QUE HIPÓTESES ESTATÍSTICAS TESTAMOS ATRAVÉS DO *SAS* EM PRESENÇA DE CASELAS VAZIAS?[1]

A.F. IEMMA
Depto. de Matemática e Estatística - ESALQ/USP, C.P. 9, CEP: 13418-900 - PIRACICABA, SP e ASSER - Centro de Ensino Superior - São Carlos - SP.

RESUMO: A interpretação das hipóteses testadas através da análise de variância de dados agropecuários balanceados pode ser feita, em geral, sem grandes problemas, mormente para experimentos bem planejados e bem conduzidos. Se, no entanto, os dados são desbalanceados e apresentam caselas vazias, então a interpretação das hipóteses testadas através das somas de quadrados fornecidas pelos pacotes estatísticos disponíveis pode ser extremamente difícil para os estatísticos e praticamente impossível para os profissionais das ciências aplicadas, usuários de pacotes estatísticos. Neste estudo, discute-se a interpretação das hipóteses mais comumente testadas através do procedimento GLM (General Linear Models) do sistema SAS (Statistical Analysis System), visando alertar os usuários sobre os problemas inerentes à opção por aquela que melhor espelha os objetivos de suas pesquisas.
Descritores: hipóteses estatísticas, caselas vazias, SAS/GLM

WHICH STATISTICAL HYPOTHESIS ARE TESTED BY *SAS* IN THE PRESENCE OF MISSING PLOTS?

ABSTRACT - The interpretation of the tested hypothesis through variance analysis of balanced agricultural data, can be made, in general, without great difficulties, specially in the case of well designed and well conducted experiments. If, however, data is not balanced and missing plots are present, the interpretation of the tested hypothesis through the sums of squares given by the available statistical packages, may be extremely difficult for statisticians, and practically impossible for profissionals of applied sciences, which use these packages. In this study, the interpretation of the most common tested hypothesis is discussed through the General Linear Models (GLM) procedure of the Statistical Analysis System (SAS), with the objective of alerting users about the problems related to the choice of the hypothesis that best reflects the objectives of his research.
Key Words: Cell Mean Models; Empty - Cell Data; Statistical Hyphotesis; SAS/GLM

LOCALIZAÇÃO DO PROBLEMA

Como é sabido dos iniciados na área dos modelos lineares, o modelo de Gauss-Markov tem sido, por sua simplicidade, o mais utilizado na análise de dados das mais diversas áreas do saber. Naturalmente as pesquisas do setor agropecuário não fogem à regra. Um fato, entretanto, deve ser realçado: a interpretação das verdadeiras hipóteses que são testadas através das somas de quadrados obtidas pelos vários métodos disponíveis, que é feita sem grandes problemas com amostras equilibradas, pode tornar-se extremamente difícil para os estatiscistas e praticamente impossível para os pesquisadores das ciências aplicadas, se há caselas vazias. Assim, não raro, pode ocorrer que o pesquisador julgue estar testando uma certa Hipótese $H_0^{(1)}$ quando na verdade a estrutura de desbalanceamento dos dados pode induzir ao teste de uma outra hipótese $H_0^{(2)}$, sem qualquer sentido prático para ele.

Um sério agravante para esse problema consiste no facil acesso, e portanto na utilização muitas vezes inadequada dos pacotes estatísticos. Assim, por exemplo, como esperar que um pesquisador faça uma boa interpretação da hipótese que esta sendo testada por uma das quatro somas de quadrados fornecidas pelo SAS. Como esperar que ele possa fazer uma comparação sensata de seus resultados com os resultados de outros pesquisadores que utilizaram, digamos, o SPSS, o MINITAB, o BMDP ou o GENTAST, entre outros, para a análise de seus dados? E o problema das amostras desconexas, em presença do qual o pesquisador pode julgar estar testando um hipótese quando na verdade as funções lineares que a definem nem mesmo são estimáveis?

Essas questões, embora possam parecer elementares, são ainda de difícil solução, sobretudo para pesquisadores que mesmo sendo verdadeiros expoentes em suas áreas de pesquisa, desconhecem os processos mais refinados da análise de dados.

[1] Pesquisa Realizada com Auxílio do CNPq.

O objetivo, aqui, não é o de apresentar todas as soluções para tais problemas mesmo porque, queremos crer, alguns não estão ainda satisfatoriamente resolvidos. Pretende-se apenas apresentar idéias básicas sobre o tema, de modo a alertar os pesquisadores das ciências aplicadas sobre sua complexidade, visando estimulá-los a planejar, conduzir e analisar suas futuras pesquisas com a orientação de estatíscistas consagrados ao planejamento e à análise de experimentos. Nesse contexto são feitos alguns comentários iniciais:

i) Não são discutidas aqui as técnicas de estimação de parcelas perdidas, largamente abordadas na literatura e que não resolveram satisfatoriamente o problema, principalmente se existem "muitas"parecelas perdidas. Referências sobre o tema podem ser obtidas em FREUND (1980), DODGE (1985) e MURRAY (1986) entre outros.

ii) O problema é discutido de modo geral e não são considerados casos particulares, como por exemplo, o caso de frequências proporciais conforme abordado em URQUARDT & WEEKS (1978), mesmo porque os contra exemplos publicados em BURDICK & HERR (1980) contestam fortemente as apregoadas vantagens de certos casos particulares.

iii) Aborda-se o problema das amostras desequilibradas, do ponto de vista da robustez estrutural e não da robustez estocástica. Boas noções sobre as implicações probabilisticas podem ser obtidas em HERR (1986).

iv) Considerando que estas notas são dirigidas aos pesquisadores de ciências aplicadas evitam-se, tanto quanto possível, as demonstrações e as generalizações e utilizam-se exemplos numéricos que simplificam sobremaneira a visualização do problema. Ademais, discute-se o tema com base em modelos de efeitos fixos com dois fatores cruzados. Essa opção pode ser justificada pelo fato de que esse modelo é o mais simples que já apresenta os graves problemas e dificuldades dos modelos com grandes números de parâmetros.

UM EXEMPLO NUMÉRICO

Conforme comentado anteriormente, as idéias centrais deste estudo estão sendo apresentadas com base num exemplo numérico, visando a simplificar sua interpretação.

Nesse contexto, seja o Quadro 1, no qual estão os dados resultantes de um experimento com dois fatores: O Fator A: Dois Tipos de Solo e o Fator B: Três Tipos de Sondas Pedológicas.

Com base nesse conjunto de dados é natural que o pesquisador venha a optar entre um modelo

TABELA 1: Dados em mg de $P_2 0_5$ por 100g de terra seca, segundo o solo e a sonda utilizada.

	j=1 Sonda 1	j = 2 Sonda 2	j = 3 Sonda 3
i = 1 Solo 1	43* 45	41 -	42 44
	88(2) 44	41(1) 41	86(2) 43	215(5) 43
i = 2 Solo 2	40 40 -	35 37 33	- - -
	80(2) 40	105(3) 35	-	185(5) 37
	168(4) 42	145(4) 36,5	86(2) 43	400(10) 40

(*) Dados adaptados de DAGNELIE (1973)

"com"e um modelo "sem" interação. Qual será mais adequado para analisar seus dados? Assim, como é quase um padrão de conduta na análise de dados, o pesquisador em geral estuda a interação e com base no resultado opta por um dos dois modelos, com vistas a verificar a igualdade entre solos e a igualdade entre sondas. Nesse momento, se há caselas vazias aparecem os primeiros problemas, dentre os quais alguns são citados a seguir:

i) O teste de interação não deve ser utilizado como um critério de escolha entre um modelo "com" e um modelo "sem" interação.

ii) Hipóteses sobre "apenas" efeitos de solos ou de sondas não são testáveis. Em outras palavras, se $\alpha_1 e \alpha_2$ denotam os efeitos de solos e β_1, β_2 e β_3, os efeitos de sondas, as hipóteses do tipo H_o: $\alpha_1 = \alpha_2$ e H_o: $\beta_1 = \beta_2 = \beta_3$, não são testáveis, a menos de restrições paramétricas e/ou reparametrizações convenientes.

iii) As formas das hipóteses utilizadas pelos principais pacotes estatísticos para "substituir"as descritas em (ii) são, em geral, extremamente complicadas e de difícil interpretação, mesmo para o pessoal de estatística.

NOÇÕES SOBRE MODELOS LINEARES

Faz-se aqui uma breve introdução de algumas idéias sobre modelos lineares, apenas para propiciar aos leitores das ciências aplicadas um modo simples e rápido de interpretação das hipóteses mais comumente testadas. Não se pretende discutir aqui as regras utilizadas para calcular as somas de quadrados. Informações sobre o tema podem ser obtidas por SEARLE (1987); IEMMA (1991, 1993 a, 1993 b); IEMMA et al. (1993 b) ; entre outros.

Suponha-se, para tanto, que uma observação qualquer, dentre as 10 existentes no Quadro 1, possa ser denotada por y_{ijk}, onde i = 1, 2 é o índice de solos, j = 1, 2, 3 é o índice de sondas e k = 1, ..., n_{ij} é o índice de repetições. Assim, por exemplo $y_{223} = 33$ significa que a terceira medida efetuada pela sonda 2, no solo 2 resultou em 33 mg de P_2O_5 por 100g de terra seca.

Há vários modos para modelar y_{ijk}. Um deles, através do modelo superparametrizado (Modelo - S),exprime os verdadeiros anseios do pesquisador, pois exibe explicitamente um parâmetro para cada efeito considerado no estudo:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

(Modelo - S)

onde:

y_{ijk} é o valor da k-ésima resposta observada na j-ésima sonda, no i-ésimo solo;

μ é uma constante inerente a cada observação. Em alguns casos particulares de restrições paramétricas e/ou de reparametrizações convenientes, pode ser tomada como média geral;

α_i é o efeito do solo i;

β_j é o efeito da sonda j;

γ_{ij} é o efeito da interação entre o solo i e a sonda j;

e_{ijk} é o erro aleatório atribuído à observação y_{ijk}, admitido com a estrutura de Gauss-Markov;

$$e_{ijk} \cap N(\phi, \sigma^2).$$

O Modelo - S é parte integrante da história da estatística e é, conforme já dito, de grande utilidade aos pesquisadores das ciências aplicadas, pois exibe explicitamente os parâmetros $\mu, \alpha, \beta, \gamma$ etc, sobre os quais estão concentradas as hipóteses de interesse. Por outro lado, segundo HOCKING (1985) e SEARLE (1987), entre outros, eles apresentam o inconveniente de conter um número de parâmetros maior que o número disponível de médias de caselas para estimá-los. No exemplo, há cinco médias de caselas $\bar{y}_{ij}$., na TABELA 1, e 1+2+3+5 = 11 parâmetros. Por essa razão, muitos autores preferem adotar o modelo de médias de caselas (Modelo - M).

Uma vantagem incontestável do Modelo - M, sobre o Modelo - S, diz respeito à grande simplificação que propicia na descrição das hipóteses, como será visto adiante.

Conforme SPEED *et al.* (1978) será tomada a forma

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

(Modelo - M)

onde y_{ijk} e e_{ijk} são como no Modelo - S e μ_{ij} é a média da população da qual foi extraída a amostra que compõe a casela (i,j). Pode-se observar que $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.

Como base no Modelo - M, os dados da TABELA 1 podem ser esquematizados na TABELA2:

TABELA 2: A TABELA 1 em relação ao Modelo - M

	SONDAS (B)		
SOLOS (A)	j = 1	j = 2	j = 3
i = 1	μ_{11} (2)	μ_{12} (1)	μ_{13} (2)
i = 2	μ_{21} (2)	μ_{22} (3)	

OBSERVAÇÃO: O número entre parenteses descreve a frequência da casela.

HIPÓTESES MAIS COMUNS SOBRE LINHAS, COLUNAS E INTERAÇÃO

Há vários tipos de hipóteses que têm sido utilizadas, ao longo do tempo, para avaliar a igualdade, digamos, entre os efeitos de linhas e entre os efeitos de colunas. Sem dúvida, é possível conviver muito bem com elas. Os problemas começam a surgir quando se observa que em geral, até por uma tradição histórica, o pesquisador pensa que está testando hipóteses do tipo

$$H_o : \alpha_1 = \alpha_2 = \ldots = \alpha_a$$

(Igualdade de Linhas)

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_b$$

(Igualdade de Colunas)

quanto na verdade, se há interação elas nem mesmo são estimáveis, quanto mais testáveis. Naturalmente pode-se utilizar de modelos com restrições adequadas e/ou reparametrizações, mas isto não será aqui discutido e pode ser visto em IEMMA (1987). A presença de caselas vazias agrava sensivelmente o problema, no sentido de que afeta severamente a forma, e portanto, a interpretação das hipóteses. Parece que o problema maior não está no fato de não se estar testando essas duas hipóteses, mas sim no fato de não saber o que se está testando.

Dentre os vários tipos de hipóteses existentes o procedimento GLM (General Linear Model) do sistema SAS (Statistical Analysis System), incorporou, em relação ao modelo em estudo, quatro tipos para testar efeitos de linhas, quatro para efeitos de colunas e um para a interação, como será visto a seguir. Detalhes sobre o assunto, bem como sobre os métodos utilizados na obtenção das somas de quadrados podem ser obtidos em IEMMA (1991, 1993 a, 1993 b).

Hipóteses Mais Comuns Sobre Linhas

a) Hipóteses do Tipo I

Nas hipóteses do tipo I, a igualdade entre os efeitos de linhas (solos), é verificada através de suas médias ponderadas.

Utilizando-se o Modelo - M, tem-se para os dados da TABELA 1, ou para o esquema da TABELA 2:

$$H_o^{(I)} = \frac{2\mu_{11} + \mu_{12} + 2\mu_{13}}{5} = \frac{2\mu_{21} + 3\mu_{22}}{5}$$

Já, para o Modelo - S, lembrando que $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, tem-se:

$$H_o^{(1)}: \alpha_1 - \alpha_2 - 2/5\beta_2 + 2/5\beta_3 + 2/5\gamma_{11} + \\ + 1/5\gamma_{12} + 2/5\gamma_{13} - 2/5\gamma_{21} - 3/5\gamma_{22} = 0$$

que, sem dúvida não é uma hipótese simples de ser interpretada e que é bem diferente da hipótese $H_o: \alpha_1 = \alpha_2$ ($ou\ H_o: \alpha_1 - \alpha_2 = 0$) que, em geral, o pesquisador julga estar testando.

A soma de quadrados fornecida pelo SAS, para $H_0^{(1)}$ pode ser obtida como em IEMMA (1987, 1991); IEMMA et ali. (1993 b), entre outros:

$$SQ\ H_0^{(1)} = R(\alpha|\mu) = 90{,}000$$

conforme pode ser observado na TABELA 3.1.

b) Hipóteses do Tipo II

Com as hipóteses do tipo II, a igualdade entre efeitos de linhas (solos) é verificada através de médias ponderadas de linhas ajustadas para colunas.

No Modelo - M, sua forma geral é:

$$H_o^{(2)}: \sum_i n_{ij}\,\mu_{ij} = \sum_{i'}\sum_j \frac{n_{ij}\,n_{i'j}\,\mu_{i'j}}{n.j}$$

que, naturalmente não é de fácil compreensão, mesmo para os profissionais da estatística

Para o exemplo em questão, tem-se:

$$H_0^{(2)}: 1/5\mu_{11} + 3/20\mu_{12} = 1/5\mu_{21} + 3/20\mu_{22}$$

E, com o Modelo - S.

$$H_o^{(2)}: \alpha_1 - \alpha_2 + 4/7\gamma_{21} + 3/7\gamma_{12} - \\ - 4/7\mu_{21} - 3/7\gamma_{22} = 0$$

É importante ressaltar que, no exemplo em questão, $H_o^{(2)}$ não considera as observações da casela (1;3), isto é, não considera a sonda 3. Assim, se as três sondas devem ser comparadas, a hipótese $H_o^{(1)}$ deve ser a escolhida e $H_o^{(2)}$ deve ser descartada. Até aí tudo bem O grande problema consiste no fato de que, em geral, o pesquisador não sabe que isto está ocorrendo com suas hipóteses.

A soma de quadrados fornecida pelo SAS para testar $H_o^{(2)}$ pode ser obtida como em IEMMA (1991):

$$SQ\ H_o^{(2)} = R(\alpha|\mu,\beta) = 41{,}2857$$

conforme consta das TABELAS 3 2 e 4.1.

c) Hipóteses do Tipo III

Com as hipóteses do tipo III, a igualdade entre efeitos de linhas (solos) é verificada através de suas médias não ponderadas. Pode ser observado que agora, as frequências das caselas deixam de ser importantes.

Para o exemplo em questão, tem-se no Modelo - M:

$$H_o^{(2)}: \frac{\mu_{11} + \mu_{12}}{2} = \frac{\mu_{21} + \mu_{22}}{2}$$

que, como $H_o^{(2)}$, não utiliza as observações da casela (1;3).

No Modelo - S, vem:

$$H_o^{(3)}: \alpha_1 - \alpha_2 + 1/2\gamma_{11} + 1/2\gamma_{12} - \\ - 1/2\gamma_{21} - 1/2\gamma_{22} = 0$$

A soma de quadrados correspondente pode ser obtida conforme IEMMA (1991) e IEMMA, et al. (1993 b), entre outros:

$$SQ\ H_o^{(3)} = R(\dot{\alpha}|\dot{\mu}, \dot{\beta}, \dot{\gamma}) = 42{,}8571$$

como pode ser observado nas TABELAS 3.3 e 4.3.

d) Hipóteses do Tipo IV

Com as hipóteses do tipo IV, a igualdade entre efeitos de linhas (solos) é verificada através de médias não ponderadas. Nesse contexto elas podem ser semelhantes às do tipo III. Se, no entanto, existem caselas vazias e mais de dois níveis do fator então, em geral, as hipóteses dos tipos III e IV são diferentes.

Ademais, as hipóteses do tipo IV não são únicas, em geral, quando há caselas vazias, pois elas dependem da quantidade e da posição de tais caselas.

As hipóteses do tipo IV podem ser obtidas construindo-se contrastes entre médias de caselas que estão na mesma coluna, iniciando-se o processo sempre pela última linha.

Para o exemplo em questão tem-se:

$$\mu_{11}\ \mu_{12}\ \mu_{13}$$

$$\mu_{21}\ \mu_{22}\ -\qquad\qquad \leftarrow \text{início}$$

e então:

$$H_o^{(4)}: \mu_{12} - \mu_{22} + \mu_{11} - \mu_{21} = 0$$

ou seja

$$H_o^{(4)}: \mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}$$

que neste exemplo, em particular, coincide com a hipótese $H_o^{(3)}$.

Então:

$$SQ\,H_o^{(4)} = 42{,}8571$$

como pode ser visto nas TABELAS 3.4 e 4.4.

Detalhes sobre a não unicidade das hipóteses, e portanto das somas de quadrados, do tipo IV podem ser obtidos em IEMMA (1991, 1993 a, 1993 b).

Hipóteses Mais Comuns Sobre Colunas

De modo análogo ao discutido no caso das linhas, obtem-se as hipóteses sobre colunas. Deve ser ressaltado que agora, como há dois graus de liberdade para sondas, então as hipóteses respectivas apresentam duas partes e não mais uma única como no caso dos solos, que apresentam um só grau de liberdade.

a) Hipóteses do Tipo I

Aqui, os efeitos de colunas (sondas), são testados através de suas médias ponderadas.

$$H_o^{(5)}: \begin{cases} \dfrac{2\mu_{11} + 2\mu_{21}}{4} = \mu_{13} \\ \dfrac{2\mu_{11} + 2\mu_{21}}{4} = \mu_{13} \end{cases}$$

E no Modelo - S:

$$H_O^{(5)}: \begin{cases} \beta_1 - \beta_3 - 1/2\alpha_1 + 1/2\alpha_2 + 1/2\gamma_{11} - \\ \quad -\gamma_{13} + 1/2\gamma_{21} = 0 \\ \beta_2 - \beta_3 - 3/4\alpha_1 + 3/4\alpha_2 + 1/4\gamma_{12} - \\ \quad -\gamma_{13} + 3/4\gamma_{22} = 0 \end{cases}$$

que sem dúvida, é bem difícil de ser interpretada e é muito diferente da hipótese que, em geral, o pesquisador supõe estar testando:

$$H_o: \beta_1 = \beta_2 = \beta_3.$$

A soma de quadrados fornecida pelo SAS, para $H_o^{(5)}$ é, conforme referências do item anterior:

$$SQ\,H_o^{(5)} = R(\beta|\mu) = 83{,}0000$$

como pode ser observada na TABELA 4.1.

b) Hipóteseses do Tipo II

Que testa os efeitos de colunas (sondas), através das médias ponderadas de colunas ajustadas para linhas.

Conforme IEMMA (1991, 1993 a, 1993 b)

$$H_o^{(6)}: \begin{cases} 6/5\mu_{11} - 2/5\mu_{12} - 4/5\mu_{13} + \\ +6/5\mu_{21} - 6/5\mu_{22} = 0 \\ -2/5\mu_{11} + 4/5\mu_{12} - 2/5\mu_{13} + \\ +6/5\mu_{21} + 6/5\mu_{22} = 0 \end{cases}$$

E no Modelo - S:

$$H_o^{(6)}: \begin{cases} \beta_1 - \beta_3 - 11/14\gamma_{11} + 3/14\gamma_{12} - \\ -\gamma_{13} + 3/4\gamma_{21} - 3/14\gamma_{22} = 0 \\ \beta_2 - \beta_3 + 3/7\gamma_{11} + 4/7\gamma_{12} - \\ -\gamma_{13} - 3/7\gamma_{21} + 3/7\gamma_{22} = 0 \end{cases}$$

que mesmo não contendo os efeitos de solos (pois nesta hipótese os efeitos de sondas estão ajustados para solos), ainda é bem difícil de ser interpretada e o pesquisador deve estar consciente disso.

De modo análogo aos anteriores, o SAS fornece.

$$SQ\,H_o^{(6)}: R(\beta|\mu,\alpha) = 34{,}2857$$

como pode ser visto nas TABELAS 3.2 e 4 2.

c) Hipóteses do Tipo III

As hipóteses do tipo III testam efeitos de colunas (sondas), através de suas médias não ponderadas

$$H_o^{(7)}\begin{cases}\dfrac{\mu_{11}+\mu_{21}}{2} = \mu_{13}\\[2ex] \dfrac{\mu_{12}+\mu_{22}}{2} = \mu_{13}\end{cases}$$

E no Modelo - S:

$$H_o^{(7)}:\begin{cases}\beta_1 - \beta_3 + 3/4\gamma_{11} + 1/4\gamma_{12} -\\ -\gamma_{13} + 1/4\gamma_{21} - 1/4\gamma_{22} = 0\\ \beta_2 - \beta_3 + 1/4\gamma_{11} + 3/4\gamma_{12} -\\ -\gamma_{13} - 1/4\gamma_{21} + 1/4\gamma_{22} = 0\end{cases}$$

que, conforme as TABELAS 3.3. e 4.3 , fornece

$$SQH_0^{(7)} = R(\alpha,|\mu,\beta,\gamma) = 27{,}5789$$

d) Hipóteses do Tipo IV

De modo análogo ao desenvolvido para linhas (solos), testam os efeitos de colunas (sondas) através de suas médias não ponderadas. Para simplificar o processo, pode-se trocar as linhas pelas colunas, na TABELA 2 e proceder conforme descrito em 4.1.d., obtendo-se·

$$\begin{matrix}\mu_{11} & \mu_{21}\\ \mu_{12} & \mu_{22}\\ \mu_{13} & -\end{matrix} \Rightarrow H_o^{(8)}:\begin{cases}\mu_{11} = \mu_{13}\\ \mu_{12} = \mu_{13}\end{cases}$$

pode ser observado que essa hipótese não considera qualquer observação verificada na linha 2 (solo 2). Será que o usuário está consciente disso?

No Modelo - S, vem:

$$H_o^{(6)}:\begin{cases}\beta_1 - \beta_3 + \gamma_{11} - \gamma_{13} = 0\\ \beta_2 - \beta_3 + \gamma_{12} - \gamma_{13} = 0\end{cases}$$

e, conforme pode ser visto nas TABELAS 3,4 e 4.4, o SAS fornece:

$$SQ\,H_o^{(8)} = 6{,}0000$$

Hipóteses sobre a Interação

Como é bem sabido, se não há caselas vazias, tem-se, para o modelo em questão $\frac{1}{4}$ ab(a-1) (b-1) possíveis interações, das quais (a-1) (b-1) são linearmente independentes. Em presença de caselas vazias, no entanto, essa regra deixa de ser válida. No exemplo em questão, tem-se apenas uma interação que pode ser estimada·

$$\Delta = \mu_{11} + \mu_{22} - \mu_{12} - \mu_{21}$$

$$\begin{matrix}\mu_{11} & \mu_{12} & \mu_{13}\\ \mu_{21} & \mu_{22} & -\end{matrix}$$

e não (a-1) (b-1) = (1) (2) = 2 interações estimáveis como seria de se esperar se não houvesse casela vazia. Esse fato pode ser verificado nas TABELAS de 3.1. a 3.4. e de 4.1. a 4.4., nas quais pode-se observar "um"grau de liberdade para a interação e não "dois" como poderia supor o pesquisador.

Nesse contexto, a hipótese testável é:

$$H_o^{(9)}: \mu_{11} + \mu_{22} - \mu_{12} - \mu_{21} = 0$$

e não

$$H_o: \mu_{ij} + \mu_{ij'} + \mu_{i'j} - \mu_{i'j'} = 0, \forall (i,j,i',j').$$

bem mais geral e que é testável se não há caselas vazias. Sendo assim, o teste da interação, isto é, o teste de $H_o^{(9)}$, não pode ser utilizado como um critério de escolha entre um modelo com ou sem interação, já que não corresponde à hipótese de que "todas"as interações são nulas.

Em termos do Modelo - S, tem-se:

$$H_o^{(9)}: \gamma_{11} + \gamma_{22} - \gamma_{12} - \gamma_{21} = 0$$

Para o qual o SAS fornece:

$$SQ\,H_o^{(9)} = R(\gamma|\mu,\alpha,\beta) = 1{,}7143$$

como pode ser observado nas TABELAS de 3.1 a 3.4 e de 4.1 a 4.4..

TABELA 3· Análise de Variância dos Dados da TABELA 1, fornecida pelo PROC GLM do do SAS, com Base noModelo de Dois Fatores de Efeitos Fixos e Interação. Ordem A-B-A*B.

3 1 - S Q do Tipo I

VC	G L	H_o	R( )	S Q I
A(não ajustado)	1	$H_o^{(1)}$	$R(\alpha\|\mu)$	90,0000
B(ajustado)	2	$H_o^{(6)}$	$R(\beta\|\mu,\alpha)$	34,2857
AB	1	$H_0^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143

3 2 - S Q do Tipo II

VC	G L	H_o	R( )	S Q II
A(ajustado)	1	$H_o^{(2)}$	$R(\alpha\|\mu,\beta)$	41,2857
B(ajustado)	2	$H_o^{(6)}$	$R(\beta\|\mu,\alpha)$	34,2857
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143

3 3 - S Q do Tipo III

VC	G L	H_o	R( )	S Q III
A	1	$H_o^{(3)}$	$R(\alpha\|\mu,\beta,\gamma)$	42,8571
B	2	$H_o^{(7)}$	$R(\beta\|\mu,\alpha,\gamma)$	27,5789
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143

3 4 S Q do Tipo IV

VC	G L	H_o	R( )	S Q III
A	1*	$H_o^{(4)}$	$SQ\,H_o^{(4)}$	42,8571
B	2*	$H_o^{(8)}$	$SQ\,H_o^{(8)}$	6,0000
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143

NOTE: Other Type IV Testable Hypothesis exist which may yield different SS. (Mensagem Fornecida)

TABELA 4: Análise de Variância dos Dados da TABELA 1, Fornecida pelo PROC GLM do SAS, com Base noModelo de Dois Fatores de Efeitos Fixos e Interação. Ordem B-A-B*A.

	4.1. - S.Q. do Tipo I			
VC.	G L.	H_o	R( )	S.Q I
B(não ajustado)	2	$H_o^{(5)}$	$R(\beta\|\mu)$	83,0000
A(ajustado)	1	$H_o^{(2)}$	$R(\alpha\|\mu,\beta)$	41,2857
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143
	4.2. - S Q. do Tipo II			
VC	G L.	H_o	R( )	S Q II
B(ajustado)	2	$H_o^{(6)}$	$R(\beta\|\mu,\alpha)$	34,2857
A(ajustado)	1	$H_o^{(2)}$	$R(\alpha\|\mu,\beta)$	41,2857
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143
	4 3 - S Q. do Tipo III			
V.C.	G L.	H_o	R( )	S Q III
B	2	$H_o^{(7)}$	$R(\beta\|\mu,\alpha,\gamma)$	27,5789
A	1	$H_o^{(3)}$	$R(\alpha\|\mu,\beta,\gamma)$	42,8571
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143
	4 4. - S Q. do Tipo IV			
V.C.	G L	H_o	R( )	S Q IV
B	2*	$H_o^{(8)}$	$SQ\ H_o^{(8)}$	6,0000
A	1*	$H_o^{(4)}$	$SQ\ H_o^{(4)}$	42,8571
AB	1	$H_o^{(9)}$	$R(\gamma\|\mu,\alpha,\beta)$	1,7143

NOTE: Other Type IV Testable Hypothesis exist which may yield different SS (Mensagem Fornecida)

CONSIDERAÇÕES FINAIS

Conforme combinado anteriormente, este é um texto escrito para profissionais das ciências aplicadas, que fazem da análise de dados e dos pacotes estatísticos meras ferramentas de trabalho. Sendo assim, embora apresentem necessidade vital de conhecer e bem interpretar os aspectos práticos da análise de dados, tais pesquisadores não têm, em geral, interesse nos fundamentos teóricos e nas provas requintadas de sua validade.

Nesse contexto procurou-se omitir os aspectos teóricos e apenas apresentar resultados considerados úteis do ponto de vista prático. A cada momento ficou registrado, como objetivo básico, o grito de alerta para a interpretação das hipóteses que realmente são testadas através do PROC GLM do SAS, quando há caselas vazias. Novamente aqui,

recomendam-se aos interessados em maiores detalhes, os textos de IEMMA (1991, 1993 a, 1993 b); IEMMA e PALM(1992 a, 1992 b); IEMMA et al. (1993 a, 1993 b), entre outros.

Algumas considerações finais, no entanto, se fazem necessárias para um melhor aproveitamento do texto:

Porque duas tabelas de análises de variância?

Com a apresentação das TABELAS 3 e 4, pretende-se mostrar ao leitor que as somas de quadrados do tipo I, fornecidas pelo PROC GLM do SAS, dependem da ordem de entrada dos parâmetros no modelo, pois elas são obtidas sequencialmente. Assim, por exemplo, se como na TABELA 3, a ordem é A-B-A*B, isto é, se o modelo é

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

, então as somas de quadrados são obtidas supondo-se inicialmente o modelo

$y_{ijk} = \mu + \alpha_i + e_{ijk}$, a seguir o modelo $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$ e finalmente o modelo completo, obtendo-se respectivamente $R(\alpha|\mu)$ = 90.0000; $SQ(\beta|\mu,\alpha)$ = 34,2857 e $SQ(\gamma|\mu,\alpha,\beta)$ = 1,7143. De modo análogo, se como na TABELA 4, a ordem de entrada dos parâmetros é B-A-B*A, isto é, se o modelo adotado é

$y_{ijk} = \mu + \beta_i + \alpha_j + \gamma_{ij} + e_{ijk}$, então sequencialmente são obtidas

$SQ(\beta|\mu)$ = 83,0000, $SQ(\alpha|\mu,\beta)$ = 41,2857 e $SQ(\gamma|\mu,\alpha,\beta$ = 1,7143,

respectivamente através dos modelos

$$Y_{ijk} = \mu + \beta_i + e_{ijk}\,; Y_{ijk} = \mu + \beta_i + \alpha_j + e_{ijk}$$

e

$$y_{ijk} = \mu + \beta_i + \alpha_j + \gamma_{ij} + e_{ijk}.$$

Agora, as somas de quadrados do tipo II, são obtidas diretamente das tabelas das somas de quadrados do tipo I. Basta que se tomem apenas as ajustadas, isto é, $R(\alpha|\mu,\beta)$ e $R(\beta|\mu,\alpha)$ como nas TABELAS 3.2 e 4.2. Pode ser observado também que apenas as S.Q. do tipo I tem seus valores "alterados" pela ordem de entrada dos parâmetros.

Como estão relacionados os quatro tipos de hipóteses?

Conforme pode ser visto em IEMMA (1991) os quatro tipos de somas de quadrados fornecidas pelo PROC GLM do SAS, são obtidos com base nos métodos de YATES(1934); HENDERSON(1953); OVERALL e SPIEGEL (1969) que, de certo modo, pretendem generalizar os casos nos quais não há desbalanceamento. Nesse contexto, tem-se que:

i) Se os dados são balanceados os quatro tipos de somas de quadrados testam a mesma hipótese e, nesse caso

SQ Tipo I = SQ Tipo II = SQ Tipo III = SQ Tipo IV

ii) Mesmo no caso de dados desbalanceados, se não há caselas vazias, então as hipóteses dos tipos III e IV são iguais, e portanto

SQ Tipo III = SQ Tipo IV

iii) Mesmo em presença de caselas vazias, se o modelo não contém interação, as hipóteses dos tipos II, III e IV são iguais, e então

SQ Tipo II = SQ Tipo III = SQ Tipo IV

Novamente em alerta

Conforme visto anteriormente, o valor da SQ Interação em todas as oito análises apresentadas nas TABELAS de 3.1 a 3.4 e de 4.1 a 4.4 foi $SQ(\gamma|\mu,\alpha,\beta)$ = 1,7143, invariante para qualquer tipo de soma de quadrados, dentre os apresentados. Este fato vem confirmar o postulado de que "modelo com interação é bom para testar interação". Assim, a menos de restrições paramétricas e/ou reparametrizações convenientes, ver IEMMA(1993 a , 1993 b), não são testáveis nesse modelo, hipóteses que contenham apenas efeitos principais. Na verdade as funções paramétricas correspondentes a tais hipóteses, nem mesmo são estimáveis. Um sério agravante consiste na ocorrência de caselas vazias, pois conforme já visto, nesse caso o teste da interação não testa que todas as interações são nulas, mas sim apenas aquelas que são estimáveis e portanto testáveis. Desse modo, o teste da interação não deve ser utilizado como um critério para a escolha entre um modelo "com" ou "sem" interação.

Num contexto mais geral, registra-se aqui o alerta de que as hipóteses testadas em presença de caselas vazias devem ser estudadas com cautela por parte do pesquisador, pois uma interpretação

descuidada poderá comprometer resultados obtidos a custa de muita dedicação, pesquisa e investimento. Sem dúvida o planejamento a priori e o auxílio de estaticista habituado com a análise de dados, em geral minimiza esse problema.

REFERÊNCIAS BIBLIOGRÁFICAS

BURDICK, D.S; HERR, D.G. Counterexemples in unbalanced two way analysis of variance. **Communications in Statistics**. Part A: theory and methods, New york, v.9 nº .2, p. 231 - 241, 1980

DAGNELIE, P. **Estatística:** teoria e métodos I. Gembloux: Faculdade de Ciências Agronômicas, 1973 439 p.

DODGE, Y. **Analysis of experiments with missing data.** New York. John Wiley, 1985. 499 p

FREUND, R.J. The case of missing cell **The American Statistician,** Alexandria, v.34, nº 2, p. 94-98, 1980

HENDERSON, C.R. Estimation on variance and covariance components. **Biometrics.** *Alexandria, v 9, p 226-252, 1953*

HERR, D.G. On the history of Anova in unbalanced, factorial Designs The first 30 years. **The American Statician, Alexandria,** v. 40 , nº 4, p 265-270, 1986

HOCKING, R.R. **The analysis of linear models.**. Monterey, Brooks/Cole, 1985. 385 p.

IEMMA, A.F. Testes de hipóteses nos experimentos com parcelas subdivididas em blocos incompletos . **Desarollo Rural en Las Américas,** Costa Rica, v. 15, nº 2, p. 143 - 151, 1983

IEMMA, A.F. Modelos Lineares: **uma introdução para profissionais da pesquisa agropecuária.** Londrina. Imprensa Oficial do Estado do Paraná, 1987. 263 p.

IEMMA, A.F. **Testes de hipóteses em modelos lineares com amostras desequilibradas.** Gembloux. Faculdade de Ciências Agronômicas, 1991. 105 p.

IEMMA, A.F. PALM, R. Les inverses généralisées et leur utilisation dans le modéle linéaire. **Notes de statistique et informatique**,Gembloux, v. 92, nº 1, p.1 - 25, 1992 - a.

IEMMA, A.F. PALM R. Orthogonal projections for unbalanced data. **Congreso iberoamericano de estadística.** 1, 1992, Caceres. Acta...Espana, 11-16 1992 b.

IEMMA, A.F. **Análisis de varianza con datos desbalanceados.** Colombia Universidad Nacional. 1993 a. 101 p.

IEMMA, A.F. Analyse de la variance des modèles linéaires à effects fixes avec echantillons déséquilibrées. Nancy. INRA, 1993 b. 46 p.

IEMMA, A.F.; PALM, R.; ALVES, M.I.F. Statistical hypothesis and ortogonal projections for unbalanced data. **Biometrics Bulletin**, Alexandria, v. 10, nº 1, p 18, 1993 b.

IEMMA, A.F.; PALM, R; CLAUSTRIAUX, J.J. Sobre a construção de projetores ortogonais. **Revista de Matemática e Estatística,** São Paulo, v. 11, p.133 - 142, 1993 c.

IEMMA, A.F. **Análise de variância de dados desbalanceados.** 4° Congresso Brasileiro de Usuários do "SAS". USP/ESALQ Piracicaba, 1995. 111 p

LEVY, K.J.; NARULA, S ; ABRAMI, P. An empirical comparison of the methods of least squares in unweighted means for the analysis of disproportionate Cell Data **International statistical review,** Hague, v 3, p 335-338, 1973.

MURRAY, L.W. Estimation of missing cells in randomized block and latin square designs **The American Statistician,** Alexandria, v 40 , nº 4, p 289 - 293, 1986.

OVERALL, J E.; SPIEGEL, D K. Concerning least squares analysis of experimental data. **Psychologica Bulletin,** Lancaster, v. 12, nº 5, p 311 - 322, 1969

SAS USER"s GUIDE: Statistics. 6ª ed. Cary. SAS Institute 1990. 846 p

SEARLE, S.R. **Linear models for unbalanced data.** New York. John Wiley, 1987, 536 p.

SPEED, F.M.; HOCKING, R.R., Hackney, O.P. methods of analysis of linear models with unbalanced data. **Journal of the American Statistical Association,** Alexandria, v 13. p. 105 - 112, 1978.

URQUARDT, N.S.; Weeks. D.L. Linear models in messy data: Some problems and alternatives. **Biometrics,**Alexandria, v. 34, p. 696 - 705, 1978.

YATES, F. The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association,** Alexandria, v. 29, p.51-66, 1934.

[1]Recebido para publicação em 30.06.94
[2]Aceito para publicação em 10.01.95