OCORRÊNCIA DE VALORES ATÍPICOS EM EXPERIMENTOS AGRONÔMICOS

Nilva Maria Prestes de Toledo[1]

INTRODUÇÃO

Ao examinar os dados de uma determinada amostra, pode ocorrer que poucos valores constantes dessa amostra chamem a atenção por serem extremos, tendo um maior afastamento da maioria dos dados. Supondo-se que se tenha uma distribuição normal, essas observações estariam muito distantes da média, causando sérias distorções na análise e deixando o pesquisador com o problema de decidir o que fazer com elas.

Denominados valores atípicos, extremos, aberrantes ou "outliers", essas observações ocorrem com certa probabilidade de uma massa de dados, tornando os modelos probabilísticos pré-estabelecidos ineficientes.

As causas dessas ocorrências são diversas, e algumas facilmente visíveis, como erros primários na tomada de dados, que podem ser substituídos por valores corretos, ou recalculados como parcela perdida.

Mas, o verdadeiro "outlier" aparece por razões inexplicáveis, sem uma causa que nos possa parecer lógica, reflexo de uma variabilidade intrínseca dos dados. Pode conduzir o experimentador a novas idéias e descobertas, induzindo-o a pesquisar novos materiais e, muitas vezes, a chegar a novas e surpreendentes conclusões. Inúmeros exemplos de fatos como esse estão na literatura sobre o assunto.

No entanto, é importante que o pesquisador tenha sensibilidade suficiente para "sentir" quando está ocorrendo algo inusitado, não confundindo possíveis erros na condução do experimento com a ocorrência de valores anômalos.

---

[1] Instituto Agronômico, Campinas, SP
Com bolsa de suplementação do CNPq.

BARNETT & LEWIS (1979) estudaram em profundidade o problema. Relatam que as primeiras publicações sobre o assunto datam de 1755, quando BOSCOVICH tentava determinar a elipticidade da terra, tomando por base uma amostra de 10 medidas de graus polares relacionados com graus equatoriais. Apareceram 2 determinações de valores excessivos, que ele decidiu rejeitar como tendo sido um erro, tendo reduzido sua amostra para 8 observações.

Em 1777, DANIEL BERNOULLI detectou valores estranhos em observações astronômicas e questionou o procedimento a ser adotado sobre esse tipo de observação. Comentou que seria difícil estabelecer uma linha divisória entre a normalidade e a anormalidade dos dados.

Durante todo o século XIX, o problema foi exaustivamente discutido, vários testes para avaliar observações, como "outliers", foram criados e com a evolução da estatística, em nosso século, inúmeras publicações interessantes apareceram.

RIDER (1933) publicou uma revisão do problema dos testes de discordância para valores extremos até 1933. Os critérios adotados para tratar o problema apontavam como condição inicial que o desvio padrão da amostra fosse conhecido e que ela tivesse uma distribuição normal.

THOMPSON (1935) foi um dos primeiros autores a propor um teste matemático para verificar se uma observação é um valor atípico. Esse trabalho foi comentado e analisado, em 1936, por PEARSON & CHANDRA SEKAR (1936).

Em 1950, GRUBBS publicou um histórico dos métodos utilizados para detectar "outliers" a partir de 1937, época em que o problema começou a assumir maior importância.

Em nossa literatura, MENDES & CONAGIN (1960), estudando a produtividade e rendimento de duas classes de plantas em cafeeiro Mundo Novo, encontraram dois valores atípicos referentes ao rendimento dos frutos, e, após um teste estatístico específico, decidiram eliminá-los do conjunto de dados.

STEFANSKY (1975) e posteriormente JOHN (1978) estudaram a ocorrência de valores atípicos em experimentos fatoriais e as distorções provocadas na análise de variância.

DIXON (1950, 1953) analisou os tipos de erros que causam o aparecimento de valores anômalos, apresentan-

do testes de discordância para, a partir da hipótese de que o valor em questão é atípico, aceitar ou rejeitar essa hipótese. Para isso, ele criou tabelas de significância específicas para diversos níveis de probabilidade.

Técnicas exploratórias de dados, como um estudo preliminar para detectar esse tipo de problema foram recomendadas por TUKEY (1977). Ele também recomenda um estudo minucioso dos resíduos, que sempre fornecem "pistas" sobre o comportamento das variáveis.

MATERIAL E MÉTODOS

Os dados utilizados referem-se a um experimento que visava estudar o efeito do cloreto de mepiquat como regulador de crescimento em algodoeiro herbáceo, Dos dados obtidos, empregaremos os resultados do rendimento, em porcentagem, de fibras de algodão, onde o delineamento estatístico foi de 4 blocos ao acaso, com 5 tratamentos (doses crescentes de cloreto de mepiquat e uma testemunha). Esse experimento foi conduzido em 6 localidades diferentes do Estado de São Paulo, e executado pelos técnicos da Seção de Fisiologia do Instituto Agronômico.

Com permissão do autor, esses dados foram analisados para verificar o que ocorre com a análise de variância na ausência e na presença de valores atípicos. Esses valores foram introduzidos no tratamento 2, bloco II e no tratamento 4, bloco IV, em substituição aos valores verdadeiros.

No quadro I, estão os dados do experimento, onde 2' e 4' são os tratamentos contendo "outliers".

As análises de variância para os dados verdadeiros e para os que contém os 2 "outliers" são os que constam do quadro II.

As observações anômalas introduzidas artificialmente tem valores tais que não houve alteração na média geral dos dois blocos de dados.

As distorções nos desvios padrão e nos coeficientes de variação foram grandes, além de alterar a significância do teste de F, que passou de significativo a 5% para os dados verdadeiros a não significativo na presença das duas observações anômalas ($F_{4,12} = 3,26$).

Quadro I - Dados do experimento de tratamento de algodoeiros pelo cloreto de Mepiquat.

Tratamentos \ Blocos	I	II	III	I-V	Totais tratamentos	Médias
1	40,8	39,8	39,6	38,9	159,1	39,77
2	39,1	39,1	39,0	39,7	156,9	39,22
2'	39,1	49,1	39,0	39,7	166,9	41,72
3	38,7	39,0	38,4	39,5	155,6	39,90
4	39,7	37,9	37,4	39,6	154,6	38,65
4'	39,7	37,9	37,4	29,6	144,6	36,15
5	37,8	37,6	38,1	38,8	152,3	38,07
Totais blocos	196,1	193,4 203,4	192,5	196,5 186,5	778,50	38,92

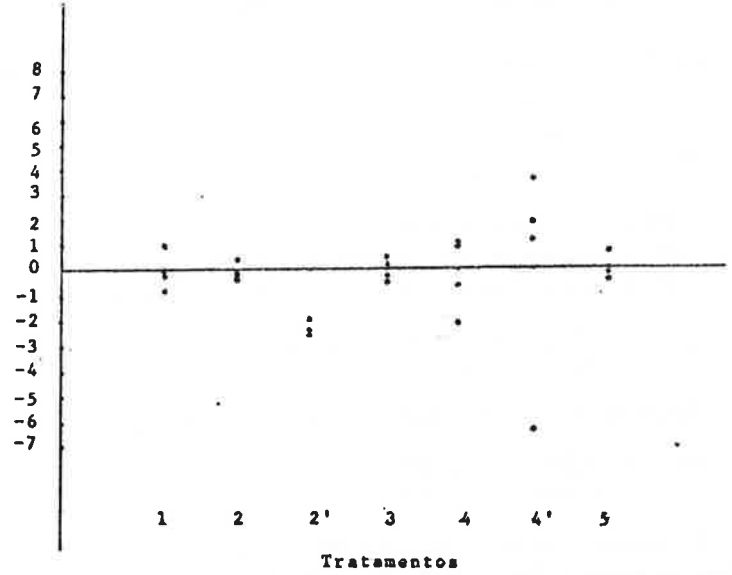
Quadro II - Análise de variância

Dados verdadeiros					Com 2 observações alteradas		
F.V.	G.L.	SQ	QM	F	SQ	QM	F
Trat.	4	6,4450	1,6112	3,56*	67,9450	16,9862	1,92ns
Blocos	3	2,3415	0,7805		29,9415	9,9805	
Res.	12	5,4310	0,4526		106,3310	8,8609	
Total	19	14,2175			204,2175		
s = 0,6727 cv = 1,73%					s = 2,9767 cv = 7,65%		

Os valores dos desvios com relação às médias por tratamento são:

Tratamento	1	1,0	0,0	-0,1	-0,9
	2	-0,1	-0,1	-0,2	0,4
	2'	-2,6	7,3	-2,7	-2,0
	3	-0,2	0,1	-0,5	0,6
	4	1,0	-0,7	-1,2	0,9
	4'	3,5	1,8	1,2	-6,5
	5	-0,2	-0,5	0,0	0,7

Um gráfico nos dá uma idéia mais clara do que ocorre com os desvios quando temos valores estranhos. No eixo das abcissas estão indicados os tratamentos, e no das ordenadas os valores dos desvios:



Os pontos +7,3 e -6,5 correspondem aos valores anômalos introduzidos. Por esse gráfico pode-se ter uma idéia clara da dispersão dos desvios causada por eles.

DIXON (1950, 1953) propôs um teste de discordância, com uma tabela de significância para diversos níveis de α, a fim de se determinar se uma observação é um "outlier" ou não. Há 2 fórmulas que podem ser usadas:

a) quando a observação atípica é o maior valor da amostra:

$$T = \frac{Xn - Xn-1}{Xn - X_1}$$

onde Xn é o maior valor da amostra (outlier)
Xn-1 é o valor abaixo dela
X_1 é o menor valor da amostra.

No caso em estudo temos:

$$T = \frac{49,1 - 40,8}{48,1 - 29,6} = 0,426$$

Para N = 20, os valores tabelados são:

$T(\alpha = 0,10) = 0,401$
$T(\alpha = 0,05) = 0,450$

Portanto, com uma probabilidade entre 10 e 5%, pode se considerar o valor 49,1 como uma observação anômala.

b) quando a observação atípica é o menor valor da amostra:

$$U = 10\,\frac{(X2 - X1)}{(Xn - X1)}$$

onde Xn é o maior valor da amostra
X_1 é o menor valor da amostra (outlier)
X_2 é o penúltimo menor valor da amostra

$$U = 10\,\frac{37,4 - 29,6}{49,1 - 29,6} = 0,400$$

Para N = 20, os valores tabelados são:

$U(\alpha = 0,20) = 0,340$
$U(\alpha = 0,10) = 0,401.$

O valor encontrado 0,400 tem o seu nível de significância entre 20 e 10% de probabilidade.

DISCUSSÃO E CONCLUSÕES GERAIS

As distorções causadas na análise de variância pela presença de dois valores atípicos atingem a significância do teste de F, os quadrados médios dos resíduos e, consequentemente, os desvios padrão e os coeficientes de variação. Se fôssemos considerar como correta a análise com as duas observações atípicas, concluiríamos que não existe diferença entre os tratamentos, portanto, estaríamos incorrendo em um erro.

Antes de iniciar uma análise, interessante seria observar cuidadosamente os dados e a qualquer suspeita de um valor anômalo, fazer um estudo prévio para verificar a causa do seu aparecimento e como manejá-lo. Muitas vezes, excluir esse valor é a solução, mas isso só deve ser feito após cuidadoso exame.

Há, ainda, outros testes específicos, além do apresentado aqui, para estabelecer se o valor em questão é ou não um "outlier". Após um estudo assim mais acurado, decide-se o que fazer com o valor discordante.

A substituição dos modelos geralmente utilizados por outros, como inferência robusta, muitas vezes é aconselhável.

SUMMARY

This paper shows how outliers may distort the variance analysis, and how it is possible to detect them by an appropriate statistical test, as proposed by DIXON. Consequences caused by extreme values my be analysed in a plot of deviations of the treatment means.

Statistical design was randomized blocks, and 5 treatments (increasing doses of mepiquat chloride and an untreated check). Fiber rates were the data analysed in this paper (herbaceous cotton).

LITERATURA CITADA

BARNETT, V. & T. LEWIS, 1979. **Outliers in statistical data**, John Willey & Sons.

DIXON, W.J., 1950. Analysis of extreme values. **Annals of Math. Stat.**, vol. 21.

DIXON, W.J., 1953. Processing data for outliers. **Biometrics**, vol. 9.

GRUBBS, F.E., 1950. Sample criteria for testing outlying observations. **Annals of Math. Stat.**, vol. 21.

JOHN, J.A., 1978. Outliers in factorial experiments. **Applied Statistics**, vol. 27.

MENDES, A.J.T. & A. CONAGIN, 1960. Produtividade e rendimento de duas classes de plantas existentes no café "Mundo Novo". **Bragantia**, vol. 2.

PEARSON, E.S. & C. CHANDRA SEKAR, 1933. The efficiency of statistical tools and a criterion for the rejection of outlying observations. **Biometrika**, vol.28.

RIDER, P.R., 1933. **Criteria for rejection of observations**. Washington Univ. Studies, New Series, Science and Technology, 8.

SNEDECOR, G.W. & W.G. COCHRAN, 1980. **Statistical Methods**, The Iowa State University Press, Ames, Iowa.

STEFANSKY, W., 1975. Rejecting outliers in factorial designs. **Technometrichs**, vol. 17.

THOMPSON, W.R., 1935. On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. **Annals of Math. Stat.**, vol. 6.

TUKEY, J.W., 1977. **Exploratory data analysis.**