5 \\ 4,6443 & -0,9626 & 4,3047 & | & 2,2172 & -0,8207 & -1,4441 & 0,6821 & -0,0743 & F6 \\ 2,1540 & 0,2680 & -0,2958 & | & -1,1342 & -1,9958 & 1,4980 & -1,4577 & -0,7254 & F7 \\ -2,7205 & 2,0173 & 3,5341 & | & 0,0964 & -1,1755 & 1,4722 & -0,5587 & 0,7452 & F8 \\ -3,0488 & -4,0508 & -1,4991 & | & 4,7167 & 0,9508 & 0,0844 & -0,5623 & -0,0585 & F9 \end{bmatrix}$$

2-Matriz  $H'_{(8 \times 9)}$  de marcadores de colunas (genitores masculinos  $M1, M2, \dots, M9$ )

$$M1 \quad M2 \quad M3 \quad M4 \quad M5 \quad M6 \quad M7 \quad M8 \quad M9$$

$$H' = \begin{bmatrix} -2,3198 & 3,0311 & 3,2249 & 0,1409 & -3,0482 & -4,6443 & -2,1540 & 2,7205 & 3,0488 \\ -1,4001 & 2,8750 & -2,0848 & -4,5983 & 2,4803 & 0,9626 & -0,2680 & -2,0173 & 4,0508 \\ -1,2579 & -0,4565 & 1,9591 & -3,9856 & -2,3030 & 4,3047 & -0,2958 & 3,5341 & -1,4991 \\ - & - & - & - & - & - & - & - & - \\ 0,3811 & -3,0524 & -1,9692 & 1,8411 & -3,0966 & 2,2172 & -1,1342 & 0,0964 & 4,7167 \\ 3,0903 & -2,3502 & 3,1374 & -1,9398 & 1,1035 & -0,8207 & -1,9958 & -1,1755 & 0,9508 \\ -2,3389 & -0,0834 & 1,6595 & 1,3960 & 0,9773 & 1,4441 & -1,4980 & -1,4722 & -0,0844 \\ 1,2392 & 1,5617 & -0,3646 & 0,2928 & -0,8225 & 0,6821 & -1,4577 & -0,5587 & -0,5623 \\ 0,0276 & 0,1585 & 0,4379 & -0,0998 & -0,6372 & 0,0743 & 0,7254 & -0,7452 & 0,0585 \end{bmatrix}$$

Os tracejados nas matrizes  $\mathbf{G}$  e  $\mathbf{H}'$  delimitam na esquerda e acima as matrizes  $\tilde{\mathbf{G}}$  e  $\tilde{\mathbf{H}'}$ , respectivamente, que contêm as coordenadas correspondentes a genitores masculino e femininos, para a construção de um **biplot – AMMI3** (tridimensional).

Este biplot captura 74,15% da  $SQ_{CEC}$  original (Tabela 1), o que representa o padrão predominante da  $CEC$ . Então, deve ser esperado que a predição da capacidade específica de combinação obtida pela aproximação  $\hat{\mathbf{B}} = \tilde{\mathbf{G}}\tilde{\mathbf{H}'}$  seja mais realística do que a apresentada na matriz  $\mathbf{B}$  que é a original do dialelo. Os valores  $\hat{s}_{ij}$  preditos pela aproximação  $\hat{\mathbf{B}} = \tilde{\mathbf{G}}\tilde{\mathbf{H}'}$  da matriz  $\mathbf{B}$  são expressos por:

$$\hat{\mathbf{B}} = \begin{bmatrix} M1 & M2 & M3 & M4 & M5 & M6 & M7 & M8 & M9 \\ -5,76 & 11,63 & 2,10 & -1,10 & -0,70 & -14,84 & -5,00 & -0,96 & 14,63 & F1 \\ 11,63 & -17,24 & -4,68 & 14,61 & 3,16 & 9,35 & 7,43 & -4,06 & -20,20 & F2 \\ 2,10 & -4,68 & -10,91 & -17,85 & 10,49 & 25,42 & 5,81 & -6,06 & -4,32 & F3 \\ -1,10 & 14,61 & -17,85 & -5,28 & 21,01 & -12,08 & 0,25 & -23,74 & 24,17 & F4 \\ -0,70 & 3,16 & 10,49 & 21,01 & -10,14 & -26,46 & -5,22 & 5,16 & 2,70 & F5 \\ -14,84 & 9,35 & 25,42 & -12,08 & -26,46 & -3,97 & -11,02 & 29,79 & 3,81 & F6 \\ -5,00 & 7,43 & 5,81 & 0,25 & -5,22 & -11,02 & -4,62 & 4,27 & 8,10 & F7 \\ -0,96 & -4,06 & -6,06 & -23,74 & 5,16 & 29,79 & 4,27 & 1,02 & -5,42 & F8 \\ 14,63 & -20,20 & -4,32 & 24,17 & 2,70 & 3,81 & 8,10 & -5,42 & -23,46 & F9 \end{bmatrix}$$

Neste exemplo, a correlação entre  $\mathbf{B}$  e  $\hat{\mathbf{B}}$  é  $r = \sqrt{0,7415} = 0,861$ . Então, o erro associado com a predição baseada em  $\hat{\mathbf{B}}$  ao invés de  $\mathbf{B}$  não é considerado grande. Isto pode ser justificado porque com uma escolha adequada do número de termos multiplicativos na análise AMMI, um valor baixo da correlação entre  $\mathbf{B}$  e  $\hat{\mathbf{B}}$  está relacionado com a predominância de ruído na matriz  $\mathbf{B}$ , o que torna errada a interpretação de valores  $\hat{s}_{ij}$  como estimativas confiáveis dos verdadeiros valores  $(s_{ij})$  da  $CEC$ .

Algumas interpretações da matriz  $\hat{\mathbf{B}}$  são as seguintes:

1-Os cruzamentos que se destacam como combinações favoráveis são  $6 \times 8$ ,  $3 \times 6$ ,  $4 \times 9$  e  $4 \times 5$ .

2-A maioria dos valores  $\hat{s}_{ij}$  foram negativos, indicando que este tipo de cruzamento não é adequado para a extração de híbridos.

3-Observe que a **CEC** de cada genitor com ele mesmo  $(\hat{s}_{jj})$ , possibilitando inferências sobre heterose varietal. Os genitores de números **8**, **6** e **7** apresentaram maior heterose varietal. Na Tabela 2 mostrada a seguir, estão apresentadas as coordenadas de cada genitor, que são usadas para a construção de gráficos:

Tabela 2. Coordenadas dos genitores para os três primeiros eixos principais de interação de construção de um gráfico biplot, na análise AMMI de um dialelo

Pontos	Feminino/Masculino	<b>IPCA1</b>	<b>IPCA2</b>	<b>IPCA3</b>
1	F1	2,3197	1,4001	-1,2579
2	F2	-3,0311	-2,8749	-0,4565
3	F3	-3,2249	2,0848	1,9591
4	F4	-0,1408	4,5982	-3,9855
5	F5	3,0481	-2,4802	-2,3030
6	F6	4,6443	-0,9625	4,3046
7	F7	2,1540	0,2680	-0,2958
8	F8	-2,7205	2,0172	3,5341
9	F9	-3,0488	-4,0507	-1,4990
10	M1	-2,3197	-1,4001	-1,2579
11	M2	3,0311	2,8749	-0,4565
12	M3	3,2249	-2,0848	1,9591
13	M4	0,1408	-4,5982	-3,9855
14	M5	-3,0481	2,4802	-2,3030
15	M6	-4,6443	0,9625	4,3046
16	M7	-2,1540	-0,2680	-0,2958
17	M8	2,7205	-2,0172	3,5341
18	M9	3,0488	4,0507	-1,4990

No caso de um **biplot – AMMI3**, que é um gráfico tridimensional projetado no plano, não é possível uma avaliação da distância relativa entre os genitores. Para esta avaliação, é necessário projetar os pontos em, no mínimo, dois planos complementares, ou seja, **IPCA1 vs IPCA2** e **IPCA1 vs IPCA3**.

De acordo com o conceito da análise AMMI, o verdadeiro padrão de interação, neste caso a **CEC**, é descrito principalmente pelo eixo correspondente ao primeiro valor singular. Cada eixo singular é uma combinação linear de vetores linhas e colunas que formam a matriz  $\hat{B}$ . Como todos estes vetores contêm valores  $\hat{s}_{ij}$ , os eixos representam funções lineares da **CEC** construídas com todos os dados da matriz. Observe também que como os eixos sucessivos capturam sempre menos padrão e mais ruído, então, o descarte dos últimos eixos torna-se necessário para evitar a incorporação de erros na descrição da interação.

Exemplo de Aplicação 2 (Adaptado de Maeda, 2016)

Considere os resultados de capacidade específica de combinação de 21 híbridos experimentais, obtidos por meio de um dialelo (método 2 de Griffing, 1956) envolvendo sete progêneres de meios irmãos de milho (213, 205, 233, 225, 30, 232 e 128) selecionadas para eficiência no uso de fósforo (P).

Tabela 3. Efeito de capacidade específica de combinação para produtividade de grãos (PG), em genótipos de milho cultivados em Dourados-MS e Caarapó-MS, em nível alto e baixo P.

Código	Genótipo	Alto P			Baixo P		
		Dourados	Caarapó	Média	Dourados	Caarapó	Média
Capacidade específica de combinação $(\hat{s}_{ij})$							
8	H <sub>30x128</sub>	-230,22	929,73	349,75	-1201,84	250,51	-475,66
9	H <sub>30x232</sub>	-417,65	271,434	-73,11	-126,15	-297,54	-211,84
10	H <sub>225x30</sub>	-60,12	276,25	108,06	619,02	234,00	426,52
11	H <sub>233x30</sub>	-1511,41	-1536,11	-1523,76	-1523,97	-757,69	-1140,83
12	H <sub>205x30</sub>	190,37	49,40	119,88	321,72	-534,72	-106,49
13	H <sub>213x30</sub>	1581,62	-421,61	580,00	1083,73	948,91	1016,32
14	H <sub>232x128</sub>	-903,22	-401,19	-652,21	173,49	-486,33	-156,42
15	H <sub>225x128</sub>	285,98	-359,81	-36,91	-1178,81	-389,77	-784,29
16	H <sub>233x128</sub>	354,12	162,89	258,51	535,95	-95,39	220,27
17	H <sub>205x128</sub>	551,43	473,26	512,34	490,21	100,75	295,48
18	H <sub>213x128</sub>	1147,97	-218,30	464,83	-699,83	28,83	-355,49
19	H <sub>225x232</sub>	-1416,01	-318,06	-867,04	857,57	264,07	560,82
20	H <sub>233x232</sub>	840,29	-383,41	228,44	-163,14	719,37	278,11
21	H <sub>205x232</sub>	1749,92	-621,88	564,02	-274,28	399,55	62,63
22	H <sub>213x232</sub>	-646,69	45,80	-300,44	41,26	-521,65	-240,19
23	H <sub>233x225</sub>	272,64	-186,59	43,02	-479,65	74,54	-202,55
24	H <sub>205x225</sub>	246,17	-365,13	-59,48	1608,42	35,69	822,06
25	H <sub>213x225</sub>	-507,83	24,63	-241,60	-1048,86	491,88	-278,49
26	H <sub>205x233</sub>	-1903,43	174,80	-864,31	-683,53	-91,37	-387,45
27	H <sub>213x233</sub>	392,01	515,18	453,60	1107,507	232,62	670,06
28	H <sub>213x205</sub>	-553,85	424,99	-64,42	-678,88	41,80	-318,54

Os híbridos que tiveram o maior e menor efeito médio de  $\hat{s}_{ij}$  para a situação de alto e baixo P foram os mesmos. Para alto P (aplicação de P na semeadura) os valores variaram de 580,00 a -1523,76, enquanto que para baixo P (não aplicação de P na semeadura), o efeito médio  $\hat{s}_{ij}$  variou de 1016,32 a -1140,83, sendo os menores valores para a combinação H<sub>233x30</sub> e os maiores para H<sub>213x30</sub>.

Observa que a progénie 205 participa nos cruzamentos que obtiveram o segundo e terceiro maior valor de efeito  $\hat{s}_{ij}$ , ou seja, os híbridos H<sub>205x232</sub> e H<sub>205x128</sub>, respectivamente em alto P e em baixo P. O cruzamento H<sub>205x225</sub>, também apresentou um elevado valor de  $\hat{s}_{ij}$ . No

cruzamento  $H_{205 \times 128}$  em que participam as duas melhores progênies para baixo P, se observa um valor positivo para  $\hat{s}_{ij}$ .

A Tabela 4 apresentada a seguir contém os componentes principais, obtidos á partir da decomposição em valores singulares da matriz de capacidade específica de combinação  $(\hat{s}_{ij})$ , a proporção de variação total retida em cada componente principal e a porcentagem acumulada, com base na metodologia AMMI-Biplot. Nas representações gráficas biplot são utilizados os três primeiros componentes principais, que são os que guardam em si a maior parte da variação que as variáveis originais possuem.

TABELA 4. Componentes principais (CP) obtidos por decomposição em valores singulares da matriz de capacidade específica de combinação, proporção retida em cada componente principal e a porcentagem acumulada, em uma análise AMMI-Biplot para dialelos.

Dourados Alto P				Dourados Baixo P			
CP	Autovalor	Proporção	% Acumulada	Autovalor	Proporção	% Acumulada	
1	11739440,46	0,302	30,27	12733037,14	0,407	40,72	
2	8811165,14	0,227	53,00	8490170,40	0,271	67,87	
3	7992158,06	0,206	73,61	5650898,14	0,180	85,95	
4	7931791,86	0,204	94,06	4016794,24	0,128	98,79	
5	1912325,02	0,049	99,00	306197,94	0,009	99,77	
6	343673,66	0,008	99,88	52467,48	0,010	99,94	
7	44076,50	0,001	100,00	17356,96	0,001	100,00	
Total	38774630,70	1,00	-	31266922,30	1,00	-	
Caarapó Alto P				Caarapó Baixo P			
CP	Autovalor	Proporção	% Acumulada	Autovalor	Proporção	% Acumulada	
1	5778414,72	0,476	47,65	3573058,62	0,446	44,60	
2	2936797,52	0,242	71,87	2966676,84	0,370	81,63	
3	1713214,44	0,141	86,00	718251,60	0,089	90,59	
4	924598,98	0,076	93,63	512101,40	0,063	96,99	
5	657725,44	0,054	99,05	231290,74	0,028	99,87	
6	111881,80	0,009	99,97	7865,64	0,001	99,97	
7	2523,90	0,000	100,00	1926,70	0,000	100,00	
Total	12125156,80	1,00	-	8011171,54	1,00	-	
Alto P				Baixo P			
CP	Autovalor	Proporção	% Acumulada	Autovalor	Proporção	% Acumulada	
1	5103080,03	0,376	37,62	3988911,35	0,332	33,27	
2	3839682,05	0,283	65,94	3815456,19	0,318	65,10	
3	2641451,81	0,194	85,42	1794609,58	0,149	80,06	
4	1463287,19	0,107	96,21	1421456,10	0,118	91,92	
5	370400,00	0,027	98,94	777617,68	0,064	98,41	
6	140829,28	0,010	99,97	186556,74	0,156	99,96	
7	2747,32	0,000	100,00	3652,70	0,000	100,00	
Total	38774931,00	1,00	-	11988260,00	1,00	-	

Na Figura 1 é apresentado o gráfico biplot para o local Dourados, nos dois níveis contrastantes de P, para as progénies de meios-irmãos, que são identificadas como Fêmeas (F) e Machos (M) e codificadas de 1 a 7, totalizando 14 pontos no gráfico. As melhores combinações híbridas são aquelas que se encontram mais próximas, considerando F e M, nos eixos CP1, CP2 e CP3.

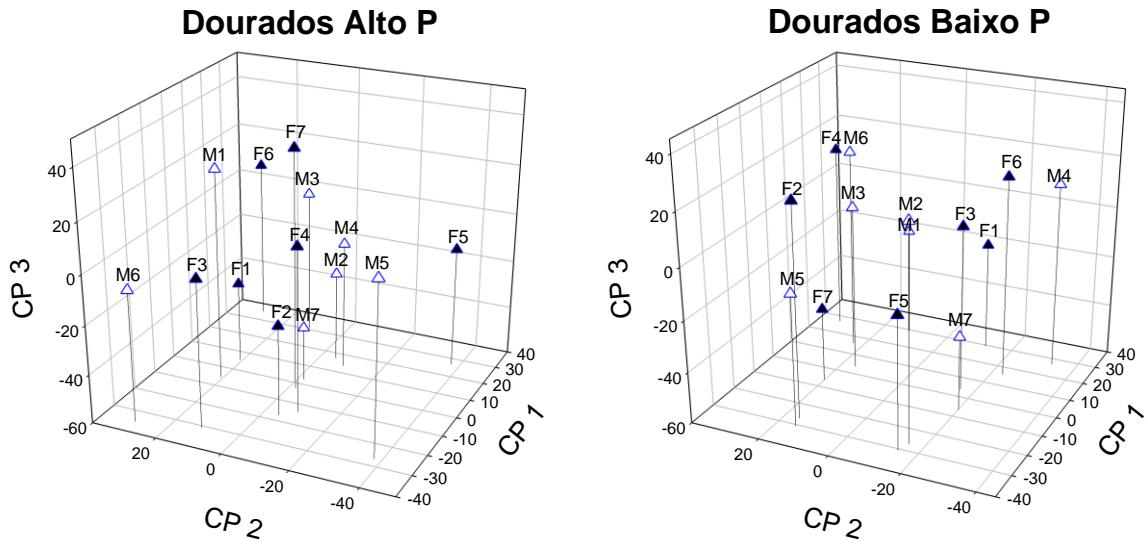


Figura 1. Biplot AMMI para capacidade específica de combinação (CEC) de produtividade de grãos de milho dos cruzamentos dialélicos em Dourados-MS, para alto e baixo nível de P. Os triângulos identificam os genótipos Fêmea (F) e Macho (M).

No gráfico Dourados Alto P, os pontos (F/M) mais próximos estão representados pelas combinações F7/M3 ( $H_{213 \times 30}$ ) e F2/M7 ( $H_{213 \times 128}$ ), que são as que tiveram os dois maiores valores  $\hat{s}_{ij}$ , indicando que são híbridos promissores para este ambiente. A combinação F5/M6 ( $H_{205 \times 233}$ ) apresentou a maior distância entre os pontos F/M, indicando que não foi uma boa combinação, com o valor de  $\hat{s}_{ij}$  de -1903,43, o menor para este local e nível de P.

No gráfico Dourados Baixo P, as combinações F4/M6 ( $H_{205 \times 225}$ ) foram as que mais se aproximaram, com valor de  $\hat{s}_{ij}$  igual a 1608,42, que foi o maior valor dentre as combinações híbridas. Para este híbrido foi obtida a produtividade de 5897,38 kg ha<sup>-1</sup>, que foi a maior obtida em condição de baixo P, considerando os dois locais. Além deste híbrido, outras duas combinações se destacaram com valores de  $\hat{s}_{ij}$  maiores que 1000, que foram as combinações F7/M1 ( $H_{213 \times 30}$ ) e F7/M5 ( $H_{213 \times 233}$ ). A pior combinação obtida foi a F1/M5 ( $H_{225 \times 30}$ ), justificado pela distância entre os pontos F/M.

O biplot para Alto e Baixo P em Caarapó está representado na Figura 2. No alto P, observa-se que o cruzamento F1/M2 ( $H_{30 \times 128}$ ) foi o de maior aproximação entre os pontos F/M e o F5/M1 ( $H_{233 \times 30}$ ) de maior distância, em que o maior e menor valor de  $\hat{s}_{ij}$  foi respectivamente de

929,73 e -1536,11. O híbrido  $H_{30 \times 128}$  foi o mais produtivo, com média de 4363,92 kg ha<sup>-1</sup> e o híbrido  $H_{233 \times 30}$  o de menor produtividade, com média de 1752,50 kg ha<sup>-1</sup>.

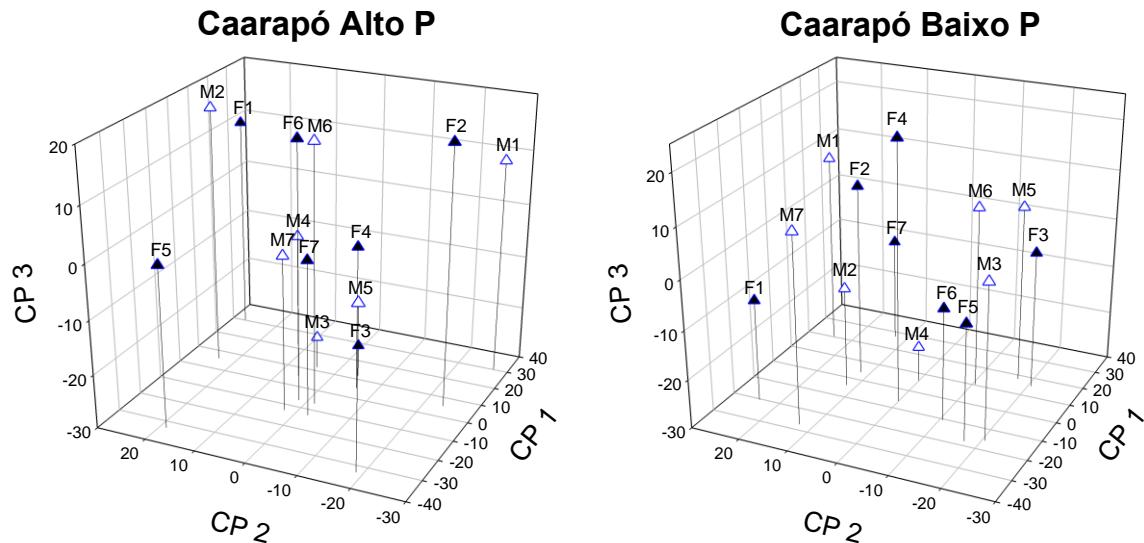


Figura 2. Biplot AMMI para capacidade específica de combinação (CEC) de produtividade de grãos de milho dos cruzamentos dialélicos em Caarapó-MS, para alto e baixo nível de P. Os triângulos identificam os genótipos Fêmea (F) e Macho (M).

No gráfico de baixo P se destacam as combinações F2/M1 ( $H_{30 \times 128}$ ), F3/M5 ( $H_{233 \times 232}$ ), F6/M3 ( $H_{205 \times 232}$ ) e F1/M7 ( $H_{213 \times 30}$ ), que apresentaram valores positivos para efeitos de CEC e produtividades semelhantes, variando de 3793,10 a 3292,46 kg ha<sup>-1</sup>, as maiores para o local. A maior distância observada no gráfico foi entre os pontos F1/M5 ( $H_{30 \times 233}$ ), sendo esta a combinação que apresentou o menor valor de  $\hat{s}_{ij}$  e menor produtividade de grãos.

Na figura 3 é apresentado o biplot, considerando os níveis de P na média dos dois ambientes, sendo utilizados os valores médios de  $\hat{s}_{ij}$  na análise AMMI- biplot.

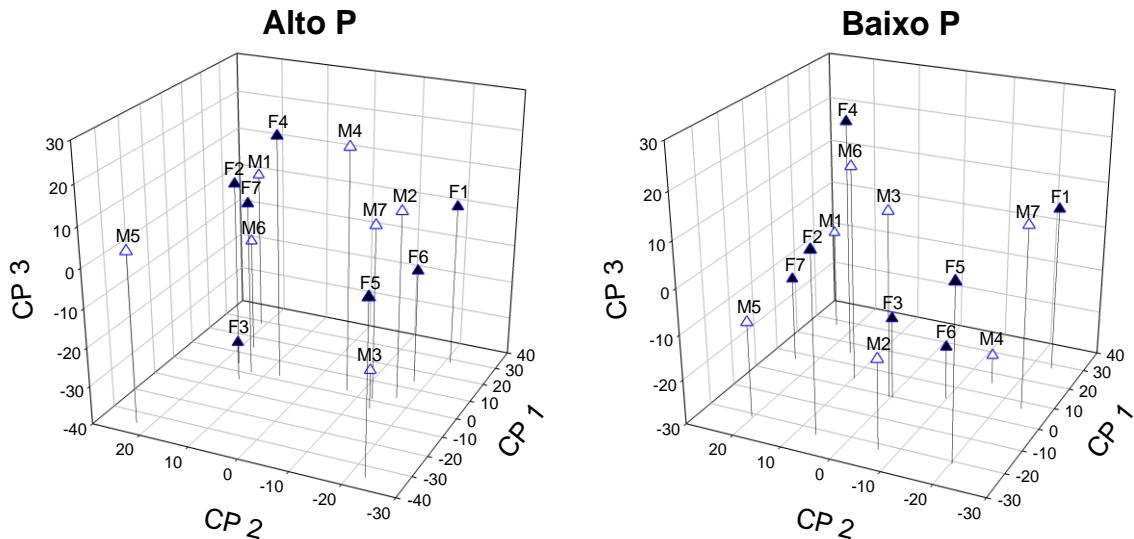


Figura 3. Biplot AMMI conjunta para capacidade específica de combinação (CEC) de produtividade de grãos de milho dos cruzamentos dialélicos em Caarapó-MS e Dourados-MS, para alto e baixo nível de P. Os triângulos identificam os genótipos Fêmea (F) e Macho (M).

Em Alto P, a proximidade entre os pontos F/M nas combinações F2/M1 e F7/M1, híbridos  $H_{30 \times 128}$  e  $H_{213 \times 30}$ , respectivamente, se mostraram como combinações mais estáveis, ou seja, apresentam maiores chances de apresentarem características semelhantes em um futuro cruzamento envolvendo as mesmas progênies. Em baixo P destacam com maior estabilidade nos cruzamentos as combinações F4/M6 ( $H_{205 \times 225}$ ) e F1/M7 ( $H_{213 \times 30}$ ), que foram as melhores médias de produtividade para baixo P, de 4543,83 e 4486,00 kg ha<sup>-1</sup>, respectivamente. Tanto em alto quanto em baixo fósforo, observa-se que a combinação F1/M5 ( $H_{233 \times 30}$ ) foi a pior, apresentando estimativas negativas de maior magnitude para efeitos de CEC em todos os locais e níveis estudados, indicando que o cruzamento entre as progênies deste híbrido não foi favorável, caracterizando baixa heterose.

### 9. Análise GGE-Biplot

O método GGE (*Genotype + genotype-by-environment*) é similar ao método AMMI. No entanto, o GGE subtrai da matriz de médias apenas os efeitos ambientais e após é realizada a análise de componentes principais (ACP) via decomposição em valores singulares (DVS). No método AMMI, apenas interação GxE é tratada como efeito multiplicativo enquanto que no método GGE-Biplot, o conjunto de genótipos e da interação GxE é tratado como efeito multiplicativo. Nesse método apenas o genótipo e a interação GxE são importantes e devem ser considerados de maneira simultânea.

Os gráficos apresentados no método GGE-Biplot permitem considerar três aspectos: 1) o relacionamento entre genótipos e ambientes, agrupando os genótipos e ambientes com comportamento semelhante, além de identificar o genótipo com maior potencial em cada subgrupo de ambiente (mega-ambientes); 2) o relacionamento entre os ambientes, facilitando a identificação de ambientes mais semelhantes entre si e a seleção de ambientes mais favoráveis e também os desfavoráveis para a avaliação de genótipos; e 3) promove o conhecimento da relação entre os genótipos, identificando àqueles mais semelhantes entre si e o ordenamento para os parâmetros de

produtividade e estabilidade (Yan, 2011) Como desvantagem, não deve ser utilizado em experimentos com dados desbalanceados e a interpretação gráfica pode ser dificultada quando o número de ambientes e de genótipos é muito elevado.

A análise pelo método GGE-Biplot, proposta por Yan et al. (2007) pode ser descrita a seguir.

O modelo GGE é expresso como:

$$Y_{ij} - \mu - E_j = y_1 e_{i1} \rho_{j1} + y_2 e_{i2} \rho_{j2} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : desempenho do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\mu$ : média geral das observações;

$E_j$ : efeito principal do  $j$ -ésimo ambiente;

$y_1$  e  $y_2$ : valores singulares associados ao IPCA1 e IPCA2, respectivamente;

$e_{i1}$  e  $e_{i2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $i$ -ésimo genótipo;

$\rho_{j1}$  e  $\rho_{j2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : efeito residual não explicado por nenhum dos fatores (“ruído”).

Então, com base nos escores associados aos ambientes e genótipos, são construídos gráficos que permitem as seguintes análises: “which-won-where” (qual-vence-onde), média x estabilidade, discriminativo x representativo e genótipo ideal (YAN e TINKER, 2006).

Para avaliar a adequação de um biplot em exibir os padrões de uma tabela de dupla-entrada, Yan e Tinker (2006) propuseram a relação de informação (IR). Esta tabela de dupla-entrada tem  $g$  genótipos e  $e$  ambientes. O número máximo de componentes principais (CP) demandado para apresentar essa tabela completamente é  $k = \min(e, g - 1)$ .

O IR é a proporção da variação total explicada para cada CP multiplicado por  $k$ . Se não há correlação entre os ambientes, todos os valores  $k$  dos CPs serão completamente independentes, e a proporção da variação total explicada para cada CP será exatamente igual a  $1/k$ . Se houver correlação entre os ambientes, a proporção da variação explicada para os primeiros CPs deverá ser maior ou igual a  $1/k$ , e a variação explicada por outros CPs deverá ser inferior ou igual a  $1/k$ . (YAN e TINKER, 2006). Um CP com  $IR > 1$  contém padrão (associação entre os ambientes), um CP com  $IR = 1$  não contém padrão, mas pode conter alguma informação independente e um PC com  $IR < 1$  não contém nenhum padrão ou informação (YAN, 2011).

Exemplo de Aplicação (Adaptado de Yamamoto, 2018 e Yamamoto et al., 2021)

Considere os resultados de médias de produtividade de grãos e de escores dos dois primeiros componentes principais de interação genótipos x ambientes (IPCA1 e IPCA2) de 36 genótipos de milho, apresentados a seguir:

Genótipo	Média	IPCA1	IPCA2	Genótipo	Média	IPCA1	IPCA2
G1	10817,72	41,56	-10,03	G19	6558,24	-6,33	14,39
G2	6294,13	-12,22	-12,64	G20	7026,20	9,38	15,37
G3	6082,55	0,55	-6,22	G21	7305,50	26,13	-2,02
G4	6568,73	-13,29	3,07	G22	6784,99	1,28	30,24
G5	7932,03	-6,65	5,29	G23	5387,09	-10,54	-20,59
G6	6909,75	0,25	-27,48	G24	6266,63	-14,76	12,60
G7	7395,88	-8,30	7,64	G25	5565,02	-8,86	11,36
G8	7701,63	12,01	-0,15	G26	6188,53	-15,11	11,17
G9	6994,61	-4,91	25,93	G27	6674,15	15,40	-16,21
G10	6303,22	-21,30	30,24	G28	6180,25	-19,83	-0,19
G11	7066,88	0,73	-8,29	G29	7043,75	14,52	4,03
G12	7251,99	17,32	27,86	G30	7556,21	10,86	1,68
G13	6657,36	9,61	-4,62	G31	7780,03	19,41	-0,99
G14	6871,67	19,78	0,12	G32	6734,88	5,22	-3,70
G15	7033,31	-5,58	-10,14	G33	5921,91	-11,45	-25,12
G16	6686,04	-3,68	-1,73	G34	7915,18	35,65	-5,46
G17	6193,47	-5,91	-6,28	G35	6807,73	-17,52	-8,09
G18	5353,55	-31,61	-14,87	G36	6902,21	-21,83	-16,16

Os genótipos foram avaliados, na safra 2012/2013, em nove ambientes (locais de cultivo): E1: Sete Lagoas/MG, E2: Londrina/PR, E3: Goiânia/GO, E4: Janaúba/MG, E5: Planaltina/DF, E6: Paragominas/PA, E7: Altamira/PA, E8: Campo Grande/MS, and E9: Manduri/SP.

Com base na decomposição em valores singulares (DVS), os dois primeiros componentes principais (IPCA1 e IPCA2) explicaram 71,88% da variação total das médias de produtividade de grãos dos genótipos de milho.

No método GGE-Biplot, por meio do gráfico do tipo qual-vence-onde (Figura 1), são criados setores que são limitados por linhas vermelhas. Nesse gráfico é possível agrupar os ambientes avaliados em mega-ambientes, que representa o conjunto entre os ambientes mais semelhantes entre si, em relação à produtividade de grãos de milho. Permite também indicar o genótipo com melhor desempenho para cada mega-ambiente, sendo este denominado o genótipo vencedor (YAN, 2011). Genótipos localizados no vértice do polígono são mais distantes da origem e classificados como os mais responsivos aos estímulos do ambiente. Estes genótipos podem ter alto ou baixo desempenho em alguns ou em todos os ambientes. Os genótipos alocados no interior do polígono são os menos responsivos.

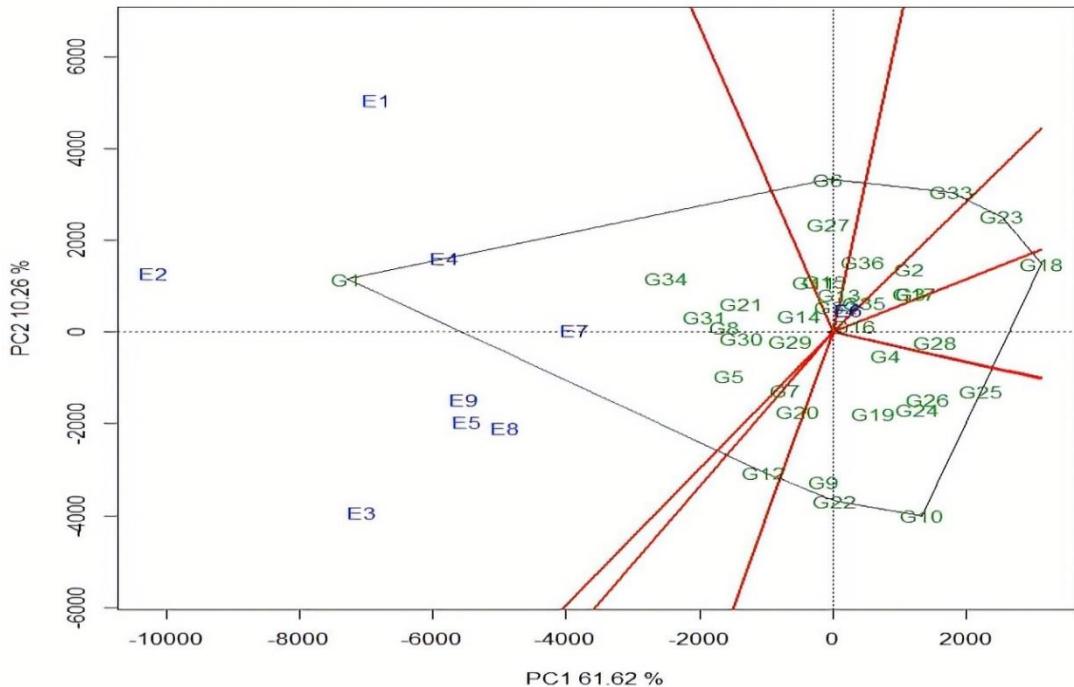


Figura 1. Gráfico GGE Biplot: qual-vence-onde referente à produtividade de grãos ( $\text{kg ha}^{-1}$ ) de 36 genótipos de milho avaliados em nove ambientes da região Central do Brasil na safra 2012/2013.

O genótipo vencedor G1 (BRS 1055) foi localizado no vértice do polígono, onde oito dos nove ambientes foram agrupados no setor 1 (Figura 1). Esse apresentou a maior média de produtividade de grãos dentre todos os genótipos avaliados. No que se refere aos ambientes, o E6 (Paragominas) demonstrou ser o ambiente mais diferente dos demais e por essa razão diferentes genótipos poderão serem semeados e selecionados nesse mega-ambiente.

Os genótipos G6 (Sint. 10717), G33 (Guepa), G23 (AL 2008), G18 (BRS Gorutuba), G10 (Sint. 10805) e G22 (AL Avaré) que deram origem a vértices, mas não contém ambientes nos seus setores tiveram baixo desempenho em todos os ambientes, o que significa que não demonstraram adaptação específica nos ambientes avaliados. Por esse método, os genótipos citados não devem ser utilizados para o cultivo nos ambientes avaliados devido à baixa produtividade de grãos alcançada.

O método GGE-Biplot, permite também avaliar a relação entre a produtividade de grãos e a estabilidade do genótipo, representada pelo gráfico média x estabilidade (Figura 2). Neste gráfico, o eixo que liga a linha horizontal que representa a coordenada ambiental média (*Average Environmental Coordination – AEC*), com a linha da média geral (linha vertical – com seta indicando valor de maior produtividade), indica quais genótipos foram superiores ou inferiores à média geral. Observe que, quanto maior a projeção do genótipo no IPCA2, maior a instabilidade do genótipo e maior interação deste com o ambiente (YAN e TINKER, 2006).

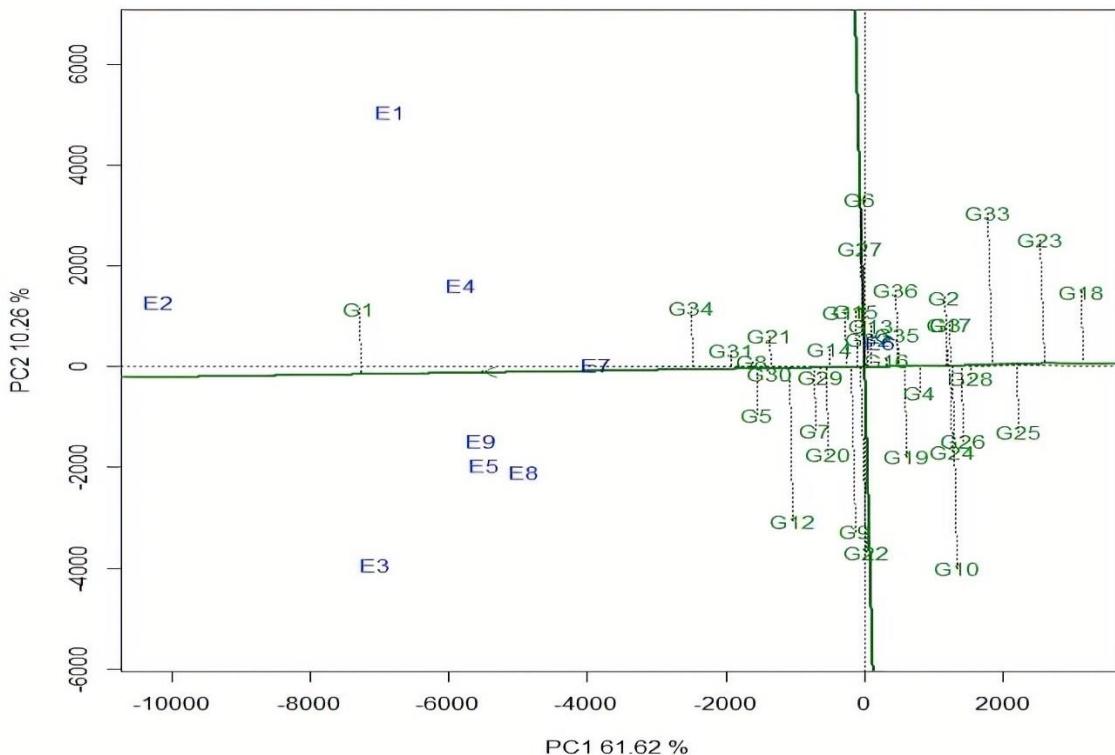


Figura 2. Gráfico GGE Biplot: média x estabilidade, referente à produtividade de grãos ( $\text{kg ha}^{-1}$ ) de 36 genótipos de milho avaliados em nove ambientes da região Central do Brasil na safra 2012/2013.

Observe no gráfico que, em relação a produtividade de grãos, os genótipos de milho com valores superiores à média geral foram G1 (BRS 1055), G34 (Bio 4), G31 (PC 0905), G8 (Sint. 10697), G5 (Sint. 10771), G30 (PC 0904), G21 (AL 2010), G12 (Sint. 10781), G7 (Sint. 10795), G20 (VSL BS 42 C 60), G29 (PC 0903), G14 (Sint. 10699), G11 (Sint. 10731), G15 (Sint. Mult TL) e G9 (Sint. 10707) (Figura 2).

Pode-se verificar no gráfico que os genótipos G8 (Sint. 10697) e G16 (Sint. RxS Spod) foram os mais estáveis, seguidos de G30 (PC 0904), G29 (PC 0903), G28 (1I934), G4 (BR106), G31 (PC 0905) e G14 (Sint. 10699). Considerando simultaneamente a produtividade e a estabilidade, os genótipos G8 (Sint. 10697) e G31 (PC 0905) foram os mais estáveis e os mais produtivos dentre os genótipos avaliados. O G10 (Sint. 10805) foi o genótipo mais instável, seguido do G22 (AL Avaré), G6 (Sint. 10717) e G33 (Guepa).

O genótipo ideal, que deve ter simultaneamente alta produtividade e elevada estabilidade entre os ambientes, no gráfico da análise GGE-Biplot é definido pelo centro dos círculos concêntricos (Figura 3) e SERVE como um modelo representativo do que seria um ideótipo de milho. Neste sentido, genótipos localizados mais próximos ao ideótipo são mais desejáveis para a seleção (YAN e TINKER, 2006). O G1 (BRS 1055) foi alocado no terceiro círculo concêntrico e apresenta como o mais próximo ao ideal, em termos de produtividade e estabilidade fenotípica.

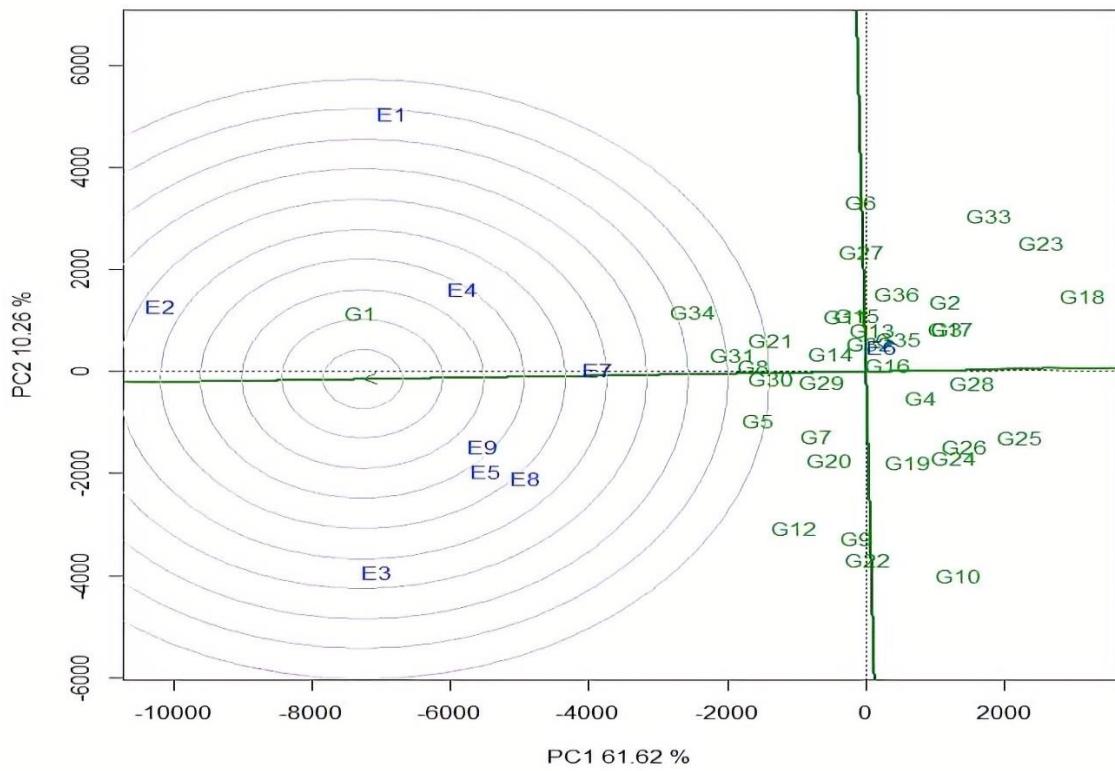


Figura 3. Gráfico GGE-Biplot: estimação do genótipo ideal, referente à produtividade de grãos ( $\text{kg ha}^{-1}$ ) de 36 genótipos de milho avaliados em nove ambientes da região Central do Brasil na safra 2012/2013.

O método GGE-Biplot ainda permite avaliar a relação existente entre os ambientes analisados (Figura 4). Um ambiente-teste ideal deve discriminar bem os genótipos e ser representativo de todos os ambientes testados. No gráfico, esse ambiente deve apresentar um alto valor de escore para o IPCA1, pois este valor expressa a sua capacidade em discriminar os genótipos, e valor baixo para o escore IPCA2, pois este valor demonstra uma maior capacidade de representar todos os outros ambientes. A percentagem da variação total explicada pelo IPCA1 (61,62%) e pelo IPCA2 (10,26%) permite realizar uma seleção confiável dos ambientes mais representativos.

Os ambientes discriminativos e representativos são úteis para selecionar genótipos com adaptação ampla a um ambiente. Os ambientes discriminativos, mas não representativos podem auxiliar no descarte de genótipos instáveis. Os ambientes não discriminativos e não representativos podem ser descartados, pois não são úteis em redes de experimentos (YAN e TINKER, 2006).

As diferenças ambientais são indicadas por vetores que se originam no centro do biplot (Figura 4). O valor do cosseno do ângulo entre os vetores de dois ambientes demonstra que há correlação entre eles. De maneira geral, todos os ambientes estão correlacionados entre si pois apresentam ângulos agudos entre si e menor que  $90^\circ$ . Os ambientes Planaltina e Campo Grande foram semelhantes entre si, pois apresentaram o menor ângulo agudo entre os ambientes. O ambiente Londrina foi o que teve maior capacidade em discriminar os genótipos, seguido por Sete Lagoas e Goiânia.

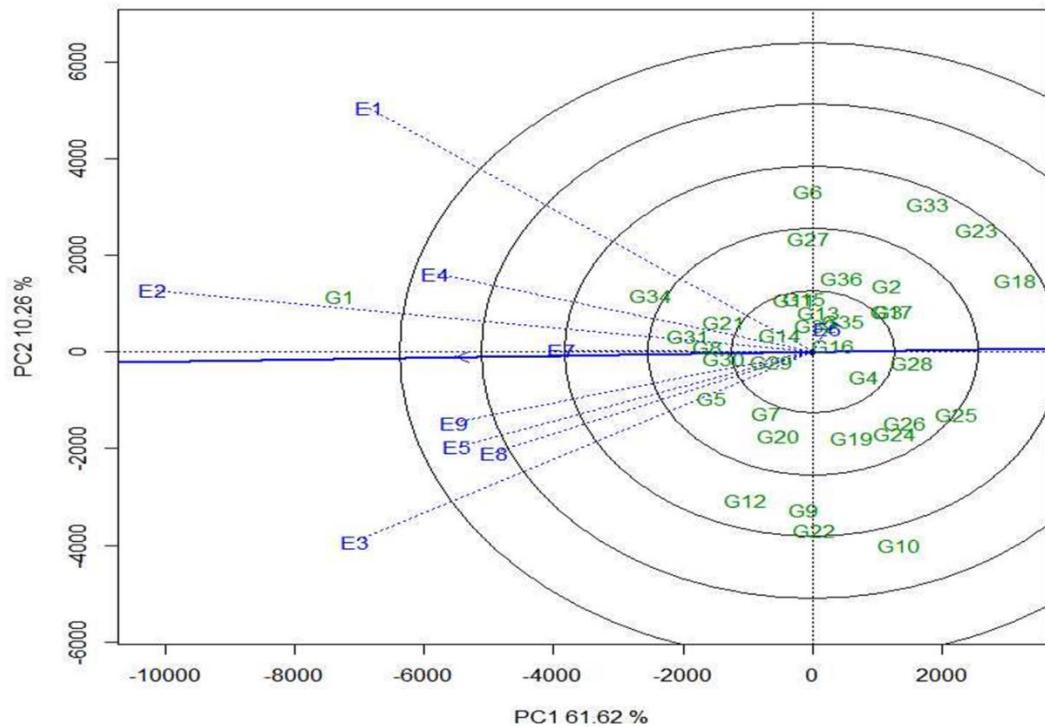


Figura 4. Gráfico GGE-Biplot: discriminativo x representativo, referente à produtividade de grãos ( $\text{kg ha}^{-1}$ ) de 36 genótipos de milho avaliados em nove ambientes da região Central do Brasil na safra 2012/2013.

# CAPITULO 24

## Análise de Correlações Canônicas

### Introdução

Tem como objetivo principal avaliar as relações lineares existentes entre dois conjuntos de variáveis respostas. A ideia básica é resumir a informação de cada conjunto em combinações lineares, sendo que a determinação dos coeficientes dessas combinações é feita tendo-se como critério a maximização da correlação entre os dois conjuntos. As combinações lineares que podem ser construídas são chamadas de variáveis canônicas, enquanto que a correlação entre elas é chamada de correlação canônica.

### Construção de variáveis e correlações canônicas

A forma de construção das variáveis canônicas com as respectivas correlações canônicas, tanto para a decomposição da matriz de covariâncias quanto para a matriz de correlação é como a seguir:

Considere dois vetores aleatórios  $X$  e  $Y$ , de dimensões  $px1$  e  $qx1$ , respectivamente. Os vetores de médias e matrizes de covariâncias dos vetores  $X$  e  $Y$  são representados por:

$$E(X) = \mu_X ; E(Y) = \mu_Y$$

$$Var(X) = \Sigma_{XX} ; Var(Y) = \Sigma_{YY}$$

$$Cov(X, Y) = \Sigma_{XY} ; Cov(Y, X) = \Sigma_{YX} = \Sigma'_{XY}$$

As variâncias das variáveis aleatórias que estão nos vetores  $X$  e  $Y$  se encontram nas matrizes  $\Sigma_{XX}$  e  $\Sigma_{YY}$ , de dimensões  $pxp$  e  $qxq$ , respectivamente. As covariâncias entre pares de variáveis de  $X$  e  $Y$  se encontram na matriz  $\Sigma_{XY}$ , de dimensão  $pxq$ .

Hotelling (1935, 1936) propôs que as relações existentes entre os vetores  $X$  e  $Y$  sejam avaliadas por meio da análise de combinações lineares destes vetores, construídas de modo que estejam fortemente correlacionadas entre si. Essas combinações lineares (novas variáveis) são chamadas de variáveis canônicas.

Em cada estágio do procedimento, duas combinações lineares são construídas, uma relativa às variáveis do vetor  $X$  e outra relativa às variáveis do vetor  $Y$ , ou seja, em cada estágio é construído um par de variáveis canônicas, ou seja, um par canônico. O procedimento assegura que as variáveis canônicas de um par não são correlacionadas com as variáveis canônicas de outro par. O número de pares canônicos que pode ser obtido é igual ao mínimo valor entre  $p$  e  $q$ .

## Variáveis canônicas populacionais

O primeiro par canônico é definido como sendo o par que contém as combinações lineares seguintes:

$U_1 = a'_1 X$  e  $V_1 = b'_1 Y$ , em que  $a_1$  e  $b_1$  são vetores de coeficientes (constantes) de dimensões  $px1$  e  $qx1$ , respectivamente, obtidos de modo que a correlação entre as variáveis  $U_1$  e  $V_1$  seja máxima e que essas variáveis canônicas tenham variâncias iguais a 1, ou seja:

$$Var(U_1) = Var(V_1) = 1$$

O segundo par canônico é definido como sendo o par que contém as combinações lineares seguintes:

$$U_2 = a'_2 X \text{ e } V_2 = b'_2 Y$$

, em que  $a_2$  e  $b_2$  são vetores de coeficientes de dimensões  $px1$  e  $qx1$ , respectivamente, obtidos de modo que a correlação entre as variáveis canônicas  $U_2$  e  $V_2$  seja máxima no conjunto das combinações lineares de  $X$  e  $Y$ , sendo não correlacionadas com o primeiro par de variáveis  $V_1$  e  $U_1$ . Além disso,  $a_2$  e  $b_2$  são tais que:

$$Var(U_2) = Var(V_2) = 1$$

De um modo geral, o  $k$ -ésimo par canônico é definido como sendo o par contendo as combinações lineares  $U_k$  e  $V_k$ , tais que:

$U_k = a'_k X$  e  $V_k = b'_k Y$ , em que  $a_k$  e  $b_k$  são vetores de coeficientes de dimensões  $px1$  e  $qx1$ , respectivamente, obtidos de modo que a correlação entre as variáveis canônicas  $U_k$  e  $V_k$  seja maximizada no conjunto das combinações lineares de  $X$  e  $Y$  que têm variâncias iguais a 1 e que são não correlacionadas com as variáveis canônicas dos  $k-1$  primeiros pares canônicos. A correlação entre as variáveis  $U_k$  e  $V_k$  é chamada de correlação canônica,  $k=1,2,\dots,\min(p,q)$ .

Os vetores de coeficientes (constantes)  $a_k$  e  $b_k$ ,  $k=1,2,\dots,\min(p,q)$  que satisfazem os critérios estabelecidos para a construção dos pares de variáveis canônicas são obtidos por meio de soluções do sistema de equações seguinte:

$$\begin{cases} \left( \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} - \lambda_k \Sigma_{XX} \right) a_k = 0 \\ \left( \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} - \lambda_k \Sigma_{YY} \right) b_k = 0 \end{cases}$$

, em que  $\lambda_k$  satisfaz as seguintes equações características:

$$\begin{cases} \left| \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \lambda_k \Sigma_{XX} \right| = 0 \\ \left| \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \lambda_k \Sigma_{YY} \right| = 0 \end{cases}$$

, ou seja,  $\lambda_k$  é o  $k$ -ésimo maior autovalor da matriz

$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$  ou, equivalentemente, da matriz

$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$  (Anderson, 2003).

A correlação canônica é a correlação em valor absoluto entre  $U_k$  e  $V_k$  e é igual a  $\sqrt{\lambda_k}$ , ou seja:

$$\rho_k^{*2} = \lambda_k = \left[ \text{Corr}(U_k, V_k) \right]^2 = \frac{(a'_k \Sigma_{XY} b_k)^2}{(a'_k \Sigma_{XX} a_k)(b'_k \Sigma_{YY} b_k)} \text{As variáveis canônicas também}$$

podem ser construídas para as variáveis padronizadas, ou seja, por meio da análise das matrizes de correlações das variáveis originais que estão nos vetores  $X$  e  $Y$ . Nesse caso, o sistema de equações torna-se:

$$\begin{cases} \left( P_{XY} P_{YY}^{-1} P_{YX} - \lambda_k P_{XX} \right) a_k = 0 \\ \left( P_{YX} P_{XX}^{-1} P_{XY} - \lambda_k P_{YY} \right) b_k = 0 \end{cases}$$

, enquanto que o sistema de equações seguinte torna-se:

$$\begin{cases} \left| P_{XY} P_{YY}^{-1} P_{YX} - \lambda_k P_{XX} \right| = 0 \\ \left| P_{YX} P_{XX}^{-1} P_{XY} - \lambda_k P_{YY} \right| = 0 \end{cases}$$

, em que  $P_{XX}$  e  $P_{YY}$  são as matrizes de correlações populacionais das variáveis dos vetores  $X$  e  $Y$ , respectivamente,  $P_{XY}$  é a matriz de correlações entre as variáveis que estão no vetor  $X$  e aquelas que estão no vetor  $Y$  e  $P_{YX} = P'_{XY}$ .

### Estimação de variáveis canônicas

Dada uma amostra aleatória de tamanho  $n$  dos vetores  $X$  e  $Y$ , as matrizes de covariâncias populacionais  $\Sigma_{XX}$ ,  $\Sigma_{YY}$ ,  $\Sigma_{XY}$  e  $\Sigma_{YX}$  são estimadas pelas correspondentes matrizes de covariâncias amostrais  $S_{XX}$ ,  $S_{YY}$ ,  $S_{XY}$  e  $S_{YX}$ , respectivamente e os sistemas de equações (6) e (7) são resolvidos utilizando-se essas matrizes amostrais.

Da mesma forma, as matrizes de correlações populacionais  $P_{XX}$ ,  $P_{YY}$ ,  $P_{XY}$  e  $P_{YX}$  são estimadas pelas respectivas matrizes de correlações amostrais  $R_{XX}$ ,  $R_{YY}$ ,  $R_{XY}$  e  $R_{YX}$ , sendo os sistemas de equações (9) e (10) resolvidos utilizando-se essas matrizes amostrais.

### Exemplo de Aplicação 1

Considere como ilustração uma análise de correlações canônicas feitas com base nas matrizes de correlações amostrais estimadas a partir de uma amostra de tamanho  $n=44$ . Para cada um dos indivíduos (genótipos) foram medidos os vetores  $X$  e  $Y$  definidos por:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \text{ e } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} \text{ constituídos das variáveis (caracteres) que representam dois grupos relativamente distintos.}$$

As matrizes das correlações amostrais dos vetores  $X$  e  $Y$  e as matrizes de correlações cruzadas são dadas por:

$$R_{XX} = \begin{bmatrix} 1 & 0,923 & 0,894 \\ 0,923 & 1 & 0,849 \\ 0,894 & 0,849 & 1 \end{bmatrix}$$

$$R_{YY} = \begin{bmatrix} 1 & 0,602 & 0,194 & 0,366 \\ 0,602 & 1 & 0,457 & 0,649 \\ 0,194 & 0,457 & 1 & 0,606 \\ 0,366 & 0,649 & 0,606 & 1 \end{bmatrix}$$

$$R_{XY} = \begin{bmatrix} 0,529 & 0,757 & 0,715 & 0,932 \\ 0,492 & 0,783 & 0,502 & 0,949 \\ 0,674 & 0,718 & 0,687 & 0,860 \end{bmatrix}$$

$$R_{YX} = R'_{XY}$$

Os autovalores obtidos por meio da solução da equação:

$$\left| R_{XY} R_{YY}^{-1} R_{YX} - \lambda_k R_{XX} \right| = 0, \text{ são: } \hat{\lambda}_1 = 0,9892, \hat{\lambda}_2 = 0,7531 \text{ e } \hat{\lambda}_3 = 0,1591.$$

Com base nesses autovalores são obtidos os vetores de coeficientes,  $a_k$  e  $b_k$  ( $k=1,2,3$ ) e os pares de variáveis canônicas seguintes:

$$U_1 = 0,4371Z_1 + 0,2087Z_2 + 0,3909Z_3$$

$$V_1 = 0,2563W_1 + 0,1031W_2 + 0,1846W_3 + 0,6417W_4$$

$$U_2 = -1,2654Z_1 + 2,4435Z_2 - 1,1139Z_3$$

$$V_2 = -0,6793W_1 + 0,6238W_2 - 1,1025W_3 + 0,7209W_4$$

$$U_3 = -2,7968Z_1 + 0,9514Z_2 + 1,9094Z_3$$

$$V_3 = 1,0198W_1 - 0,4017W_2 - 0,5963W_3 + 0,1138W_4$$

, em que  $Z_i = \frac{(X_i - \bar{X}_i)}{s_{X_i}}, i = 1, 2, 3$  e  $W_j = \frac{(Y_j - \bar{Y}_j)}{s_{Y_j}}, j = 1, 2, 3, 4$  são as variáveis originais na forma padronizada, com base em suas medidas descritivas.

Quadro 10.1. Medidas descritivas das variáveis originais

Variável	Média	Desvio Padrão
X <sub>1</sub>	24,76	1,87
X <sub>2</sub>	26,69	2,61
X <sub>3</sub>	25,69	1,19
Y <sub>1</sub>	5,65	1,78
Y <sub>2</sub>	7,00	1,65
Y <sub>3</sub>	5,27	1,08
Y <sub>4</sub>	15,00	5,53

A variável canônica  $U_1$  pode ser interpretada como um índice de desempenho e a variável canônica  $V_1$  como outro índice de desempenho em relação ao comportamento dos indivíduos avaliados.

Por exemplo, um indivíduo que tivesse os seguintes valores observados das variáveis nos vetores  $X$  e  $Y$ :  $x_1 = 26$ ,  $x_2 = 29$ ,  $x_3 = 27$ ,  $y_1 = 7$ ,  $y_2 = 9$ ,  $y_3 = 7$  e  $y_4 = 20$  teria seus valores padronizados iguais a:  $z_1 = 0,662$ ,  $z_2 = 0,885$ ,  $z_3 = 1,094$ ,  $w_1 = 0,758$ ,  $w_2 = 1,212$ ,  $w_3 = 1,606$  e  $w_4 = 0,904$ . Então, os valores numéricos (escores) para o primeiro par de variáveis canônicas seriam iguais a  $U_1 = 0,902$  e  $V_1 = 1,195$ .

A correlação canônica entre  $U_1$  e  $V_1$  é igual a  $\sqrt{0,9892} = 0,9945$ . Dessa forma,  $U_1$  seria a melhor combinação linear para ser utilizada na predição de  $V_1$  e vice-versa.

Como estratégia de classificação de novos indivíduos, cada indivíduo candidato poderia ser avaliado em relação aos caracteres do vetor  $Y$  e com base nestas medidas obtidas calcular o seu escore na variável canônica  $V_1$ . A classificação seria feita com base nos valores dessa nova variável. Como a correlação entre  $U_1$  e  $V_1$  é positiva, os indivíduos com maiores escores teriam melhores

desempenhos em relação aos caracteres do vetor  $X$ , dado que, pela análise de correlação canônica  $V_1$  é a combinação linear mais correlacionada com  $U_1$ .

As outras duas correlações canônicas são:  $\sqrt{0,7531} = 0,8678$  para  $U_2$  e  $V_2$  e  $\sqrt{0,1591} = 0,3988$  para  $U_3$  e  $V_3$ .

### Cargas canônicas e proporção da variância explicada pelas variáveis canônicas

As correlações das variáveis canônicas com as variáveis originais, chamadas de cargas canônicas, são definidas por:

$$R_{U_k X}^* = R_{XX} a_k ; R_{V_k Y}^* = R_{YY} b_k$$

$$R_{U_k Y}^* = R_{YX} a_k ; R_{V_k X}^* = R_{XY} b_k$$

O cálculo da proporção de variância total que é explicada pelas variáveis canônicas separadamente é feito da seguinte forma:

$$PVTE_{U_k} = \frac{\sum_{i=1}^p \text{Corr}(U_k, X_i)^2}{p} \times 100$$

$$PVTE_{V_k} = \frac{\sum_{j=1}^q \text{Corr}(U_k, Y_j)^2}{q} \times 100$$

### Exemplo de Aplicação 2

Com base nos resultados do Exemplo 1 tem-se:

Quadro 10.2. Variáveis canônicas do vetor  $X$

Variável	Cargas Canônicas			Cargas Canônicas		
	U <sub>1</sub>	U <sub>2</sub>	U <sub>3</sub>	V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>
X <sub>1</sub>	0,979	-0,005	-0,202	0,974	-0,004	-0,081
X <sub>2</sub>	0,944	0,329	-0,001	0,939	0,286	-0,001
X <sub>3</sub>	0,959	-0,170	0,227	0,954	-0,148	0,091
PVTE (%)	92,34	4,58	3,08			

Quadro 10.3. Variáveis canônicas do vetor  $Y$

Variável	Cargas Canônicas			Cargas Canônicas		
	V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>	U <sub>1</sub>	U <sub>2</sub>	U <sub>3</sub>
Y <sub>1</sub>	0,601	-0,254	0,705	0,598	-0,220	0,281
Y <sub>2</sub>	0,779	0,179	0,016	0,775	0,156	-0,007
Y <sub>3</sub>	0,689	-0,512	-0,510	0,686	-0,445	0,204
Y <sub>4</sub>	0,947	0,209	-0,131	0,942	0,182	-0,052
PVTE (%)	58,53	10,07	19,36			

### Teste de significância das correlações canônicas

Admitindo que os vetores  $X$  e  $Y$  são normais multivariados, as seguintes hipóteses são consideradas:

$$H_0^m : \rho_1^{*2} \neq 0, \rho_2^{*2} \neq 0, \dots, \rho_m^{*2} \neq 0, \rho_{m+1}^{*2} = 0, \dots, \rho_k^{*2} = 0$$

$H_a^m : \rho_i^{*2} \neq 0$ , para algum  $i \geq m+1$ , ou seja, testa-se se as  $m$  primeiras correlações canônicas são significativas e, portanto, que as variáveis canônicas correspondentes seriam as mais importantes para caracterização da informação dos dois conjuntos de dados analisados, sendo  $m < k = \min(p, q)$ .

O teste estatístico rejeita a hipótese nula se o valor numérico da estatística de teste, definida por:

$$\begin{aligned} & -\left(n-1-\frac{1}{2}(p+q+1)\right) \ln \left( \prod_{i=m+1}^k \left(1-\hat{\rho}_i^{*2}\right) \right) \\ &= -\left(n-1-\frac{1}{2}(p+q+1)\right) \ln \left( \prod_{i=m+1}^k \left(1-\lambda_i\right) \right) \end{aligned}$$

, for maior ou igual ao valor crítico de uma distribuição qui-quadrado com  $(p-m)(q-m)$  graus de liberdade e com o nível de significância adotado no teste.

### Exemplo de Aplicação 3

Considere novamente os dados do Exemplo 1.

As hipóteses de interesse são:

$$H_0^2 : \rho_1^{*2} \neq 0, \rho_2^{*2} \neq 0 \text{ contra } H_a^2 : \rho_3^{*2} \neq 0$$

O valor da estatística de teste é igual a:

$$-(44-1-\frac{1}{2}(3+4+1)) \ln(1-0,1591) = 6,76.$$

Para um nível de significância de 5% o valor crítico da distribuição qui-quadrado com  $(3-1)(4-1)=6$  graus de liberdade é igual a 12,59, concluindo-se pela não rejeição da hipótese nula. Portanto, as duas primeiras correlações canônicas e, consequentemente, os dois primeiros pares canônicos são os mais importantes para a análise realizada.

### Relação entre correlações canônicas e regressão linear múltipla

Considere o vetor aleatório  $\begin{bmatrix} Y & X_1 & X_2 & \dots & X_p \end{bmatrix}$ , onde  $Y$  é a variável resposta e  $\begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$  são as variáveis explicativas. Tendo-se uma amostra aleatória de tamanho  $n$  deste vetor, o objetivo, em regressão linear múltipla, é encontrar a combinação linear de  $\begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$  que tenha a maior correlação amostral com a variável resposta  $Y$ , ou seja, deseja-se encontrar a variável  $\hat{Y}$  que tenha máxima correlação com  $Y$ , sendo  $\hat{Y}$  denotada por:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Esse problema pode ser resolvido por meio de análise de correlações canônicas. Neste caso, haveria apenas um par de variáveis canônicas representadas por  $U=Y$  e  $V=\hat{Y}$ . A solução obtida utilizando-se a metodologia de análise de correlações canônicas, isto é a obtenção dos valores dos parâmetros  $(\beta_1, \beta_2, \dots, \beta_p)$ , seria a mesma encontrada pela metodologia de quadrados mínimos ordinários, e:

$$\begin{aligned} \left( \text{Corr}(Y, \hat{Y}) \right)^2 &= \frac{\sum_{j=1}^n (Y_j - \bar{Y})(\hat{Y}_j - \bar{Y})}{\left( \sum_{j=1}^n (Y_j - \bar{Y})^2 \right) \left( \sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2 \right)} \\ &= R^2 = \frac{SQModelo}{SQTotal} \end{aligned}$$

### Exemplo de Aplicação 4

Suponha que, no caso do Exemplo 1, haja interesse em construir um modelo de regressão linear considerando a variável  $X_1$  padronizada, ou seja,  $Z_1$ , como variável resposta e as variáveis do vetor  $Y$  padronizadas, ou seja,  $W_1, W_2, W_3$  e  $W_4$ .

O modelo de regressão linear múltipla ajustado pelo método dos quadrados mínimos ordinários fornece o resultado seguinte:

$$\hat{Z}_1 = 0,17W_1 + 0,13W_2 + 0,24W_3 + 0,65W_4$$

, com um coeficiente de determinação igual a  $R^2 = 0,9552$ .

Os resultados obtidos usando a análise de correlações canônicas para as matrizes de correlações dos vetores  $X = X_1$  e  $Y = \begin{pmatrix} Y_1 & Y_2 & Y_3 & Y_4 \end{pmatrix}'$  são semelhantes aos obtidos pela regressão linear múltipla, sendo que as variáveis canônicas nesse caso são:

$$U_1 = Z_1 \text{ e } V_1 = 0,17W_1 + 0,13W_2 + 0,24W_3 + 0,65W_4$$

, com correlação canônica ao quadrado igual a 0,9552.

Portanto, o melhor preditor da variável resposta  $U_1 = Z_1$  é a variável canônica  $V_1$ , que é a mesma solução encontrada por regressão linear múltipla. Logo, a regressão linear múltipla é um caso particular da análise de correlações canônicas.

### Orientações Gerais para Análise de Correlações Canônicas (ACC)

A palavra **canônica** significa: reduzida à forma mais simples ou mais clara possível. Significa também: o estado ou a maneira usual ou padrão de alguma coisa. **Correlação canônica** é um procedimento multivariado para analisar o relacionamento entre dois conjuntos de variáveis, cada conjunto podendo conter várias variáveis. A correlação canônica é uma variação do conceito de análise de regressão e correlação múltipla.

Em regressão e correlação múltipla analisa-se o relacionamento entre uma combinação linear de um conjunto de variáveis  $X$  e uma única variável  $Y$ . Na correlação canônica, analisa-se o relacionamento entre uma combinação linear do conjunto de variáveis  $X$  com uma combinação linear do conjunto de variáveis  $Y$ . Correlação simples e múltipla são casos especiais de correlação canônica nos quais um dos conjuntos ou ambos contêm uma única variável.

**O modelo da correlação canônica** é o seguinte: considere dois conjuntos de variáveis  $Y_1, Y_2, Y_3, \dots, Y_p$  e  $X_1, X_2, X_3, \dots, X_q$ . Constroem-se as combinações lineares seguintes:

$U = u_1 Y_1 + u_2 Y_2 + \dots + u_p Y_p$  e  $V = v_1 X_1 + v_2 X_2 + \dots + v_q X_q$ , determinando-se  $u_j$  e  $v_j$  de forma que  $r_{UV}$  seja um máximo. Em termos de modelo geral tem-se:

$$u_1 Y_1 + u_2 Y_2 + \dots + u_p Y_p = v_1 X_1 + v_2 X_2 + \dots + v_q X_q.$$

Cada uma das combinações lineares (somas ponderadas) anteriores são chamadas de **variáveis canônicas**. A correlação ao quadrado entre duas variáveis canônicas é chamada de **raiz canônica**. **Analisando-se os valores dos coeficientes** (pesos) nas variáveis canônicas, podem-se descrever variáveis latentes subjacentes aos conjuntos de combinações lineares (variáveis canônicas).

**Pode-se extrair mais de uma raiz canônica** de dois conjuntos de variáveis, pois se pode ter em um mesmo conjunto mais de uma **dimensão latente**. Então, o número possível de raízes canônicas extraídas será igual

ao menor número de variáveis em um dos dois conjuntos estudados. Quando se extrai mais de uma raiz, cada par sucessivo de variáveis canônicas explicará uma **proporção adicional única da variabilidade** existente nos dois conjuntos de variáveis iniciais. Os sucessivos pares de variáveis canônicas serão **ortogonais** entre si e com um poder de explicação da variabilidade cada vez menor.

A ACC é um procedimento exploratório, de redução de dimensionalidade, mas também é preditivo. Os verdadeiros métodos de redução de dimensionalidade fazem mais que isto – eles realmente substituem a **variável original** por uma **variável nova** que sumariza a **redundância** e a **variância compartilhada** nestas variáveis originais. No caso da ACC, ela não faz exatamente isto, mas ela calcula uma série de funções canônicas que sumarizam o relacionamento entre dois conjuntos de variáveis.

Então, na análise de correlação canônica o que se faz é calcular (estimar, extrair ou determinar) uma série de funções canônicas que faz o melhor trabalho de sumarização do relacionamento entre uma combinação linear de variáveis consideradas dependentes e uma combinação linear de variáveis consideradas independentes. Ela extraírá tantas funções quanto for o menor número de variáveis dentre os grupos, isto é, se há cinco variáveis independentes e três dependentes, serão obtidas três funções canônicas.

Cada função descreve uma quantidade menor de variação, isto é, a primeira função descreverá a maior parte da variação, então será computada outra função que trabalha na variância restante e assim por diante. Geralmente a função terciária é de uso e valor questionável. Pode-se extraí-la e um programa computacional o fará, mas isto não a torna significativa ou útil. Cada função tem um coeficiente de determinação associado com ela, e em geral estes reduzirão rapidamente após a primeira função.

**O procedimento geral da análise de correlação canônica é o seguinte:**

1. Cada conjunto de variáveis é combinado em uma função linear (variável ou dimensão canônica), de forma a maximizar a correlação entre as variáveis canônicas (par de variáveis canônicas);
2. Vários pares de variáveis canônicas, relacionando os dois conjuntos de variáveis, podem ser extraídos (até o número total de variáveis no menor conjunto);
3. Os pares de variáveis canônicas (pares canônicos) são extraídos em ordem decrescente de correlação entre as variáveis canônicas;
4. O segundo par canônico não é correlacionado com o primeiro e assim por diante, ou seja, os pares canônicos são ortogonais entre si;
5. O primeiro passo da análise consiste na obtenção da matriz de correlação canônica, dada por:  $R = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{YY}$ ;
6. A matriz de correlação canônica é então submetida a uma análise característica, para a extração dos autovalores e dos autovetores;
7. A análise característica redistribui a variação dos dados originais em poucas variáveis que são combinações lineares das variáveis originais;
8. Os autovalores indicam a proporção da variação total explicada em cada combinação linear, enquanto que os autovetores fornecem os coeficientes das variáveis em cada função linear;
9. Desta forma,  $\lambda_j = r_{Cj}^2$ , ou seja, o autovalor é igual ao coeficiente de correlação canônica elevado ao quadrado (raiz canônica);

10. Uma vez obtidos os autovalores, procede-se o cálculo dos coeficientes das variáveis canônicas (pesos canônicos).

A análise de correlação canônica pode ser interpretada da seguinte forma:

Os **autovalores** podem ser interpretados como a **proporção da variância explicada** pela correlação entre as respectivas variáveis canônicas (combinações lineares). Tomando-se a raiz quadrada dos autovalores têm-se os coeficientes de correlação canônica, ou seja, as **correlações canônicas**. Como índice geral de correlação canônica é comum considerar-se aquela extraída em primeiro lugar.

**Raiz canônica** é a correlação entre duas variáveis canônicas ao quadrado. A significância de cada uma das **raízes canônicas** pode ser testada e se elas forem significativas serão interpretadas. Primeiro são testadas todas as raízes em conjunto, depois se retira a primeira e testa-se novamente e assim por diante. Com isso avalia-se a falta de uma raiz para a significância do conjunto.

**Pesos canônicos** são os coeficientes das variáveis originais nas combinações lineares (variáveis canônicas) e podem ser interpretados em suas magnitudes. Para facilitar a interpretação, eles são **padronizados** (semelhante aos betas da Regressão Linear Múltipla e aos coeficientes dos escores fatoriais na Análise de Fatores). Podem ser calculados escores canônicos usando-se os pesos canônicos (padronizados), para utilização em análises posteriores.

A **estrutura canônica** pode ser interpretada com base na obtenção das correlações entre as variáveis canônicas e as variáveis originais, sendo estas correlações chamadas de **cargas canônicas**. As variáveis originais com altas cargas canônicas numa variável canônica têm mais em comum com a respectiva variável canônica. As **interpretações de pesos canônicos e cargas canônicas** podem ser vistas da forma seguinte: pode ocorrer que o peso canônico seja próximo de zero e a carga canônica seja alta, ou o contrário. Isso pode acontecer se houver multicolinearidade entre as variáveis originais. Lembre-se que os pesos canônicos, como as correlações parciais, representam as contribuições únicas das variáveis, enquanto que as cargas canônicas representam correlações gerais.

A **variância extraída** pelas variáveis canônicas pode ser interpretada como: os quadrados das cargas canônicas representam a proporção de variância explicada em cada variável. Para cada raiz canônica (quadrado da correlação entre duas variáveis canônicas) pode-se tomar a média dessas proporções para se ter uma idéia de quanta variabilidade está explicada (ou quanta variância é extraída), em média, pela respectiva variável canônica naquele conjunto de variáveis. Multiplicando-se a variância extraída de uma variável canônica pelo quadrado do coeficiente de correlação canônica obtém-se uma medida de **redundância**, que indica o quanto é redundante um conjunto de variáveis dada a presença do outro conjunto. A redundância é dada pela expressão seguinte:

$$Red_{Eesq} = \left( \sum c \arg a_{Eesq}^2 / p \right) r_c^2, \text{ em que:}$$

$Red_{Eesq}$  : redundância do conjunto de variáveis da esquerda;

$p$  : número de variáveis do conjunto da esquerda;

$r_c^2$  : coeficiente de correlação canônica ao quadrado.

A medida de **redundância** é útil para se acessar a **significância prática** das **raízes canônicas**. Ocorre que, com grandes amostras, correlações canônicas de  $r_C = 0,30$ , por exemplo, podem ser significativas.

Nesse caso  $r_C^2 = 0,09$ , o qual usado no cálculo da redundância irá mostrar a pouca importância dessa raiz canônica.

**Em resumo**, na interpretação da ACC considera-se o seguinte:

1. Correlações entre pares de variáveis canônicas;
2. Correlações entre variável canônica e variáveis originais do mesmo conjunto de variáveis;
3. Correlações entre variável canônica e variáveis originais de outro conjunto de variáveis;
4. Gráficos de escores canônicos;
5. Análise de redundância canônica.

**Cuidados na análise de correlação canônica:**

Verificar as pressuposições, que são: normalidade multivariada, linearidade, homocedasticidade na relação entre pares de variáveis. A normalidade é necessária por causa dos testes de significância. Entretanto, a análise é robusta para grandes amostras. Recomenda-se que o número de casos (observações) seja bem maior que o número de variáveis.

Associações não lineares entre variáveis são problemáticas. As variáveis dentro de conjuntos e entre conjuntos não devem ser muito correlacionadas (multicolinearidade). Os valores discrepantes (outliers) afetam a análise e devem ser examinados. Os testes das pressuposições podem, alternativamente, ser feitos nos escores das variáveis canônicas.

# CAPITULO 25

## Análise de Fatores

### Introdução

Como na Análise de Componentes Principais (ACP), aqui também são obtidas novas variáveis a partir das variáveis originais, mas de forma diferente, ou seja, dentro das variáveis originais é evidenciada uma estrutura latente de fatores (comuns às variáveis e específicos de cada variável). Na ACP é feita a redistribuição da variância das variáveis originais de tal forma que seja obtido um conjunto de eixos não correlacionados. Na ACP procura-se reduzir o conjunto de variáveis sem levar em conta a estrutura de relações entre as variáveis.

Na AFA procura-se obter um novo conjunto de variáveis (fatores) com base na estrutura de relações entre as variáveis. Pode-se dizer que a ACP é a análise da variação, enquanto que a AFA é a análise da covariação.

### Modelo da análise de fatores

Na análise de fatores (AFA) considera-se que a variação total observada de determinada variável tem origem em dois tipos de fatores: fatores que influenciam simultaneamente duas ou mais variáveis do conjunto, denominados de fatores comuns ou communalidades e fatores que contribuem para a variação de uma única variável do conjunto, denominados de fatores específicos ou unicidade (ou especificidade). Além destes fatores existe o termo erros ou variância residual. Então, Variação Total = Comunalidade + Especificidade + Erros.

Então, na AFA cada variável pode ser modelada como:

$$\begin{aligned} X_1 &= c_1 a_{11} + c_2 a_{21} + \cdots + c_r a_{r1} + E_1 \\ X_2 &= c_1 a_{12} + c_2 a_{22} + \cdots + c_r a_{r2} + E_2 \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \\ X_p &= c_1 a_{1p} + c_2 a_{2p} + \cdots + c_r a_{rp} + E_p. \end{aligned}$$

Em notação matricial tem-se:  $X = CA + E$ , em que:

$X$  : vetor das variáveis originais;

$C$  : vetor dos fatores comuns (comunalidades);

$A$  : matriz de conexão entre variáveis aleatórias e communalidades;

$E$  : vetor dos fatores específicos (especificidades).

Observe que na ACP são construídas combinações lineares das variáveis originais observadas e na AFA as variáveis originais são representadas por combinações lineares de um conjunto de fatores comuns não observados.

## Determinação ou extração dos fatores

Para extrair os fatores, primeiro deve-se estimar a comunalidade e a especificidade de cada variável. Tem-se que:

$$zz' = (n-1)R_X \text{ ou } R_X = \frac{1}{n-1}zz', \text{ em que:}$$

$z$ : matriz das variáveis originais na forma padronizada;

$n$ : número de observações (grupos, objetos, indivíduos);

$R_X$ : matriz de correlação.

Assim, pode-se representar cada variável  $z_j$  por  $z_j = c_1 a_{1j} + c_2 a_{2j} + \dots + c_r a_{rj} + e_j$ , em

que:

$z_j$ : variável na forma padronizada (ou escore reduzido de  $X_j$ )

$a_{kj}$ : coeficiente de conexão entre o fator comum de ordem  $k$  e a variável de ordem  $j$ , sendo

$k=1,2,\dots,r$  ( $r < p$ ) e  $j=1,2,\dots,p$ ;

$c_k$ : fator comum (comunalidade) da variável de ordem  $k$ ;

$e_j$ : fator específico (especificidade) da variável de ordem  $j$ .

Em notação matricial, para cada variável tem-se:

$$z_j = CA_j + e_j, \text{ em que:}$$

$CA_j$ : parte do escore reduzido de  $X_j$  (ou variável padronizada  $z_j$ ) explicada por fatores comuns;

$e_j$ : parte devida ao fator específico do escore reduzido de  $X_j$ .

A equação matricial geral é expressa como:  $z = CA + E$ . A expressão  $z'z = A'C'CA + E'E$  significa que a variância comum é decomposta em fatores estruturais ortogonais entre si e padronizados.

A matriz diagonal, que representa a parcela da variância que é específica de cada variável é dada

$$\text{por: } D_e = \frac{1}{n-1}E'E, \text{ em que:}$$

$D_e$ : matriz de dispersão dos fatores de especificidade;

$$E'E = \begin{bmatrix} \sum_{i=1}^n e_{i1}^2 & 0 & \dots & 0 \\ 0 & \sum_{i=1}^n e_{i2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_{i=1}^n e_{ip}^2 \end{bmatrix}.$$

Tem-se que:  $z'z = A'(n-1)IA + E'E$ . Então, segue que

$$z'z = A'(n-1)IA + (n-1)D_e \text{ ou } z'z = (n-1)[A'A + D_e].$$

Sabe-se também que:  $z'z = (n-1)R_X$ . Logo, segue que

$$(n-1)R_X = (n-1)[A'A + D_e] \text{ e } R_X = A'A + D_e, \text{ ou seja:}$$

$A'A = R_X - D_e$ , em que:  $A$ : matriz dos coeficientes de conexão, a serem estimados.

A expressão anterior pode ser reescrita como:  $R^* = R_X - D_e$ , em que:

$R^*$ : matriz de correlação modificada;

$R_X$ : matriz de correlação das variáveis originais;

$D_e$ : matriz de dispersão dos fatores de especificidade.

A matriz  $R^*$  é obtida a partir da matriz  $R_X$  substituindo-se os valores 1,0 da diagonal principal de

$R_X$  por estimativas de comunalidades das variáveis originais. Note então que na AFA é feita a diagonalização de uma matriz de correlação, modificada pelas comunalidades, ou seja, valores 1,0 –

$$\text{Especificidade} = 1 - \frac{1}{n-1} \sum_{i=1}^n e_{ij}^2.$$

Desta forma, a AFA admite a existência de uma estrutura latente (oculta) num conjunto de variáveis originais observadas, que é representada por fatores comuns às variáveis. Admite também que a determinação dos fatores comuns (ou fatores estruturais) é de grande importância no estudo de um conjunto de variáveis. No processo de obtenção (extração) de fatores, primeiro deve-se estimar a comunalidade de cada variável, que é a parte da variação de cada variável que é comum às outras variáveis. Para estimar a comunalidade de cada variável  $X_j$ , pode ser utilizado o método dos

coeficientes de correlações múltiplas de acordo com os seguintes passos:

- a) Toma-se como estimativa da comunalidade o quadrado do coeficiente de correlação múltipla de cada variável com as demais. Por exemplo, para a variável  $X_1$  é obtido como:

$$r_1^2(2,3,\dots,p) = 1 - \frac{|R_X|}{|R_1|}, \text{ em que:}$$

$r_1^2(2,3,\dots,p)$ : quadrado do coeficiente de correlação múltipla entre a variável  $X_1$  e as demais variáveis;

$|R_X|$ : determinante da matriz de correlação;

$|R_1|$ : menor de  $r_{11}$ , que é o determinante da matriz obtida de  $R_X$  eliminando-se a primeira linha e a primeira coluna;

- b) Substituem-se os valores 1,0 da diagonal principal da matriz  $R_X$  pelos quadrados dos coeficientes de correlação estimados, obtendo-se a matriz de correlação modificada  $R^*$ . Os quadrados dos coeficientes de correlações múltiplas são as estimativas das comunalidades.

Na estimação de comunalidades, pode ser usado também o método da raiz quadrada – esse método tem como base as raízes quadradas dos autovalores de  $R^*$  e estabelece uma base para o cálculo da matriz dos coeficientes de conexão.

Observe que, se  $V = [V_1 | V_2 | \cdots | V_r]$  é a matriz ortogonal formada pelos autovetores de  $R^*$ , então tem-se:  $R^* = V \Delta V'$  e  $\Delta = V' R^* V$ , em que:

$\Delta$ : matriz diagonal formada pelos autovalores de  $R^*$ . Dado que  $A'A = R^*$  e  $\Delta = V' R^* V$ , então:  $V'(A'A)V = \Delta$  e  $A = \Delta^{1/2} V'$ .

Sendo  $V$  uma matriz de transformação linear que representa rotação rígida de eixos, tem-se:

$$\text{tr}(A'A) = \text{tr}(\Delta) = \sum_{k=1}^r \lambda_k = \text{tr}(R^*), \text{ em que:}$$

$\lambda_k$ : autovalor de  $R^*$ .

Por exemplo, considerando que os fatores  $[C_1, C_2]$  representam a estrutura de relações de  $[X_1, X_2, X_3]$ , as matrizes  $A$  e  $A'A$  podem ser explicitadas como:

$$\begin{bmatrix} z_1, z_2, z_3 \end{bmatrix} = \begin{bmatrix} C_1, C_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \text{ e}$$

$$A'A = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \text{ ou}$$

$$A'A = \begin{bmatrix} a_{11}^2 + a_{21}^2 & \cdots & a_{11}a_{13} + a_{21}a_{23} \\ \vdots & a_{12}^2 + a_{22}^2 & \vdots \\ a_{13}a_{11} + a_{23}a_{21} & \cdots & a_{13}^2 + a_{23}^2 \end{bmatrix}$$

Note que na diagonal de  $A'A$  estão expressões do tipo  $\sum_{j=1}^r a_{jk}^2$ , que correspondem às variâncias de cada  $CA_k$  e, portanto, indicam que a communalidade de cada variável pode ser calculada como a

soma de seus coeficientes de conexão, ou seja:  $h_k^2 = \sum_{j=1}^r a_{jk}^2$ .

### Observações sobre Análise de Fatores

1) A decisão sobre quais fatores considerar não deve basear-se apenas em testes estatísticos, mas também na natureza dos dados, o que enfatiza a importância de formulação prévia de hipóteses sobre a estrutura latente;

2) Na análise de componentes principais (ACP) os componentes e as variáveis originais são características dos indivíduos (ou objetos) observados, enquanto que na análise de fatores (AFA), os fatores são características das variáveis analisadas, os quais procuram evidenciar a estrutura latente das relações entre as mesmas;

3) Na ACP procura-se determinar a verdadeira dimensionalidade do espaço multivariado e redistribuir as “observações” das variáveis no sistema de eixos principais. Na AFA calculam-se os fatores com base na matriz de correlação modificada,  $R^*$ , e o primeiro fator corresponde à maior parcela da variância que as variáveis têm em comum. O segundo fator representa a segunda maior parcela da variância que é comum às variáveis e assim sucessivamente;

4) Na AFA os eixos do espaço comum de variação podem não representar adequadamente a estrutura das relações entre variáveis devido à presença de erros nas variáveis e também porque

não existe razão para admitir “a priori” que os eixos de variância máxima nesse espaço representem efetivamente a estrutura latente que se procura evidenciar;

5) A rotação de fatores e o número de fatores são aspectos estreitamente relacionados na AFA porque se referem à interpretação de resultados.

### Rotação de Eixos na Análise de Fatores (AFA)

- 1) Princípio da Estrutura Simples – apesar de os fatores influenciarem simultaneamente todas as variáveis, um pequeno número deles influencia marcadamente uma ou poucas variáveis. O que diferencia as várias modalidades de rotação, baseadas neste princípio, é a forma como se procura fazer evidenciar a chamada estrutura simples.

#### 2) Modalidades de Rotação de Fatores

2.1) Ortogonais – procura-se evidenciar a estrutura latente simples mantendo a ortogonalidade entre os fatores;

2.2) Oblíquas – admite-se a possibilidade de correlações entre os fatores, mas correlações altas dificultam a interpretação da análise.

As rotações ortogonais e oblíquas são classificadas como:

- a) Ortogonais: Varimax, Quartimax e Equimax;
- b) Oblíquas: Oblimin, Promax e Tandem.

A rotação Varimax – é a mais utilizada; tem como base a comunalidade de uma variável, que é igual à soma de quadrados dos coeficientes de conexão da variável com os fatores, ou seja,

$$h_k^2 = \sum_{j=1}^r a_{jk}^2, \text{ em que:}$$

$r$  : número de fatores, sendo  $r < p$ ;

$p$  : número de variáveis;

A parcela de comunalidade da variável de ordem  $k$  devida ao fator de ordem  $j$  é igual a  $\frac{a_{jk}^2}{h_k^2}$ . A

parcela média de comunalidade devida ao fator de ordem  $j$  é igual a  $\frac{1}{p} \sum_{k=1}^p a_{jk}^2 / h_k^2$ , em que:

$p$  é o número de variáveis.

A variância da parcela de comunalidade devida ao fator de ordem  $j$  é igual a

$$\frac{1}{p} \sum_{k=1}^p \left( \frac{a_{jk}^2}{h_k^2} \right)^2 - \frac{1}{p^2} \left[ \sum_{k=1}^p \left( \frac{a_{jk}^2}{h_k^2} \right) \right]^2. \text{ A variância total da parcela de comunalidade de}$$

cada variável que é atribuída a cada fator, para todas as  $p$  variáveis e todos os  $r$  fatores é igual a

$$\frac{1}{p} \sum_{j=1}^r \left\{ \sum_{k=1}^p \left( \frac{a_{jk}^2}{h_k^2} \right)^2 - \frac{1}{p} \left[ \sum_{k=1}^p \left( \frac{a_{jk}^2}{h_k^2} \right) \right]^2 \right\}$$

e a rotação varimax consiste em maximizar

essa variância. A rotação varimax tem maior uso porque objetiva facilitar a interpretação dos fatores em termos das variáveis originais.

Os tipos de rotação ortogonal podem ser considerados como critérios de obtenção de uma matriz  $V$  de rotação rígida a ser aplicada sobre a configuração inicial dos fatores.

### Exemplo de Aplicação 1

Considere os escores reduzidos de três variáveis ( $X_1$ ,  $X_2$  e  $X_3$ ):

Observação	$Z_1$	$Z_2$	$Z_3$
1	0,24	-0,96	-1,26
2	0,46	1,62	0,70
3	-0,42	-0,59	-0,42
4	1,13	1,25	0,14
5	-2,18	0,88	-1,26
6	-0,20	0,52	-0,42
7	1,13	-0,22	1,82
8	0,68	-0,59	0,98
9	-0,86	-1,33	-0,70
10	0,02	-0,59	0,42

Obtenção da matriz de correlações  $(R_X)$ :

$$z'z = \begin{bmatrix} 8,99 & 0,635 & 6,52 \\ & 9,01 & 1,14 \\ & & 8,98 \end{bmatrix}$$

$$R_X = \frac{1}{n-1} z'z = \frac{1}{9} \begin{bmatrix} 8,99 & 0,635 & 6,52 \\ & 9,01 & 1,14 \\ & & 8,98 \end{bmatrix} = \begin{bmatrix} 1,00 & 0,07 & 0,72 \\ & 1,00 & 0,13 \\ & & 1,00 \end{bmatrix}$$

### Obtenção das estimativas de communalidades

Comunalidade é a parte da variação observada de uma variável que é comum às outras variáveis do conjunto. A estimativa da communalidade de cada variável é tomada como sendo o coeficiente de determinação da regressão linear múltipla de cada variável sobre as demais, ou o quadrado do coeficiente de correlação múltipla da variável em relação às demais.

O coeficiente da determinação múltiplo é calculado pela expressão seguinte:

$r_1^2(2,3,4,\dots,p) = 1 - \frac{|R_X|}{|R_1|}$ , sendo  $r_1^2(2,3,4,\dots,p)$ , o quadrado do coeficiente de correlação múltipla entre a variável  $X_1$  e as demais,  $|R_X|$ , o determinante da matriz de correlação e  $|R_1|$ , o determinante do menor de  $R_{11}$ , ou seja, o determinante da matriz obtida eliminando-se a primeira linha e a primeira coluna de  $R_X$ .

Então se tem:

$$r_1^2(2,3) = 1 - \frac{0,473}{0,983} = 0,52$$

$$r_2^2(1,3) = 1 - \frac{0,473}{0,482} = 0,02$$

$$r_3^2(1,2) = 1 - \frac{0,473}{0,995} = 0,52$$

Desta forma, a matriz de correlação modificada,  $R^*$  é:

$$R^* = \begin{bmatrix} 0,52 & 0,07 & 0,72 \\ & 0,02 & 0,13 \\ & & 0,52 \end{bmatrix}$$

Nesta matriz, as estimativas de comunalidades, são os coeficientes de determinação da regressão linear múltipla de cada variável sobre as outras duas.

#### Obtenção dos autovalores

Observe que, o polinômio característico de  $R_X$  é

$$\lambda^3 - 3\lambda^2 + 2,4598\lambda - 0,4729 = 0, \text{ cujas raízes}$$

características (autovalores) são:  $\lambda_1 = 1,746$ ;  $\lambda_2 = 0,976$ ;  $\lambda_3 = 0,278$

O polinômio característico de  $R^*$  é

$$\lambda^3 - 1,06\lambda^2 - 0,249\lambda + 0,00320 = 0, \text{ cujas raízes}$$

características (autovalores) são:  $\lambda_1 = 1,256$ ;  $\lambda_2 = 0,003$ ;  $\lambda_3 = -0,199$ .

Note que na matriz de correlação original têm-se duas raízes bem diferentes de zero, enquanto que na matriz modificada apenas uma raiz é bem diferente de zero.

## Exemplo de Aplicação 2

Dada a seguinte matriz  $R^*$ ,

$$R^* = \begin{bmatrix} 1,0 & 0,3 & -0,2 \\ 0,3 & 0,9 & -0,3 \\ -0,2 & -0,3 & 1,0 \end{bmatrix}, \text{ encontre a sua estrutura latente.}$$

1º) Determine os autovalores e suas respectivas raízes quadradas:

O polinômio característico é:

$$\lambda^3 - 2,9\lambda^2 + 2,58\lambda - 0,72 = 0, \text{ cujas raízes}$$

$$(\text{autovalores}) \text{ são } \lambda_1 = 1,5; \lambda_2 = 0,8 \text{ e } \lambda_3 = 0,6.$$

As raízes quadradas dos autovalores são:

$$\lambda_1^{1/2} = 1,2247; \lambda_2^{1/2} = 0,8944 \text{ e } \lambda_3^{1/2} = 0,7746. \text{ Logo tem-se:}$$

$$\Delta^{1/2} = \begin{bmatrix} 1,2247 & 0 & 0 \\ 0 & 0,8944 & 0 \\ 0 & 0 & 0,7746 \end{bmatrix}$$

$$2º) \text{ Construa } V = [V_1 \mid V_2 \mid V_3]:$$

$$V = \begin{bmatrix} 0,577 & 0,707 & 0,408 \\ 0,577 & 0,000 & -0,816 \\ -0,577 & 0,707 & -0,408 \end{bmatrix} \quad e \quad |V| = 0,9986 \cong 1$$

3º) Calcule  $A = \Delta^{1/2} V'$ , sendo  $A$  a matriz dos coeficientes de conexão:

$$A = \begin{bmatrix} 1,2247 & 0 & 0 \\ 0 & 0,8944 & 0 \\ 0 & 0 & 0,7746 \end{bmatrix} \begin{bmatrix} 0,577 & 0,577 & -0,577 \\ 0,707 & 0,000 & 0,707 \\ 0,408 & -0,816 & -0,408 \end{bmatrix}$$

$$A = \begin{bmatrix} 0,707 & 0,707 & -0,707 \\ 0,632 & 0,000 & 0,632 \\ 0,316 & -0,632 & -0,316 \end{bmatrix}$$

Portanto, a estrutura latente (subjacente) nas variáveis é representada pelas equações:

$$\begin{aligned} z_1^* &= 0,707C_1 + 0,632C_2 + 0,316C_3 \\ z_2^* &= 0,707C_1 + 0,000C_2 - 0,632C_3 \\ z_3^* &= -0,707C_1 + 0,632C_2 - 0,316C_3 \end{aligned}$$

### Exemplo de Aplicação 3

Se a estrutura de relações entre três variáveis for representada por três fatores comuns, como a seguir:

$$\begin{aligned} z_1^* &= 0,707C_1 + 0,632C_2 + 0,316C_3 \\ z_2^* &= 0,707C_1 + 0,000C_2 - 0,632C_3 \\ z_3^* &= -0,707C_1 + 0,632C_2 - 0,316C_3 \end{aligned}$$

Calcule a communalidade de cada variável. A communalidade de cada variável é:

$$h_1^2 = \sum_{j=1}^3 a_{j1}^2 = (0,707)^2 + (0,632)^2 + (0,316)^2 = 0,999$$

$$h_2^2 = \sum_{j=1}^3 a_{j2}^2 = (0,707)^2 + (0,000)^2 + (-0,632)^2 = 0,899$$

$$h_3^2 = \sum_{j=1}^3 a_{j3}^2 = (-0,707)^2 + (0,632)^2 + (-0,316)^2 = 0,999$$

Observe que ocorrem algumas relações importantes:

No exercício anterior tem-se  $\text{tr}(R^*) = 1,0 + 0,9 + 1,0 = 2,9$ , onde  $R^*$  é a matriz de

correlação modificada que evidenciou a estrutura latente apresentada no exercício. Tem-se também

que:  $\text{tr}(\Delta) = 1,5 + 0,8 + 0,6 = 2,9$ , onde  $\Delta = \Delta^{1/2} \Delta^{1/2}$  e  $\Delta^{1/2}$  = matriz das raízes

quadradas dos autovalores, o que implica que  $\Delta$  é a matriz dos autovalores de  $R^*$ . Tem-se ainda

que:  $\text{tr}(A'A) = \sum_k h_k^2 = 0,999 + 0,899 + 0,999 \approx 2,9$ .

Sabe-se que a parte da variação total observada que é comum às variáveis é representada por:

$A'C'CA$ , sendo  $C'C = (n-1)I$  (porque os fatores comuns são ortogonais entre si e

padronizados), o que significa que a variação comum às variáveis é então expressa por:

$(n-1)A'A$ . Então, os vetores-colunas que compõem  $CA$  e que representa o que cada variável

tem em comum com as demais (note que  $z = CA + E$ ), têm variância expressa como:

$$\frac{1}{(n-1)}(CA)'CA = \frac{1}{(n-1)}A'C'CA = A'A, \text{ porque } C'C = (n-1)I.$$

Uma vez que  $A'A = R^*$  e que  $V'R^*V = \Delta$ , implica que  $V'(A'A)V = \Delta$  e, como  $V$  é uma matriz de transformação linear que representa rotação rígida, tem-se que:

$$tr(A'A) = tr(\Delta) = \sum_{k=1}^r \lambda_k = tr(R^*), \text{ onde } \lambda_k \text{ é um valor característico (autovalor) de } R^*.$$

#### Exemplo de Aplicação 4

Considere novamente as variáveis  $X_1$ ,  $X_2$  e  $X_3$ . As raízes (autovalores) da matriz de correlação modificada foram calculadas como  $\lambda_1 = 1,256$ ;  $\lambda_2 = 0,003$  e  $\lambda_3 = -0,199$ . Desconsidere as duas últimas raízes e determine o vetor característico correspondente a  $\lambda_1$ . Então, tem-se:

$$(R^* - 1,256I) = \begin{bmatrix} -0,736 & 0,07 & 0,72 \\ 0,07 & -1,236 & 0,13 \\ 0,72 & 0,13 & -0,736 \end{bmatrix}$$

$$Adj(R^* - 1,256I) = \begin{bmatrix} 0,893 & - & - \\ 0,145 & - & - \\ 0,899 & - & - \end{bmatrix} e$$

$$V = \begin{bmatrix} 0,700 \\ 0,114 \\ 0,705 \end{bmatrix}$$

Neste caso, a matriz  $\Delta^{1/2}$  se reduz a um número (escalar), ou seja,  $256^{0,5} = 1,121$ .

$$\text{Então: } A = \Delta^{1/2} V' = 1,121 [0,700; 0,114; 0,705]$$

$$A = [0,785; 0,128; 0,790]$$

Portanto, as variáveis originais teriam um fator comum subjacente e poderiam ser escritas como:

$$z_1 = 0,785C_1 + C_1 + e_1$$

$$h_1^2 = (0,785)^2 = 0,62$$

$$z_2 = 0,128C_1 + C_2 + e_2$$

$$h_2^2 = (0,128)^2 = 0,02$$

$$z_3 = 0,790C_1 + C_3 + e_3$$

$$h_3^2 = (0,790)^2 = 0,62$$

Nesse caso,  $\text{tr}(\Delta) = \text{tr}(A'A) = \sum_k h_k^2 \neq \text{tr}(R^*)$ , porque foram desconsiderados dois

autovalores de  $R^*$  que eram próximos de zero.

### Análise de Fatores Ortogonal usando o Método dos Componentes Principais

#### Descrição do modelo

Considere o vetor  $X = (X_1, X_2, \dots, X_p)'$  de variáveis aleatórias originais (observadas), com vetor de médias  $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$ , matriz de covariâncias  $\Sigma_{p \times p}$  e matriz de correlações  $R_{p \times p}$ . Considere ainda as variáveis originais padronizadas  $Z_j = [(X_j - \mu_j)/\sigma_j]$ , onde  $\mu_j$  e  $\sigma_j$  são, respectivamente, a média e o desvio padrão da variável  $X_j$ ,  $j = 1, 2, \dots, p$ .

Neste caso, a matriz  $R_{p \times p}$  é a matriz de covariâncias do vetor  $Z = (Z_1, Z_2, \dots, Z_p)'$  de variáveis aleatórias padronizadas (escores reduzidos das variáveis aleatórias originais observadas). O modelo de análise de fatores, construído com base na matriz  $R_{p \times p}$ , combina linearmente as variáveis padronizadas e os  $r$  fatores comuns (desconhecidos). Então, o modelo ortogonal de fatores é expresso pelo sistema de equações seguinte:

$$\begin{aligned}
Z_1 &= a_{11}C_1 + a_{12}C_2 + \dots + a_{1r}C_r + E_1 \\
Z_2 &= a_{21}C_1 + a_{22}C_2 + \dots + a_{2r}C_r + E_2 \\
&\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \quad \vdots \\
Z_p &= a_{p1}C_1 + a_{p2}C_2 + \dots + a_{pr}C_r + E_p
\end{aligned}$$

Em notação matricial o modelo pode ser expresso como:

$$Z = D(X - \mu) = AC + E, \text{ onde:}$$

$$\begin{aligned}
D_{p \times p} &= \begin{bmatrix} 1/\sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/\sigma_p \end{bmatrix}; \\
(X - \mu)_{p \times 1} &= \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix}; \quad A_{p \times r} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pr} \end{bmatrix}; \quad E_{p \times 1} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_p \end{bmatrix}; \\
C_{r \times 1} &= \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_r \end{bmatrix}, \text{ em que:}
\end{aligned}$$

$C_{r \times 1}$ : vetor aleatório contendo  $r$  fatores ou variáveis latentes (não observáveis),  $1 \leq r \leq p$ ; as variáveis  $C_k$ ,  $k = 1, 2, \dots, r$  devem ser estimadas;

$E_{p \times 1}$ : vetor aleatório que corresponde aos erros de medição e à variação de  $Z_j$  que não é explicada pelos fatores comuns  $C_k$  incluídos no modelo;

$A_{p \times r}$ : matriz de posto  $r \leq p$ , que contém os coeficientes  $a_{kj}$  a serem estimados;

$a_{kj}$ : chamados de cargas fatoriais; é o coeficiente da  $j$ -ésima variável padronizada  $Z_j$  no  $k$ -ésimo fator  $C_k$  e representa o grau de relacionamento linear entre  $Z_j$  e  $C_k$ .

Desta forma, neste modelo é considerado que a informação contida nas variáveis originais padronizadas  $(Z_1, Z_2, \dots, Z_p)$  é representada por  $(p+r)$  variáveis aleatórias não observáveis, ou seja, fatores específicos  $(E_1, E_2, \dots, E_p)$  e fatores comuns  $(C_1, C_2, \dots, C_r)$ . Então, o objetivo da análise de fatores é estimar as  $r$  variáveis latentes (fatores), interpretá-las e calcular seus escores.

No processo de estimação com base no modelo de fatores ortogonal são necessárias as seguintes suposições:

(1)  $E(C_{r \times 1}) = 0$ , o que implica que  $E(C_k) = 0$ ,  $k = 1, 2, \dots, r$ , ou seja, todos os fatores  $C_k$  têm média igual à zero;

(2)  $Var(C_{r \times 1}) = I_{r \times r} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ , ou seja, todos os fatores comuns  $C_k$  têm variâncias iguais a 1 e são não correlacionados;

(3)  $E(E_{p \times 1}) = 0$ , o que implica que  $E(E_j) = 0$ ,  $j = 1, 2, \dots, p$ , ou seja, todos os fatores específicos  $E_j$  têm média igual à zero;

(4)  $Var(E_{p \times p}) = \varepsilon = \begin{bmatrix} \varepsilon_1 & 0 & \cdots & 0 \\ 0 & \varepsilon_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \varepsilon_p \end{bmatrix}$ , ou seja,

$Var(E_j) = \varepsilon_j$  e  $Cov(E_j, E_{j'}) = 0$ , para todo  $j \neq j'$ , o que significa que os fatores específicos têm variâncias diferentes e são não correlacionados;

(5) Os vetores  $E_{p \times 1}$  e  $C_{r \times 1}$  são independentes entre si, ou seja,

$Cov(E_{p \times 1}, C_{r \times 1}) = E(EC') = 0$ , o que significa que os vetores  $E$  e  $C$  representam duas fontes de variação distintas, relacionadas com as variáveis originais padronizadas.

O modelo de análise de fatores sob as suposições (1) a (5) é denominado de ortogonal porque os fatores são ortogonais entre si. Uma vez assumido este modelo, a matriz de correlações  $R_{p \times p}$

pode ser reparametrizada da seguinte forma:

$$R_{p \times p} = Var(Z) = Var(AC + E) = Var(AC) + Var(E)$$

$$Var(AC) + Var(E) = AIA' + \varepsilon = AA' + \varepsilon, \text{ logo } R_{p \times p} = AA' + \varepsilon.$$

O objetivo da análise de fatores é obter as matrizes  $A_{p \times r}$  e  $\varepsilon_{p \times p}$  que possam representar a matriz  $R_{p \times p}$  para um dado valor de  $r$  (número de fatores comuns) menor do que  $p$  (número de variáveis originais). Esta decomposição da matriz  $R_{p \times p}$  pode ser visualizada por meio das matrizes envolvidas, da seguinte forma:

$$R_{p \times p} = \begin{bmatrix} \sum_{k=1}^r a_{1k}^2 & \sum_{k=1}^r a_{1k} a_{k2} & \cdots & \sum_{k=1}^r a_{1k} a_{kp} \\ \sum_{k=1}^r a_{2k} a_{k1} & \sum_{k=1}^r a_{2k}^2 & \cdots & \sum_{k=1}^r a_{2k} a_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^r a_{pk} a_{k1} & \sum_{k=1}^r a_{pk} a_{k2} & \cdots & \sum_{k=1}^r a_{pk}^2 \end{bmatrix} +$$

$$+ \begin{bmatrix} \varepsilon_1 & 0 & \cdots & 0 \\ 0 & \varepsilon_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \varepsilon_p \end{bmatrix}$$

Com base na decomposição  $R_{p \times p} = AA' + \varepsilon$  podem-se realizar as deduções seguintes:

$$(1) \quad Var(Z_j) = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jr}^2 + \varepsilon_j, \text{ fazendo}$$

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jr}^2, \quad j = 1, 2, \dots, p, \text{ obtém-se}$$

$$Var(Z_j) = h_j^2 + \varepsilon_j.$$

A parte da variância de  $Z_j$  denotada por  $h_j^2$  é a variação de  $Z_j$  explicada pelos  $r$  fatores incluídos no modelo da análise de fatores, sendo chamada de comunalidade, devido ao fato de as variáveis  $Z_j$  terem uma fonte de variação em comum representada pelos fatores comuns  $C_k$ ,  $k=1,2,\dots,r$ . A parte da variância de  $Z_j$  denotada por  $\varepsilon_j$  é a variação específica de cada variável  $Z_j$ , sendo chamada de especificidade.

Como as variáveis  $Z_j$  têm variâncias iguais a 1, implica que  $h_j^2 + \varepsilon_j = 1$ .

$$(2) Cov(Z_j, Z_{j'}) = a_{j1}a_{j'1} + a_{j2}a_{j'2} + \dots + a_{jr}a_{j'r}, \text{ com } j, j' = 1, 2, \dots, p, j \neq j'.$$

$$(3) Cov(Z, C) = A_{p \times r}, \text{ e então}$$

$$Cov(Z_j, C_k) = Corr(Z_j, C_k) = a_{jk}; j = 1, 2, \dots, p; k = 1, 2, \dots, r$$

Observe que se pode utilizar a matriz  $A_{p \times r}$  para a interpretação dos fatores  $C_k$ ,  $k = 1, 2, \dots, r$ .

(4) A proporção da variância total explicada ( $PVE$ ) pelo fator  $C_k$  é obtida por meio da expressão seguinte:

$$PVE_{C_k} = \frac{\sum_{j=1}^p a_{jk}^2}{p}. \text{ Os fatores mais importantes no modelo são aqueles com maiores valores de } PVE.$$

Estimação do número de fatores ( $r$ ):

Para a estimação de  $r$  deve-se extrair os autovalores da matriz  $R_{p \times p}$  e ordená-los de forma decrescente. Em seguida estima-se o número de fatores  $r$  a serem retidos na análise, com base no valor numérico dos autovalores, utilizando os seguintes critérios:

(1) São considerados importantes aqueles autovalores que apresentam as maiores proporções

$$\text{da variância total das variáveis originais, obtidas como: } PVE_{\lambda_j} = \frac{\hat{\lambda}_j}{p},$$

$j = 1, 2, \dots, p$ , sendo o valor de  $r$  igual ao número de autovalores extraídos na análise;

- (2) O número de fatores retidos  $r$  é igual ao número de autovalores estimados  $\hat{\lambda}_j$  maiores ou iguais a 1;
- (3) O número de fatores retidos  $r$  é igual ao número de autovalores anteriores ao ponto que representa um decréscimo de importância dos autovalores em relação à variância total no gráfico *scree*, que dispõe os valores  $\hat{\lambda}_j$  em ordem decrescente.

Método dos componentes principais para estimação das matrizes  $A_{p \times r}$  e  $\varepsilon_{p \times p}$ :

Este método recebeu este nome simplesmente porque tem como base a utilização das matrizes de autovalores e autovetores, obtidas por meio da decomposição espectral da matriz  $R_{p \times p}$ , para a obtenção das cargas fatoriais. Porém, esta denominação tem causado confusão entre a análise de fatores e a análise de componentes principais.

A matriz  $R$  pode ser decomposta como:

$$R = V\Lambda V' = V\Lambda^{1/2}\Lambda^{1/2}V = AA'$$

onde

$$V_{p \times p} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix}$$

é a matriz constituída pelos autovetores de  $R$ , que pode

ser representada em suas colunas por  $V = [v_1 \ v_2 \ \cdots \ v_p]$ , e

$$\Lambda_{p \times p} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

é a matriz diagonal dos autovalores de  $R$ .

A matriz de cargas fatoriais é expressa como:

$$A_{p \times p} = V\Lambda^{1/2} = \left[ \sqrt{\lambda_1}v_1 \ \sqrt{\lambda_2}v_2 \ \cdots \ \sqrt{\lambda_p}v_p \right], \text{ nas suas colunas.}$$

Considerando um número de autovalores  $r < p$  da matriz  $R$  e apenas os  $r < p$  autovetores correspondentes, pode-se construir um modelo parcimonioso com o número de fatores ( $r$ ) menor que o de variáveis ( $p$ ). Para fazer isto se define a matriz seguinte:

$$A_{p \times r} = V_{p \times r} \Lambda_{p \times r}^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} v_1 & \sqrt{\lambda_2} v_2 & \cdots & \sqrt{\lambda_r} v_r \end{bmatrix},$$

sendo  $V_{p \times r} = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix}$  uma matriz formada pelos primeiros  $r$  autovetores de  $R$  em suas colunas e  $\Lambda_{r \times r}^{1/2} = \begin{bmatrix} \sqrt{\lambda_j} \end{bmatrix}$ ,  $j=1,2,\dots,r$ , uma matriz diagonal formada pela raiz quadrada dos autovalores da matriz  $R$ .

Então, a matriz  $R$  pode ser parcialmente reproduzida como  $R \cong AA'$  neste modelo, sendo desconsiderados os  $p-r$  autovalores e autovetores de  $R$ . Incorporando a contribuição dos fatores específicos no modelo a matriz  $R$  é aproximada por  $R \cong AA' + \varepsilon$ , sendo  $\varepsilon = diag(R - AA')$ .

Neste método, para cada autovalor  $\hat{\lambda}_j$ ,  $j=1,2,\dots,r$ , associado ao fator  $C_k$ ,  $k=1,2,\dots,r$  retido na análise, estima-se o autovetor normalizado correspondente  $\hat{v}_j$ , sendo

$\hat{v}_j = (\hat{v}_{j1} \quad \hat{v}_{j2} \quad \cdots \quad \hat{v}_{jp})'$ . A matriz  $A_{p \times r}$  é estimada por:

$\hat{A}_{p \times r} = \begin{bmatrix} \sqrt{\hat{\lambda}_1} \hat{v}_1 & \sqrt{\hat{\lambda}_2} \hat{v}_2 & \cdots & \sqrt{\hat{\lambda}_r} \hat{v}_r \end{bmatrix}$ . A matriz  $\varepsilon_{p \times p}$  é estimada por:

$\hat{\varepsilon}_{p \times p} = diag(R_{p \times p} - \hat{A}_{p \times r} \hat{A}'_{r \times p})$ . Observe que a matriz  $\hat{\varepsilon}_{p \times p}$  tem a diagonal principal igual aos elementos da diagonal principal da matriz  $(R - \hat{A}\hat{A}')$ .

Basicamente, este método consiste na aplicação da decomposição espectral na matriz  $R_{p \times p}$ . Pelo teorema da decomposição espectral, a matriz de correlação amostral pode ser decomposta como uma soma de  $p$  matrizes, cada uma delas relacionada com um autovalor da matriz  $R_{p \times p}$ , ou seja, para um determinado valor  $r$  tem-se:

$$R_{p \times p} = \sum_{j=1}^p \hat{\lambda}_j \hat{v}_j \hat{v}'_j = \sum_{j=1}^r \hat{\lambda}_j \hat{v}_j \hat{v}'_j + \sum_{j=r+1}^p \hat{\lambda}_j \hat{v}_j \hat{v}'_j. \text{ Desta forma, uma aproximação}$$

para a matriz  $AA'$  é obtida como:  $\hat{A}\hat{A}' = \sum_{j=1}^r \hat{\lambda}_j \hat{v}_j \hat{v}'_j$ , ou seja,

$$\hat{A}\hat{A}' = \begin{bmatrix} \sqrt{\hat{\lambda}_1} \hat{v}_1 & \sqrt{\hat{\lambda}_2} \hat{v}_2 & \cdots & \sqrt{\hat{\lambda}_r} \hat{v}_r \end{bmatrix} \begin{bmatrix} \sqrt{\hat{\lambda}_1} \hat{v}_1 & \sqrt{\hat{\lambda}_2} \hat{v}_2 & \cdots & \sqrt{\hat{\lambda}_r} \hat{v}_r \end{bmatrix}'.$$

Neste caso, a matriz  $\hat{\varepsilon}_{p \times p}$  pode ser obtida utilizando a seguinte matriz:

$\sum_{j=r+1}^p \hat{\lambda}_j \hat{v}_j \hat{v}'_j = R_{p \times p} - \hat{A}_{p \times r} \hat{A}'_{r \times p}$ , considerando apenas a sua diagonal principal. Por esta forma de estimação, a matriz de correlação  $R_{pxp}$  está sendo aproximada como:

$$R \approx \hat{A}\hat{A}' + \hat{\varepsilon}$$

A matriz residual resultante do ajuste do modelo de fatores ortogonal é estimada como:

$$E = R - (\hat{A}\hat{A}' + \hat{\varepsilon})$$
, e pode ser obtida a soma de quadrados de resíduo por meio de  $tr(E^2)$ .

Esta soma de quadrados tem limite superior dado por  $\sum_{j=r+1}^p \hat{\lambda}_j^2$ , o qual é usado como um critério para avaliação da qualidade de ajuste do modelo de fatores por meio do método de componentes principais.

Neste método de estimação das matrizes  $A_{p \times r}$  e  $\varepsilon_{p \times p}$ , a proporção da variância total

explicada pelo fator  $C_k$ , expressa por  $PVE_{C_k} = \frac{\sum_{j=1}^p a_{jk}^2}{p}$ , reduz-se a  $PVE_{C_k} = \frac{\hat{\lambda}_j}{p}$ , que

representa o quanto o fator  $C_k$  capta da variação das variáveis  $Z_j$ .

### Exemplo de Aplicação

Considere os dados de amostras de solos ( $n=13$ ) de pastagem da Amazônia, para as características Ca, Mg, SB e T (Adaptado de Ferreira, 2008).

Ca	Mg	SB	T
2,3	1,7	4,1	7,5
2,5	2,5	5,1	7,4
1,8	2,1	4,1	6,4
3,4	2,5	6,1	8,4
1,8	1,1	3,0	6,5
3,7	1,4	5,2	8,5
1,4	0,7	2,2	4,7
1,5	0,6	2,2	4,0
2,8	2,2	5,1	9,0
1,4	0,8	2,3	3,7
1,8	0,6	2,5	7,2
1,9	1,7	3,7	6,0
2,8	0,8	3,7	5,9

Ca: cálcio; Mg: magnésio; SB: Soma de Bases= K+Ca+Mg; T= SB+H+Al.

Pede-se: realizar a análise de fatores, para modelo com  $r=1$  fator, utilizando o método dos componentes principais.

A matriz de correlações estimada é:

$$R = \begin{bmatrix} 1,00000 & 0,52642 & 0,87434 & 0,79334 \\ 0,52642 & 1,00000 & 0,87238 & 0,67437 \\ 0,87434 & 0,87238 & 1,00000 & 0,83733 \\ 0,79334 & 0,67437 & 0,83733 & 1,00000 \end{bmatrix}$$

A matriz diagonal dos autovalores extraídos da matriz  $R$  é:

$$\hat{\Lambda} = \begin{bmatrix} 3,3000 & 0,0000 & 0,0000 & 0,0000 \\ 0,0000 & 0,4926 & 0,0000 & 0,0000 \\ 0,0000 & 0,0000 & 0,2070 & 0,0000 \\ 0,0000 & 0,0000 & 0,0000 & 0,0003 \end{bmatrix}$$

A matriz dos autovetores correspondentes aos autovalores da matriz  $R$  é:

$$\hat{V} = \begin{bmatrix} -0,4862 & -0,6023 & -0,4456 & 0,4498 \\ -0,4647 & 0,7622 & -0,0745 & 0,4445 \\ -0,5438 & 0,0899 & -0,3101 & -0,7746 \\ -0,5019 & -0,2196 & 0,8365 & -0,0080 \end{bmatrix}$$

Para o modelo com  $r=1$  fator, a matriz de cargas de fatores é:

$$\hat{A} = \hat{V}_1 \hat{\Lambda}_1^{1/2} = \left[ \sqrt{\hat{\lambda}_1} \hat{v}_1 \right], \text{ ou seja,}$$

$$\hat{A} = \begin{bmatrix} -0,4862 \\ -0,4647 \\ -0,5438 \\ -0,5019 \end{bmatrix} \sqrt{3,3000} = \begin{bmatrix} -0,8832439 \\ -0,8441787 \\ -0,9878814 \\ -9118011 \end{bmatrix}.$$

A matriz de especificidades (variâncias específicas) é:

$\hat{\varepsilon} = \text{diag}(R - \hat{A}\hat{A}')$ , ou seja,

$$\hat{\varepsilon} = \begin{bmatrix} 0,2198802 & 0,0000000 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,2873622 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0240904 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0000000 & 0,1686187 \end{bmatrix}$$

A matriz de comunalidades (variâncias comuns) é:

$\hat{h}^2 = I - \hat{\varepsilon}$ , ou seja,

$$\hat{h}^2 = \begin{bmatrix} 1,0000000 & 0,0000000 & 0,0000000 & 0,0000000 \\ 0,0000000 & 1,0000000 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,0000000 & 1,0000000 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0000000 & 1,0000000 \end{bmatrix} - \begin{bmatrix} 0,2198802 & 0,0000000 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,2873622 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0240904 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0000000 & 0,1686187 \end{bmatrix},$$

ou,  $\hat{h}^2 = \begin{bmatrix} 0,7801198 & 0,0000000 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,7126378 & 0,0000000 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,9759096 & 0,0000000 \\ 0,0000000 & 0,0000000 & 0,0000000 & 0,8313813 \end{bmatrix}$

Os resultados mais importantes desta análise de fatores estão resumidos no quadro a seguir:

Variáveis	Comunalidades	Especificidades	Cargas Fatoriais
Ca	0,7801198	0,2198802	-0,8832439
Mg	0,7126378	0,2873622	-0,8441787
SB	0,9759096	0,0240904	-0,9878814
T	0,8313813	0,1686187	-0,9118011

A explicação da variância total neste modelo de  $r=1$  é:

$$PVE_{C_1} = \frac{\sum_{j=1}^p \hat{\lambda}_{j1}^2}{\text{tr}(R)} = \frac{\hat{\lambda}_1}{p}, \text{ ou seja, } PVE_{C_1} = \frac{3,3000}{4} = 0,825$$

Com base nas comunalidades, a variável menos explicada pelo modelo ajustado é Mg e a mais explicada é SB. Então, o fator  $C_1$  pode ser interpretado/rotulado como um índice de saturação de bases.

A estimativa da soma de quadrados de resíduos  $E = R - (\hat{A}\hat{A}' + \hat{\varepsilon})$  do modelo ajustado é 0,1255731. O valor do limite superior da soma de quadrados de resíduos é  $\sum_{j=2}^4 \hat{\lambda}_j^2 = 0,2855101$ .

### Orientações Gerais para Análise de Fatores (AFA)

O modelo, os princípios, interpretações e conceitos para a AFA são:

- O modelo geral da análise de fatores pode ser descrito como a seguir:

$X_j = a_{j1}F_1 + a_{j2}F_2 + \dots + a_{jr}F_r + \varepsilon_j$ , em que:  $X_j$ : é a j-ésima variável observada, com  $j=1,2,\dots,p$ ;  $a_{jk}$ : é a carga fatorial para a j-ésima variável, associada ao k-ésimo fator, com  $k=1,2,\dots,r$  e  $r < p$ ;  $F_k$ : é o k-ésimo fator comum;  $\varepsilon_j$ : é o fator específico.

- Na matriz de dados, é recomendado que a relação entre número de observações e número de variáveis deve ser de no mínimo **5:1**.
- Na determinação do número de fatores interpretáveis, deve-se considerar como fatores de significância prática aqueles associados a autovalores maiores ou iguais a  $1,0$  ( $\lambda_i \geq 1,0$ ), para número de variáveis de **20** a **50**. Para número de variáveis  $< 20$  e  $> 50$ , usar critério pré-determinado:  $\geq 95\%$  da variância total explicada ou  $\geq 60\%$  da variância total explicada, para dados pouco precisos (Hair et al., 2005) ou ainda  $\geq 80\%$ , caso o número de  $\lambda_i \geq 1,0$  explique uma baixa porcentagem da variância total (Cruz et al., 2004).
- Os valores interpretáveis das cargas fatoriais (correlações entre as variáveis originais e os fatores) podem ser estabelecidos segundo os critérios seguintes: no mínimo valores  $> \pm 0,30$ , valores  $\pm 0,40$  são importantes, valores  $\geq \pm 0,50$  têm significância prática (Hair et al., 2005), valores  $\geq \pm 0,70$  têm significância prática (Cruz et al., 2004). De uma

maneira geral, quanto maior o valor da carga fatorial maior a sua importância na interpretação dos resultados.

- A carga fatorial da variável ao quadrado  $(a_j)^2$  representa a quantia da variância total da variável  $j$  explicada pelo fator.
- Uma variável com cargas elevadas em diversos fatores é uma candidata à eliminação.
- Comunalidade de cada variável  $j$  é a quantidade de variância explicada pela análise de fatores para a variável  $j$ .
- Variáveis com comunalidades menor que  $0,50$  ( $Comunalidade_j < 0,50$ ), identifica variáveis com explicação não suficiente, se for considerado (estabelecido) que pelo menos **50%** da variância de cada variável deve ser levada em conta na análise.
- É recomendada a rotação de fatores (varimax) com o objetivo de redistribuir as variâncias e possibilitar a obtenção de uma estrutura de fatores mais simples e de mais fácil interpretação.
- Os fatores podem ser nomeados ou rotulados (determinação de complexos) em função das variáveis que apresentam cargas fatoriais elevadas nos mesmos, com o objetivo de designar algum significado para o padrão de cargas fatoriais extraído. As variáveis com cargas mais altas são consideradas mais importantes e têm maior influência sobre o nome ou rótulo selecionado para representar um fator (Ferreira et al., 2005; Hair et al., 2005).
- Um fator pode ser determinado (extraído) na análise, mas se ele apresenta relações de variáveis indefinidas ele não deve ser interpretado. Apenas os fatores com relações significantes para o estudo devem ser interpretados.
- É recomendada a validação (avaliação do grau de generalidade) e a verificação da estabilidade dos resultados. A estabilidade depende do tamanho da amostra e do número de casos (observações) por variável.
- A matriz de fatores (conjunto das cargas fatoriais das variáveis nos fatores) pode ser utilizada para a construção de escalas múltiplas (complexos de variáveis), indicadores, índices e escores fatoriais, para uso em análises estatísticas posteriores (agrupamento, discriminante e outras).
- Em cada fator, sinais iguais significam que as variáveis estão positivamente relacionadas e sinais opostos significam que as variáveis estão negativamente relacionadas.
- Em soluções ortogonais, os fatores são independentes um do outro. Portanto, os sinais para cargas fatoriais relacionam-se apenas com o fator no qual elas aparecem, e não com outros fatores na solução.

- Alfa de Cronbach: é uma medida de confiabilidade que varia de **0** a **1,0**, sendo os valores de **0,60** a **0,70** considerados o limite inferior de aceitabilidade.
- Análise de fatores: aborda o problema de analisar a estrutura das inter-relações (correlações) entre um grande número de variáveis, definindo um conjunto de dimensões latentes comuns, chamadas de fatores. A análise de fatores fornece a base para a criação de um novo conjunto de variáveis que incorpore o caráter e a natureza das variáveis originais em um número muito menor de novas variáveis, usando variáveis representativas, escores fatoriais ou escalas múltiplas.
- Análise de fatores tipo Q: forma grupo de respondentes (ou casos) com base em sua similaridade em relação a um conjunto de características avaliadas. Semelhante à análise de agrupamento.
- Análise de fatores tipo R: analisa relações entre variáveis para identificar grupos de variáveis que formam dimensões (estruturas) latentes, que são os fatores.
- Autovalor: soma em coluna de cargas fatoriais ao quadrado para um determinado fator; também conhecido como raiz latente. Representa a quantidade de variância explicada por um fator.
- Cargas fatoriais: correlações entre as variáveis originais e os fatores; constituem a chave para o entendimento da natureza de um fator em particular. As cargas fatoriais ao quadrado indicam qual percentual da variância de uma variável original é explicada por um fator.
- Comunalidade: quantidade total de variância que uma variável original compartilha com todas as outras variáveis incluídas na análise.
- Escore fatorial: medida composta criada para cada observação sobre cada fator extraído na análise de fatores. Os coeficientes (pesos) fatoriais são usados em conjunção com os valores da variável original para calcular o escore de cada observação. O escore fatorial pode então ser usado para representar o(s) fator(es) em análises subsequentes. Os escores fatoriais são padronizados para que tenham média e desvio padrão **1,0**.
- Fator: combinação linear (variável estatística) das variáveis originais. Os fatores representam também as dimensões latentes (constructos) que resumem ou explicam o conjunto original de variáveis observadas.
- Matriz de fatores: quadro das cargas fatoriais de todas as variáveis originais sobre cada fator extraído na análise.
- Rotação de fatores: processo de ajuste dos eixos fatoriais para conseguir uma solução de fatores mais simples e pragmaticamente mais significativa.

- Traço: representa a quantidade total de variância na qual a solução fatorial é baseada. O traço é igual ao número de variáveis, baseado na suposição de que a variância em cada variável é igual a **1,0**.
- Variância comum: variância compartilhada com outras variáveis na análise de fatores.
- Variância específica: variância de cada variável, única àquela variável e que não é explicada ou associada com outras variáveis na análise de fatores.
- Variância do erro: variância de uma variável devida a erros na coleta de dados ou na medida da variável.
- Variável substituta: seleção de uma única variável com maior carga fatorial para representar um fator no estágio de redução de dados, em vez de usar uma escala múltipla ou um escore fatorial.
- Validade: grau em que uma medida ou um conjunto de medidas representa corretamente o conceito em estudo – o grau em que se está livre de qualquer erro sistemático ou não aleatório. A validade se refere à quão bem o conceito é definido pela(s) medida(s), ao passo que confiabilidade se refere à consistência da(s) medida(s).

# CAPITULO 26

## Análise Discriminante

### Introdução

Análise discriminante (ADI) é um procedimento que tem como objetivo a obtenção de uma combinação/função linear de duas ou mais variáveis independentes que melhor discrimina grupos estabelecidos *a priori*. Na função discriminante a variável dependente é discreta/categórica e as variáveis independentes são contínuas /métricas.

A análise discriminante é baseada na maximização da variância entre grupos em relação à variância dentro de grupos, utilizando uma matriz de variâncias-covariâncias. Então, o objetivo central da ADI é a discriminação de grupos. Os objetivos da ADI são:

- 1) Desenvolvimento de funções discriminantes;
- 2) Avaliação de existência de diferenças significativas entre grupos preestabelecidos em termos das variáveis independentes/ preditoras;
- 3) Determinação de quais variáveis preditoras contribuem mais para as diferenças entre grupos;
- 4) Avaliação da acurácia de classificação.

Esse procedimento é estreitamente relacionado com regressão linear múltipla e também com análise de variância multivariada e as funções discriminantes têm sido utilizadas como parte de outros procedimentos multivariados. A estrutura de dados para a ADI é formada por uma única variável agrupadora que é predita por uma série de outras variáveis. A variável agrupadora deve ser categórica (ou nominal), podendo ser uma reclassificação de variáveis contínuas em grupos.

Outro objetivo da análise discriminante consiste em usar as funções discriminantes para classificar uma observação (indivíduo, objeto, processo)  $\underline{X}$  com base em medidas de  $p$  variáveis/características do mesmo, em uma de várias populações  $\pi_i$ , ( $i=1,2,\dots,g$ ), buscando minimizar a probabilidade de erro de classificação, ou seja, classificar um indivíduo em uma população  $\pi_i$  quando ele de fato pertence à população  $\pi_{i'}$ , ( $i \neq i'$ ).

A combinação linear das variáveis observadas que apresenta o maior poder de discriminação entre grupos é denominada de Função Discriminante Linear de Fisher-FDLF, constituindo a base de toda a análise discriminante.

A FDLF tem a propriedade de minimizar as probabilidades de erros de classificação, quando as populações são normalmente distribuídas com parâmetros  $\underline{\mu}_i$  e  $\Sigma$  conhecidos. Na prática, são necessários métodos de estimação desses parâmetros, uma vez que eles são desconhecidos.

## Classificação em uma de duas populações

A classificação de uma observação em uma de duas populações é feita por meio de uma função linear do vetor aleatório  $\underline{X}$ , a qual se caracteriza por produzir a máxima separação entre as duas populações, sendo denominada de FDLF.

Considere que o vetor de médias  $\boldsymbol{\mu}$ -variado,  $\boldsymbol{\mu}_i$  e a matriz de covariâncias comuns das populações

$\pi_i$ , ( $i=1,2$ ) de ordem  $p, \Sigma$ , sejam parâmetros conhecidos. Demonstra-se que a função linear do vetor de médias  $\underline{X}' = [X_1, X_2, \dots, X_p]$ , que produz separação máxima entre as duas populações, é obtida por meio da seguinte expressão:

$\underline{l}'\underline{X} = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} \underline{X}$ , que é a Função Discriminante Linear de Fisher-FDLF. Observe que

$\underline{l}' = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1}$ . Outra notação para a FDLF é a

seguinte:  $D(\underline{X}) = \underline{l}'\underline{X} = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} \underline{X}$ .

Considere  $D(\underline{x}_0) = \underline{l}'\underline{x}_0 = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} \underline{x}_0$  como o valor da função discriminante para uma observação  $\underline{x}_0$  a ser classificada e seja o ponto médio entre as duas médias populacionais univariadas  $D(\underline{\mu}_1)$  e  $D(\underline{\mu}_2)$  expresso como:

$$m = \frac{1}{2} [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} [\underline{\mu}_1 + \underline{\mu}_2] = \frac{1}{2} [\underline{l}' \underline{\mu}_1 + \underline{l}' \underline{\mu}_2] \text{ ou}$$

$$m = \frac{1}{2} [D(\underline{\mu}_1) + D(\underline{\mu}_2)]$$

A regra para classificação da observação  $\underline{x}_0$  é a seguinte:

- Alocar  $\underline{x}_0$  na população  $\pi_1$  se  $D(\underline{x}_0) = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} \underline{x}_0 \geq m$ ;
- Alocar  $\underline{x}_0$  na população  $\pi_2$  se  $D(\underline{x}_0) = [\underline{\mu}_1 - \underline{\mu}_2]' \Sigma^{-1} \underline{x}_0 < m$ ;

Uma forma alternativa para a regra de classificação consiste em subtrair o valor de  $m$  do valor de  $D(\underline{x}_0)$  e comparar o resultado com o valor zero, ou seja:

- Alocar  $\underline{x}_0$  na população  $\pi_1$  se  $D(\underline{x}_0) - m \geq 0$ ;

- Alocar  $\tilde{x}_0$  na população  $\pi_2$  se  $D(\tilde{x}_0) - m < 0$ ;

Observe que na prática os parâmetros  $\tilde{\mu}_i$  e  $\Sigma$  são estimados pressupondo que as populações apresentam distribuição normal. Desta forma, os parâmetros  $\boldsymbol{l}$  e  $\boldsymbol{m}$  podem ser estimados a partir das observações que já foram corretamente classificadas.

Considere duas populações normais multivariadas  $\pi_1$  e  $\pi_2$ , com  $n_1$  observações da variável aleatória multivariada  $X' = [X_1, X_2, \dots, X_p]$  em  $\pi_1$  e  $n_2$  observações da variável aleatória multivariada  $\tilde{X}' = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p]$  em  $\pi_2$ . As matrizes de observações amostrais (dados) das duas populações são:

$$\begin{matrix} X_1 \\ (p \times n_1) \end{matrix} = \begin{bmatrix} \tilde{x}_{11}, \tilde{x}_{12}, \tilde{x}_{13}, \dots, \tilde{x}_{1n_1} \end{bmatrix} \quad \text{e} \quad \begin{matrix} \tilde{X}_2 \\ (p \times n_2) \end{matrix} = \begin{bmatrix} \tilde{x}_{21}, \tilde{x}_{22}, \tilde{x}_{23}, \dots, \tilde{x}_{2n_2} \end{bmatrix}, \quad \text{sendo} \\ \tilde{x}_{11}, \tilde{x}_{12}, \tilde{x}_{13}, \dots, \tilde{x}_{1n_1} \text{ observações com } N_p(\tilde{\mu}_1; \Sigma) \text{ e} \\ \tilde{x}_{21}, \tilde{x}_{22}, \tilde{x}_{23}, \dots, \tilde{x}_{2n_2} \text{ observações com } N_p(\tilde{\mu}_2; \Sigma).$$

Com base nas matrizes de dados obtém-se os vetores de médias e as matrizes de covariâncias amostrais da seguinte forma:

$$\begin{matrix} \bar{x}_1 \\ (p \times 1) \end{matrix} = \frac{1}{n_1} \sum_{j=1}^{n_1} \tilde{x}_{1j}; \quad \begin{matrix} \bar{x}_2 \\ (p \times 1) \end{matrix} = \frac{1}{n_2} \sum_{j=1}^{n_2} \tilde{x}_{2j};$$

$$\begin{matrix} S_1 \\ (p \times p) \end{matrix} = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\tilde{x}_{1j} - \bar{x}_1)(\tilde{x}_{1j} - \bar{x}_1)';$$

$$\begin{matrix} S_2 \\ (p \times p) \end{matrix} = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\tilde{x}_{2j} - \bar{x}_2)(\tilde{x}_{2j} - \bar{x}_2)';$$

Normalmente é assumido que as populações têm a mesma matriz de covariâncias  $\Sigma$ . Então, as matrizes de covariâncias amostrais  $S_1$  e  $S_2$  são combinadas para obter-se uma única matriz de covariâncias amostral comum, obtida da seguinte forma:

$$S_c = \left[ \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_1 + \left[ \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_2 \text{ ou}$$

$$S_c = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}$$

Se as matrizes de dados  $X_1$  e  $X_2$  são constituídas de amostras aleatórias das populações  $\pi_1$  e  $\pi_2$  então  $S_c$  é um estimador não viesado de  $\Sigma$ . Portanto, os parâmetros  $\mu_1$ ,  $\mu_2$  e  $\Sigma$  da FDLF são substituídos pelos seus estimadores amostrais  $\bar{x}_1$ ,  $\bar{x}_2$  e  $S_c$ , respectivamente, resultando na FDLF amostral com a seguinte expressão:

$D(\tilde{x}) = \hat{l}'\tilde{x} = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} \tilde{x}$ . Neste caso, é preciso se ter  $(n_1 + n_2 - 2) > p$  para que exista  $S_c^{-1}$ .

O ponto médio entre as duas médias amostrais univariadas  $D(\bar{x}_1)$  e  $D(\bar{x}_2)$  é obtido da seguinte forma:

$$\hat{m} = \frac{1}{2} [\bar{x}_1 - \bar{x}_2]' S_c^{-1} [\bar{x}_1 + \bar{x}_2] = \frac{1}{2} (\hat{l}' \bar{x}_1 + \hat{l}' \bar{x}_2), \text{ ou seja,}$$

$$\hat{m} = \frac{1}{2} [D(\bar{x}_1) + D(\bar{x}_2)].$$

A regra de classificação baseada nas estimativas dos parâmetros é a seguinte:

- Alocar  $\tilde{x}_0$  na população  $\pi_1$  se  $D(\tilde{x}_0) = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} \tilde{x}_0 \geq \hat{m}$  ou  $D(\tilde{x}_0) - \hat{m} \geq 0$ ;
- Alocar  $\tilde{x}_0$  na população  $\pi_2$  se  $D(\tilde{x}_0) = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} \tilde{x}_0 < \hat{m}$  ou  $D(\tilde{x}_0) - \hat{m} < 0$ .

### Avaliação da Função Discriminante

O teste de significância para verificar se existe diferença significativa entre as duas populações, ou seja, se os dois vetores de médias diferem significativamente, pode ser feito com base na distribuição da distância de Mahalanobis entre as duas populações. A distância de Mahalanobis amostral entre duas populações é expressa como:

$$D^2 = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} [\bar{x}_1 - \bar{x}_2] = D(\bar{x}_1) - D(\bar{x}_2)$$

Admitindo que as populações  $\pi_1$  e  $\pi_2$  são normais multivariadas com matriz de covariância comum  $\Sigma$ , o teste para as hipóteses  $H_0: \mu_1 = \mu_2$  versus  $H_a: \mu_1 \neq \mu_2$  é feito utilizando a estatística seguinte:

$$F_0 = \left( \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \right) \left( \frac{n_1 \times n_2}{n_1 + n_2} \right) D^2, \text{ em que:}$$

$n_1$ : número de observações na população  $\pi_1$ ;

$n_2$ : número de observações na população  $\pi_2$ ;

$p$ : número de variáveis;

$D^2$ : distância de Mahalanobis entre as duas populações.

A estatística de teste  $F_0$  tem distribuição  $F$  com  $v_1 = p$  e  $v_2 = n_1 + n_2 - p - 1$  graus de liberdade. Se o teste for significativo, rejeita-se  $H_0$  e conclui-se que a separação/discriminação

entre as populações  $\pi_1$  e  $\pi_2$  é significativa. Este teste corresponde ao teste  $T^2$  de Hotelling para vetores de médias de duas populações normais. A discriminação significativa não implica em classificação eficiente. A eficiência de um método de classificação pode ser avaliada independentemente de qualquer teste de discriminação.

### Métodos de Estimação das Probabilidades de Erro de Classificação

Na apresentação dos métodos de estimação das probabilidades de erro de classificação considera-se que os custos de classificação errada,  $C(i'/i)$ , é o custo de se classificar erroneamente uma observação da população  $\pi_i$  na população  $\pi_{i'}$ , ( $i \neq i'$ ) e também que as probabilidades “a priori” de uma observação ser proveniente das populações  $\pi_i$ ,  $i=1,2$ , são iguais. Os métodos são os seguintes:

- Método de Okamoto

Considere a distância de Mahalanobis amostral:

$$D^2 = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} [\bar{x}_1 - \bar{x}_2] = D(\bar{x}_1) - D(\bar{x}_2)$$

Supondo que as populações têm distribuição normal, as probabilidades aproximadas de classificação errada para as duas populações são iguais e estimadas como:

$$\hat{p}(2/1) = \hat{p}(1/2) = \phi\left(-\frac{D}{2}\right), \text{ em que:}$$

$$\phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx;$$

$\hat{p}(i'/i)$ : estimativa da probabilidade de classificar erroneamente uma observação da população  $\pi_i$  na população  $\pi_{i'}$ ,  $i \neq i'$ .

- Método de Smith

O método de Smith, que não depende da suposição de normalidade das populações, propõe que as observações usadas para a construção da função discriminante sejam reusadas para a estimação de  $p(i'/i)$ . Desta forma, se  $D(\bar{x})$  foi construída usando-se  $n_1 + n_2$  observações e se  $m_i \leq n_i$  das observações forem classificadas de forma errada na população  $\pi_i$  por meio de  $D(\bar{x})$  tem-se:

$$\hat{p}(i'/i) = \hat{p}_i = \frac{m_i}{n_i}, i = 1, 2. \text{ Então, } \hat{p}(2/1) = \hat{p}_1 = \frac{m_1}{n_1} \text{ e}$$

$$\hat{p}(1/2) = \hat{p}_2 = \frac{m_2}{n_2}.$$

### Exemplo de Aplicação

Considere os dados referentes a duas raças de insetos, para duas variáveis/características, apresentados no quadro a seguir (Extraído de Regazzi, 1999):

$X_1$ : número médio de cerdas primordiais;

$X_2$ : número médio de cerdas distais;

Raça A		Raça B	
$X_1$	$X_2$	$X_1$	$X_2$
6,36	5,24	6,00	4,88
5,92	5,12	5,60	4,64
5,92	5,36	5,64	4,96
6,44	5,64	5,76	4,80
6,40	5,16	5,96	5,08
6,56	5,56	5,72	5,04
6,64	5,36	5,64	4,96
6,68	4,96	5,44	4,88
6,72	5,48	5,04	4,44
6,76	5,60	4,56	4,04
6,72	5,08	5,48	4,20
		5,76	4,80

Observe que neste exemplo tem-se  $p=2$ ,  $n_1=11$  e  $n_2=12$ . A matriz  $X_1(p \times n_1)$  é constituída pelos elementos da Raça  $A(\pi_1)$  e a matriz  $X_2(p \times n_2)$  é constituída pelos elementos da Raça  $B(\pi_2)$ .

As estimativas do vetor de média e da matriz de covariância para as duas raças (populações) são:  
Raça  $A(\pi_1)$

$$\bar{x}_1 = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{12} \end{bmatrix} = \begin{bmatrix} 6,46545 \\ 5,32364 \end{bmatrix} \text{ e } S_1 = \begin{bmatrix} 0,091287 & 0,011258 \\ 0,011258 & 0,052625 \end{bmatrix}$$

Raça  $B(\pi_2)$

$$\bar{x}_2 = \begin{bmatrix} \bar{x}_{21} \\ \bar{x}_{22} \end{bmatrix} = \begin{bmatrix} 5,55000 \\ 4,72667 \end{bmatrix} \text{ e } S_2 = \begin{bmatrix} 0,160327 & 0,107418 \\ 0,107418 & 0,111661 \end{bmatrix}.$$

Como deve-se considerar que  $\Sigma_1 = \Sigma_2 = \Sigma$ , estima-se a matriz de covariância amostral comum da forma seguinte:

$$S_c = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}, \text{ ou}$$

$$S_c = \frac{(11-1)S_1 + (12-1)S_2}{(11+12-2)} = \begin{bmatrix} 0,127451 & 0,061627 \\ 0,061627 & 0,083549 \end{bmatrix}$$

A matriz inversa de  $S_c$  é  $S_c^{-1} = \begin{bmatrix} 12,196015 & -8,995964 \\ -8,995964 & 18,604583 \end{bmatrix}$ .

Então, a análise discriminante é realizada seguindo os seguintes passos:

- 1) Estimação da Função Discriminante Linear de Fisher

A expressão amostral da FDLF é  $D(\tilde{x}) = \hat{l}'\tilde{x} = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} \tilde{x}$ , sendo

$$\hat{l}' = \begin{bmatrix} 6,46545 - 5,55000 \\ 5,32364 - 4,72667 \end{bmatrix}' S_c^{-1}, \text{ ou}$$

$$\hat{l}' = \begin{bmatrix} 0,91545 & 0,59697 \end{bmatrix} \begin{bmatrix} 12,196015 & -8,995964 \\ -8,995964 & 18,604583 \end{bmatrix} e$$

$$\hat{l}' = \begin{bmatrix} 5,794819 & 2,871023 \end{bmatrix}. \text{ Então,}$$

$$D(\tilde{x}) = \hat{l}'\tilde{x} = \begin{bmatrix} 5,794819 & 2,871023 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} e$$

$$D(\tilde{x}) = 5,794819x_1 + 2,871023x_2$$

- 2) Classificação de novas observações/indivíduos

A questão é a seguinte: uma nova observação/indivíduo  $\tilde{x}_0$  pertence à Raça  $A(\pi_1)$  ou à Raça  $B(\pi_2)$ ?

Para responder esta questão, calcula-se:

$$\hat{m} = \frac{1}{2} [D(\bar{x}_1) + D(\bar{x}_2)], \text{ sendo}$$

$$D(\bar{x}_1) = \hat{l}'\bar{x}_1 = \begin{bmatrix} 5,794819 & 2,871023 \end{bmatrix} \begin{bmatrix} 6,46545 \\ 5,32364 \end{bmatrix}, \text{ ou}$$

$$D(\bar{x}_1) = 52,750405$$

$$D(\bar{x}_2) = \hat{l}'\bar{x}_2 = [5,794819 \quad 2,871023] \begin{bmatrix} 5,55000 \\ 4,72667 \end{bmatrix}, \text{ ou}$$

$$D(\bar{x}_2) = 45,731624. \text{ Logo,}$$

$$\hat{m} = \frac{1}{2}(52,750405 + 45,731624) = 49,241$$

Neste caso, a regra de classificação é a seguinte:

Um novo indivíduo  $\tilde{x}_0$  será classificado na raça (população) Raça  $A(\pi_1)$  se  $D(\tilde{x}_0) = \hat{l}'\tilde{x}_0 \geq 49,241$  ou  $D(\tilde{x}_0) - 49,241 \geq 0$ .

Este indivíduo  $\tilde{x}_0$  será classificado na raça (população) Raça  $B(\pi_2)$  se

$$D(\tilde{x}_0) = \hat{l}'\tilde{x}_0 < 49,241 \text{ ou } D(\tilde{x}_0) - 49,241 < 0.$$

Considere, por exemplo, um novo indivíduo com número médio de cerdas proximais e distais iguais a 6,21 e 5,31, respectivamente. Em qual das duas raças este indivíduo seria classificado? Para responder esta questão calcula-se:

$$D(\tilde{x}_0) = \hat{l}'\tilde{x}_0 = [5,794819 \quad 2,871023] \begin{bmatrix} 6,21 \\ 5,31 \end{bmatrix}, \text{ ou}$$

$$D(\tilde{x}_0) = 51,230958$$

Como  $D(\tilde{x}_0) = 51,230958 > 49,241$ , o indivíduo  $\tilde{x}_0$  deve ser classificado na Raça  $A(\pi_1)$ .

3) Teste de significância para a separação entre as populações  $\pi_1$  e  $\pi_2$

As hipóteses testadas são:  $H_0: \mu_1 = \mu_2$  e  $H_a: \mu_1 \neq \mu_2$

A estatística de teste é:

$$F_0 = \left[ \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \right] \left( \frac{n_1 \times n_2}{n_1 + n_2} \right) D^2 \square F(p; n_1 + n_2 - p - 1), \text{ sendo}$$

$$D^2 = [\bar{x}_1 - \bar{x}_2]' S_c^{-1} [\bar{x}_1 - \bar{x}_2] = D(\bar{x}_1) - D(\bar{x}_2)$$

$$D^2 = D(\bar{x}_1) - D(\bar{x}_2) = 52,750405 - 45,731624 = 7,018781$$

Então,  $F_0 = \left[ \frac{11+12-2-1}{(11+12-2)2} \right] \left( \frac{11x12}{11+12} \right) (7,018781) = 19,18$ . Tem-se o valor tabelado  $F_{0,01}(2;11+12-2-1) = F_{0,01}(2;20) = 5,85$ .

Como  $F_0 = 19,18 > F_{0,01}(2;20) = 5,85$  o teste é significativo. Portanto, rejeita-se

$H_0$  e conclui-se que a discriminação entre as duas raças (populações), Raça  $A(\pi_1)$  e Raça  $B(\pi_2)$ , é significativa ( $p < 0,01$ ).

- 4) Estimação das probabilidades de erro de classificação

Por meio do método de Okamoto tem-se: Considerando o valor  $D^2 = 7,018781$  e que as estimativas de probabilidade de erro de classificação são iguais para as duas populações tem-se que

$\hat{p}(2/1) = \hat{p}(1/2) = \phi\left(-\frac{D}{2}\right)$ . Os valores  $\phi\left(-\frac{D}{2}\right)$  são tabelados, para diferentes valores de  $D$  (Okamoto, 1963). Então tem-se:

$$\hat{p}(2/1) = \hat{p}(1/2) = \phi\left(-\frac{\sqrt{7,018781}}{2}\right) = 0,0934$$

Este resultado significa que a probabilidade de erro ao classificar um indivíduo na Raça  $A(\pi_1)$  ou na Raça  $B(\pi_2)$  é igual a 9,34%.

Segundo o método de Smith tem-se: Considere o quadro a seguir com os valores de  $D(\tilde{x})$ , para os indivíduos das duas raças de insetos, obtidos por meio da função discriminante estimada,

$$D(\tilde{x}) = 5,794819x_1 + 2,871023x_2$$

Raça A	Raça B
51,8992	48,7795
49,0049	45,7725
49,6940	46,9230
53,5112	47,1591
51,9013	49,1219
53,9769	47,6163
53,8663	46,9230
52,9497	45,5344
54,7644	41,9532
55,2507	38,0233
53,5259	43,8139
	47,1591

Observe que a média dos escores da Raça A e da Raça B são iguais a  $D(\bar{x}_1)$  e  $D(\bar{x}_2)$ , respectivamente, ou seja:

$$D(\bar{x}_1) = \frac{510,2545}{11} = 52,750405 \text{ e}$$

$$D(\bar{x}_2) = \frac{548,7792}{12} = 45,731624$$

As estimativas das probabilidades de erro de classificação são:

$$\text{Raça A } (\pi_1)$$

$$\hat{p}(2/1) = \hat{p}_1 = \frac{m_1}{n_1} = \frac{1}{11} = 0,0909, \text{ ou } \hat{p}(2/1) = \hat{p}_1 = 9,09\%$$

$$\text{Raça B } (\pi_2)$$

$$\hat{p}(1/2) = \hat{p}_2 = \frac{m_2}{n_2} = \frac{0}{12} = 0, \text{ ou } \hat{p}(1/2) = \hat{p}_2 = 0\%$$

A distribuição dos indivíduos segundo a classificação *a priori* e por meio da função discriminante é resumida no quadro a seguir:

Classificação <i>a priori</i>	Função Discriminante		Total	Classificação Errada (%)
	Raça A	Raça B		
Raça A	10	1	11	9,09
Raça B	0	12	12	0,00
Total	10	13	23	-

## Classificação em uma de várias populações – Funções Discriminantes de Anderson (1958)

### Introdução

Neste contexto, o estabelecimento das probabilidades *a priori* para as várias populações é um ponto importante, além da pressuposição de alguma distribuição para a obtenção das funções discriminantes. Além disso, é preciso considerar os custos de erros de classificação. Nesse momento a experiência do pesquisador na escolha das variáveis realmente necessárias para o uso deste método é muito importante.

### Estimação de funções discriminantes

A estimativa de funções discriminantes é feita da seguinte forma:

Considere  $\pi_1, \pi_2, \dots, \pi_g$  um conjunto de  $g$  populações. Nesta situação, tem-se uma observação (indivíduo ou objeto)  $\underline{x}$  e deseja-se classificá-la em uma dentre  $g$  populações, sendo  $g > 2$ . Considere ainda que a discriminação a ser realizada envolve custos idênticos de classificação errada e probabilidades conhecidas  $(p_i)$ , de que uma observação seja classificada na população  $i$ , para as  $g$  populações, sendo  $\sum_{i=1}^g p_i = 1$ .

Admitindo que a distribuição densidade de probabilidade associada à  $\pi_i$  é normal multivariada, a função discriminante é expressa por:

$$D_i(\underline{x}) = -\frac{1}{2} \ln |\Sigma_i| - \frac{1}{2} [\underline{x} - \underline{\mu}_i]^\top \Sigma_i^{-1} [\underline{x} - \underline{\mu}_i] + \ln(p_i)$$

Supondo a igualdade das matrizes de covariâncias ( $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$ ), os componentes constantes, para todo  $i$ , podem ser retirados da expressão anterior e a função discriminante torna-se:

$$D_i(\underline{x}) = \underline{l}_i' \underline{x} - \frac{1}{2} \underline{l}_i' \underline{\mu}_i + \ln(p_i) \quad (i = 1, 2, \dots, g), \text{ em que } \underline{l}_i = \Sigma^{-1} \underline{\mu}_i$$

Com a validade da pressuposição de que  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$ , a função discriminante, que é quadrática em  $\underline{x}$ , torna-se linear, o que faz com que a regra de classificação seja mais simples.

### Regra de classificação/discriminação

A regra de classificação/discriminação de  $\underline{x}$  quanto à população é a seguinte: Classificar  $\underline{x}$  em  $\pi_i$  se e somente se,

$D_i(\underline{x}) = \max[D_1(\underline{x}), D_2(\underline{x}), \dots, D_g(\underline{x})]$ . Como não se conhece os parâmetros  $\mu_i$  e  $\Sigma$

a construção da regra de discriminação é feita utilizando-se as estimativas destes parâmetros, da seguinte forma: Supondo a homogeneidade das matrizes de covariâncias, a função discriminante de

Anderson é expressa como  $D_i(\underline{x}) = \hat{l}'_i \underline{x} - \frac{1}{2} \hat{l}'_i \bar{x}_i + \ln(p_i)$ , ( $i=1,2,\dots,g$ ), em que:

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{bmatrix}, \bar{x}_i = \begin{bmatrix} \bar{x}_{i1} \\ \bar{x}_{i2} \\ \dots \\ \bar{x}_{ip} \end{bmatrix}, \hat{l}_i = S_c^{-1} \bar{x}_i, \text{ sendo } S_c^{-1} \text{ a matriz de covariância amostral comum a}$$

todas as populações, obtida por meio da expressão

$$S_c = \frac{(n_1-1)S_1 + (n_2-1)S_2 + \dots + (n_g-1)S_g}{(n_1-1) + (n_2-1) + \dots + (n_g-1)}, \text{ ou}$$

$$S_c = \frac{\sum_{i=1}^g (n_i-1)S_i}{\sum_{i=1}^g n_i - g} = \frac{\sum_{i=1}^g (n_i-1)S_i}{n-g}$$

A observação  $\underline{x}$  será classificada em  $\pi_i$  se e somente se

$D_i(\underline{x}) = \max[D_1(\underline{x}), D_2(\underline{x}), \dots, D_g(\underline{x})]$ . Esta regra de discriminação pode ser utilizada também para o caso particular de  $g=2$ .

### Probabilidade de Erro de Classificação

Pode-se calcular a probabilidade de classificação errada para cada grupo. É feita uma análise de consistência, que consiste em resubstituir os dados e fazer a reclassificação. A soma dos casos de classificação errada em cada grupo fornece uma estimativa da probabilidade de erro de classificação para o mesmo. Desta forma, tem-se:

$$\hat{p}_i = \frac{m_i}{n_i}, i=1,2,\dots,g, \text{ em que}$$

$m_i$ : número de observações com classificação errada em  $\pi_i$ ;

$n_i$ : número de observações na população  $\pi_i$ .

A soma de todos os casos de classificação errada fornece a taxa de erro aparente, ou seja:

$T = \frac{\sum_{i=1}^g m_i}{\sum_{i=1}^g n_i}$ . Note que esta taxa é subestimada, uma vez que os mesmos dados são utilizados para obtenção das funções discriminantes e da probabilidade de erro de classificação.

### Distância de Mahalanobis entre dois grupos

Para  $i$  e  $i'$  fixados, sendo  $i \neq i' = 1, 2, \dots, g$ , estima-se a distância de Mahalanobis entre os grupos  $i$  e  $i'$  por meio da seguinte expressão:

$$D_{ii'}^2 = (\bar{x}_i - \bar{x}_{i'})' S_c^{-1} (\bar{x}_i - \bar{x}_{i'})$$

Quanto maiores forem os  $D_{ii'}^2$ , maior será a distinção entre os grupos e, portanto, a classificação de uma nova observação em um dos grupos se fará com a máxima probabilidade de acerto.

### Exemplo de Aplicação

Considere os dados de duas espécies de abelhas e uma terceira população obtida por simulação. As três populações são (Extraído de Regazzi, 1999):

$\pi_1$ : *Partamona testacea*;  $\pi_2$ : *Partamona pseudomusarum*;  $\pi_3$ : Simulação de dados. Foram avaliadas seis características quantitativas contínuas medidas em milímetros:  $X_1$ : comprimento do clípeo;  $X_2$ : largura máxima da cabeça;  $X_3$ : distância máxima interorbital;  $X_4$ : comprimento do olho composto;  $X_5$ : comprimento da área malar;  $X_6$ : comprimento do flagelo.

Quadro 1 – Dados da população  $\pi_1$ : *Partamona testacea*

Amostra	Características					
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
1	17,0	62,0	41,5	40,0	4,0	46,0
2	17,5	59,5	40,0	39,0	4,5	45,0
3	18,0	62,0	41,0	41,0	4,5	46,5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	17,5	58,5	40,5	38,5	4,5	45,0
31	18,0	62,5	42,5	40,5	4,5	47,0
32	19,0	64,0	41,5	42,0	4,0	47,0

Quadro 2 – Dados da população  $\pi_2$ : *Partamona pseudomusarum*

Amostra	Características					
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
1	16,5	61,0	38,5	40,0	3,0	41,0
2	16,5	57,0	37,5	38,0	3,5	41,5
3	16,0	57,0	37,0	39,0	3,0	39,5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	17,0	62,0	40,0	42,0	3,5	42,5
31	15,0	56,0	36,5	37,5	3,0	39,5
32	16,5	60,5	38,5	40,5	3,5	41,5

Quadro 3 – Dados da população  $\pi_3$ : Simulação de dados

Amostra	Características					
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$
1	16,0	60,0	38,0	40,5	3,0	39,0
2	15,0	57,5	36,0	39,5	3,0	38,5
3	15,0	59,0	37,5	41,0	2,5	39,0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	14,5	56,5	36,0	39,0	3,0	38,5
31	16,0	59,0	38,0	40,5	3,5	40,5
32	15,0	56,0	35,0	39,0	2,5	37,5

Observe que neste caso tem-se:  $p = 6$  variáveis;  $g = 3$  grupos;  $n_1 = n_2 = n_3 = 32$  observações ( $\underline{x}$ ). Os vetores de médias são:

$$\bar{\underline{x}}_1 = \begin{bmatrix} 17,765 \\ 61,125 \\ 40,937 \\ 40,141 \\ 4,281 \\ 45,750 \end{bmatrix}; \bar{\underline{x}}_2 = \begin{bmatrix} 16,703 \\ 60,062 \\ 39,297 \\ 40,125 \\ 3,516 \\ 42,125 \end{bmatrix}; \bar{\underline{x}}_3 = \begin{bmatrix} 15,437 \\ 58,250 \\ 37,016 \\ 40,172 \\ 3,031 \\ 39,078 \end{bmatrix}$$

As matrizes de variâncias e covariâncias são:

$$S_1 = \begin{bmatrix} 0,90297 & 1,72379 & 0,71068 & 1,09047 & 0,05192 & 1,23790 \\ & 5,93548 & 2,78226 & 3,71573 & 0,07661 & 3,38709 \\ & & 1,65726 & 1,55746 & 0,12298 & 1,54032 \\ & & & 2,69733 & -0,04083 & 2,06855 \\ & & & & 0,07964 & 0,11290 \\ & & & & & 2,56452 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0,32031 & 0,76109 & 0,45388 & 0,50605 & 0,06930 & 0,57863 \\ & 4,99597 & 3,01310 & 3,26613 & 0,23286 & 3,45161 \\ & & 2,17515 & 1,91331 & 0,22908 & 2,33266 \\ & & & 2,33871 & 0,12702 & 2,25000 \\ & & & & 0,10459 & 0,231854 \\ & & & & & 3,04839 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} 0,44758 & 0,79839 & 0,49294 & 0,46270 & 0,01008 & 0,59375 \\ & 3,69355 & 2,22177 & 2,22177 & -0,07258 & 2,06855 \\ & & 1,71749 & 1,35207 & 0,05595 & 1,47555 \\ & & & 1,54209 & -0,07812 & 1,35711 \\ & & & & 0,11189 & 0,00554 \\ & & & & & 1,75983 \end{bmatrix}$$

A obtenção das funções discriminantes de Anderson é realizada da seguinte forma:

As pressuposições são de distribuição normal multivariada, homogeneidade das matrizes de covariâncias, custos de classificação errada idênticos e probabilidades de erro de classificação *a priori* iguais, ou seja,  $p_1 = p_2 = p_3 = 1/3$ . A matriz de covariância amostral comum é expressa

como:  $S_c = \frac{\sum_{i=1}^g (n_i - 1) S_i}{\sum_{i=1}^g n_i - g} = \frac{\sum_{i=1}^g (n_i - 1) S_i}{n - g}$ . Neste exemplo tem-se:

$$g=3, n_1=n_2=n_3=32 \text{ e } S_c = \frac{S_1+S_2+S_3}{3}.$$

Note que  $S_c$ , a matriz de covariâncias comum ou combinada é uma matriz de covariâncias dentro, sendo, portanto, uma matriz de covariâncias residuais, que pode ser obtida por meio da análise de variância multivariada, supondo  $g=3$  tratamentos e  $r=32$  repetições no modelo inteiramente casualizado. Neste caso, tem-se  $S_c = \frac{E}{n_e}$ , em que:

$E$ : matriz de somas de quadrados e de produtos residuais;

$n_e$ : número de graus de liberdade do resíduo.

A matriz de covariâncias comum ou combinada  $(S_c)$  e sua inversa  $(S_c^{-1})$  são as seguintes:

$$S_c = \begin{bmatrix} 0,55695 & 1,09442 & 0,55250 & 0,68641 & 0,04377 & 0,80342 \\ & 4,87500 & 2,67238 & 3,06788 & 0,07896 & 2,96909 \\ & & 1,84997 & 1,60761 & 0,13600 & 1,78251 \\ & & & 2,19271 & 0,00269 & 1,89188 \\ & & & & 0,09871 & 0,11677 \\ & & & & & 2,45758 \end{bmatrix}$$

$$S_c^{-1} = \begin{bmatrix} 3,97312 & -1,10879 & 1,28986 & 0,22397 & -1,47856 & -0,99704 \\ & 3,59528 & -2,29844 & -2,85749 & 1,05465 & -0,16438 \\ & & 4,02469 & 0,65556 & -3,21291 & -0,91601 \\ & & & 4,37908 & 1,81953 & -0,55401 \\ & & & & 14,99916 & -0,57379 \\ & & & & & 2,04959 \end{bmatrix}$$

A expressão utilizada para a obtenção das funções discriminantes de Anderson é:

$D_i(x) = \hat{l}_i' x - \frac{1}{2} \hat{l}_i' \bar{x}_i + \ln(p_i); (i=1,2,\dots,g)$ . Tem-se ainda que:  $\hat{l}_i = S_c^{-1} \bar{x}_i$ , obtendo-se então os seguintes vetores:

$$\hat{l}_1 = \begin{bmatrix} 12,65980 \\ -11,73583 \\ 17,83547 \\ 14,37464 \\ 17,67059 \\ 3,81394 \end{bmatrix}; \hat{l}_2 = \begin{bmatrix} 12,24306 \\ -10,77379 \\ 18,07427 \\ 16,64402 \\ 13,96003 \\ -0,43105 \end{bmatrix}; \hat{l}_3 = \begin{bmatrix} 10,04635 \\ -10,77379 \\ 15,80446 \\ 21,05622 \\ 15,81697 \\ -2,77449 \end{bmatrix} \text{ e os valores}$$

$$\hat{\bar{x}}_1 = 1064,84290; \hat{\bar{x}}_2 = 966,42084; \hat{\bar{x}}_3 = 897,11451$$

Admite-se  $p_1 = p_2 = p_3 = \frac{1}{3}$ , logo

$$\ln p_1 = \ln p_2 = \ln p_3 = \ln \frac{1}{3} = 1,0986123$$

A função discriminante associada à população  $\pi_1$  é:

$$D_1(x) = 12,65980x_1 - 11,73583x_2 + 17,83547x_3 + \\ + 14,37464x_4 + 17,67059x_5 - 3,81393x_6 - 533,52007$$

A função discriminante associada à população  $\pi_2$  é:

$$D_2(x) = 12,24306x_1 - 10,77379x_2 + 18,07427x_3 + \\ + 16,64402x_4 + 13,96003x_5 - 0,43105x_6 - 484,30903$$

A função discriminante associada à população  $\pi_3$  é:

$$D_3(x) = 10,04635x_1 - 10,78761x_2 + 15,80446x_3 + \\ + 21,05622x_4 + 15,81697x_5 - 2,77449x_6 - 449,65587$$

### Classificação de observações

Por exemplo, considere a classificação da primeira observação  $(x_1)$  da população  $\pi_1$  com base nas funções discriminantes anteriores:

Tem-se  $x'_1 = [17,0 \ 62,0 \ 41,5 \ 40,0 \ 4,0 \ 46,0]$ , então

$$D_1(\underline{x}_1) = 12,65980(17,0) - 11,73583(62,0) + \\ + \dots + 3,81394(46,0) - 533,52007, \text{ ou}$$

$$D_1(\underline{x}_1) = 515,35597$$

De forma análoga obtém-se:  $D_2(\underline{x}_1) = 507,70321$  e

$$D_3(\underline{x}_1) = 486,07521$$

Como  $\max\{D_1(\underline{x}_1), D_2(\underline{x}_1), D_3(\underline{x}_1)\} = 515,35597 = D_1(\underline{x}_1)$ , conclui-se que  $\underline{x}_1 \in \pi_1$ , o que é realmente verdadeiro.

No caso da primeira observação de  $\pi_2$  tem-se:

$$\underline{x}'_2 = [16,5 \quad 61,0 \quad 38,5 \quad 40,0 \quad 3,0 \quad 41,0], \text{ então}$$

$$D_1(\underline{x}_2) = 12,65980(16,5) - 11,73583(61,0) + \\ + \dots + 3,81394(41,0) - 533,52007, \text{ ou}$$

$$D_1(\underline{x}_2) = 430,51523$$

De forma análoga obtém-se:

$$D_2(\underline{x}_2) = 446,32786 \text{ e } D_3(\underline{x}_2) = 442,48175.$$

Como  $\max\{D_1(\underline{x}_2), D_2(\underline{x}_2), D_3(\underline{x}_2)\} = 446,32786 = D_2(\underline{x}_2)$ , conclui-se que  $\underline{x}_2 \in \pi_2$ , o que é realmente verdadeiro.

### Análise de consistência

Utilizando o mesmo procedimento anterior reclassificam-se as 96 observações das populações  $\pi_1$ ,  $\pi_2$  e  $\pi_3$ , e constata-se que apenas a observação de número 18 da população  $\pi_1$  foi classificada de forma errada, sendo alocada na população  $\pi_2$ .

Os resultados estão apresentados no quadro a seguir:

Quadro – Distribuição das observações segundo a classificação *a priori* e por meio das funções discriminantes

Classificação <i>priori</i>	$a$	Função Discriminante			Total	Classificação Errada (%)		
		Pop.1	Pop.2	Pop.3				
Pop.1		31	1	0	32	3,12		
Pop.2		0	32	0	32	0,00		
Pop.3		0	0	32	32	0,00		
Total		31	33	32	96	-		

### Taxa de Erro Aparente

Calculada por meio da expressão  $T = \sum_{i=1}^g m_i / \sum_{i=1}^g n_i$ . Neste exemplo tem-se:

$$n_1 = n_2 = n_3 = 32$$

$$\sum_{i=1}^3 n_i = 32 + 32 + 32 = 96$$

$$m_1 = 1, m_2 = 0, m_3 = 0$$

$$T = \frac{1+0+0}{96} = 0,0104 = 1,04\%.$$

# CAPITULO 27

## Análise de Variáveis Canônicas

### Conceitos e descrição de análise de variáveis canônicas

A análise por meio de variáveis canônicas pode ser utilizada como uma alternativa à análise por componentes principais, quando se dispõe de dados experimentais com repetições, uma vez que para esta análise são necessárias as matrizes de médias ( $n \times p$ ) e de variâncias – covariâncias residuais ( $p \times p$ ), sendo  $n$  o número de observações e  $p$  o número de variáveis originais. As variáveis canônicas são descritas da seguinte forma:

- A variável canônica  $Y_{ij}$  é definida por uma combinação linear, expressa como:

$$Y_{ij} = a_1 X_{i1} + a_2 X_{i2} + \dots + a_p X_{ip}; i=1,2,\dots,n \text{ e } j=1,2,\dots,p$$

- A variável canônica  $Y_{ij'}$  é definida por uma nova combinação linear, expressa como:

$$Y_{ij'} = b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip}; i=1,2,\dots,n \text{ e } j=1,2,\dots,p$$

- Dentre todas as variáveis canônicas possíveis de serem estimadas,  $Y_{i1}$  é a que apresenta a maior variância,  $Y_{i2}$  a que apresenta a segunda maior variância e assim sucessivamente;

- A questão estatística consiste em: estimar os coeficientes de ponderação das variáveis originais em cada uma das variáveis canônicas; estimar as variâncias de cada uma das variáveis canônicas;

- Por exemplo, a variância da primeira variável canônica é expressa por:

$$V(Y_{i1}) = V(Y_1) = \mathbf{a}' T \mathbf{a}, \text{ em que } \mathbf{a} \text{ é um vetor } n \times 1 \text{ com elementos } a_j \quad (j=1,2,\dots,p);$$

- O vetor  $\mathbf{a}$  é obtido de forma que a variância de  $Y_1$  seja maximizada e que seus elementos sejam estimados sem as influências das variâncias e covariâncias residuais.

- A variância da segunda variável canônica é expressa por:

$V(Y_{i2}) = V(Y_2) = b' Tb$ . Os elementos do vetor  $b$ ,  $n \times 1$ , são os coeficientes  $b_j$  ( $j = 1, 2, \dots, p$ ), que devem ser estimados de forma que  $V(Y_2)$  seja maximizada, as influências das variâncias e covariâncias residuais sejam eliminadas e as variáveis canônicas  $(Y_1 e Y_2)$  sejam não correlacionadas;

As demais variáveis canônicas são estimadas de forma análoga à descrita anteriormente.

- As variáveis canônicas são estimadas a partir de dados transformados pelo processo de condensação pivotal, que permite obter novas variáveis, com variâncias residuais iguais a 1,0 e covariâncias residuais nulas.

- A transformação por condensação pivotal é feita da seguinte forma:

Seja  $X_1, X_2, \dots, X_p$ , um conjunto de variáveis originais, com matriz de covariâncias de médias  $T$  e matriz de covariâncias residuais  $E$ . Com o processo de condensação pivotal é obtido um novo conjunto de variáveis  $z_1, z_2, \dots, z_p$ , com matriz de covariâncias de médias  $T^*$  e matriz de covariâncias residuais  $I$  (matriz identidade). Na forma matricial a transformação é expressa por:

$$z' = V X'$$

$z$ : matriz  $n \times p$  de médias transformadas de  $n$  observações (indivíduos, objetos, tratamentos) em relação à  $p$  variáveis (características avaliadas);

$X$ : matriz  $n \times p$  de médias originais de  $n$  observações em relação à  $p$  variáveis;

$V$ : matriz  $p \times p$  de transformação, obtida pelo processo de condensação pivotal.

- Os autovalores e os autovetores associados às variáveis canônicas são estimados com base na matriz de covariâncias de médias de variáveis transformadas  $T^*$  e em seguida são obtidas as combinações lineares seguintes:

$$C_1 = \alpha_{11} z_1 + \alpha_{12} z_2 + \dots + \alpha_{1p} z_p = a_{11} X_1 + a_{12} X_2 + \dots + a_{1p} X_p$$

$$C_2 = \alpha_{21} z_1 + \alpha_{22} z_2 + \dots + \alpha_{2p} z_p = a_{21} X_1 + a_{22} X_2 + \dots + a_{2p} X_p$$

$$C_p = \alpha_{p1} z_1 + \alpha_{p2} z_2 + \dots + \alpha_{pp} z_p = a_{p1} X_1 + a_{p2} X_2 + \dots + a_{pp} X_p$$

- Os coeficientes  $a_{jj'}$ , associados às variáveis originais podem ser obtidos a partir dos coeficientes  $\alpha_{jj'}$ , estimados por meio da seguinte relação:

$$\begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jp} \end{bmatrix} = \begin{bmatrix} \alpha_{j1} & \alpha_{j2} & \cdots & \alpha_{jp} \end{bmatrix} V$$

Para eliminar os efeitos de escala de medida e realizar as inferências sobre as variáveis padronizadas, os coeficientes  $a_{jj'}$ , devem ser ainda multiplicados pelos desvios padrões residuais das variáveis, obtendo-se:

$$\delta_j = a_j \sqrt{\hat{\sigma}_j^2}, \text{ em que:}$$

$\hat{\sigma}_j^2$ : estimativa do quadrado médio do resíduo associado à  $j$ -ésima variável. Os valores  $\delta_j$  constituem uma medida da importância relativa da  $j$ -ésima variável em cada variável canônica.

- Os escores das primeiras variáveis canônicas (as que explicam  $\geq 80\%$  da variância total disponível) podem ser utilizados em gráficos de dispersão que mostram as distâncias geométricas entre as observações (indivíduos, objetos, métodos). As variáveis originais mais importantes são as que apresentam uma carga elevada nas primeiras variáveis canônicas.
- O uso das variáveis canônicas tem por objetivo reduzir o espaço  $p$ -dimensional das variáveis originais para um espaço  $r$ -dimensional de variáveis canônicas,  $r < p$ , em geral bi ou tridimensional. A eficácia deste uso depende do grau de distorção das distâncias geométricas entre as observações quando se passa do espaço  $p$ -dimensional para o espaço  $r$ -dimensional.
- Para as variáveis canônicas, o grau de distorção é influenciado pelas variâncias-covariâncias entre médias das observações e pelas variâncias-covariâncias residuais. Então, o grau de distorção pode ser medido com base numa relação entre variáveis canônicas e análise de agrupamento, comparando-se o total das distâncias geométricas em relação aos eixos que representam as primeiras variáveis canônicas com o total das distâncias generalizadas de Mahalanobis, da seguinte forma:

$GD = 1 - \alpha$ , sendo  $GD$ : Grau de Distorção;

$$\alpha = \frac{\sum_{i < i'} \sum d_{vcii'}^2}{\sum_i \sum_{i' < i} D_{ii'}^2}, \text{ em que:}$$

$\alpha$ : proporção acumulada da variância explicada pelas  $r$  variáveis canônicas;

$d_{vcii'}^2$ : quadrado da distância euclidiana estimada a partir dos escores de  $p$  variáveis canônicas;

$D_{ii'}^2$ : distância generalizada de Mahalanobis estimada a partir de  $p$  variáveis originais.

- As variáveis de menor importância para a separação das observações são aquelas cujos coeficientes de ponderação, obtidos a partir de variáveis padronizadas, são as de maior valor absoluto nas últimas variáveis canônicas.
- Quando o maior coeficiente de ponderação em uma variável canônica de menor variância está associado a uma variável original já descartada, recomenda-se não fazer nenhum descarte de variável com base nos coeficientes desta variável canônica. Neste caso, a identificação da importância relativa das variáveis originais deve prosseguir na próxima variável canônica de menor variância.

### Exemplo de Aplicação

Considere o estudo da divergência genética de oito genitores com base nas variáveis canônicas. Considere também os resultados básicos das análises de variâncias e de médias das avaliações de oito genitores em relação a quatro variáveis/características  $(X_1, X_2, X_3, X_4)$ , num experimento em blocos casualizados (Extraído de Cruz et al., 2004):

Quadro 1. Resumo das análises de variâncias de quatro características e seus Produtos Médios (PM)

Caractere	QM ou PM Blocos (3) <sup>1</sup>	QM ou PM Genótipos (7)	QM ou PM Resíduo (21)	PM	F
X1	46,6992	163,9392	18,7459	8,74**	
X2	0,2018	1,1408	0,1879	6,07**	
X3	0,0045	0,3448	0,0096	35,92**	
X4	123,4010	892,8013	72,1533	12,37**	
X1 e X2	-0,6209	0,4143	0,0151	-	
X1 e X3	0,1429	-0,1370	-0,0146	-	
X1 e X4	74,6738	291,7570	30,8056	-	
X2 e X3	0,0151	-0,2929	-0,0036	-	
X2 e X4	-1,5408	0,7343	0,8517	-	
X3 e X4	0,2529	-2,6835	0,0871	-	

\*\*:Significativo a 1% de probabilidade pelo teste F.

1 :Valores entre parênteses correspondem aos graus de liberdade.

Quadro 2. Médias de oito genitores em relação a quatro características

Genitores	$X_1$	$X_2$	$X_3$	$X_4$
1	41,900	20,300	3,900	85,675
2	43,800	19,750	3,650	98,250
3	37,300	18,725	4,600	74,575
4	40,150	20,300	4,300	91,625
5	32,500	20,250	4,100	54,125
6	52,750	19,725	4,375	100,375
7	43,900	20,225	4,275	91,000
8	49,250	20,025	4,150	82,175

As variáveis originais  $X_1, X_2, X_3$  e  $X_4$  apresentam matrizes de variâncias-covariâncias de médias das características e residuais  $T$  e  $E$ , respectivamente, como apresentadas a seguir:

$$T = \frac{1}{4} \begin{bmatrix} 163,9392 & 0,4143 & -0,1370 & 2 \\ & 1,1408 & -0,2929 & \\ & & 0,3448 & \\ & & & 8 \end{bmatrix}$$

$$E = \begin{bmatrix} 18,7459 & 0,0151 & -0,0146 & 30,8056 \\ & 0,1879 & -0,0036 & 0,8517 \\ & & 0,0096 & 0,0871 \\ & & & 72,1533 \end{bmatrix} \text{ e}$$

Observe que as matrizes  $T$  e  $E$  podem ser obtidas por meio de uma análise de variância multivariada.

- A transformação das variáveis originais em variáveis padronizadas é feita usando a matriz de transformação  $V$ , que é obtida por meio de um processo denominado condensação pivotal, aplicado na matriz  $E$ . Para este exemplo a matriz  $V$  é:

$$V = \begin{bmatrix} 0,2309 & 0 & 0 & 0 \\ -0,0018 & 2,3072 & 0 & 0 \\ 0,0078 & 0,2011 & 10,2581 & 0 \\ -0,4103 & -1,1580 & -3,3395 & 0,2486 \end{bmatrix}$$

A transformação das variáveis  $X_j$  em variáveis  $z_j$  é feita por meio da expressão seguinte:

$$z = X V', \text{ em que:}$$

$z$ : matriz  $n \times p$  de médias das variáveis padronizadas;

$X$ : matriz  $n \times p$  de médias das variáveis originais.

Quadro 3. Médias de variáveis transformadas por condensação pivotal para oito parentais

Parental	$z_1$	$z_2$	$z_3$	$z_4$
1	9,6774	46,7582	44,2935	-32,4207
2	10,1163	45,4857	41,6367	-28,6017
3	8,6150	43,1329	51,1309	-33,8070
4	9,2732	46,7615	48,3830	-31,5590
5	7,5064	46,6603	46,2615	-37,0187
6	12,1834	45,4115	49,1393	-34,1375
7	10,1394	46,5815	48,1414	-33,0827
8	11,3750	46,1101	46,8622	-36,8602

As variáveis padronizadas  $z_j$  apresentam matriz de variâncias-covariâncias de médias,  $T^*$ , e matriz de variâncias-covariâncias residuais,  $E^*$ , como a seguir:

$$T^* = \frac{1}{4} \begin{bmatrix} 8,74550 & 0,14930 & -0,00893 & 1,21554 \\ & 6,06808 & -6,40402 & -0,77009 \\ & & 35,08148 & -14,62830 \\ & & & 30,31572 \end{bmatrix}$$

$$E^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}$$

- Em seguida são obtidos os autovalores e os respectivos autovetores normalizados da matriz  $T^*$ . Os autovalores correspondem às variâncias das variáveis canônicas. A partir dos autovetores normalizados são obtidos os coeficientes de ponderação das variáveis canônicas.

Neste exemplo, as variáveis canônicas estimadas com a análise das variáveis  $z_1$ ,  $z_2$ ,  $z_3$  e  $z_4$ , originadas das variáveis originais  $X_1, X_2, X_3$  e  $X_4$  por condensação pivotal são as seguintes:

Variável Canônica	Variância ( $\lambda_j$ )	Variância Acumulada (%)	Coeficiente de ponderação associado a:			
			$z_1$	$z_2$	$z_3$	$z_4$
VC <sub>1</sub>	11,9702	59,70	-0,02019	-0,10570	0,76828	-0,63100
VC <sub>2</sub>	4,9086	84,18	0,07896	-0,31388	0,57564	0,75092
VC <sub>3</sub>	2,1722	95,02	0,99373	0,09507	-0,00858	-0,05818
VC <sub>4</sub>	0,9989	100,00	-0,07651	0,93876	0,27984	0,18592

Observe que  $\sum_j \lambda_j = \text{Traço}(\mathbf{T}^*) = 20,0499$  e que  $\sum_j \alpha_{jj'}^2 = 0$  e  $\sum_j \alpha_{jj'} \alpha_{jj''} = 0$

- A dispersão gráfica dos escores dos oito parentais obtidos por meio das duas primeiras variáveis canônicas, que explicam 84,18% da variação total, pode ser utilizada para avaliar a divergência entre os parentais. Por exemplo, considerando as médias das variáveis transformadas (Quadro 3), estima-se o escores da primeira variável canônica, para o primeiro parental da seguinte forma:

$$\begin{aligned} VC_1 &= -0,02019z_1 - 0,10570z_2 + 0,76828z_3 - 0,63100z_4 \\ VC_1 &= -0,02019(9,6774) - 0,10570(46,7582) + \\ &+ 0,76828(44,2935) - 0,63100(-32,4207) \\ VC_1 &= 49,2481 \end{aligned}$$

Os escores para todos os parentais para as duas primeiras variáveis canônicas estão no quadro a seguir:

Quadro 4. Escores de oito parentais para as duas primeiras variáveis canônicas

Parentais	VC <sub>1</sub>	VC <sub>2</sub>
1	49,2481	-12,6703
2	44,9301	-10,9040
3	55,7635	-8,7123
4	51,8453	-9,9671
5	53,7093	-15,1257
6	54,1336	-10,5424
7	52,6223	-10,8550
8	54,0488	-14,1822
Variância	11,9702	4,9086
Covariância	0,00	

Pode ser realizada a dispersão gráfica dos escores das duas primeiras variáveis canônicas. A distância gráfica entre cada par de parentais é obtida por meio da distância euclidiana, ou seja:

$$dVC_{ii'} = \sqrt{[VC_{1i} - VC_{1i'}]^2 + [VC_{2i} - VC_{2i'}]^2}, \text{ para } i' > i; i = 1, 2, \dots, 8.$$

Por exemplo, obtém-se a distância gráfica entre os parentais 1 e 2 da seguinte forma:

$$dVC_{12} = \sqrt{[49,2481 - 44,9301]^2 + [(-12,6703) - (-10,9040)]^2} = 4,6653$$

As distâncias (dissimilaridades) entre todos os outros pares de parentais, baseadas nas distâncias euclidianas, obtidas a partir dos escores das duas primeiras variáveis canônicas, são calculadas de forma análoga e estão apresentadas no quadro a seguir:

Quadro – Dissimilaridade entre parentais com base nas duas primeiras variáveis canônicas

Parentais	2	3	4	5	6	7	8
1	4,66	7,62	3,95	5,09	5,33	3,83	5,03
2		11,05	7,02	9,74	9,21	7,69	9,69
3			4,04	6,73	2,45	3,80	5,73
4				5,74	2,44	1,39	5,00
5					4,60	4,41	1,00
6						1,54	3,64
7							3,62

O grau de distorção das medidas de dissimilaridade apresentadas no quadro anterior pode ser

calculado por meio da estatística  $\alpha = \left( \sum_i \sum_{i' < i} d^2 VC_{ii'} / \sum_i \sum_{i' < i} D^2_{ii'} \right) (100)$ . Neste exemplo

tem-se  $\alpha = \frac{945,2152}{1125,3940} (100) = 84\%$ . Então, o grau de distorção  $(1-\alpha)$  é igual a 16%.

Como não existe relação direta entre as variáveis transformadas e as variáveis originais. Então, para avaliação da importância relativa das variáveis originais (características avaliadas), é necessária a

obtenção do vetor  $\alpha$  (coeficientes de ponderação das variáveis originais) a partir do vetor  $\alpha'$  (coeficientes de ponderação das variáveis transformadas), com base na seguinte expressão:  $\alpha' = \alpha'V$ , em que:

$V$ : matriz de transformação das variáveis originais em variáveis com covariâncias residuais nulas e variâncias residuais iguais a 1,0.

Por exemplo, tem-se:

$$VC_1 = -0,02019z_1 - 0,10570z_2 + 0,76828z_3 - 0,63100z_4. \quad \text{Deve-se obter:}$$

$$VC_1 = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4. \text{ Então se obtém:}$$

$$\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix} = \begin{bmatrix} -0,02019 & -0,10570 & 0,76828 & -0,63100 \end{bmatrix}^T$$

$$\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix} = \begin{bmatrix} 0,2602 & 0,6364 & 9,9636 & -0,1568 \end{bmatrix}$$

Os demais coeficientes de ponderação das variáveis originais, obtidos de forma análoga, estão contidos no quadro a seguir:

Quadro – Variáveis canônicas  $(VC_1, VC_2, VC_3, VC_4)$  obtidas pela combinação linear de quatro variáveis originais  $(X_1, X_2, X_3, X_4)$

Variáveis canônicas	Coeficientes de ponderação $(a_j)$ associados a:			
	$X_1$	$X_2$	$X_3$	$X_4$
$VC_1$	0,2602	0,6364	9,9636	-0,1568
$VC_2$	-0,2845	-1,4800	3,4131	0,1866
$VC_3$	0,2531	0,2849	0,1047	-0,0144
$VC_4$	-0,0934	2,0054	2,2524	0,0462

Observe que no quadro anterior tem-se:

$$\sum_j \sum_{j'} a_j a_{j'} \hat{\sigma}_{jj'} = \sum_j \sum_{j'} b_j b_{j'} \hat{\sigma}_{jj'} = 1 \text{ e } \sum_j \sum_{j'} a_j b_{j'} \hat{\sigma}_{jj'} = 0$$

Os coeficientes de ponderação das características nas variáveis canônicas são utilizados para quantificar a importância relativa das características para o estudo da divergência genética dos parentais. Como esses coeficientes, que são elementos de autovetores, são afetados pelas escalas de avaliação das características, recomenda-se que a avaliação da importância relativa das características sobre a divergência seja feita utilizando coeficientes de ponderação associados a

variáveis padronizadas, obtidos por meio da seguinte expressão:  $\delta_j = a_j \sqrt{\hat{\sigma}_j^2}$ , em que:

$\hat{\sigma}_j^2$ : quadrado médio do resíduo associado à  $j$ -ésima característica.

Os coeficientes de ponderação associados às variáveis padronizadas são apresentados no quadro a seguir:

Quadro – Variáveis canônicas  $(VC_1, VC_2, VC_3, VC_4)$  obtidas pela combinação linear de quatro variáveis padronizadas  $(x_1, x_2, x_3, x_4)$

Variáveis canônicas	Coeficientes de ponderação $(\delta_j)$ associados a:			
	$x_1$	$x_2$	$x_3$	$x_4$
$VC_1$	1,1266	0,2759	0,9762	-1,3317
$VC_2$	-1,2318	-0,6416	0,3344	1,5848
$VC_3$	1,0959	0,1235	0,0102	-0,1227
$VC_4$	-0,4046	0,8693	0,2207	0,3924

Observe que as características de menor importância foram a  $X_2$ , com maior coeficiente de ponderação na  $VC_4(0,8693)$  e a  $X_1$ , com maior coeficiente de ponderação na  $VC_3(1,0959)$ . As características de menor importância são aquelas que são invariantes (baixos valores da estatística  $F$ ) ou que apresentam redundância (apresentam alta correlação com outras características ou combinação de características).

# CAPITULO 28

## Análise Espacial

### Introdução

Os delineamentos experimentais tradicionais podem não ser suficientes para considerar / controlar a heterogeneidade espacial que surge devido a posição do experimento, condições de solo, doenças, pragas e plantas daninhas. Esta heterogeneidade não aleatória gera uma autocorrelação espacial, que se não for considerada numa análise, pode resultar em estimativas de efeitos de tratamentos incorretas, erros correlacionados e baixo poder experimental, o que viola as pressuposições dos modelos lineares e invalida as análises. A incorporação da correlação espacial entre parcelas experimentais pode melhorar a acurácia e a precisão das estimativas. Aplicações de análise estatística espacial a experimentos agrícolas de campo requerem moderado conhecimento de ferramentas de Geoestatística e de análise de variância de modelos autoregressivos.

### Geoestatística na análise de experimentos

O trabalho de Kridge (1951) com dados de concentração de ouro, permitiu concluir que somente a informação de variância seria insuficiente para explicar a variação dos dados. Então, foi admitido que seria necessário considerar a distância entre as observações, a qual deve levar em consideração a posição geográfica e a dependência espacial. Para resolver este problema, foi desenvolvida a teoria das variáveis regionalizadas, também denominada de Geoestatística devido sua aplicação inicial nos estudos de Geologia e Mineralogia.

Variáveis regionalizadas são descritas por uma função que varia de um local para outro no espaço, com certa aparência de continuidade. Portanto, são variáveis cujos valores estão relacionados de alguma forma com a posição espacial que ocupam. A continuidade das variáveis regionalizadas ocorre devido a tendência de dois pontos amostrais terem valores mais próximos, quanto menos afastados estejam um do outro (Guerra, 1988).

Uma medida tomada em um ponto contém relações de dependência com medidas tomadas em pontos adjacentes, sugerindo a existência de uma estrutura de correlações. As características dessas variáveis violam os pressupostos dos métodos de análise de dados da estatística clássica, que pressupõem a independência entre as observações (Clark, 1979).

Dessa forma, a Geoestatística é utilizada com os seguintes objetivos (Journel; Huijbregts, 1978):

- 1) Extrair da aparente desordem dos dados, uma imagem da variabilidade dos mesmos e uma medida da correlação entre valores tomados em dois pontos do espaço, que é obtida por meio do semivariograma;
- 2) Medir a precisão da predição ou estimativa com base em dados fragmentados, obtidos por um processo discreto de amostragem, ou seja, obter uma estimativa contínua da área amostrada por meio da técnica de *krigagem*.

## Modelagem de variabilidade espacial

Na análise espacial, a pressuposição para a modelagem de relações entre distâncias e associação espacial, é que a associação de variáveis em pontos distintos é maior à medida que estes pontos estejam mais próximos um do outro. Em Geoestatística, a medida que descreve essa relação é a semivariância, que é uma medida de dissimilaridade, ou seja, o valor da semivariância é maior à medida que as variáveis estão menos associadas. Na teoria Geoestatística, uma, dentre três hipóteses seguintes devem ser assumidas: Estacionariedade de primeira ordem; Estacionariedade de segunda ordem; Hipótese intrínseca.

Um processo é estacionário se o seu desenvolvimento no tempo ou espaço ocorrer de maneira homogênea, com oscilações aleatórias contínuas em torno de um valor médio, em que nem a amplitude média e nem as oscilações mudam bruscamente no tempo ou no espaço. Por exemplo, define-se uma variável aleatória  $Z(x)$ , onde  $x$  denota uma posição em duas dimensões com componentes  $(x_i, y_i)$ , como estacionária, se todos os momentos estatísticos são invariantes para toda mudança de origem. Então, pode-se dizer que, se o processo é estacionário de ordem  $k$ , então:

$$E[Z(x)] = m_1(x) = c \forall x$$

$$E[Z(x)] = m_2(x) = c \forall x$$

⋮

, em que:  $c$ : constante;  $m_k(x)$ :  $k$ -ésimo momento da variável

$$E[Z(x)] = m_4(x) = c \forall x$$

aleatória  $Z(x)$ .

Em Geoestatística exige como restrição máxima Estacionariedade de segunda ordem. Se um processo é estacionário de ordem  $k$  ele também será estacionário para ordens inferiores a  $k$ .

A hipótese intrínseca é a mais frequentemente usada em Geoestatística por ser menos restritiva do que estacionariedade de primeira ou segunda ordem. Na hipótese intrínseca tem-se:

1-A esperança de  $Z(x)$  existe e não depende do ponto  $x$ , sendo

$$E[Z(x)] = \mu \quad (1)$$

2-Para todo vetor de distância  $h$ , a variância da diferença  $[Z(x+h) - Z(x)]$  existe e não depende do ponto  $x$ , tal que:

$$Var[Z(x+h) - Z(x)] = E\left\{[Z(x+h) - Z(x)]^2\right\} = 2\gamma(h) \quad (2)$$

, sendo  $\gamma(h)$  denominada de semivariância.

A semivariância é a medida do grau de dependência espacial entre duas amostras, cuja magnitude depende da distância entre elas. O gráfico de semivariâncias em função de distâncias a um ponto é chamado semivariograma. Este gráfico permite representar a variação de um fenômeno regionalizado no espaço (JOURNEL; HUIJBREGTS, 1978). Por exemplo, considere duas variáveis regionalizadas  $X$  e  $Y$ , onde  $X = Z(x)$  e  $Y = Z(x+h)$ , referentes à mesma característica, mas medidas em duas posições diferentes, em que  $h$  é um vetor distância que separa os dois pontos veja Figura 1.

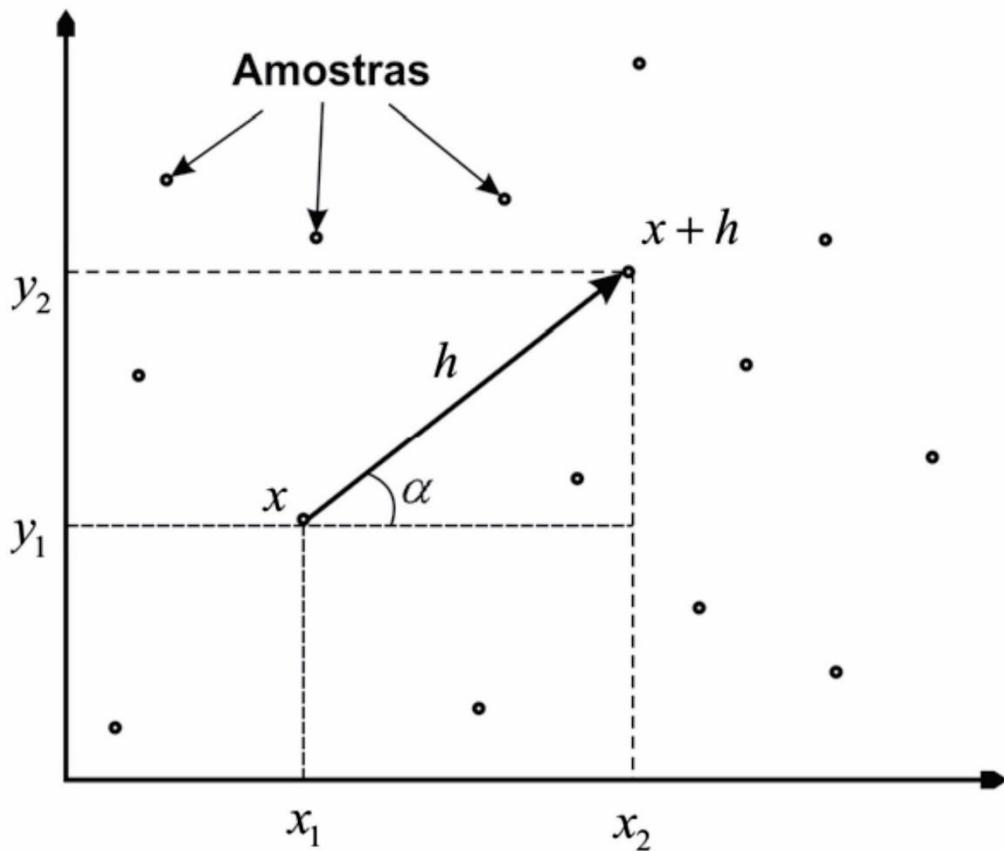


Figura 1. Amostragem em duas dimensões. Fonte: Rossoni, 2011.

O grau de dependência entre duas variáveis regionalizadas,  $X$  e  $Y$  é representado pela variância  $2\gamma(h)$ , que é definida como a esperança matemática do quadrado da diferença entre os valores de pontos no espaço, separados pelo vetor distância  $\mathbf{h}$ , ou seja,  

$$2\gamma(h)=E\left\{\left[Z(x+h)-Z(x)\right]^2\right\}.$$
 Por conveniência matemática, o valor 2 passou a dividir a expressão e  $\gamma(h)$  chamada de semivariância. Então, o estimador da semivariância de uma amostra  $Z(x_i)$ , com  $i=1,2,\dots,n$ , é definido como (CRESSIE, 1993):

$$\hat{\gamma}(h)=\frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i+h)-z(x_i)]^2 \quad (3)$$

, em que:

$N(h)$ : número de pares de valores medidos para  $Z(x_i)$  e  $Z(x_i+h)$ ;

$z(x_i)$  e  $z(x_i+h)$ : valores da  $i$ -ésima observação da variável regionalizada, coletados nos pontos amostrais  $x_i$  e  $x_i+h$ , com  $i=1,2,\dots,n$ , separados pelo vetor distância  $\mathbf{h}$ . O gráfico

das semivariâncias estimadas entre o vetor distância máximo e mínimo é chamado de semivariograma empírico.

Segundo Isaacks e Srivastava (1989) os parâmetros de um semivariograma, mostrados na Figura 2 e descritos a seguir são:

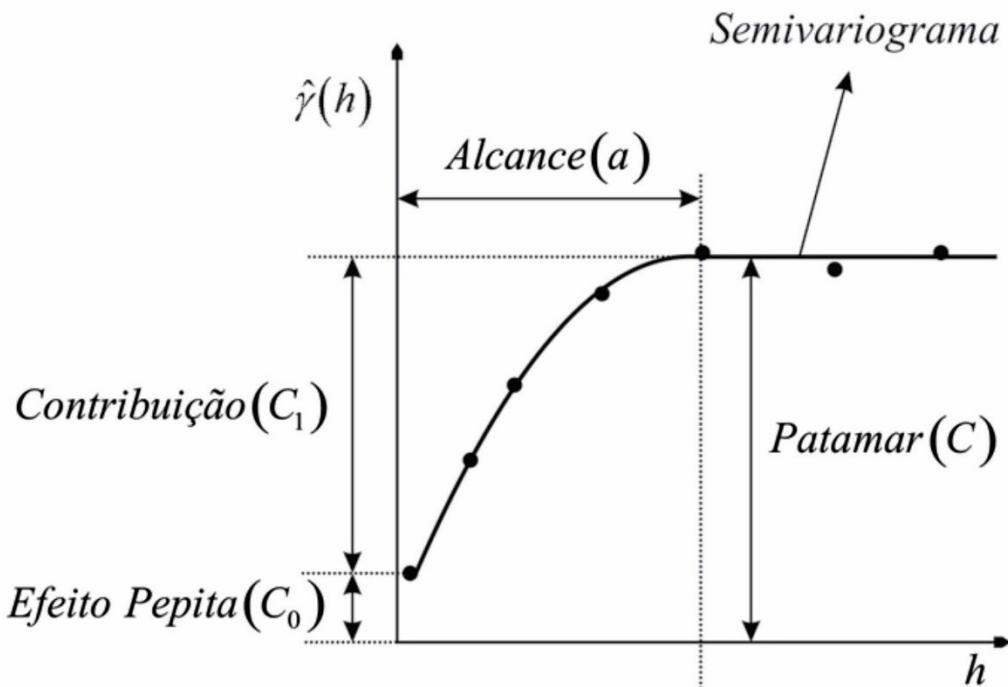


Figura 2. Exemplo de semivariograma. Fonte: Rossoni, 2011

1-Alcance( $a$ ): distância dentro da qual as amostras apresentam-se correlacionadas espacialmente;

2-Patamar( $C$ ): valor da semivariância correspondente ao valor do alcance, ou seja,  $\gamma(a) = C$ . A partir deste ponto, considera-se que não existe mais dependência espacial entre as amostras, o que significa que a variância da diferença entre pares de amostras, estimada como  $Var[Z(x+h) - Z(x)]$ , torna-se invariante com a distância;

3-Pepita( $C_0$ ): à medida que  $h$  tende para zero,  $\gamma(h)$  se aproxima de um valor positivo chamado efeito pepita ( $C_0$ ), que revela a descontinuidade do semivariograma para distâncias menores que a menor distância entre as amostras. Esta descontinuidade pode ser devida a erros de medição ou a variabilidade de pequena escala não captada pela amostragem;

4-Contribuição( $C_1$ ): é a diferença entre o patamar( $C$ ) e o efeito pepita( $C_0$ ).

O ajuste de modelos teóricos de semivariogramas ao semivariograma empírico é utilizado para avaliar a capacidade de semivariogramas em detectar a variabilidade espacial da amostra. Neste processo, os modelos teóricos de semivariogramas são superpostos aos pontos obtidos no semivariograma empírico, de modo que o modelo que melhor se ajusta aos pontos representa a

magnitude, o alcance e a intensidade da variabilidade espacial da variável estudada. Os modelos mais utilizados são (Journel; Huijbregts, 1978):

1-Modelo Esférico – é um dos modelos mais utilizados:

$$\gamma(h) = \begin{cases} 0 & , \quad |h|=0 \\ C_0 + C_1 \left[ 1,5 \left( \frac{|h|}{a} \right) - 0,5 \left( \frac{|h|}{a} \right)^3 \right] & , \quad 0 < |h| \leq a \\ C_0 + C_1 & , \quad |h| > a \end{cases} \quad (4)$$

2-Modelo Exponencial – modelo bastante utilizado:

$$\gamma(h) = \begin{cases} 0 & , \quad |h|=0 \\ C_0 + C_1 \left[ 1 - \exp \left( - \frac{|h|}{a} \right) \right] & , \quad |h| \neq 0 \end{cases} \quad (5)$$

3-Modelo Gaussiano – modelo utilizado no ajuste de fenômenos que demoram certa distância para aumentar:

$$\gamma(h) = \begin{cases} 0 & , \quad |h|=0 \\ C_0 + C_1 \left[ 1 - \exp \left( - \frac{|h|^2}{a^2} \right) \right] & , \quad |h| \neq 0 \end{cases} \quad (6)$$

Os fenômenos de isotropia e anisotropia são definidos da seguinte forma: Se são construídos semivariogramas direcionais, geralmente nas direções 0º, 45º, 90º e 135º, e estes apresentam o mesmo comportamento espacial, então diz-se que a variável é isotrópica. Quando a variável apresenta diferentes padrões de dependência espacial, obtidos por meio de semivariogramas direcionais, diz-se que a variável é anisotrópica. A anisotropia é classificada da seguinte forma (Isaaks e Srivastava, 1989):

1-Geométrica – os semivariogramas direcionais apresentam o mesmo patamar ( $C$ ) com diferentes alcances ( $a$ );

2-Zonal – os semivariogramas direcionais apresentam o mesmo alcance ( $a$ ) com diferentes patamares ( $C$ );

3-Combinada – os semivariogramas direcionais apresentam diferentes patamares ( $C$ ) e alcances ( $a$ ).

O método de krigagem, termo derivado do nome de Daniel G. Krige, que introduziu o uso de médias móveis para evitar a superestimação sistemática de reservas minerais, foi desenvolvido para solucionar problemas de mapeamentos geológicos. A krigagem é utilizada nos casos em que além de um modelo de dependência espacial existe o interesse de se obter valores em pontos não amostrados ou obter uma malha de pontos interpolados que permita uma melhor visualização do comportamento da variável na região, por meio de mapas ou gráficos de superfície.

No método de interpolação denominado krigagem ordinária, os pesos atribuídos às diferentes amostras são determinados a partir da análise espacial, baseada no semivariograma empírico que foi ajustado com base em um modelo teórico. Em média, a krigagem fornece estimativas não tendenciosas e com variância mínima. Um dos métodos de estimação, denominado krigagem ordinária, é a krigagem que envolve apenas o semivariograma e cuja função média não é conhecida, mas estimada a partir dos dados, tornou-se a mais conhecida.

O estimador de krigagem ordinária é descrito da seguinte forma (Isaaks e Srivastava, 1989):

Considere que existe interesse em estimar  $Z(x_0)$ , que é uma variável a ser avaliada para um local qualquer  $x_0$  onde não se tem valores medidos. O estimador de krigagem ordinária é definido pela seguinte expressão:

$$\hat{Z}(x_0) = \sum_{i=1}^N \lambda_i Z(x_i) \quad (7)$$

, em que:

$Z(x_i)$ : valores medidos;

$\lambda_i$ : pesos associados a cada valor medido;

$N$ : número de valores medidos.

Neste método de krigagem, os pesos variam de acordo com a variabilidade espacial expressa no semivariograma. Então, o estimador  $\hat{Z}(x_0)$  nada mais é do que uma média móvel ponderada.

Para que a krigagem ordinária seja um interpolador ótimo, o estimador de krigagem deve ser não tendencioso e ter variância mínima, ou seja:

$$E\{\hat{Z}(x_0) - Z(x_0)\} = 0 \quad (8)$$

$$\text{e } Var\{\hat{Z}(x_0) - Z(x_0)\} = E\left\{\left[\hat{Z}(x_0) - Z(x_0)\right]^2\right\} = \text{mínimo} \quad (9)$$

Substituindo a expressão (7) na expressão (8) e aplicando as propriedades de esperança matemática, obtém-se:

$$E\{\hat{Z}(x_0) - Z(x_0)\} = E\left\{\sum_{i=1}^N \lambda_i Z(x_i) - Z(x_0)\right\} = 0 \quad (10)$$

$$\text{e } E\left\{\sum_{i=1}^N \lambda_i Z(x_i) - Z(x_0)\right\} = \sum_{i=1}^N \lambda_i E\{Z(x_i)\} - E\{Z(x_0)\} = 0 \quad (11)$$

Aplicando a primeira condição de estacionaridade,  $E\{z(x_i)\} = \mu$ , tem-se:

$$E\{\hat{Z}(x_0) - Z(x_0)\} = \sum_{i=1}^N \lambda_i \mu - \mu = \mu \left( \sum_{i=1}^N \lambda_i - 1 \right) = 0 \quad (12)$$

Para qualquer valor de  $\mu$ , a expressão (12) é verdadeira se  $\sum_{i=1}^N \lambda_i - 1 = 0$  ou  $\sum_{i=1}^N \lambda_i = 1$ , o que significa que, para qualquer distribuição dos valores dos pesos, a estimativa  $E\{\hat{Z}(x_0) - Z(x_0)\}$  é não tendenciosa para soma de pesos igual a 1.

Sob a condição  $\sum_{i=1}^N \lambda_i = 1$ , a minimização da variância do erro,  $Var\{\hat{Z}(x_0) - Z(x_0)\}$ ,

permite a obtenção do seguinte sistema de equações, denominado de sistema de krigagem ordinária:

$$\begin{cases} \sum_{j=1}^N \lambda_j \gamma(x_i, x_j) - \alpha = \gamma(x_i, x_0) \\ \sum_{j=1}^N \lambda_j = 1 \end{cases} \quad (13)$$

, em que:

$\gamma(x_i, x_j)$ : semivariância entre os pontos  $x_i$  e  $x_j$ , para  $i = 1, 2, \dots, n$ ;

$\gamma(x_i, x_0)$ : semivariância entre os pontos  $x_i$  e  $x_0$ ;

$\alpha$ : multiplicador de Lagrange, necessário para a minimização da variância do erro.

A variância do erro minimizada, denominada de variância de krigagem ordinária, é expressa como (CRESSIE, 1993):

$$\begin{aligned} \sigma_{ko}^2 &= Var[\hat{Z}(x_0) - Z(x_0)] \\ &= Cov(x_0, x_0) - \sum_{i=1}^N \lambda_i Cov(x_0, x_0) - \alpha \end{aligned} \quad (14)$$

Conforme as expressões (13) e (14), a krigagem ordinária é um interpolador exato, no sentido de que os valores interpolados irão coincidir com os valores de pontos amostrais.

Quando um modelo teórico de semivariograma é ajustado a um conjunto de dados, o método de validação cruzada é utilizado para verificar se a modelagem foi adequada. A ideia básica do método

consiste em excluir uma observação e por meio das observações remanescentes estimar a observação excluída. Em seguida, o valor estimado é comparado com o valor observado. Então, a validação cruzada de um método de krigagem pode ser descrita da seguinte forma:

1-Suprimir do conjunto de dados um ponto qualquer  $Z(x_i)$ ;

2-Obter a estimativa  $\hat{Z}(x_i)$  do ponto suprimido, com base nos pontos remanescentes;

3-Estimar o erro de estimação por meio da seguinte expressão:

$$\frac{[Z(x_i) - \hat{Z}(x_i)]}{\sigma_{x_i}} \quad (15)$$

4-Repetir os passos anteriores para cada um de todos os pontos da amostra;

5-Estimar o Erro Médio Reduzido ( $ER$ ) e o Desvio Padrão do Erro Reduzido ( $S_{ER}$ ), por meio das seguintes expressões:

$$ER = \frac{1}{N} \sum_{i=1}^N \left( \frac{[Z(x_i) - \hat{Z}(x_i)]}{\sigma_{x_i}} \right) \quad (16)$$

$$\text{e } S_{ER} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \frac{[Z(x_i) - \hat{Z}(x_i)]}{\sigma_{x_i}} \right)^2} \quad (17)$$

, em que:

$Z(x_i)$ : valor observado no ponto  $x_i$ ;

$\hat{Z}(x_i)$ : valor estimado para o ponto  $x_i$ ;

$\sigma_{x_i}$ : desvio padrão da krigagem no ponto  $x_i$ .

Para valor de  $ER$  próximo de zero e valor de  $S_{ER}$  próximo de um, o modelo é considerado como bem ajustado (CRESSIE, 1993).

A modelagem Geoestatística do erro pode ser descrita da seguinte forma:

Considere o modelo linear geral, expresso como:  $Y = X\beta + \varepsilon$ , em que:

$Y$ : vetor  $n \times 1$  de valores observados;

$X$ : matriz  $n \times p$  das variáveis independentes;

$\beta$ : vetor  $p \times 1$  de parâmetros;

$\boldsymbol{\varepsilon}$  : vetor  $n \times 1$  de erros associados a cada observação, assumidos como independentes, com distribuição normal, média zero e idênticos, ou seja,  $\boldsymbol{\varepsilon} \sim N(\boldsymbol{\phi}, \mathbf{I}\sigma^2)$ , em que  $\boldsymbol{\phi}$  é um vetor nulo,  $\mathbf{I}$  é a matriz identidade e  $\sigma^2$  a variância amostral.

Quando os erros não são independentes, como no caso de autocorrelação espacial, são assumidos como (DUARTE, 2000):

$\boldsymbol{\varepsilon} \sim N(\boldsymbol{\phi}, \mathbf{R})$ , em que  $\boldsymbol{\phi}$  é um vetor nulo e  $\mathbf{R}$  é a matriz de variâncias e covariâncias dos erros, definida como:

$$\mathbf{R} = \begin{bmatrix} \sigma^2 & Cov(h_i) & \cdots & Cov(h_{máx}) \\ Cov(h_i) & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ Cov(h_{máx}) & \cdots & \cdots & \sigma^2 \end{bmatrix}, \text{em que:}$$

$Cov(h_i)$ : covariância entre duas parcelas  $(x_i, x_i + h_i)$  separadas por uma distância  $h_i$ , obtida por meio de  $Cov(h_i) = 1 - \gamma(h_i)$ , considerando as pressuposições de estacionaridade de primeira e segunda ordem;

$h_{máx}$ : distância máxima entre duas parcelas no experimento.

### Análise de variância usando modelos autoregressivos

A modelagem autoregressiva (AR) lida com modelos lineares de predição que procuram prever uma saída  $y_n$  de um sistema, com base nas saídas anteriores  $(y_{n-1}, y_{n-2}, \dots)$ . Quando são acrescentadas informações de localização a esses modelos, são obtidos os modelos autoregressivos que consideram a dependência espacial, como por exemplo os modelos SAR e CAR (GRIFFITH, 1988). O modelo SAR (spatial auto regressive model), descreve a variação espacial para o vetor resposta  $Y_{n \times 1}$ , sendo definido como:

$$Y = \rho W Y + X \beta + \varepsilon \quad (18)$$

, em que:

$Y$  : vetor  $n \times 1$  de valores observados;

$\rho$  : parâmetro espacial autoregressivo;

$W$  : matriz  $n \times n$  com atribuições de pesos da vizinhança espacial;

$X$  : matriz  $n \times p$  de incidência dos efeitos fixos;

$\beta$  : vetor  $p \times 1$  de parâmetros;

$\varepsilon$  : vetor  $n \times 1$  de erros associados a cada observação.

A matriz  $\mathbf{W}$  é obtida como  $\mathbf{W} = \mathbf{D} \times \mathbf{C}$ , ou seja, por meio da multiplicação das matrizes  $\mathbf{D}$  e  $\mathbf{C}$ , definidas a seguir. A matriz  $\mathbf{C}$   $n \times n$  é binária e descreve a vizinhança das parcelas experimentais. Considere quatro padrões de proximidade para definir a região de vizinhança, para uma amostragem com grid regular, como proposto por Gumpertz et al. (1997):

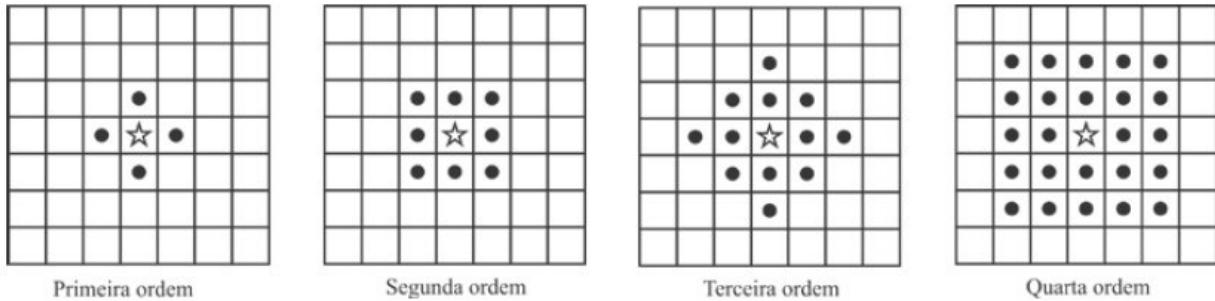


Figura 3. Padrões de proximidade, segundo Gumpertz et al. (1997), em que  $\star$  representa a parcela de referência e  $\bullet$  representa as parcelas vizinhas consideradas na comparação e  $\square$  representa as demais parcelas. Fonte: Rossoni, 2011.

Na Figura 3, considera-se que cada observação adjacente está distante uma da outra na proporção de uma unidade de medida ( $1u.m.$ ). Logo, os centros de cada uma das parcelas também estão distantes  $1u.m.$  dos outros centros adjacentes. Então, pode-se definir a vizinhança de uma parcela como as parcelas que são abrangidas pela circunferência de raio  $\mathbf{r}$ , partindo da parcela de origem. Dessa forma, pode-se descrever os padrões de proximidade em função dos raios, centrados na parcela de referência, como apresentado no Quadro 2 a seguir.

Quadro 2. Padrões de proximidade em função do raio da circunferência

Padrão de proximidade	Raio da circunferência
Primeira ordem	$1u.m.$
Segunda ordem	$1\sqrt{2}u.m.$
Terceira ordem	$2u.m.$
Quarta ordem	$2\sqrt{2}u.m.$
:	:
n-ésima ordem ímpar	$n u.m.$
n-ésima ordem par	$n\sqrt{2}u.m.$

Dessa forma, cada elemento  $c_{ij}$  da matriz  $\mathbf{C}$  pode ser definido como:

$c_{ij} = 1$ , para as parcelas contidas na circunferência de raio  $\mathbf{r}$ , centradas em  $c_{ij}$  e  $c_{ij} = 0$ , caso contrário, ou seja:

$$c_{ij} = \begin{cases} 1 & : c_{ij}, \text{adjacente}, c_{i+1,j} \\ 1 & : c_{ij}, \text{adjacente}, c_{i-1,j} \\ 1 & : c_{ij}, \text{adjacente}, c_{i,j+1} \\ 1 & : c_{ij}, \text{adjacente}, c_{i,j-1} \\ 0 & : \text{caso contrário} \end{cases}$$

Para exemplificar a composição da matriz  $C$ , considere um experimento com 9 observações, como apresentado na Figura 4 a seguir:

1	2	3
4	5	6
7	8	9

Figura 4. Grid regular com nove observações. Fonte: Rossoni, 2011.

Admitindo que a parcela número 1 seja a parcela de referência e considerando um padrão de proximidade de primeira ordem, as parcelas 2 e 4 seriam consideradas vizinhas. Então, na matriz  $C$  os elementos  $c_{12}$  e  $c_{14}$  receberiam o valor 1 e os demais  $c_{1i} = 0$ , para  $i = 3, 5, 6, 7, 8, 9$ .

Considerando agora a parcela número 5 como referência e mesmo padrão de proximidade, tem-se os elementos  $c_{52} = c_{54} = c_{56} = c_{58} = 1$  correspondentes às parcelas vizinhas e os demais  $c_{5i} = 0$ , para  $i = 1, 3, 7, 9$ .

Portanto, para o exemplo da Figura 2, a matriz de vizinhança ( $C$ ) considerando um padrão de proximidade de primeira ordem é expressa como:

$$C = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

A matriz  $D$  é diagonal e formada por elementos  $1/k_i$ , sendo  $k_i$  a soma dos valores da  $i$ -ésima linha da matriz  $C$ , neste exemplo expressa como:

$$D = \begin{bmatrix} 1/2 & & & & & & & & \\ & 1/3 & & & & & & & \\ & & 1/2 & & & & & & \\ & & & 1/3 & & & & & \\ & & & & 1/4 & & & & \\ & & & & & 1/3 & & & \\ & & & & & & 1/2 & & \\ & & & & & & & 1/3 & \\ & & & & & & & & 1/2 \end{bmatrix}$$

Portanto, tem-se que:

$$W = \begin{bmatrix} 1/2 & & & & & & & & \\ & 1/3 & & & & & & & \\ & & 1/2 & & & & & & \\ & & & 1/3 & & & & & \\ & & & & 1/4 & & & & \\ & & & & & 1/3 & & & \\ & & & & & & 1/2 & & \\ & & & & & & & 1/3 & \\ & & & & & & & & 1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

O parâmetro  $\rho$  do modelo SAR pode ser estimado por meio do método de máxima verossimilhança, que tem como base o princípio de selecionar valores dos parâmetros que maximizem a probabilidade de serem obtidos os dados observados. A solução da estimação por máxima verossimilhança de um modelo espacial autoregressivo, proposta por Ord (1975), consiste em explorar a decomposição do Jacobiano  $|I - \rho W|$  em termos dos autovalores  $\omega_i$  da matriz  $W$ , expressa como:

$$|I - \rho W| = \prod_{i=1}^n (1 - \rho \omega_i) \text{ ou } \ln |I - \rho W| = \sum_{i=1}^n \ln(1 - \rho \omega_i) \quad (19)$$

O polinômio obtido a partir do uso de qualquer uma das equações anteriores não tem solução única e, portanto, deve ser solucionado por meio de métodos iterativos computacionais.

A análise de variância de um modelo autoregressivo (ANOVA-AR) consiste basicamente em transformar observações consideradas autocorrelacionadas em observações não-correlacionadas. Para realizar esta transformação deve-se ajustar os dados observados, após estimar  $\rho$ , por meio da seguinte expressão (Long, 1996):

$$Y_{adj} = Y - (\hat{\rho} W Y - \hat{\rho} \beta_0) \quad (20)$$

, em que:

$\hat{Y}_{adj}$ : vetor  $n \times 1$  de valores ajustados;

$\hat{\rho}$ : vetor  $n \times 1$  de valores observados;

$\hat{\rho}$ : estimativa do parâmetro espacial autoregressivo;

$W$ : matriz  $n \times n$  contendo as atribuições de peso da vizinhança espacial;

$\beta_0$ : média dos valores observados.

Em seguida, é realizada a análise de variância com os valores ajustados, que é utilizada como base para a construção do quadro da ANOVA-R, como apresentado a seguir:

	$FV$	$GL$	$SQ$	$QM$	$F$
Fator ( $\rho$ )	1		$SQ_{\rho}$	-	-
Parâmetros	$k$		$SQP_{adj}$	$QMP_{adj}$	$QMP_{adj} / QME_{adj}$
Resíduo	$n - k - 2$		$SQE_{adj}$	$QME_{adj}$	
Total		$n - 1$	$SQT$		

$$QMP_{adj} = SQP_{adj} / k$$

A soma de quadrados do parâmetro espacial autoregressivo é obtida da seguinte forma:

$$SQ_{\rho} = SQT - SQT_{adj}, \text{ em que:}$$

$SQT$ : Soma de Quadrados Total da análise de variância dos dados não ajustados;

$SQT_{adj}$ : Soma de Quadrados Total da análise de variância dos dados ajustados.

Outro tipo de modelo espacial autoregressivo é o modelo de erros espaciais ou modelo condicional autoregressivo (CAR), que é especificado da seguinte forma (YWATA; ALBUQUERQUE, 2011):

$$Y = X\beta + u \quad (21)$$

, em que:

$Y$ : vetor  $n \times 1$  de valores observados;

$X$ : matriz  $n \times p$  de incidência das variáveis explicativas;

$\beta$ : vetor  $p \times 1$  de parâmetros;

$u$ : vetor  $n \times 1$  de erros espacialmente dependentes.

Neste caso, os erros possuem uma estrutura autoregressiva da seguinte forma:

$$u = \lambda W u \quad (22)$$

, em que:

$u$ : vetor  $n \times 1$  de erros espacialmente dependentes;

$\lambda$ : parâmetro espacial autoregressivo;

$W$ : matriz  $n \times n$  de atribuições de peso da vizinhança espacial;

$\varepsilon$ : vetor  $n \times 1$  de erros associados a cada observação.

Observe que o vetor de erros  $\varepsilon$  possui distribuição normal multivariada, com média nula e matriz de covariância  $\sigma^2 I$ . O coeficiente  $\lambda$  indica o grau da autocorrelação espacial entre os erros do

vetor de erros  $\mathbf{u}$ , ou seja, mede o efeito médio dos vizinhos sobre o erro de determinada região. Note então, que a autocorrelação espacial nos modelos CAR aparece nos termos de erro, diferentemente dos modelos SAR em que a variável resposta é uma função direta dos vizinhos.

Os coeficientes do vetor  $\beta$  podem ser estimados usando o método de quadrados mínimos ordinários (ols), mas devido aos erros serem correlacionados, a matriz de covariância dos estimadores de  $\hat{\beta}_{ols}$  não é obtida por meio de  $\sigma^2(X'X)^{-1}$ , mas com base na seguinte expressão:

$$Var(\hat{\beta}_{ols}) = X'X \left( X'\Omega^{-1}X \right)^{-1} X'X \quad (23)$$

$$\text{, sendo } \Omega = Var(u) = \sigma^2(I - \lambda W)^{-1} \left[ (I - \lambda W)^{-1} \right]'$$

Neste caso, o coeficiente  $\lambda$  e a variância  $\sigma^2$  podem ser estimados a partir de um modelo SAR usando o método de máxima verossimilhança, sendo os resíduos  $\hat{u} = Y - X\hat{\beta}_{ols}$ . Em seguida pode-se obter uma estimativa para a matriz de covariância de  $\hat{\beta}_{ols}$  por meio de:

$$Var(\hat{\beta}_{ols}) = X'X \left( X'\hat{\Omega}^{-1}X \right)^{-1} X'X \quad (24)$$

$$\text{, sendo } \hat{\Omega} = \hat{\sigma}^2(I - \hat{\lambda}W)^{-1} \left[ (I - \hat{\lambda}W)^{-1} \right]'$$

O estimador linear de variância mínima para o modelo CAR é o de quadrados mínimos generalizados (GLS), expresso como:

$$\hat{\beta}_{ols} = \left( X'\Omega^{-1}X \right)^{-1} X'\Omega^{-1}Y \quad (25)$$

$$\text{, sendo } \Omega = Var(\hat{u}) = \sigma^2(I - \lambda W)^{-1} \left[ (I - \lambda W)^{-1} \right]'$$

Como a matriz desconhecida  $\Omega$  depende dos parâmetros desconhecidos  $\lambda$  e  $\sigma^2$ , utiliza-se o seguinte estimador de quadrados mínimos generalizados:

$$\hat{\beta}_{ols} = \left( X'\hat{\Omega}^{-1}X \right)^{-1} X'\hat{\Omega}^{-1}Y \quad (26)$$

$$\text{, sendo } \hat{\Omega} = \hat{\sigma}^2(I - \hat{\lambda}W)^{-1} \left[ (I - \hat{\lambda}W)^{-1} \right]', \text{ com } \hat{\lambda} \text{ e } \hat{\sigma}^2 \text{ estimados utilizando o método}$$

de máxima verossimilhança do modelo SAR, a partir dos resíduos  $\hat{u} = Y - X\hat{\beta}_{ols}$ .

A estimação de parâmetros do modelo CAR por meio do método de máxima verossimilhança pode ser realizada com base na seguinte expressão:

$$Y = X\beta + (1 - \lambda W)^{-1} + \varepsilon \quad (27)$$

Na expressão anterior, o vetor de variável resposta  $Y$  possui distribuição normal multivariada com média condicional e matriz de variância condicional expressas como:

$$E[Y/X] = X\beta \text{ e } \Sigma_{Y/X} = \sigma^2(I - \lambda W)^{-1}[(I - \lambda W)^{-1}]'$$

A função log-verossimilhança condicional, obtida a partir de  $Y$ , é expressa como:

$$\begin{aligned} \ln[L(\beta, \sigma, \lambda / Y, X)] &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) + \ln(I - \lambda W) \\ &\quad - \frac{1}{2\sigma^2} (Y - X\beta)'(I - \lambda W)'(I - \lambda W)(Y - X\beta) \end{aligned} \quad (28)$$

Esta função deve ser maximizada em relação aos parâmetros do modelo para obtenção das estimativas dos coeficientes e da variância dos resíduos. O processo de maximização é feito utilizando métodos iterativos, como os de Gauss-Newton e Newton-Raphson.

A análise de variância usando o modelo autoregressivo CAR dentro do conceito de modelo linear geral, em situações em que os dados apresentam dependência espacial, pode ser realizada com base nos seguintes procedimentos (Griffith, 1992):

- 1- Estimar o valor do parâmetro espacial autoregressivo;
- 2- Ajustar a variável dependente com o parâmetro autoregressivo;
- 3- Submeter a variável dependente ajustada ao procedimento de análise de variância clássico.

### Diagnóstico da dependência espacial

Para detectar e caracterizar a dependência espacial, por são utilizados os seguintes índice e testes:

#### 1) Índice de Moran

É uma medida para avaliar a natureza e o grau de autocorrelação de variáveis georreferenciadas e com arranjo regular de parcelas em experimentos agronômicos de campo (Long, 1996). O índice de Moran é obtido comparando-se os pares de observações adjacentes com o seu desvio em relação a média de todas as observações, por meio da seguinte expressão:

$$I = \frac{n}{\sum_{i,j} w_{ij}} \times \frac{\sum_i \sum_j w_{ij} (Y_j - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \quad (29)$$

, em que:

$I$  : índice de Moran;

$n$  : número de parcelas experimentais;

$w_{ij}$  :  $ij$ -ésima entrada binária na matriz de proximidade espacial;

$Y_i$  :  $i$ -ésima observação;  $Y_j$  :  $j$ -ésima observação;

$\bar{Y}$  : média dos valores observados

O índice de Moran pode apresentar qualquer valor no conjunto dos números reais, mas em geral os valores encontram-se entre -1 e 1. Se o valor estiver entre 0 e 1, indica correlação direta e se estiver entre 0 e -1, indica correlação inversa. Se as parcelas próximas forem similares, o  $I$  tende a ser positivo e se forem dissimilares, o  $I$  tende a ser negativo (Plant, 2012).

A significância do índice de Moran deve ser testada sob a hipótese nulidade de existência de independência espacial. Então, a rejeição da hipótese nula implica em evidência de dependência espacial. Geralmente, a distribuição associada ao índice  $I$  é a distribuição normal.

O diagnóstico da presença de dependência espacial também pode ser realizado calculando o índice de Moran dos resíduos, por meio da seguinte equação:

$$\hat{I}_{res} = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \times \frac{\mathbf{u}' \mathbf{W} \mathbf{u}}{\mathbf{u}' \mathbf{u}} \quad (30)$$

, em que:

$\mathbf{u}$  : vetor de resíduos observados;

$w_{ij}$  : elementos da matriz de vizinhança  $\mathbf{W}$ ;

$n$  : número de parcelas experimentais.

O índice de Moran dos resíduos também segue uma distribuição normal. A estatística de teste para a hipótese nula de independência espacial é expressa como:

$$Z = \frac{I_{res} - E(\hat{I}_{res})}{\sqrt{Var(\hat{I}_{res})}} \quad (31), \text{ sendo}$$

$$E(\hat{I}_{res}) = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \times \frac{tr(MW)}{n-p} \text{ e}$$

$$Var(\hat{I}_{res}) = \left( \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \right)^2 \times \frac{tr(MWMW') + tr(MW)^2 + [tr(MW)]^2}{(n-p)(n-p+2)}$$

$$-\left[ I_{res} \right]^2$$

, em que:

$p$  : número de parâmetros;  $tr(\cdot)$  : traço da matriz;

$M = I - X(X'X)^{-1}X'$ , matriz de projeção;  $I_n$  : matriz identidade.

O índice de Moran é considerado significativo quando o valor-p do quantil da distribuição normal padronizada for menor que a probabilidade de erro escolhida.

## 2) Teste de Wald

O teste de Wald, que é baseado nos estimadores de máxima verossimilhança, é estruturado de forma a ter a hipótese nula formulada sobre a existência de independência espacial, com base nos parâmetros autoregressivos do modelo, ou seja, considera-se o parâmetro autoregressivo do modelo SAR ou CAR são iguais a zero ( $\rho = 0$  ou  $\lambda = 0$ ).

A expressão da estatística do teste de Wald, para o modelo CAR com  $H_0: \lambda = 0$ , é a seguinte:

$$Wald = \hat{\lambda}^2 \left[ t_2 + t_3 - \frac{(t_1)^2}{n} \right] \quad (32)$$

, sendo:

$$t_1 = tr(W)\hat{B}^{-1}; \quad t_2 = tr\left[\left(W\hat{B}^{-1}\right)'\right]^2;$$

$$t_3 = tr\left[\left(W\hat{B}^{-1}\right)' \left(W\hat{B}^{-1}\right)\right] \text{ e } \hat{B} = (I - \hat{\lambda}W), \text{ onde } tr(\cdot) \text{ é o traço da matriz e } I \text{ é a}$$

matriz identidade.

A estatística de Wald segue uma distribuição  $\chi_1^2$ , ou seja, distribuição qui-quadrado com um grau de liberdade.

## 3) Teste de razão de verossimilhança

O teste de razão de verossimilhança para presença de dependência espacial no erro, com base nos modelos autoregressivos, pode ser descrito da seguinte forma:

Considere a função baseada na diferença entre as expressões usadas para estimação de parâmetros de um modelo CAR por máxima verossimilhança e por quadrados mínimos ordinários:

$$\ln(L(\beta, \sigma, \lambda | Y, X)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) + \ln(I - \lambda W) \quad (33)$$

$$-\frac{1}{2\sigma^2} (Y - X\beta)' (I - \lambda W)' (I - \lambda W) (Y - X\beta)$$

Sob  $H_0: \lambda = 0$ , pode-se reescrever a equação (33) como:

$$\begin{aligned} \ln(L(\beta, \sigma, \lambda | Y, X)) &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) \\ &\quad - \frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \end{aligned} \quad (34)$$

Neste caso, a variância é expressa como:

$$S_1^2 = \frac{\varepsilon' \varepsilon}{n} = \frac{u'(I - \lambda W)'(I - \lambda W)}{n}$$

Então, a equação (33) baseada nos parâmetros estimados por máxima verossimilhança com  $S_1^2$  no lugar de  $\sigma^2$  é transformada em:

$$\ln(L(\beta, \sigma, \lambda | Y, X)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln S_1^2 + \ln(I - \lambda W) - \frac{n}{2} \quad (35)$$

De forma semelhante, a equação (34) baseada nos parâmetros estimados por quadrados mínimos ordinários com  $S_0^2$  no lugar de  $\sigma^2$  é transformada em:

$$\ln(L(\beta, \sigma, \lambda | Y, X)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln S_0^2 - \frac{n}{2} \quad (36)$$

A estatística do teste de razão de verossimilhança, obtida tomando duas vezes a diferença entre as equações (35) e (36), é expressa como:

$$LR = -n(\ln S_0^2 - \ln S_1^2) + 2\ln|I - \lambda W| \quad (37)$$

Tem-se que  $LR \square \chi_1^2$ , ou seja, estatística  $LR$  tem distribuição qui-quadrado com um grau de liberdade.

### Exemplo de Aplicação (Extraído de Piaskowski e Price, 2021)

Neste exemplo, os dados de um experimento de variedades de trigo realizado em Alliance Nebraska (Stroup et al., 1994), foi utilizado para demonstrar o ajuste de dependência espacial em modelos lineares. Os dados são referentes a produtividade de 56 variedades usando o delineamento experimental blocos casualizados. Segundo Stroup (2013), a topografia do local combinado com o inverno induziu uma variabilidade espacial que não corresponde ao delineamento blocos casualizados e que produziu estimativas viesadas numa análise padrão não ajustada.

Os dados foram preparados usando o procedimento PROC FORMAT do SAS porque o experimento original continha parcelas em branco separando os blocos. Os índices de linhas e colunas foram multiplicados por constantes para converte-los para as dimensões métricas das parcelas. São mostradas as seis primeiras observações, contendo as variáveis para variedade (entry), repetição (rep) e identificadores de coluna (col) e linha (row). O procedimento SGPlot do SAS foi utilizado

para construir um mapa das quatro repetições (blocos) do experimento original. Estes resultados são apresentados a seguir:

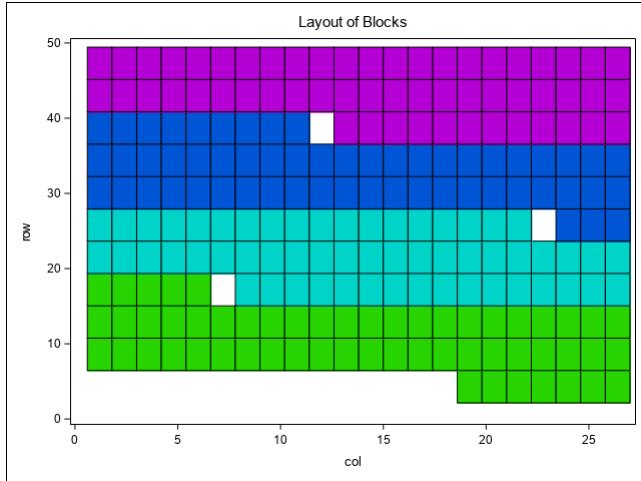


Figura 5. Disposição dos blocos no experimento.

Como primeiro passo para verificar a existência de variabilidade espacial, examinar padrões espaciais por meio de um mapa de calor para produtividade e tendências ao longo de repetições, colunas e linhas usando gráficos de caixa.

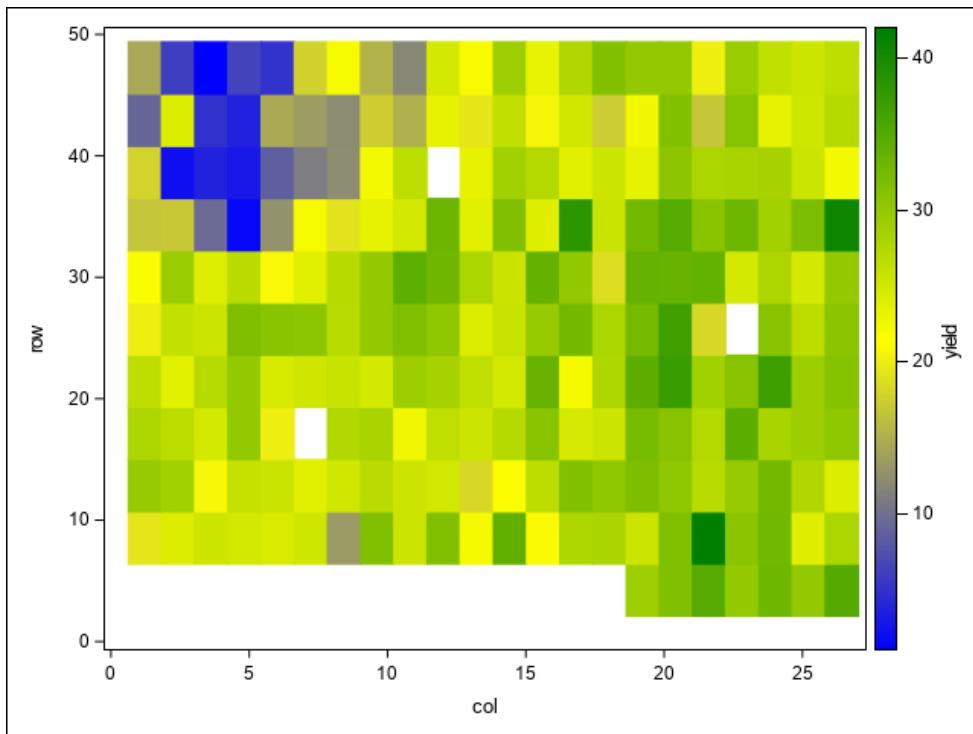


Figura 6. Padrões espaciais obtidos por meio de mapas de calor.

Observe que na área do canto noroeste do mapa de calor, a produtividade é notavelmente bem diferente no experimento. Esta área está posicionada ao longo do final dos dois blocos do topo, o que torna estes blocos não homogêneos, devido baixa cobertura e alta exposição a baixas temperaturas de inverno (Stroup, 2013). Os gráficos de caixa também demonstram este padrão ao

longo dos blocos, colunas e linhas. Portanto, fica claro a existência de padrões e tendências espaciais identificáveis.

O processo de modelagem da variabilidade espacial consiste em estimar e testar a correlação espacial e começa com a obtenção dos resíduos da análise convencional do experimento, que cria um conjunto de dados de resíduos. Em seguida a estimação e avaliação da variabilidade espacial é realizada em vários passos. No primeiro, são resumidas as distâncias potenciais (lags) entre posições linha/coluna. Aqui, o número de classes lags estabelecido deve ser grande o suficiente para cobrir a faixa de possíveis distâncias entre dados de coordenadas linha e coluna. Nesses passos os resultados obtidos foram os seguintes:

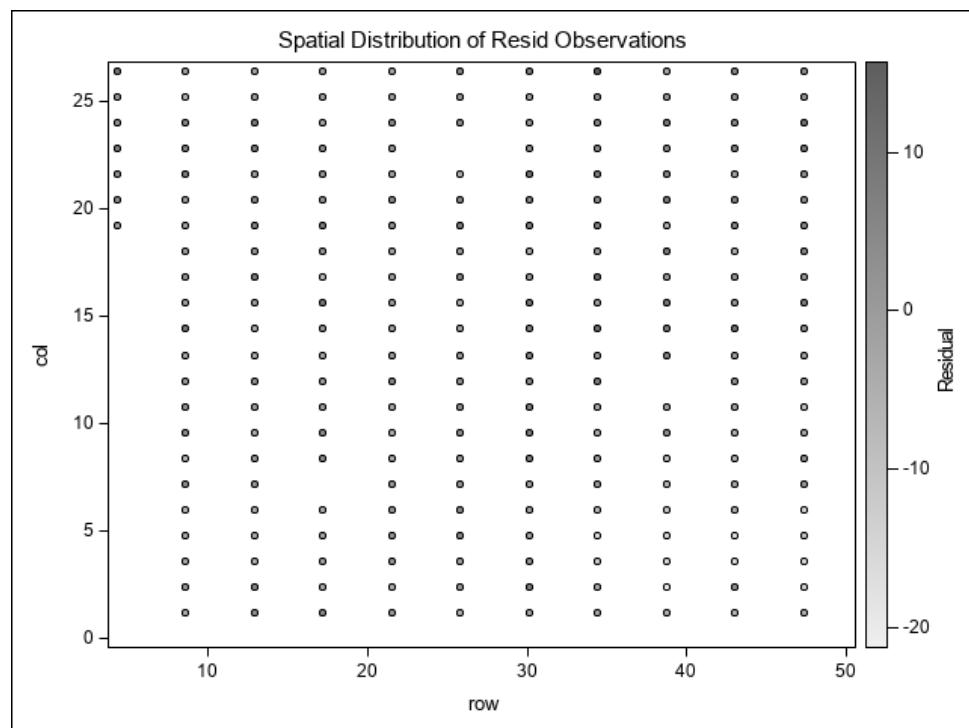


Figura 7. Distribuição espacial dos resíduos das observações.

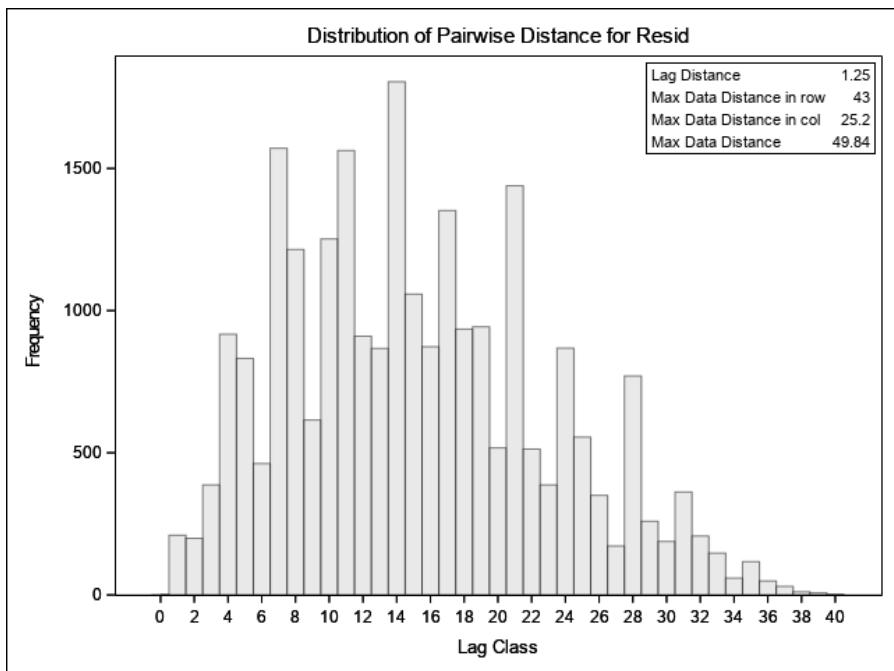


Figura 8. Distribuição de pares de distâncias para os resíduos.

Observe que, com a escolha de 40 classes lag, a distância lag mínima é 1,25 m e a máxima varia entre 25 a 43 em linhas e colunas, respectivamente. O histograma mostra o número de pares de cada distância lag. Aqui os dados são suficientes para estabelecer um número de pares de distâncias lag maior que 30, o que possibilitará uma estimativa acurada da semivariância empírica. Os índices I de Moran e C de Geary medem as correlações ponderadas entre pares de observações, que correspondem a medidas de variabilidade global e local, respectivamente. Se não existe correlação espacial, é esperado que o valor I seja próximo de zero e o valor C próximo de um.

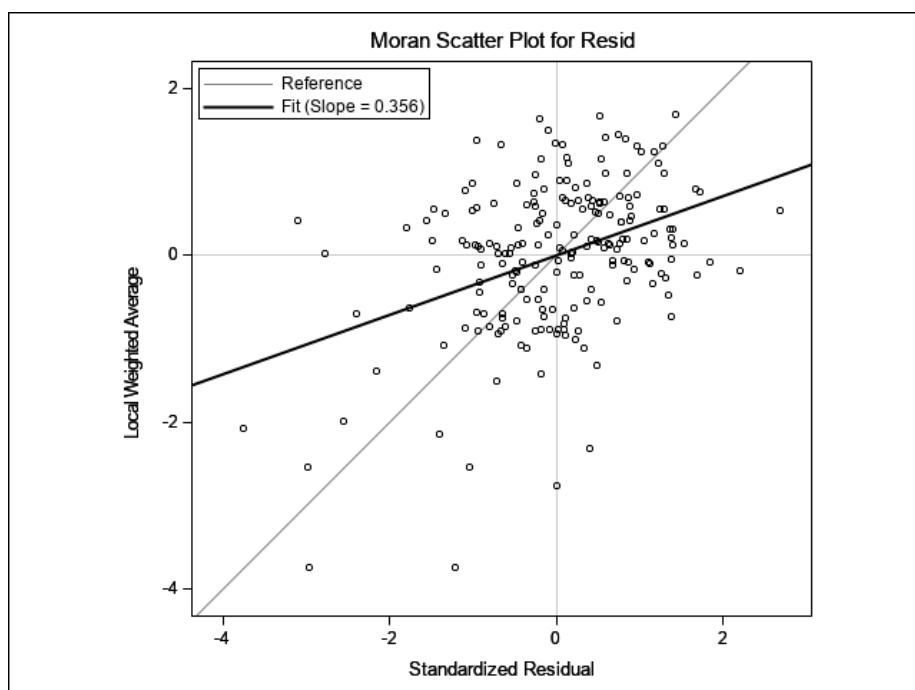


Figura 9. Diagrama de dispersão do Índice de Moran para os resíduos.

Neste conjunto de dados tanto o índice I de Moran quanto o C de Geary foram significativos, o que indica a presença de correlação espacial. O diagrama de dispersão de resíduos versus distância lag também mostrou uma correlação positiva, onde os resíduos aumentam em magnitude com o aumento da distância entre os pontos.

Os valores de distância lag obtidos nos passos anteriores são agora utilizados para estimação e modelagem de semivariância empírica:

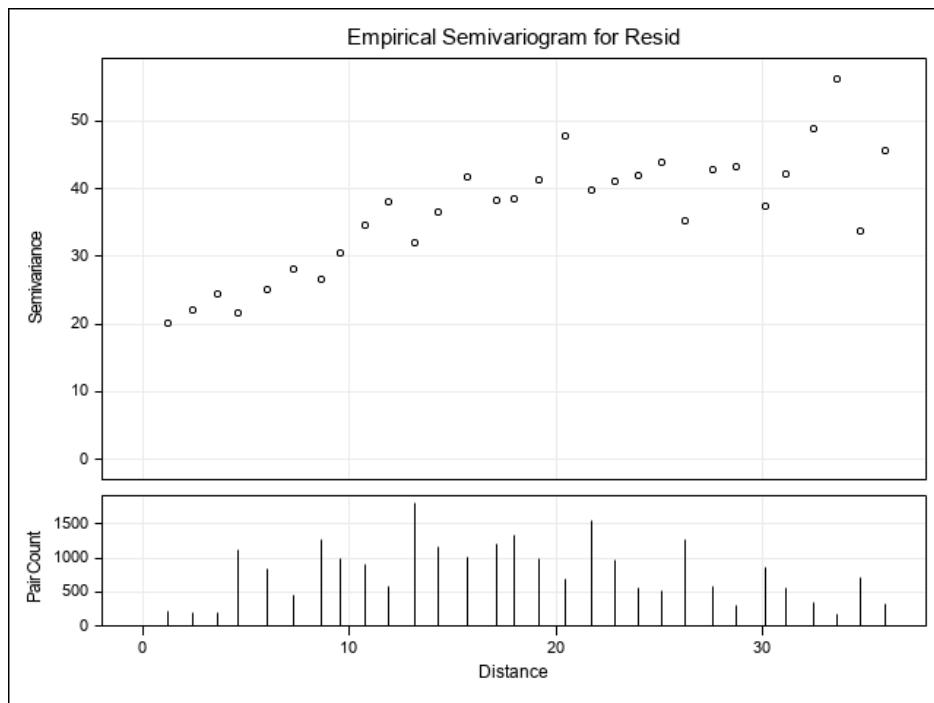


Figura 10. Semivariograma empírico para os resíduos.

Observe no gráfico que a semivariância aumenta à medida que a distância entre pontos aumenta até aproximadamente 20m onde ela começa a decrescer. Este tipo de padrão é muito comum para relacionamentos espaciais. Também é possível comparar vários modelos teóricos de variograma, tais como Gaussiano, Exponencial, Potência e Esférico:

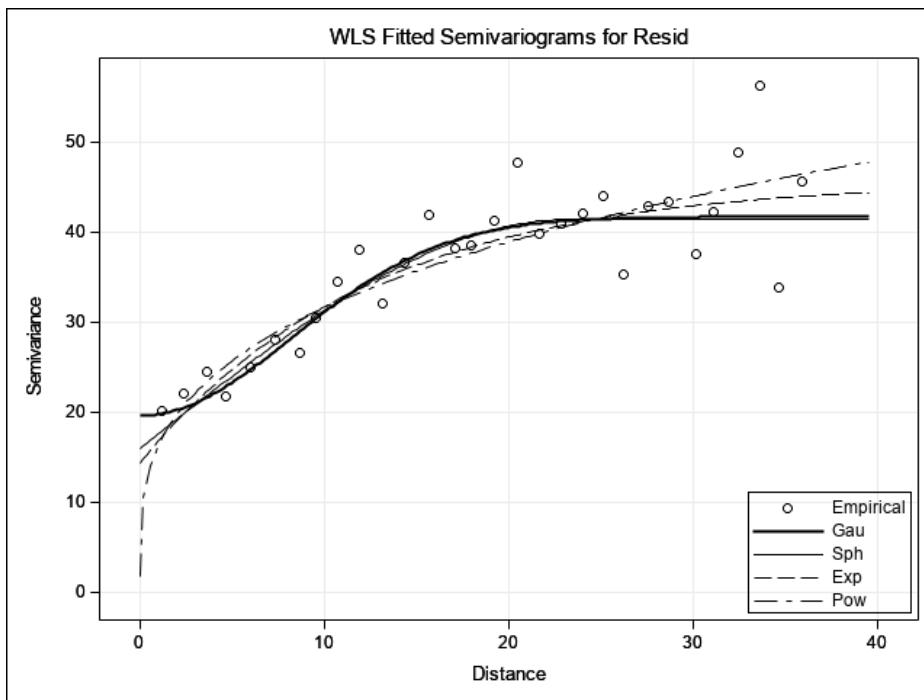


Figura 11. Semivariogramas ajustados pelo Método de Quadrados Mínimos Ponderados para os resíduos.

As estatísticas AIC ou SSE, fornecidas para o ajuste de cada modelo de variograma, podem ser usadas como guia para a seleção de modelo, onde menores valores indicam um melhor ajuste. Neste exemplo, o modelo Gaussiano foi classificado como o melhor, embora o modelo esférico seja relativamente similar. O gráfico apresentado ao final destes resultados fornece uma comparação visual entre os modelos variograma com o modelo Gaussiano em destaque.

Após diagnosticar a presença e o padrão da variabilidade espacial e estimar um modelo para descrever a variabilidade, o último passo consiste em utilizar o variograma estimado numa análise ajustada. Para fazer esta análise será usado o procedimento de modelo linear misto, incorporando a variabilidade espacial na matriz variância-covariância do modelo linear, assumindo o modelo de variograma identificado no passo anterior. Com o propósito de fazer comparação, primeiro é realizada a análise não ajustada, em seguida com o ajuste ou a inclusão da informação espacial.

Para fins de comparação de modelos a estatística AIC (descrita em capítulos anteriores) pode ser usada, sendo o menor AIC indicador de melhor ajuste. Neste conjunto de dados especificamente observamos os valores de 1221,7 e 1073,1, para os modelos não-ajustado e ajustado, respectivamente.

Outro ponto crítico são as médias ajustadas dos cultivares podem ser ordenadas incorretamente sob o modelo não ajustado, uma vez que a presença de variabilidade espacial afeta a precisão na estimativa das médias. Isto ocorre sensivelmente neste conjunto de dados, conduzindo a graves erros de seleção e, consequentemente, compromete os ganhos genéticos.

# CAPITULO 29

## Análise de Estabilidade no Melhoramento de Plantas

Existem diversos métodos de análise de estabilidade e adaptabilidade destinadas à avaliação de um grupo de genótipos em diversos ambientes. A diferença entre elas se dá, basicamente, pelos parâmetros adotados em sua avaliação, nos procedimentos biométricos empregados para medi-las e na informação ou detalhamento de sua análise, uma vez que todas são fundamentadas na existência de interações de genótipos com ambientes (Vencovsky e Barriga, 1992; Cruz et al., 2012).

### Métodos clássicos de avaliação de estabilidade e adaptabilidade

#### Método de Wricke (1964)

Este método, também conhecido como ecovalência é um dos que avalia a estabilidade fenotípica de genótipos utilizando apenas a análise de variância. O parâmetro  $W_i$ , estimado com base na decomposição da soma de quadrados da interação G x E, é obtido por meio da seguinte expressão:

$$W_i = r \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2, \text{ em que:}$$

$r$  : número de repetições;

$Y_{ij}$  : média do genótipo  $i$  no ambiente  $j$ ;

$\bar{Y}_{i\cdot}$  : média do genótipo  $i$ ;

$\bar{Y}_{\cdot j}$  : média do ambiente  $j$ ;

$\bar{Y}_{\cdot\cdot}$  : média geral.

O estimador da estabilidade,  $W_i$ , é uma medida da contribuição de cada genótipo para a interação G x E, em que os genótipos que menos contribuem para a interação são considerados os mais estáveis. O termo ecovalência se refere à capacidade do genótipo responder às variações ambientais. Então, uma alta ecovalência implica em baixo  $W_i$  e significa que o genótipo é estável (Wricke, 1964; Wricke e Weber, 1986)

#### Método de Shukla (1972)

Neste método, o componente de variância de cada genótipo ao longo dos ambientes é proposto como uma medida de estabilidade fenotípica. Ele mede a estabilidade ao invés do desempenho do

genótipo. Portanto, na obtenção da estabilidade de variância  $(\sigma_i^2)$  a soma de quadrados da interação G x E é particionada em componentes, um para cada genótipo e estimada por meio da seguinte expressão:

$$\sigma_i^2 = \frac{1}{(g-1)(g-2)(e-1)} \times \left[ g(g-1) \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2 - \sum_i \sum_j \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2 \right], \text{ em que:}$$

$g$ : número de genótipos;  $e$ : número de ambientes;

$\bar{Y}_{ij}$ : média de produtividade do genótipo  $i$  no ambiente  $j$ ;

$\bar{Y}_{i\cdot}$ : média de produtividade do genótipo  $i$  em todos os ambientes;

$\bar{Y}_{\cdot j}$ : média de produtividade de todos os genótipos no ambiente  $j$ :

$\bar{Y}_{\cdot\cdot}$ : média geral.

Outra forma de expressar o estimador da estabilidade de variância é a seguinte:

$$\sigma_i^2 = g / [(g-1)(e-1)] W_i - QMGE / g - 2, \text{ onde } QMGE \text{ é o quadrado médio da interação genótipos x ambientes e } W_i \text{ obtido como } W_i = \sum_{j=1}^e \left( Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot} \right)^2.$$

A interpretação de  $\sigma_i^2$  é realizada da seguinte forma: se a estabilidade de variância de um genótipo for igual a variância ambiental, o que implica em  $\sigma_i^2 = 0$ , então o genótipo é identificado como estável; um valor relativamente grande de  $\sigma_i^2$  indica maior instabilidade para o genótipo  $i$ ; valor significativo de  $\sigma_i^2$  significa que o desempenho do genótipo ao longo dos ambientes foi instável; genótipos com valor de  $\sigma_i^2$  não significativo ou negativo são considerados estáveis ao longo dos ambientes.

Observe que, como  $\sigma_i^2$  é obtido por meio da diferença entre duas somas de quadrados pode ocorrer valor negativo, que pode ser considerado igual a zero. Observe também que  $\sigma_i^2$  não pode ser estimado para dados desbalanceados.

### Método de Annicchiarico (1992)

Os métodos que avaliam a estabilidade com base na análise de variância das são os mais antigos e consistem na análise de grupos de experimentos sendo a variação de ambientes, dentro de cada genótipo, usada como estimador do parâmetro de estabilidade, de forma que o genótipo que apresentar menor quadrado médio, ou seja, menor variância, será o mais estável (Cruz et al., 2012).

Um dos métodos mais utilizados é o de Annicchiarico (1992). Este método vem sendo muito utilizado pelos melhoristas na análise da estabilidade fenotípica, pois o mesmo apresenta uma relativa facilidade de aplicação e de interpretação dos resultados gerados. Esta é baseado em análise de variância e considera a estimativa de um índice de confiança ( $W_i$ ) que representa a chance de um genótipo  $i$  apresentar desempenho fenotípico superior à média geral do conjunto de genótipos que está sendo avaliado.

Neste método, os valores absolutos da variável analisada são convertidos para valores em porcentagem relativa à média de cada ambiente e depois são calculados os desvios relativos de cada genótipo nos diversos ambientes. Posteriormente, a média e os desvios relativos são então utilizados no cálculo do índice de confiança/recomendação. Assim quanto maior o índice de confiança maior será a estabilidade e adaptabilidade da cultivar, traduzida na confiança da indicação do genótipo.

De acordo com Annicchiarico (1992), quanto maior a estimativa de  $W_i$ , mais estável é considerado o genótipo, sendo preferidos os genótipos que apresentem estimativa superior a 100%. Por essa proposta, as cultivares que apresentarem valor de  $W_i$  superior a 100% não deverão apresentar médias fenotípicas inferiores à média geral.

Este método estima o índice de confiança ( $W_i$ ) de um determinado genótipo apresentando desempenho abaixo da média do ambiente, de acordo com o seguinte modelo estatístico:

$$W_i = Y_i - Z_{(1-\alpha)} S_i, \text{ em que:}$$

$W_i$ : índice de confiança, em percentagem;

$Y_i$ : media do genótipo  $i$ , em porcentagem;

$Z$ : percentil  $(1-\alpha)$  da função de distribuição normal acumulada;

$\alpha$ : probabilidade de erro Tipo I;

$S_i$ : desvio padrão dos valores, em percentagem.

### Método de Eberhart e Russell (1966)

O método proposto por Eberhart e Russell (1966), dentre os que se baseiam na regressão linear, tem sido um dos mais utilizadas na recomendação de genótipos em função da simplicidade dos cálculos, facilidade de interpretação e informações fornecidas. Neste método o comportamento de cada genótipo, diante das variações ambientais, é estimado por meio de uma análise de regressão linear simples da variável dependente sobre um índice ambiental, definido como a diferença entre a média de cada ambiente e a média de todos os ambientes. É estimada uma equação de regressão para cada genótipo sob avaliação.

No método de Eberhart e Russel (1966), considera-se o coeficiente de interseção  $(\beta_{0i})$ , o coeficiente de regressão  $(\beta_{1i})$ , os desvios da regressão  $(\delta_{ij})$ , o índice ambiental  $(I_j)$ , ou qualidade do ambiente, que pode ser obtido por meio da média do ambiente menos a média geral, e o erro experimental médio  $(\varepsilon_{ij})$ . Além disso, são estimados os coeficientes de determinação  $(R^2)$  para as equações obtidas para cada um dos genótipos, para verificar o ajuste da equação, para o genótipo em questão.

O modelo de regressão linear simples utilizado é o seguinte:

$$Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \delta_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\beta_{0i}$ : média geral do  $i$ -ésimo genótipo;

$\beta_{1i}$ : coeficiente de regressão linear, que mede a resposta do  $i$ -ésimo genótipo à variação do ambiente;

$I_j$ : índice ambiental;

$\delta_{ij}$ : desvio da regressão do  $i$ -ésimo genótipo sobre o  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : erro experimental médio associado a observação  $Y_{ij}$ .

A estimativa do parâmetro de estabilidade  $(S_{d_i}^2)$  é obtida de acordo com a seguinte expressão:

$S_{d_i}^2 = QMD_i - QMR/r$ , onde:  $QMD_i$  é o quadrado médio dos desvios de regressão do  $i$ -ésimo genótipo,  $QMR$  é o quadrado médio do resíduo e  $r$  é o número de repetições.

A estimativa do parâmetro de adaptabilidade  $(\beta_i)$  foi obtida por meio da seguinte expressão:

$$\beta_i = \frac{\sum_{j=1}^n Y_{ij} I_j}{\sum_{j=1}^n I_j^2}, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$I_j$ : índice ambiental, sendo  $I_j = (Y_j/p) - (Y_\infty/pn)$ , onde  $Y_j$  é a média de todos os genótipos no  $j$ -ésimo ambiente,  $Y_\infty$  é a média geral,  $n$  é o número de genótipos e  $p$  o número de ambientes.

O coeficiente de determinação  $(R^2)$  é obtido de acordo com a seguinte expressão:

$R_i^2 = \left[ (SQRL_{Linear_i}) / (SQ(E/G_i)) \right] 100$ , onde  $SQRL_{Linear_i}$  é a soma de quadrados da regressão linear do  $i$ -ésimo genótipo e  $SQ(E/G_i)$  é a soma de quadrados de ambientes dentro do  $i$ -ésimo genótipo. As estimativas para  $\beta_i$  são testadas segundo a hipótese  $H_0: \beta_i = 1$  e hipótese alternativa  $H_1: \beta_i \neq 1$ , avaliada por meio da estatística  $t$ .

Segundo Eberhart e Russell (1966), os genótipos podem ser classificados quanto à adaptabilidade em três grupos:

- a) adaptabilidade geral, com  $\beta_i = 1$ , que apresenta média acima da média geral, sendo o tipo desejável em ambientes com muitas variações imprevisíveis;
- b) adaptabilidade específica a ambientes favoráveis, com  $\beta_i > 1$ , que agrupa os genótipos com alto desempenho em ambientes favoráveis;
- c) adaptabilidade específica a ambientes desfavoráveis, com  $\beta_i < 1$ , que agrupa os genótipos que se destacam em ambientes desfavoráveis.

Os genótipos podem ser classificados quanto à estabilidade em genótipos de alta estabilidade  $(S_{d_i}^2 = 0)$  e genótipos de baixa estabilidade  $(S_{d_i}^2 \neq 0)$ .

### Método de Cruz, Torres e Vencovsky (1989)

Por ser baseado em apenas uma regressão linear para cada genótipo, o método de Eberhart e Russel (1966) é considerado conservador, uma vez que favorece apenas genótipos com desempenho médio em relação ao conjunto analisado. Diante da hipótese de se identificar genótipos com performance desejável nos ambientes considerados desfavoráveis e favoráveis, considerou-se a alternativa de modificar este método empregando-se dois segmentos de reta, ou seja, regressão linear bissegmentada (Silva e Barreto, 1986; Silva, 1995a; Silva, 1995b).

O método de Cruz et al. (1989) baseia-se na análise de regressão bissegmentada, possuindo três parâmetros de adaptabilidade: a média  $(\beta_{0i})$ , a resposta linear aos ambientes desfavoráveis  $(\beta_{1i})$  e a resposta linear aos ambientes favoráveis  $(\beta_{2i})$ . A estabilidade é avaliada pelo desvio da regressão  $(\delta_{ij})$  de cada cultivar em função das variações ambientais.

Por esse método as estimativas  $\beta_{1i}$  e  $\beta_{1i} + \beta_{2i}$  são não correlacionadas entre si, o que leva a independência dos dois segmentos de reta. O genótipo ideal é aquele que apresenta alta média de

produtividade ( $\beta_{0i}$ ), baixo  $\beta_{1i} < 1$  (adaptabilidade a ambientes desfavoráveis),  $\beta_{1i} + \beta_{2i} > 1$  (responsividade à melhoria ambiental), e  $\delta_{ij} = 0$  (estabilidade fenotípica).

Portanto, nesse método, a média ( $\beta_0$ ), a resposta linear a ambientes desfavoráveis ( $\beta_1$ ) e favoráveis ( $\beta_1 + \beta_2$ ) são os parâmetros que estimam a adaptabilidade dos genótipos, e os desvios da regressão ( $\delta^2$ ) de cada genótipo e o coeficiente de determinação ( $R^2$ ), constituem os parâmetros que estimam a estabilidade. O modelo estatístico deste método é o seguinte:

$$Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \beta_{2i} T(I_j) + \delta_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : média do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\beta_{0i}$ : média geral do  $i$ -ésimo genótipo;

$\beta_{1i}$ : coeficiente de regressão linear, que mede a resposta do  $i$ -ésimo genótipo nos ambientes desfavoráveis;

$I_j$ : índice ambiental;

$\beta_{1i} + \beta_{2i}$ : mede a resposta do  $i$ -ésimo genótipo nos ambientes desfavoráveis;

$T(I_j)$ : variável do eixo das abcissas;

$\delta_{ij}$ : desvio da regressão do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : erro experimental médio associado a observação  $Y_{ij}$ .

### Método de Lin e Binns (1988)

O método proposto por Lin e Binns (1988), baseado em métodos não paramétricos, é simples e fácil de interpretar, possibilitando identificar um ou mais genótipos com desempenho próximo ao máximo nos vários ambientes testados. Neste método, os genótipos superiores são identificados por meio de um simples parâmetro ( $P_i$ ), associado à estabilidade e à produtividade, e define um genótipo superior como aquele que apresenta performance próxima do máximo nos vários ambientes testados. A estimativa  $P_i$  é o quadrado médio da distância em relação à resposta máxima em cada ambiente. O genótipo ideal é aquele com menor valor de  $P_i$ .

A medida de estabilidade é definida como o quadrado médio da distância entre a média do genótipo e a resposta média máxima para todos os ambientes. Essa definição é detalhada por Rocha (2002), mostrando que o parâmetro  $P_i$  representa de fato o quadrado médio da distância entre a resposta de um determinado genótipo em relação à resposta do genótipo que apresenta produtividade

máxima, dentre todos os genótipos, em um determinado ambiente. Então, quanto menor a distância entre a resposta do genótipo e a resposta do genótipo de produtividade máxima nos diversos ambientes, menor será o valor de  $P_i$  e mais estável o genótipo. O valor de  $P_i$  é estimado por meio da seguinte expressão:

$$P_i = \sum_{j=1}^n \left( Y_{ij} - M_j \right)^2 / 2n, \text{ em que:}$$

$P_i$ : índice de superioridade do  $i$ -ésimo genótipo;

$Y_{ij}$ : produtividade do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$M_j$ : resposta máxima obtida dentre todos os genótipos no  $j$ -ésimo ambiente;

$n$ : número de ambientes.

A expressão anterior pode ser desdobrada da seguinte forma:

$$P_i = \left[ n(\bar{Y}_i - \bar{M})^2 + \sum_{j=1}^n (Y_{ij} - \bar{Y}_i - M_j + \bar{M})^2 \right] / 2n, \text{ onde } \bar{Y}_i \text{ é a média de}$$

produtividade do  $i$ -ésimo genótipo nos  $n$  ambientes, sendo expressa como:

$$\bar{Y}_i = \sum_{j=1}^n Y_{ij} / n \text{ e } \bar{M} \text{ é a média das produtividades máximas de todos os genótipos em todos os}$$

ambientes, sendo expressa como:

$$\bar{M} = \sum_{j=1}^n M_j / n. \text{ Estas estimativas são utilizadas para testar a hipótese de nulidade da}$$

estimativa  $P_i$ , para cada genótipo por meio do teste F. A significância do teste indica que o genótipo difere estatisticamente do máximo ao longo dos ambientes.

### Método de Lin e Binns modificado por Carneiro (1998)

As modificações no método de Lin e Binns (1988) propostas por Carneiro (1998) têm como objetivo a recomendação de genótipos dividindo o conjunto de ambientes em favoráveis e desfavoráveis, de forma a refletir ambiente onde há emprego de alta e baixa tecnologia, respectivamente, além de introduzir um referencial mais apropriado do que a distância do genótipo a pontos máximos, como preconizado na metodologia de Lin e Binns (1988). Com isso, o método torna-se de aplicação mais ampla tanto em relação aos caracteres avaliados como aos genótipos disponíveis. A classificação dos ambientes segundo essa metodologia é baseada nos índices ambientais que nada mais são do que a diferença da média dos genótipos em cada ambiente e a média geral.

Neste método, a decomposição da estimativa  $P_i$ , para ambientes favoráveis e desfavoráveis é estimada da seguinte forma:

$$P_{if} = \sum_{j=1}^f (Y_{ij} - M_j)^2 / 2f \text{ e } P_{id} = \sum_{j=1}^d (Y_{ij} - M_j)^2 / 2d, \text{ em que:}$$

$P_{if}$ : índice de superioridade do  $i$ -ésimo genótipo nos  $j$  ambientes favoráveis;

$P_{id}$ : índice de superioridade do  $i$ -ésimo genótipo nos  $j$  ambientes desfavoráveis;

$Y_{ij}$ : produtividade do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$M_j$ : resposta máxima obtida dentre todos os genótipos no  $j$ -ésimo ambiente favorável ou desfavorável;

$f$ : número de ambientes favoráveis;

$d$ : número de ambientes desfavoráveis;

## Métodos novos de avaliação de estabilidade e adaptabilidade

### Método AMMI-Biplot

O método denominado efeitos aditivos e interação multiplicativa (AMMI-Biplot) é uma combinação entre a análise de variância e a análise de componentes principais (ACP). Neste caso, os componentes aditivos são utilizados para estudar os efeitos principais (genótipos e ambientes) e os componentes multiplicativos, para estudar a interação genótipos x ambientes. Na ACP a variação contida nos componentes principais significativos é denominada padrão, e a contida nos componentes não significativos, denomina-se de ruído (Zobel et al., 1988). O modelo estatístico associado a este método é o seguinte:

$$\bar{Y}_{ij} = \mu + g_i + a_j + \sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk} + r_{ij} + \varepsilon_{ij}, \text{ em que:}$$

$\bar{Y}_{ij}$ : média de desempenho do genótipo  $i$  ( $i=1,2,\dots,g$ ) no ambiente  $j$  ( $j=1,2,\dots,a$ );

$\mu$ : média geral dos experimentos;

$g_i$ : efeito fixo do genótipo  $i$ ;

$a_j$ : efeito fixo do ambiente  $j$ ;

$\sum_{k=1}^n \lambda_k \gamma_{ik} \alpha_{jk}$ : efeito fixo da interação genótipos x ambientes multiplicativa, onde  $\lambda_k$  é o valor singular,  $\gamma_{ik}$  e  $\alpha_{jk}$  são os escores do eixo k da ACP, para genótipo e ambiente, respectivamente, e  $n$  é o número de eixos ou de componentes principais retidos para descrever o padrão da interação GxA na análise AMMI;

$r_{ij}$ : efeito residual do modelo AMMI (ruído);

$\varepsilon_{ij}$ : erro experimental considerado de efeito aleatório.

Para a definição do número de eixos as serem retidos na análise para explicar o padrão relacionado a interação GxA, pode ser adotado o critério proposto por Gauch e Zobel (1988), que considera a proporção da soma de quadrados da interação genótipos x ambientes acumulada nos primeiros eixos. A significância do teste F de Gollob pode ser utilizado como determinação do modelo para cada família AMMI (AMMI1, AMMI2, ..., AMMIn) e os resíduos dos eixos testados pelo teste F de Cornelius.

Também pode ser calculado o valor de estabilidade AMMI ("AMMI stability value-ASV"). O ASV é obtido com base no cálculo da distância euclidiana entre a origem do plano cartesiano e a coordenada do ponto do genótipo ou ambiente (PURCHASE et al., 2000). Como o valor do IPCA1 contribui mais para a interação GxA, é necessário o uso de um valor ponderado. Este valor é calculado para cada genótipo e ambiente, de acordo com a contribuição relativa do IPCA1 e o IPCA2 para a interação GxA, por meio da seguinte expressão:

$$ASV = \sqrt{\left[ \frac{SQ_{IPCA1}}{SQ_{IPCA2}} (Escore_{IPCA1})^2 \right] + (Escore_{IPCA2})}, \text{ em que:}$$

$SQ_{IPCA1}$ : soma de quadrados do IPCA1;

$SQ_{IPCA2}$ : soma de quadrados do IPCA2.

Quanto menor o valor de ASV, maior será a estabilidade do genótipo.

### Método GGE-Biplot

O modelo associado ao método GGE, proposto por Yan et al. (2007), é expresso como:

$$Y_{ij} - \mu - E_j = y_1 e_{i1} \rho_{j1} + y_2 e_{i2} \rho_{j2} + \varepsilon_{ij}, \text{ em que:}$$

$Y_{ij}$ : desempenho do  $i$ -ésimo genótipo no  $j$ -ésimo ambiente;

$\mu$ : média geral das observações;

$E_j$ : efeito principal do  $j$ -ésimo ambiente;

$y_1$  e  $y_2$ : valores singulares associados ao IPCA1 e IPCA2, respectivamente;

$e_{i1}$  e  $e_{i2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $i$ -ésimo genótipo;

$\rho_{j1}$  e  $\rho_{j2}$ : escores do IPCA1 e IPCA2, respectivamente, referente ao  $j$ -ésimo ambiente;

$\varepsilon_{ij}$ : efeito residual não explicado por nenhum dos fatores (“ruído”).

Neste método, após obtenção dos escores associados aos ambientes e genótipos, são construídos os gráficos biplot e feitas as análises de “which-won-where”, média x estabilidade, discriminativo x representativo e genótipo ideal (Yan; Tinker, 2006).

### Método MHPRVG via REML/BLUP

A análise por meio da metodologia de modelos lineares mistos utiliza a máxima verossimilhança restrita (REML) para estimar os componentes de variância do modelo utilizado e o melhor preditor linear não viesado (BLUP) para estimar o valor genotípico dos genótipos. Esse método é conhecido como REML/BLUP (Resende, 2007).

No método BLUP, o efeito de genótipos pode ser considerado como aleatório, tendo como vantagem a facilidade de implementação nas análises quando comparado a outros estimadores do tipo “shrinkage” e o uso em experimentos desbalanceados. O efeito aleatório da interação genótipos x ambientes também permite inferências para uma população de ambientes.

Com base no método REML/BLUP, a medida simultânea de adaptabilidade e estabilidade para cada genótipo é obtida por meio da média harmônica do desempenho relativo dos valores genotípicos (MHPRVG). Este método tem como princípio que quanto menor o valor do desvio-padrão do comportamento genotípico entre os ambientes, maior será a média harmônica dos valores genotípicos. A seleção pelos maiores valores da média harmônica dos valores genotípicos (MHVG) determina, simultaneamente a produtividade e a estabilidade. O desempenho relativo associado aos valores genotípicos (PRVG) nos diferentes ambientes exprime a adaptabilidade (Resende, 2002).

Por exemplo, para o método da máxima verossimilhança restrita/melhor predição linear não viesada (REML/BLUP) pode ser utilizado o modelo 52 – Delineamento em blocos incompletos e vários locais e uma observação por parcela (Resende, 2007), associado ao seguinte modelo estatístico:

$$y = X_r + Z_g + W_b + T_i + e, \text{ em que:}$$

$y$ : vetor de observações da característica;

$r$ : vetor de efeitos fixos de repetição somados com a média geral;

$g$ : vetor de efeitos genotípicos aleatórios;

$b$ : vetor de efeitos aleatórios de blocos;

$i$ : vetor de efeitos aleatórios da interação genótipo x ambiente ( $ga$ );

$e$ : vetor aleatório de erros ou resíduos.

As matrizes  $X$ ,  $Z$ ,  $W$  e  $T$  representam as incidências dos efeitos dos fatores  $r$ ,  $g$ ,  $b$  e  $i$ , respectivamente. A média e as variâncias deste modelo são estruturadas e distribuídas da seguinte forma:

$$y/r, V \square N(Xr, V); g/\sigma_g^2 \square N(0, I\sigma_g^2); b/\sigma_b^2 \square N(0, \sigma_b^2);$$

$$i/\sigma_i^2 \square N(0, I\sigma_i^2); e/\sigma_e^2 \square N(0, I\sigma_e^2)$$

Os valores genotípicos são obtidos por meio da solução das equações de modelos mistos a seguir:

$$\begin{bmatrix} X'X & X'Z & X'W & X'T \\ Z'X & Z'Z + I \frac{\sigma_e^2}{\sigma_g^2} & Z'W & Z'T \\ W'X & W'Z & W'W + I \frac{\sigma_e^2}{\sigma_b^2} & W'T \\ T'X & T'Z & T'W & T'T + I \frac{\sigma_e^2}{\sigma_{ga}^2} \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{g} \\ \hat{b} \\ \hat{i} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \\ T'y \end{bmatrix}$$

A partir dos valores genotípicos preditos ( $\hat{g}$ ), as percentagens relativas dos valores genotípicos ( $PRVG$ ) são estimados para cada genótipo nos diferentes ambientes. A medida simultânea de adaptabilidade e estabilidade para cada genótipo é obtida por meio da Média Harmônica da Performance Relativa dos Valores Genotípicos ( $MHPRVG$ ) com base na seguinte expressão:

$$MHPRVG_i = n \sqrt[k]{\sum_{j=1}^k \frac{1}{PRVG_{ij}}}, \text{ em que:}$$

$n$ : número de ambientes;

$PRVG_{ij} = VG_{ij} / VG_j$ , sendo:

$VG_{ij}$ : valor genotípico do genótipo  $i$  no ambiente  $j$  e

$VG_j$ : média genotípica no ambiente  $j$ .

Para melhorar a interpretação dos resultados, os valores de MHPRVG são multiplicados pela média geral ( $MG$ ) de todos os ambientes ( $MHPRVG \times MG$ ), representando os resultados na mesma magnitude da característica estudada. Depois os ambientes são agrupados em favoráveis e desfavoráveis de acordo com a média geral da característica e são conduzidas análises separadas para cada grupo. Os ambientes com média acima da média geral são considerados favoráveis e os ambientes com média abaixo da média geral são considerados desfavoráveis (Mendes et al., 2012). Então, para analisar em conjunto a adaptabilidade e a estabilidade, os resultados dos valores  $MHPRVG \times MG$  de cada grupo são comparados em gráfico de dispersão, que contrasta ambientes favoráveis (eixo das abscissas) e ambientes desfavoráveis (eixo das ordenadas). O plano cartesiano do gráfico é dividido em quatro quadrantes, da seguinte forma: I(Inférieur Esquerdo)-genótipos com baixo desempenho em ambos grupos de ambientes ; II(Inférieur Direito)- genótipos com adaptabilidade específica a ambientes favoráveis; III(Superior Direito)- genótipos com desempenho superior em ambos grupos de ambientes; IV(Superior Esquerdo)-genótipos com adaptabilidade específica a ambientes desfavoráveis (Mendes et al., 2012; Yamamoto et al., 2021).

# CAPITULO 30

## Exemplos de Análises no R

Os exercícios e exemplos em R podem ser encontrados no seguinte endereço online <https://github.com/rfn-qtl/livro-biomeria-no-melhoramento/tree/main>. Periodicamente os scrips serão atualizados.

## BIBLIOGRAFIA

- AGUIAR, A. M.; RAMALHO, M. A. P.; SOUZA, E. A. de. Comparação entre látice e blocos aumentados na avaliação de famílias segregantes em um programa de melhoramento do feijoeiro. Ciênc. Agrotec., v. 24, n. 4: 857-860, 2000.
- AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: B. N. Petrox and F. Caski. Second International Symposium on Information Theory. Akademiai Kiado, Budapest. 1973. pp. 267-281.
- ALTINISIK, Y. Intrablock, interblock and combined estimates in incomplete block designs: a numerical study. Pennsylvania State University, 2013. 72 p. (Master of Science Thesis).
- ALVES, B. M.; CARGNELUTTI-FILHO, A.; BURIN, C. Multicollinearity in canonical correlation analysis in maize. Genetics and Molecular Research, 16(1): 1-14, 2017.
- ALVES, F. C.; GRANATO, I. S. C.; GALLI, G.; LYRA, D.H.; FRITSCHE-NETO, R.; CAMPOS, G. de los. Bayesian analysis and prediction of hybrid performance. Plant Methods, 15: 14. doi.org/10.1186/s13007-019-0388-x.
- ANDERSON, T. W. An introduction to multivariate statistical analysis. Third Edition. New Jersey: John Wiley & Sons Inc., 2003. 721p.
- AWATA, L. A. D.; TONGOONA, P.; DANQUAH, E.; EFIE, B.E.; MARCHELO-DRAGGA, P. W. Common mating designs in agricultural research and their reliability in Estimation of genetic parameters. Journal of Agriculture and Veterinary Science, 11(7), 2018:16-36.
- BAILEY, R. A.; SPEED, T. P. Rectangular lattice designs: efficiency factors and analysis. The Annals of Statistics, Vol. 14, No. 8: 874-895, 1986.
- BALZARINI, M. Applications of mixed models in plant breeding. Fac. Cs. Agropecuarias, Estadística y Biometría, Univ. Nac. de Córdoba, Argentina., 2015. 16 p.
- BARBIN D. Componentes de variância: teoria e aplicações. Piracicaba: FEALQ, 1993. 120 p.
- BARNETT, V. Comparative statistical inference. Third Edition. West Sussex: John Wiley & Sons, 1999. 381 p.
- BARRETO, H.; EDMEADES, G. O.; CHAPMAN, S. C.; CROSSA, J. The alpha lattice design in plant breeding and agronomy: generation and analysis. In: EDMEADES, G. O.; BÄNZIGER, M.; MICKELOSON, H. R.; PEÑA-VALDIVIA (Technical Editors). Developing Drought – and Low N – Tolerant Maize – Proceedings of a Symposium. March 25 – 29, CIMMYT, El Batán, Mexico, 1996. p. 544 – 551.
- BERNARDO, R. Breeding for quantitative traits in plants. Minnesota: Stemma Press, 2002, 369p.
- BERNARDO, R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. Heredity, 125: 375-385, 2020.
- BERTOIA, L.; LOPEZ, C.; BURAK, R. Biplot analysis of forage combining ability in maize landraces. Crop Sci., v. 46, p. 1346–1353, 2006.
- BOLSTAD, W. M. Understanding computacional bayesian statistics. New Jersey: John Wiley & Sons, 2010. 315 p.
- BOS, I.; CALIGARI, P. Selection methods in plant breeding. Netherlands: Springer, 2008. 461p.
- BROWNE, M. W. Cross-validation methods. Journal of Mathematical Psychology, v.44, p.108-132, 2000.
- BURDEN, R. L.; FAIRES, J. D.; BURDEN, A. M. Análise numérica. Tradução da 10<sup>a</sup> edição norte-americana. São Paulo: Cengage, 2017. 879 p.

- BURNHAM, K.; ANDERSON, D. R. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, v. 33, n. 2, p. 261-304, 2004.
- CAETANO, E. do R. R. Análise de variância utilizando modelos autoregressivos em experimentos com dependência especial. Dissertação (Mestrado em Estatística e Experimentação Agropecuária). Universidade Federal de Lavras, 2013. 115 p.
- CAMARINHA FILHO, J. A. Nota metodológica sobre modelos lineares mistos. Curitiba: UFPR, 2005. 27p.
- CAMPOS, J. F.; CARNEIRO, A. P.; PETERNELLI, L. A.; CARNEIRO, J. E. S.; SILVA, M. J.; CECON, P. R. Classificação de famílias do feijoeiro sob diferentes cenários de dependência espacial e precisão experimental. *Pesquisa Agropecuária Brasileira*, v. 51, n. 2, p. 105-111, 2016.
- CARROLL, J. D.; GREEN, P. E. Tools for applied multivariate analysis. Revised Edition. Berkeley: Elsevier Science, 1997. 367p.
- CASELLA, G.; BERGER, R. L. Inferência estatística. Tradução da 2<sup>a</sup> edição norte-americana. São Paulo: Cengage, 2010. 588 p.
- CECCON, F. Seleção de genótipos de milho eficientes na interação com *Azospirillum brasiliense*. Dourados: UFGD, 2019, 82 p (Dissertação de Mestrado em Agronomia)
- CLARK, I. Practical geostatistics. London: Applied Science, 1979. 129 p.
- COCHRAN, W. G.; COX, G. M. Experimental designs. New York: John Wiley & Sons Inc., 1950. 459 p.
- CONAGIN, A. Látices retangulares. *Bragantia*, vol. 13, n. 16, p. 187-197, 1954.
- CORRÊA, A. M.; TEODORO, P. E.; GONÇALVES, M. C.; BARROSO, L. M. A.; NASCIMENTO, M.; SANTOS, A.; TORRES, F. E. Adaptability and phenotypic stability of common bean genotypes through Bayesian inference. *Genetics and Molecular Research*, 15(2): 1-11, 2016.
- COX, D. R. Planning of experiments. New York: John Wiley & Sons Inc., 1958. 308 p.
- COX, D. R.; REID, N. The theory of the design of experiments. New York: Chapman & Hall/Crc, 2000. 311 p.
- CRESSIE, N. Statistics for spatial data. Second Edition. New York: John Wiley, 1993. 900 p.
- CROSSA, J. From genotype x environment to gene x environment Interaction. *Current Genomics*, 13: 225-244, 2012.
- CROSSA, J.; CORNELIUS, P. I.; YAN, W. Biplots of linear-bilinear models for studying crossover genotype x environment interaction. *Crop Science*, v. 42, p. 619-633, 2002.
- CROSSA, J.; PEREZ-ELIZALDE, S.; JARQUIN, D.; COTES, J. M.; VICLE, K.; LIU, G.; CORNELIUS, P. L. Bayesian Estimation of the Additive main Effects and Multiplicative Interaction model. *Crop Science*, 51: 1458-1469, 2011.
- CRUZ, C. D. Programa GENES: análise multivariada e simulação. Viçosa: Editora UFV, 2006. 175 p.
- CRUZ, C. D. Programa GENES: biometria. Viçosa: Editora UFV, 2006. 382p.
- CRUZ, C. D.; CARNEIRO, P.C.S. Modelos biométricos aplicados ao melhoramento genético: v. 2. 3<sup>a</sup> Edição. Viçosa: Editora UFV, 2014. 668p.
- CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. Biometria aplicada ao estudo da diversidade. Viçosa-MG: Universidade Federal de Viçosa-UFV, 2020. 614 p.
- CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. Modelos biométricos aplicados ao melhoramento genético: v. 1. 4<sup>a</sup> Edição. Viçosa: Editora UFV, 2012. 514p.
- CRUZ, C. D.; TORRES, R. A.; VENCOVSKY, R. An alternative approach to the stability analysis proposed by Silva and Barreto. *Revista Brasileira de Genética*, v. 12, n. 2, p.567-580, 1989.
- CRUZ, C.D. Princípios de genética quantitativa. Viçosa: Editora UFV, 2005.394p.

CRUZ, C.D.; VENCOVSKY, R. Comparação de alguns métodos de análise dialética. *Revista Brasileira de Genética*, v.12, p.425-438, 1989.

CYRILUS, O. W.; SAMSON, O. O.; OUKO, O. E. Screening of new strains of sugarcane using augmented block designs. *Mathematical Theory and Modeling*, Vol. 4, No. 8: 153-161, 2014.

DALBOSCO, E. Z. Progresso genético a partir de índices de seleção aplicado ao melhoramento intrapopulacional do maracujazeiro azedo. Tangará da Serra (MT): Universidade do Estado de Mato Grosso (UNEMAT), 2015. 77f. (Dissertação de Mestrado).

DEVORE, J. L. Probabilidade e estatística para engenharia e ciências. Tradução da 9<sup>a</sup> edição norte-americana. São Paulo: Cengage, 2019, 656 p.

DEY, A. Incomplete block designs. New Jersey: Hindustan Book Agency, 2010. 290 p.

DIXON, P. Should blocks be fixed or random? In: Conference on Applied Statistics in Agriculture, 2016 – 28th Annual Conference Proceedings. Kansas State University Libraries. New Prairie Press, 2016. pp. 23-39.

DUARTE, J. B. Biometria em genética e melhoramento de plantas: tendências e inquietações. In: SIMPÓSIO DE ATUALIZAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS: A GENÉTICA QUANTITATIVA E DE POPULAÇÕES NO BRASIL, Universidade Federal de Lavras, 2010. pp. 47-60.

DUARTE, J. B. Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal. Piracicaba, 2000, 293 p. (Tese de Doutorado-ESALQ/USP).

DUARTE, J. B. Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal. Tese (Doutorado em Genética e Melhoramento de Plantas). Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2000. 293 p.

DUARTE, J. B.; PINTO, R. M. C.; Biplot AMMI graphic representation of specific combining ability. *Crop Breeding and Applied Biotechnology*, v. 2, n. 2, p. 161-170, 2002.

DUARTE, J. B.; VENCOVSKY, R. Spatial statistical analysis and selection of genotypes in plant breeding, *Pesquisa Agropecuária Brasileira*, v. 40, p. 107-114, 2005.

DUARTE, J.B.; VENCOVSKY, R. Interação genótipos x ambientes: Uma Introdução à Análise “AMMI” Série Monografias, n 9. Ribeirão Preto, Sociedade Brasileira de Genética, 1999. 60 p.

EFRON, B.; HASTIE, T. Computer age statistical inference: algorithms, evidence, and data science. Cambridge: Cambridge University Press, 2017, 475 p.

EL-MOHSEN, A. A. A.; ABO-HEGAZY, S. R. Comparing the relative efficiency of two experimental designs in wheat field trials. *Sci. Res. & Rew. J.*, 1(3): 101-109, 2013.

EL-SHAFI, M. A. A. Efficiency of classical complete and incomplete block design in yield trial on bread wheat genotypes. *Res. J. Agric. & Biol. Sci.*, 10(1): 17-23, 2014.

ES, H. M. V.; ES, C. L. V. Spatial nature of randomization and its effect on the outcome of field experiments. *Agronomy Journal*, v. 85, p. 420 – 428, 1993.

FALCONER, D.S.; MACKAY, F. C. Introduction to quantitative genetics. Fourth Edition. Essex: Longman. 1996. 464p.

FEDERER, W. T. Construction and analysis of an augmented lattice square design. BU-1485-M, 2000. 12 p.

FEDERER, W. T. Experimental design: theory and application. Calcutta: Oxford and IBH Publishing, 1967. 544 p.

FEDERER, W. T. Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments. Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853. BU-1315-MB, 1997. 18 p.

FERREIRA, D. F. Estatística multivariada. 3<sup>a</sup> Edição. Lavras: Universidade Federal de Lavras-UFLA, 2018. 624 p.

FERREIRA, D. F. Fundamentos de probabilidade. Lavras: Universidade Federal de Lavras-UFLA, 2020. 707 p.

FOSS, A. H.; MARKATOU, M.; RAY, B. Distance metrics and clustering methods for Mixed-type data. *International Statistical Review*, 0, 0: 1-30, 2018.

FOULLEY, J-L. Mixed model methodology, Part I: Linear mixed models. Montpellier: Université de Montpellier-Technical Report, 2015. 191 p.

FRITSCH NETO, R.; VIEIRA, R.A.; SCAPIM, C. A.; MIRANDA, G.V.; REZENDE, L.M. 2012. Updating the ranking of the coefficient of variation from maize experiments. *Acta Scientiarum Agronomy*, v.34, p.99-101.

GAUCH-JR, H. G. A simple protocol for AMMI analysis of yield trials. *Crop Science*, 53: 1860-1869, 2013.

GAUCH, H. G., and ZOBEL, R. W. Predictive and postdictive success of statistical analyses of yield trials. *Theor. Appl. Genet.* 76: 1–10, 1988.

GAUCH, H.G. Statistical analysis of regional yield trials: AMMI analysis of factorial designs, Amsterdam: Elsevier, 1992. 278 p.

GAUCH, H.G.; PIEPHO, H.P.; ANNICCHIARICO, P. Statistical analysis of yield trials by AMMI and GGE: Further considerations. *Crop Science*, v.48, n.3, p.866-889, 2008.

GEISSER, J.; GREENHOUSE, S. W. An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of the Mathematical Statistics*, v. 29, p. 855-891, 1958.

GENTLE, J. E. Matrix algebra: theory, computations and applications in statistics. New York: Springer Science Business Media, 2007, 528p.

GEZAN, S. A. Generation of experimental designs with CycDesigN. Harpenden-UK: Rothamsted Research, 2009. 49 p.

GOMES, G. P.; BABA, V. Y.; SANTOS, O. P. dos; SUDRÉ, C. P.; BENTO, C. dos S.; RODRIGUES, R.; GONÇALVES, L. S. A. Combinations of distance measures and clustering algorithms in pepper germplasm characterization. *Horticultura Brasileira*, 37: 172-179, 2019.

GOWER, J. C. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-874, 1971.

GRIFFITH, D. A. Estimating spatial autoregressive model parameters with commercial statistical package. *Geographical Analysis*, v. 20, n. 1, p. 176-186, 1988.

GRIFFITH, D. A. Spatially N-way ANOVA model. *Regional Science and Urban Economics*, v. 22, p. 347-369, 1992.

GUERRA, P. A. G. Geoestatística operacional. Brasília: Ministério de Minas e Energia, 1988. 145 p.

GUIMARÃES, R. C.; CABRAL, J. A. S. Estatística. Edição Revista. Lisboa: McGraw Hill, 1999. 621p.

GUMPERTZ, M. L.; GRAHAM, J. M.; RISTAINO, J. B. Auto logistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics*, v. 2, n. 1, p. 131-156, 1997.

GUPTA, V. K.; PARSAD, R.; BHAR, L. M.; MANDAL, B. N. Statistical analysis of agricultural experiments. Part-I: Single factor experiments. New Delhi: ICAR, 2016. 396 p.

GUPTA, V. K.; PARSAD, R.; MANDAL, B. N. Significance of experimental designs in agricultural research. New Delhi: ICAR, 2015. 38 p.

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. Multivariate data analysis. Seventh Edition. London: Pearson, 2009. 734p.

HALLAUER, A.R; CARENA, M.J.; MIRANDA FILHO, J.B. Quantitative genetics in maize breeding, New York: Springer, 680 p., 2010.

- HARLOW, L. L. *The essence of multivariate thinking: basic themes and methods*. Second Edition. New York: Routledge-Taylor & Francis Ltd., 2014. 396p.
- HARVILLE, D. A. *Matrix algebra from a statistician's perspective*. New York: Springer-Verlag, 1997. 630p.
- HINKELMANN, K.; KEMPHTORNE, O. *Design and analysis of experiments. Vol. 2 – Advanced experimental design*. New Jersey: John Wiley & Sons, 2005. 766p.
- HINKELMANN, K.; KEMPHTORNE, O. *Design and analysis of experiments. Vol. 1 – Introduction to experimental design*, Second Edition. New Jersey: John Wiley & Sons, 2008. 631p.
- HONGYU, K.; SILVA, F; L; OLIVEIRA, A. C. S.; SARTI, D. A.; ARAÚJO, L. B.; DIAS, C. T. S. Comparação entre os modelos AMMI e GGE biplot para os dados de ensaios multiambientes. *Revista Brasileira de Biometria*, v. 33, n. 2, p.139-155, 2015.
- HUYNH, H.; FELDT, L. S. Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, v. 65, p. 1582-1589, 1970.
- HUYNH, H.; FELDT, L. S. Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics*, v. 1, p. 69-82, 1976.
- ISAAKS, E. H.; SRIVASTAVA, R. M. *An introduction to applied geostatistics*. New York: Oxford University, 1989. 561 p.
- JOHN, P. W. M. *Statistical design and analysis of experiments*. Philadelphia: Society for industrial and Applied Mathematics-SIAM, 1998, 378 p.
- JOHNSON, R.A.; WICHERN, D.W. *Applied multivariate statistical analysis*. Sixth Edition. New Jersey: Pearson Prentice Hall, 2007. 773p.
- JOLLIFFE, I. T. *Principal component analysis*. Second Edition. New York: Springer Verlag, 2002. 487p.
- JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining geostatistics*. San Diego: Academic, 1978. 600 p.
- KANG, M.S. *Quantitative genetics, genomics and plant breeding*. New York: CABI Publishing. 2002. 400p.
- KATSILEROS, A.; KOUKOUVINOS, C. Evaluation of experimental designs in durum wheat trials. *Biometrical Letters*, Vol. 52, No. 2: 105-114, 2015.
- KAUSHIK, M.; MATHUR, B. Comparative study of k-means and hierarchical clustering techniques. *International Journal of Software & Hardware Research in Engineering*, Volume 2, Issue 6: 93-98, 2014.
- KELECHI, A. C. Symmetric and unsymmetric balanced incomplete block designs: A comparative analysis. *International Journal of Statistics and Applications*, 2(4):33-39, 2012.
- KEMPTON, R. A.; FOX, P. N. (Editors). *Statistical methods for plant variety evaluation*. London: Chapman & Hall, 1977.191 p.
- KHAN, M. I.; SHAH, S. A. A.; KHAN, M.; ULLAH, K.; ULLAH, R.; KHATAK, S. I. Comparative efficiency of alpha lattice design and complete randomized block design in wheat and potato field trials. *Journal of Resources Development and Management*, vol. 11:115-117, 2015.
- KHATTREE, R.; NAIK, D. N. *Applied multivariate statistics with SAS software*. Second Edition. Cary, NC: SAS Institute and Wiley, 2003. 338p.
- KHATTREE, R.; NAIK, D. N. *Multivariate data reduction and discrimination with SAS software*. Cary, NC: SAS Institute Inc., 2000. 558p.
- KIESEPPÄ, I. A. Statistical model selection criteria and the philosophical problem of underdetermination. *Brit. J. Phil. Sci.*, v. 52, p. 761-794, 2001.
- KUTNER, M.H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied linear statistical models*. Fifth Edition. New York: McGraw-Hill/Irwin, 2005,1396 p.
- LAMB, E. G.; SHIRTLIFFE, S. J.; MAY, W. E. Structural equation modeling in plant sciences: an example using yield components in oat. *Can. J. Plant Sci.*, 91: 603-619, 2011.

- LECLERG, E. L.; LEONARD, W. H.; CLARK, A. G. Field plot technique, Second Edition. Minnesota: Burgess Publishing Company, 1962. 373p.
- LEGENDRE, P.; DALE, M. R. T.; FORTIN, M. J.; GUREVITCH, J.; HOHN, M.; MYERS, D. The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, n. 25, p. 601-615, 2002.
- LITTEL, R. C.; HENRY, P. R.; AMMERMAN, C. B. Statistical analysis of repeated measures data using SAS® procedures. *Journal of Animal Science*, v. 76, n. 4, p. 1216-1231, 1998.
- LITTEL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D.; SCHABENBERGER, O. SAS® for mixed models, Second Edition, Cary, NC: SAS® Institute Inc., 2006, 814p.
- LITTEL, R.C.; FREUND, R.J. & SPECTOR, P.C. SAS system for linear models, Third Edition. Cary (NC): SAS Institute Inc. 1991. 329 p.
- LONG, D. S. Practical handbook of spatial statistics. London: CRC Press, 1996. 307 p.
- LYNCH, M.; WALSH, B. Genetics and analysis of quantitative traits. Massachusetts: Sinauer Associates Inc., 1998.
- MACKAY, I.; PIEPHO, H. P.; GARCIA, A. A. F. Statistical methods for plant breeding. In: BALDING, D. J.; MOLTKE, I.; MARIONI, J. (Editors). Handbook of statistical Genomics, Fourth Edition. New York: John Wiley & Sons, 2019. pp. 501-530.
- MAEDA, A. K. M. Seleção de genótipos de milho eficientes no uso de fósforo por meio de análise dialélica e análise AMMI-Biplot. Dourados: UFGD, 2016. 91 p. (Dissertação de Mestrado em Agronomia).
- MAIA, E.; SIQUEIRA, D. L.; CARVALHO, S. A.; PETERNELLI, L. A.; LATADO, R. R. Aplicação da análise espacial na avaliação de experimentos de seleção de clones de laranjeira Pêra. *Ciência Rural*, v. 43, n. 1, 2013.
- MALOSETTI, M.; RIBAUT, J-M.; EEUWIJK, F. van. The statistical analysis of multi-environment data: modeling genotype-by-environment Interaction and its genetic basis. *Frontiers in Physiology*, Volume 4, Article 44: 1-17, 2013.
- MATHER, K, & JINKS, J.L. Biometrical genetics. New York: Chapman and Hall, 1977. 381 p.
- MCINTOSH, M. S. Can analysis of variance be more significant? *Agronomy journal*, v. 107, n. 2, 2015.
- MENDES, F. F.; GUIMARÃES, L. J. M.; SOUZA, J. C.; GUIMARÃES, P. E. O.; PACHECO, C. A. P.; MACHADO, J. de A.; MEIRELLES, W. F.; SILVA, A. R. da; PARENTONI, S. N. Adaptability and Stability of maize varieties using mixed models methodology. *Crop Breeding and Applied Biotechnology*, 12: 111-117, 2012.
- MIGON, H. S.; GAMERMAN, D.; LOUZADA, F. Statistical inference: An integrated approach. Second Edition. Boca Raton: CRC Press, 2015. 342 p.
- MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada. Belo Horizonte: Editora UFMG, 2005. 297 p.
- MIRANDA FILHO, J.B.; GORGULHO, E.P. Cruzamentos com testadores e dialelos. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S.; VALADARES-INGLIS, M.C. (eds.). Recursos genéticos e melhoramento de plantas. Rondonópolis: Fundação MT, p. 649-671, 2001.
- MÖHRING, J.; WILLIAMS, E.; PIEPHO, H. P. Inter-block information: to recover or not to recover it? *Theor. Appl. Genet.*, 128: 1541-1534, 2015.
- MUTHONI, J.; SHIMELIS, H. Mating designs commonly used in plant breeding: A review. *Australian Journal of Crop Science*, 14(22): 1855-1869, 2020.
- MYUNG, I. J.; PITTS, M. A.; KIM, W. Model evaluation, testing and selection. Ohio State University: Ohio, 2003, 45p.

NAKAMURA, L. R.; BAUTISTA, E. A. L.; QUARESMA, E. de S.; DIAS, C. T. dos S.; MIRANDA, E. F. O. Seleção de genótipos promissores de café: uma abordagem multivariada. *Rev. Bras. Biom.*, v. 31, n. 4: 516-528, 2013.

NDUWUMUREMYI, A.; TONGOONA, P.; HABIMANA, S. Mating designs: helpfull tool for quantitative plant breeding analysis. *Journal of Plant Breeding and Genetics*, 01(03): 117-129, 2013.

NGUYEN, N-K; FEDERER, W. T. Incomplete block designs. *Encyclopedia of Environmetrics*, Volume 2: 1039-1042, 2002.

NORMAN, P. E.; DZIDZIENYO, D. K.; KARIM, K. Y. Assessing mating designs utilized in cassava population improvement. Chapter IntechOpen, 2021, pp. 1-21.

NUNES, J. A. R. Incorporação da informação de parentesco no método genealógico pelo enfoque de modelos mistos. Lavras: UFLA, 2006. 113 p. (Tese – Doutorado em Agronomia/Genética e Melhoramento de Plantas).

NUNES, J. A. R. Minicurso: Modelos mistos aplicados ao melhoramento de plantas. 4º Congresso Brasileiro de Melhoramento de Plantas. São Lourenço: SBMP, 2007.

NUNES, R. de P. Métodos para a Pesquisa Agronômica. Fortaleza: UFC/Centro de Ciências Agrárias, 1998. 564 p.

O'Rourke, N.; HATCHER, L.; STEPANSKI, E. J. A step-by-step approach to using SAS for univariate and multivariate statistics. Second Edition. Cary, NC: SAS Institute and Wiley, 2005. 548p.

OLIVEIRA, R. L.; PINHO, R. G. V.; FERREIRA, D. F.; PIRES, L. P. M.; MELO, W. M. C. Selection index in the study of adaptability and Stability in maize. *The Scientific World Journal*, Volume 2014, Article ID 360570, 6 pages, <http://dx.doi.org/10.1155/2014/360570>

OLIVOTO, T. Índices de estabilidade genotípica e seleção simultânea multivariada: uma nova abordagem. Santa Maria: UFSM, 2019, 147 p (Tese de Doutorado em Agronomia).

OLIVOTO, T.; NARDINO, M. MGIDI: toward na effective multivariate selection in biological experiments. *Bioinformatics*, Volume 37, Issue 10, 15 May 2021, Pages 1383-1389, <http://doi.org/10.1093/bioinformatics/btaa981>

OLIVOTO, T.; SOUZA, V. Q. de; NARDINO, M.; CARVALHO, I. R.; FERRARI, M.; PELEGRI, A. J. de; SZARESKI, V. J.; SCHMIDT, D. Multicolinearity in path analysis: a simple method to reduce its effects. *Agronomy Journal*, 109 (1): 131-142, 2017.

OMER, S. O.; ABDALLA, A. W. H.; SINGH, N. P.; KUMAR, H.; SINGH, M. Bayesian Estimation of variance componentes, heritability and genetic advance from multi-year and location chickpea trials in Indian environments. *International Journal of Statistics and Applications*, 10(6): 150-159, 2020.

ORD, J. K. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, v. 70, n.3, p.120-126, 1975.

PANSE, V. G.; SUKHATME, P. V. Statistical methods for agricultural workers. New Delhi: Indian Council of Agricultural Research, 1957. 361 p.

PARSAD, R.; GUPTA, V. K.; BATRA, P. K.; SATPATI, S. K.; BISWAS, P. Monograph on  $\alpha$ -designs. New Delhi: Indian Agricultural Statistics Research Institute, 2007.203 p.

PATTERSON, H. D.; WILLIAMS, E. R. A new class of resolvable incomplete block designs. *Biometrika*, 63 (1), 83-92, 1976.

PAULENAS, V. P. Análise de experimentos em látice quadrado no melhoramento vegetal utilizando modelos mistos. Piracicaba: ESALQ/USP, 2016. 77 p. (Dissertação de Mestrado).

PEREZ-ELIZALDE, S.; JARQUIN, D.; CROSSA, J. A general bayesian Estimation method of linear-bilinear models Applied to plant breeding trials with genotype x environment Interaction. *Journal of Agricultural, Biological, and Environmental Statistics*, Volume 17, Number 1: 15-37, 2011.

PERRI, S. H. V.; IEMMA, A. F. Procedimento “MIXED” do SAS® para análise de modelos mistos. *ScientiaAgrícola*, v. 56, p. 959-967, 1999.

PETERNELLI, L. A.; RESENDE, M. D. de. Experimental designs for next generation phenotyping. In: FRITSCHÉ-NETO, R.; BORÉM, A. (Eds.). Phenomics, 2015. pp. 15-32.

PIASKOWSKI, J.; PRICE, W. Incorporating spatial analysis into agricultural field experiments. Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), November 05, 2021.

PIEPHO, H. P.; BUCHSE, A.; EMRICH, K. A Hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science*, v. 189, p. 310-322, 2003.

PIEPHO, H. P.; BUCHSE, A.; RICHTER, C. A mixed modeling approach for randomized experiments with repeated measures. *Journal of Agronomy and Crop Science*, v. 190, p. 230-247, 2004.

PIEPHO, H. P.; BÜCHSE, A.; TRUBERG, B. On the use of multiple lattice designs and  $\alpha$ -designs in plant breeding trials. *Plant Breeding*, 125: 523-528, 2006.

PIEPHO, H. P.; MÖHRING, J.; MELCHINGER, A. E.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161: 209-228, 2008.

PIMENTEL GOMES, F.; GARCIA, C.H. Experimentos em látice: planejamento e análise por meios de pacotes estatísticos, Séries técnica IPEF, Piracicaba, v.7, ed.23 p. 1-69, dez,1991.

PIMENTEL-GOMES, F. Curso de estatística experimental. 15<sup>a</sup> Edição. Piracicaba: FEALQ, 2009. 451 p.

PIMENTEL-GOMES, F.; GARCIA, C. H. Estatística aplicada a experimentos agronômicos e florestais: exposição com exemplos e orientações para o uso de aplicativos. Piracicaba: FEALQ, 2002. 309 p.

PLANT, R. E. Spatial data analysis in ecology and agriculture in R. New York: CRC Press, 2012. 617 p.

POOLE, D. Linear algebra: a modern introduction. Fourth Edition. Stanford: Cengage Learning, 2015. 720 p.

PRESS, S. J. Subjective and objective bayesian statistics. Second Edition. New Jersey: John Wiley & Sons, 2003. 552 p.

RAMALHO, M.A.P.; ABREU, A. F. B; SANTOS, J. B.; NUNES, J. A. R. Aplicações da genética quantitativa no melhoramento de plantas autógamas. Lavras: Editora UFLA, 2012. 522p.

RAMALHO, M.A.P.; FERREIRA, D.F. & OLIVEIRA, A.C. de. Experimentação em genética e melhoramento de plantas. Lavras: Editora UFLA, 2000. 326 p.

RAO, C. R. Statistical inference and its applications. Second Edition. New York: John Wiley & Sons, 2002. 625 p.

REGAZZI, A. J; SILVA, H. D; VIANA, J. M. S; CRUZ, C. D. Análises de experimentos em látice quadrado com ênfase em componente de variância, II, Análise Conjunta, Pesquisa agropecuária brasileira, Brasília, v,34, n,11, p,1987-1997, nov, 1999.

REGAZZI, A. J. Análise multivariada. Viçosa-MG: Universidade Federal de Viçosa-UFV, 1999 (Apostila da disciplina INF 766-Análise Multivariada).

REGAZZI, A. J.; CRUZ, C. D. Análise multivariada aplicada. Edição revista e ampliada. Viçosa-MG: Universidade Federal de Viçosa-UFV, 2020. 401 p.

RENCHER, A. C. Methods for multivariate analysis. Second Edition. New York: John Wiley & Sons Inc., 2002. 708p.

RENCHER, A. C.; SCHAALE, G. B. Linear models in statistics, Second Edition. New Jersey: John Wiley and Sons, 2008, 672p.

RESENDE, M. D. V. de; ALVES, R. S. Linear, generalized, hierarchical, bayesian and random regression mixed models in genetics/genomics in plant breeding. *Functional Plant Breeding Journal*, v. 2, n. 2, a 1: 1-31, 2020.

RESENDE, M. D. V. de. Genética quantitativa e de populações. Viçosa-MG: Suprema, 2015.463 p.

RESENDE, M. D. V. de. Matemática e estatística na análise de experimentos e no melhoramento genético. Brasília: Embrapa Informação Tecnológica, 2007. 561p.

- RESENDE, M. D. V. de. Software Selegen-REML/BLUP: sistema estatístico e seleção computadorizada via modelos lineares mistos. Colombo: Embrapa Florestas, 2007a, 350 p.
- RESENDE, M. D. V. Genética biométrica e estatística no melhoramento de plantas perenes. Brasília: Embrapa Informação Tecnológica, 2002. 975p.
- RESENDE, M. D. V. Métodos estatísticos ótimos na análise de experimentos de campo. Colombo: Embrapa Florestas, 2004. 57 p
- RESENDE, M. D. V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesquisa Agropecuária Tropical, v. 37, n. 3, p. 182-194, 2007.
- RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. Estatística matemática, biométrica e computacional. Viçosa, MG: Suprema, 2014. 881 p.
- RESENDE, M. D. V.; STURION. Análise estatística espacial de experimentos via modelos mistos individuais com erros modelados por processos arima em duas dimensões. Revista de Matemática e Estatística, v. 21, n.1, p. 7-33, 2003.
- RESENDE, M.D.V.; DUARTE, J.B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesquisa Agropecuária Tropical, 37:182-194, 2007.
- ROCHA, J. R. A. S. C.; MACHADO, J. C.; CARNEIRO, P. C. S. Multitrait index based on factor analysis and ideotype-design: proposal and applications to elephant grass breeding for bioenergy. GCB Bioenergy, 10, 2018: 52-60, doi: 10.1111/gcbb.12443.
- ROCHA, M. de M. Seleção de linhagens experimentais de soja para adaptabilidade e estabilidade fenotípica. Piracicaba: ESALQ/USP, 2002.173 p. (Tese de Doutorado).
- ROSA, G. J. M.; VALENTE, B. D.; CAMPOS, G. I.; WU, X. L.; GIANOLA, D. Inferring causal phenotype networks using structural equation models. Genet. Sel. Evol., 43: 6, 2011.
- ROSSONI, D. F. 2011. Análise de variância para experimentos com dependência espacial. Dissertação (Mestrado em Estatística e Experimentação Agropecuária). 108 p. 2011. Universidade Federal de Lavras.
- SCHWARZ, G. Estimating the dimension of a model. Annals of Statistics, v.6, p.461-464, 1978.
- SCHWARZBACH, E. Einige anwendungsmöglichkeiten elektronischer daten verarbeitung (EDV) für die beurteilung von zuchtmaterial. Arb. Tag. Oesterr. Pflanzenz Gumpenstein, 277-287, 1972.
- SCOLFORO, H. F.; SCOLFORO, J. R. S.; MELLO, J. M.; FERRAZ FILHO, A. C.; ROSSONI, D. F.; ALTOÉ, T. F.; OLIVEIRA, A. D.; LIMA, R. R. Autoregressive spatial analysis and individual tree modeling as strategies for the management of *Eremanthus erythropappus*. Journal of Forest Research, v. 27, n. 3, p. 595-603, 2016.
- SEARLE, S. R. Matrix algebra useful for statistics. New York: John Wiley and Sons, 1982.
- SEARLE, S. R.; CASELLA, G.; McCULLOCH, C. E. Variance components. New York: John Wiley & Sons, 1992. 501p.
- SEBER, G. A. F. A matrix handbook for statisticians. New Jersey: John Wiley & Sons Inc., 2008. 559p.
- SHAO, J. An asymptotic theory for model selection. Statistica Sinica, v.7, p.221-264, 1997.
- SHUKLA, G. K. Some statistical aspects of partitioning genotype-environment components of variability. Heredity 29 (2): 237-245, 1972.
- SILVA, H. D.; REGAZZI, A. J.; CRUZ, C. D.; VIANA, J. M. S. Análise de experimentos em látice quadrado com ênfase em componentes de variância: I. Análises individuais. Pesq. Agrop. Bras., v. 34, n. 10: 1811-1822, 1999.
- SILVA, J. G. C. da; BARRETO, J. N. An application of segmented linear regression to the study of the study of genotype x environment Interaction. Biometrics, v. 41, n. 4, p. 1093, 1986.
- SILVA, J. G. C. da. Análise da adaptabilidade através de regressão linear segmentada: I. Fundamentos. Pesq. Agropec. Bras., v. 30, n. 4, p. 435-448, 1995a.

SILVA, J. G. C. da. Análise da adaptabilidade através de regressão linear segmentada: II. Aplicação. *Pesq. Agropec. Bras.*, v. 30, n. 4, p. 449-462, 1995b.

SILVA, K. E. F.; VALE, J. C. do; FRITSCHE-NETO, R.; MARQUES, J. N. GGE biplot projection in adaptability and stability inference of soybean in an agricultural center Paraná, Brazil. *Revista Ciência Agronômica*, v. 52, n. 1, e20207131, 2021.

SILVA, M. J.; CARNEIRO, A. P. S.; FERES, A. L. G.; CARNEIRO, J. E. S.; SANTOS, N. T.; CECON, P. R. Spatial dependence in experiments of progeny selection for bean (*Phaseolus vulgaris* L.) yield. *Revista Ceres*, v. 63, n. 4, p. 477-485, 2016.

SILVA, R. R.; BENIN, G. Análises Biplot: conceitos, interpretações e aplicações. *Ciência Rural*, v. 42, n. 8: 1404-1412, 2012.

SIMÕES-PEREIRA, J. M. Introdução à matemática combinatória. Rio de Janeiro: Interciênciac, 2013. 338 p.

SINGH, M.; GUPTA, S.; PARSAD, R. Genetic crosses experiments. In: HINKELMANN, K. (Editor). Design and analysis of experiments: special designs and applications. New Jersey: John Wiley & Sons, 2012. p.1-71.

SINGH, P.; BHATIA, D. Incomplete block designs for plant breeding experiments. *Agric. Res. J.*, 54(4): 607-611, 2017.

SINGH, Y. On the analysis of simple rectangular lattice design. Department of Statistics, Central University of Rajasthan. Kishangarh (Ajmer)-305802. India. 11 p.

STEEL, R.G.D.; TORRIE, J.H.; DICKEY, D. A. Principles and procedures of statistics: a biometrical approach. Third Edition. New York: McGraw-Hill, 1997. 666 p.

STONE, M. Asymptotics for and against cross-validation. *Biometrics*, v.64, p.29-35, 1977.

STONE, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistical Society, Series B*, v.36, p.111-147, 1974.

STROUP, W. W; BAENZIGER, S. B.; MULITZE, D. K. Removing spatial variation from wheat yield trials: a comparison of methods. *Crop Science*, 34, p. 62-66, 1994.

STROUP, W. W. Generalized linear mixed models: modern concepts, methods, and applications. Boca Raton, FL: Chapman & Hall/CRC Press, 2013. 529 p.

TIMM, N. H. Applied multivariate analysis. New York: Springer-Verlag, 2002. 718p.

TIMPANI, V. D.; NASCIMENTO, T. E. C. do. Uma breve introdução à estatística bayesiana aplicada ao melhoramento genético animal. Belém-PA: Embrapa Amazônia Oriental, 2015. 57 p.

VALCHEVA, P.; OLIVEIRA, T. A. Some combinatorial structures in experimental design: overview, statistical models and applications. *Biom. Biostat. Int. J.*, 7(4): 346-351, 2018.

VALENTE, B. D.; ROSA, G. J. M.; CAMPOS, G. de los; GIANOLA, D.; SILVA, M. A. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, 185: 633-644, 2010.

VALENTE, B. D.; ROSA, G.J.M.; GIANOLA, D.; WU, X. L.; WEIGEL, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*, 194: 561-572, 2013.

VARGAS, M.; CROSSA, J. The AMMI analysis and the graph of the Biplot in SAS®. México: CIMMYT, 2000. 42p.

VENCOVSKI, R. & BARRIGA, P. Genética biométrica no fitomelhoramento. Ribeirão Preto: SBG, 1992. 496 p.

VIDIGAL, B. C. Avaliação de agrupamentos em mistura de variáveis. Viçosa-MG: UFV, 2013. 56 p. (Dissertação de Mestrado).

WALL, M. E.; RECHTSTEMER, A.; ROCHA, L. M. Singular value decomposition and principal component analysis. In: Berrar, D. P.; Dubitzky, W. Granzow, M. (eds). A practical approach to microarray data analysis. Norwell, MA: Kluwer, 2003. pp. 91-109. LANL LA-UR-02-4001.

- WEAKLIEM, D. L. Introduction to the special issue on model selection. *Sociological Methods & Research*, v. 33, n. 2, p. 167-187, 2004.
- WILLIAMS, E. R. A new class of resolvable block designs. Edinburg: University of Edinburg, 1975. 173p. (PhD Thesis).
- WOLFINGER, R.; Chang, M. Comparing the SAS® GLM and MIXED procedures for repeated measures. In: Proceedings of the Twentieth Annual SAS® Users Group Conference. Cary, NC: SAS® Institute Inc, 1995. p. 1-11.
- WRICKE, G. Zur berechnung der ökovalenz bei sommerweizen und hafer. *Zeitschrift für Pflanzenzüchtung*, v. 52, n. 2, p. 127-138, 1964.
- WRICKE, G.; WEBER, W. E. Quantitative genetics and selection in plant breeding. New York: Walter de Gruyter, 1986. 406 p.
- XAVIER, A.; MUIR, W. M.; CRAIG, B.; RAINES, K. M. Walking through the statistical black boxes of plant breeding. *Theor. Appl. Genet.*, 129: 1933-1949, 2016.
- YAMAMOTO, E. L. M. Interação genótipos x ambientes e variação espacial em experimentos de avaliação de genótipos de milho no Brasil central. Dourados: UFGD, 2018. 125 p. (Tese de Doutorado em Agronomia).
- YAMAMOTO, E. L. M.; GONÇALVES, M. C.; DAVIDE, L. M. C.; SANTOS, A. dos; CANDIDO, L. S. Adaptability and Stability of maize genotype in growing regions of central Brazil. *Rev. Ceres*, v. 68, n. 3: 163-168, 2021.
- YAN, W.; CORNELIUS, P. L.; CROSSA, J.; HUNT, L. A. Two types of GGE biplots for analyzing multi-environment trial data. *Crop Science*, 41, p. 656-663, 2001
- YAN, W. GGE Biplot vs. AMMI Graphs for Genotype-by-Environment Data Analysis. *Journal of the India Society of Agricultural Statistics*, v.65, n.2, p.181-193, 2011.
- YAN, W., HUNT, L. A., SHENG, Q.; SZLAVNICS, Z. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, v. 40, p. 597-605, 2000.
- YAN, W.; HOLLAND, J. B. A heritability adjusted GGE biplot for test environment evaluation. *Euphytica*, v.171, n. 3, p. 355-369, 2010.
- YAN, W.; HUNT, L.A. Biplot analysis of diallel data. *Crop Sci.*, v. 42, p. 21-30, 2002.
- YAN, W.; KANG, M. S. GGE biplot analysis: a graphical tool for breeders, geneticists, and agronomists. New York: CRC Press, 2003.
- YAN, W.; KANG, M. S.; MA, B.; WOODS, S.; CORNELIUS, P. L. GGE Biplot vs. AMMI analysis of genotype-by-environment data. *Crop science*, v.47, n.2, p.643-655, 2007
- YAN, W.; TINKER, N. A. Biplot analysis of multi-environment trial data: Principles and applications. *Canadian Journal of Plant Science*, v. 86, n. 3, p. 623-645, 2006.
- YANG, R-C; JUSKIW, P. Analysis of covariance in agronomy and crop research. *Can. J. Plant Sci.*, 91: 621-641, 2011.
- YANG, R-C. Towards understanding and use of mixed-model analysis of agricultural experiments. *Can. J. Plant Sci.*, 90: 605-627, 2010.
- YANG, R-C. Why is mixed analysis underutilized? *Can. J. Plant Sci.*, 88(3):563-567, 2008.
- YOUNG, G. A.; SMITH, R. L. Essentials of statistical inference. Cambridge: Cambridge University Press, 2005, 225 p.
- YWATA, A. X. C.; ALBUQUERQUE, P. H. M. Métodos e modelos em econometria especial: uma revisão. *Revista Brasileira de Biometria*, v. 29, n. 2, p. 273-306, 2011.
- ZEFFA, D. M.; MODA-CIRINO, V.; MEDEIROS, I.A.; FREIRIA, G. H.; NETO, J. S; IVAMOTO-SUZUKI, S. T.; DELFINI, J.; SCAPIM, C. A.; GONÇALVES, L. S. A. Genetic progress of seed yield and nitrogen use

efficiency of Brazilian carioca common bean cultivars using Bayesian approaches. *Front. Plant Sci.*, 11: 1168, doi: 10.3389/fpls.2020.01168.

ZIMMERMANN, F. J. P. Estatística aplicada à pesquisa agrícola. Segunda Edição. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2014. 582 p.

ZOBEL R.W.; WRIGHT M.J.; GAUCH H.G., 1988. Statistical analysis of a yield trial. *Agron J.* v.80, p. 388–393, 1988.

ZUCCHINI, W. An introduction to model selection. *Journal of Mathematical Psychology*, v. 44, p. 41-61, 2000.