BRAGANTIA

Boletim Técnico do Instituto Agronômico do Estado de São Paulo

Vol. 14 Campinas, março de 1955 N.º 15

TRANSFORMAÇÕES DOS DADOS EXPERIMENTAIS (*)

ARMANDO CONAGIN

Engenheiro agrônomo, Seção de Técnica Experimental, Instituto Agronômico

RESUMO

Neste artigo apresentamos a análise estatística de um experimento de sistema — época de plantio de milho, fatorial 2x4, em que a heterogeneidade das variâncias dos tratamentos foi eliminada pelo uso da transformação logarítmica dos dados de produção. Os dados originais representam a produção, em pêso (quilos), de grãos, por canteiro. A transformação logarítmica trouxe a necessária independência entre as médias dos tratamentos e os desvios padrão respectivos, assegurando, dessa forma, a validez da análise da variância.

INTRODUÇÃO

A "análise da variância" é um processo estatístico que foi desenvolvido por R. A. Fisher a fim de fornecer um método de análise e interpretação dos dados obtidos em experimentação de campo e de laboratório.

Para que os experimentos nos conduzam a inferências válidas são necessárias as seguintes pressuposições na análise da variância (2):

a) **variáveis alcatórias** — os valores $\mathbf{x}_{ij}$ (resultados numéricos obtidos no experimento) são valores observados de variáveis aleatórias distribuídas ao redor da média verdadeira, $\mathbf{m}_{ij}$;

b) **aditividade** — o parâmetro $\mathbf{m}_{ij}$ é composto de três componentes (modêlo matemático da análise da variância, apropriado para planos do tipo blocos ao acaso) $\mathbf{m}_{..}$, $(\mathbf{m}_{i.} - \mathbf{m}_{..})$ e $(\mathbf{m}_{.j} - \mathbf{m}_{..})$ que são aditivos e representam, respectivamente, a média da população, o efeito de tratamentos e o efeito de blocos; temos:

$$x_{ij} = m_{ij} + \varepsilon_{ij} \qquad \text{onde}$$
$$m_{ij} = m_{..} + (m_{i.} - m_{..}) + (m_{.j} - m_{..})$$

c) **variâncias iguais e correlação nula** — as variáveis $\mathbf{x}_{ij}$ são homocedásticas e independentes, isto é, têm uma variância comum σ^2, a covariância entre elas é nula, como também o é a covariância entre essa variável e as suas componentes;

(*) Recebido para publicação em 19 de janeiro de 1955.

d) **normalidade** — as variáveis x_{ij} distribuem-se conjuntamente em uma distribuição normal multidimensional.

A distribuição normal caracteriza-se por apresentar dois parâmetros: **m** (média) e σ (desvio padrão), os quais são independentes.

Os experimentos efetuados em agricultura são de competição de variedades ou tratamentos e instalados visando o julgamento dos tratamentos, pela produtividade.

A hipótese posta em prova pelo experimento é a de nulidade, a qual supõe a igualdade das produções.

$H_0 \therefore m_1 = m_2 = ... = m_k = m$

no caso de k tratamentos.

Antes de pôr em prova essa hipótese, necessitamos admitir a igualdade das variâncias dos tratamentos (homogeneidade dessas variâncias). Nem sempre, porém, essas variâncias podem ser tomadas como homogêneas. Dentre êstes casos, alguns evidenciam uma relação constante entre as médias e os desvios padrão; outras vêzes obtemos amostras que nos dão a idéia de serem marcadamente assimétricas as distribuições das quais provieram. Em casos como êsses, o uso de certas transformações apropriadas, onde a escala de medida dos valores é mudada, conduz-nos a processos de análise da variância perfeitamente válidos, pois as condições antes enumeradas passam a ser satisfeitas.

Bartlett (1) enumerou alguns ítens que a nova variável, obtida a partir da primitiva por uma transformação, deve apresentar:

a') a variância da variável transformada não deve ser afetada pelas mudanças no valor da média (fica assegurada, portanto, a independência entre a média e a variância);

b') a variável transformada deve ser normalmente distribuida;

c') a média aritmética, na nova escala, deve ser uma estimativa eficiente do valor real da média para qualquer grupo particular de medidas;

d') a transformação deve assegurar a linearidade e a aditividade dos componentes.

Se bem que essas quatro condições sejam relacionadas até certo grau (por exemplo, asseguradas as condições a', b' e d' na nova transformação, c' também fica implicitamente aí contida), isso não quer dizer que a satisfação de a' ùnicamente, nos conduza à obtenção implícita de b', c' e d'.

A transformação da variável, visando obter a condição a' apresenta, em muitos casos, a vantagem de aproximar os dados da normalidade; também pode ser eliminada por essa transformação uma certa correlação entre a variabilidade e o valor da média na escala original e ainda a assimetria excessiva encontrada.

Requer-se a condição c' porque as estimativas efetuadas na análise da variância dizem respeito às médias aritméticas dos tratamentos e nós desejamos que essas estimativas sejam eficientes.

Essas e outras razões levaram os estatísticos matemáticos a recomendar o uso de transformações, mas só naqueles casos em que são justificáveis pelas considerações já feitas.

2-TRANSFORMAÇÃO LOGARÍTMICA DOS DADOS

Em certas circunstâncias os dados parecem evidenciar uma relação linear estreita entre as médias e os desvios padrão.

É fácil provarmos que uma transformação logarítimica desses dados é a espécie de transformação que conduz à satisfação dos itens já mencionados.

Suponhamos que a variável **x** tenha uma variância dependente da média :

$$\sigma_x^2 = f(m)$$

onde **x** são os valores dos dados na escala original, **m** é a média da população e σ_x^2 a variância dos valores medidos.

Se fizermos uma transformação **y = g(x)** teremos agora

$$\sigma_y^2 = f(m)\left[\frac{dy}{dm}\right]^2$$

Se quizermos tornar σ_y^2 constante, isto é, C^2, deveremos fazer :

$$\frac{dy}{dm} = \frac{C}{\sqrt{f(m)}}$$

donde

$$y = \int \frac{C}{\sqrt{f(m)}}\, dm$$

Se

$\sigma_x = K.m$, isto é, $\sigma_x/m = K$ (constante)

então :

$$y = \int \frac{C}{\sqrt{K^2m^2}}\, dm = \log f_1(m) + K_1$$

isto é, a transformação logarítmica é a mais apropriada. Com essa transformação a variância da variável transformada passa a ser constante e independente da média **m**. Se os dados são tais que a transformação garante os itens enumerados de forma satisfatória, a inferência efetuada através da análise da variância dos dados transformados é perfeitamente válida.

3-ANÁLISE DE UM EXPERIMENTO

Vamos dar, com detalhes, um exemplo por nós analisado[1] em que a transformação logarítmica foi eficiente.

O quadro 1 contém os dados de produção em pêso de grãos de um Ensaio de sistema — época de plantío de milho, fatorial do tipo 2 x 4 onde o sistema de plantío consistiu em plantar-se em covas ou linhas.

(1) Dados de um Ensaio de época-modo de plantío de milho, realizado na Estação de Cereais e Leguminosas de Serrinha, Estado da Bahia, analisados pela Seção de Técnica Experimental.

QUADRO 1.—Ensaio de época de plantio de milho. Dados de produção em pêso de grãos por canteiro

Tratamento	Blocos					Total	Médias	Desvio Padrão
	1.°	2.°	3.°	4.°	5.°			
	kg	*kg*	*kg*	*kg*	*kg*	*kg*	*kg*	
D_1	6,50	7,10	8,50	4,25	1,85	28,20	5,64	2,61
D_2	15,40	21,70	14,65	10,30	8,00	70,05	14,01	5,28
C_1	5,60	5,90	4,55	6,30	2,75	25,10	5,02	1,42
C_2	10,50	12,20	12,90	13,95	5,70	55,25	11,05	3,24
B_1	3,30	2,75	3,35	1,75	2,55	13,70	2,74	0,65
B_2	3,50	4,50	5,80	5,05	5,05	23,90	4,78	0,85
A_1	0,70	1,10	0,75	0,25	0,45	3,25	0,65	0,32
A_2	0,70	0,80	1,55	1,65	1,25	5,95	1,19	0,43

Uma representação gráfica dos valores desvios padrão — médias de tratamentos, evidenciou uma relação funcional do tipo linear entre as médias e o desvio padrão.

A heterogeneidade das variâncias é devida a essa relação. O item c das condições necessárias e suficientes para a validez da análise da variância não foi encontrado. Precisamos fazer uma transformação logarítimica dos dados, que são os apresentados no quadro 2.

QUADRO 2.—Ensaio de época de plantio de milho. Transformação logarítmica dos dados de produção

Tratamento	Blocos					Totais
	1.°	2.°	3.°	4.°	5.°	
D_1	0,81291	0,85126	0,92941	0,62839	0,26717	3,48915
D_2	1,18752	1,33646	1,16584	1,01284	0,90309	5,60575
C_1	0,74819	0,77085	0,65801	0,79934	0,43933	3,41572
C_2	1,02119	1,08636	1,11059	1,14457	0,75585	5,11856
B_1	0,51851	0,43933	0,52504	0,24304	0,40654	2,13246
B_2	0,54407	0,65321	0,76343	0,70329	0,70329	3,36729
A_1	—0,15490	0,04139	—0,12494	—0,60206	—0,34679	—1,18730
A_2	—0,15490	—0,09691	0,19033	0,21748	0,09691	0,25291
Totais	4,52259	5,08195	5,21772	4,14689	3,22539	22,19454

Efetuamos, a seguir, a análise da variância cujos resultados estão contidos no quadro 3.

QUADRO 3.—Ensaio de época de plantio de milho. Análise da variância dos dados transformados

F. V.	S. Q.	G. L.	Q. M.	F
Total	8,461212	39	0,21695	
Tratamentos	7,450075	7	1,06430	43,22**
Blocos	0,323119	4	0,08078	3,29*
Êrro	0,688018	28	0,02457	

Se quizéssemos calcular as diferenças mínimas significativas entre dois dos oito tratamentos, faríamos :

$$\bar{x}_i - x_j \geq t \sqrt{\frac{2(0{,}02457)}{5}}$$

$$\bar{x}_i - \bar{x}_j \geq 2{,}05\ (0{,}09914)$$

Duas médias de tratamentos diferindo por mais de 0,2032 serão estatìsticamente diferentes.

O ensaio sendo fatorial 2 x 4 permite a decomposição da variação total com 7 graus de liberdade entre tratamentos, em 7 partes com 1 grau de liberdade cada (3).

Para facilitar a decomposição ortogonal, organizamos o quadro 4.

QUADRO 4.—Ensaio de época de plantio de milho. Quadro da decomposição ortogona dos tratamentos

	D–1.ª época	C–2.ª época	B–3.ª época	A–4.ª época	Total
1 — Cova	3,48915	3,41572	2,13246	—1,18730	7,85003
2 — Linha	5,60575	5,11856	3,36729	0,25291	14,34451
Total	9,09490	8,53428	5,49975	—0,93439	22,19454

Os 7 graus de liberdade vão ser decompostos assim :

$$\text{tratamentos (g.l. = 7)} \left[\begin{array}{l} \text{Épocas (g.l. = 3)} \\ \text{Plantio (g.l. = 1)} \\ \text{Interação época-plantio (g.l. = 3)} \end{array} \right.$$

Entre épocas

a) **efeito linear** (com 1 grau de liberdade)

$$3(D_1 + D_2) + (C_1 + C_2) - (B_1 + B_2) - 3(A_1 + A_2)$$

$$a = \frac{(33{,}12240)^2}{20 \times 10} = 5{,}48547 \qquad F = \frac{5{,}48547}{0{,}02457} = 223{,}26^{**}$$

b) **efeito quadrático**

$$(D_1 + D_2) - (C_1 + C_2) - (B_1 + B_2) + (A_1 + A_2)$$

$$b) = \frac{(-5{,}87352)^2}{40} = 0{,}86246 \qquad F = \frac{0{,}86246}{0{,}02457} = 35{,}10^{**}$$

c) **efeito cúbico**

$$(D_1 + D_2) - 3(C_1 + C_2) + 3(B_1 + B_2) - (A_1 + A_2)$$

$$c) = \frac{(-0{,}94308)^2}{200} = 0{,}00445 \qquad F = \frac{0{,}00445}{0{,}02457} = 0{,}18$$

Entre modos de plantio

$$\frac{(\text{linha} - \text{cova})^2}{2 \times 20} = \frac{(6{,}49448)^2}{40} = 1{,}05446 \qquad F = 42{,}92^{**}$$

Cálculo das interações

$$a)\ 3(D_1 - D_2) + (C_1 - C_2) - (B_1 - B_2) - 3(A_1 - A_2)$$
$$b)\ (D_1 - D_2) - (C_1 - C_2) - (B_1 - B_2) + (A_1 - A_2)$$
$$c)\ (D_1 - D_2) - 3(C_1 - C_2) + 3(B_1 - B_2) - (A_1 - A_2)$$

$$a) = \frac{(2{,}49718)^2}{200} = 0{,}03118 \qquad F = 1{,}27$$

$$b) = \frac{(0{,}61914)^2}{40} = 0{,}00958 \qquad F = 0{,}39$$

$$c) = \frac{(-0{,}72764)^2}{200} = 0{,}00265 \qquad F = 0{,}11$$

4-CONCLUSÕES

1. A transformação logarítmica foi apropriada, pois assegurou a independência entre as médias de tratamentos e os desvios padrão respectivos, e também a homogeneidade das variâncias. A validez da análise da variância ficou, destarte, assegurada.

2. Como resultado da análise efetuada concluiu-se haver um grande decréscimo de produção das duas primeiras para as duas últimas épocas.

3. O sistema de plantio em linhas foi grandemente superior ao sistema de plantio em covas.

4. As interações foram não significativas.

TRANSFORMATIONS OF EXPERIMENTAL DATA

SUMMARY

The conditions required for a valid analysis of variance are enumerated. A 2x4 factorial experiment with corn in randomized blocks, designed to study data of sowing and types of planting for corn, is discussed. The experiment was performed at the Estação Experimental de Serrinha, Estado da Bahia.

The original data showed that the variances within treatments were not homogenous. It was found by means of a graphical presentation that the relation — standard deviation / mean of treatment — was almost constant for the different treatments.

A log transformation brought the desirable homogeneity to the variances ; the usual analysis of variance was then carried out.

LITERATURA CITADA

1. **BARTLETT, M. S.** The use of transformations. Biometrics 3(1) : 39-52. 1947.
2. **EISENHART, C.** The assumptions underlying the analysis of variance. Biometrics 3(1) : 1-21. 1947.
3. **GOULDEN, C. H.** Methods of statistical analysis. New York, John Wiley & Sons Inc., 1939. p. 160-169.