ANÁLISE ESTATÍSTICA DO ENSAIO DE VARIEDADES DE CAFÉ (1)

W. L. STEVENS

Professor da Faculdade de Ciências Econômicas e Administrativas da Universidade de São Paulo

1—INTRODUÇÃO

Êste trabalho descreve a análise estatística de um ensaio de variedades que apresenta dois caraterísticos excepcionais :

a) a planta (café) é de uma espécie que apresenta valores máximos de colheita de dois em dois anos ;

b) o arranjo do ensaio no campo é sistemático.

As alternações de grande e pequena produção são especialmente notáveis no café e, quando não tomadas em conta na análise estatística, são capazes de esconder ou prejudicar as outras comparações estudadas no experimento. Pode-se dizer que a planta fica esgotada depois de uma produção elevada e mostra, porisso, tendência para produzir pouco no ano seguinte. Se as condições climáticas fôssem uniformes, tôdas as plantas poderiam ser classificadas permanentemente em dois tipos : as que produzem bem nos anos pares e as que produzem bem nos anos ímpares. As variações climáticas podem, contudo, quebrar essa regularidade. Por exemplo, o tempo desfavorável em dois anos seguidos pode diminuir as colheitas nesses anos, provocando, assim, uma grande colheita no terceiro ano em tôdas as plantas. O efeito das variações de clima é, porisso, o de sincronizar as plantas da mesma região, de modo que tôdas mostrem os máximos de produção nos mesmos anos, ou nos anos pares ou nos anos ímpares. É, todavia, muito comum descobrir, até no mesmo cafèzal, algumas plantas que apresentam os máximos de produção nos anos em que a maioria das plantas apresenta os mínimos (1). É claro que não podemos estudar o fenômeno de oscilação de colheita a partir dos dados de um experimento isolado ; na análise do experimento atual estamos mais interessados na eliminação dos efeitos dêste fenômeno do que no seu estudo.

As razões contra o emprêgo de delineamentos sistemáticos são, hoje, bastante conhecidas. Lembramos, a propósito, que êste experimento foi iniciado em 1933, em Campinas, quando os princípios de casualização não eram tão

(1) Esta análise refere-se ao ensaio em estudo pela Secção de Café do Instituto Agronômico de Campinas, discutido no artigo anterior desta revista.

bem compreendidos como hoje o são. Além disso, o estatístico moderno, treinado nos métodos de delineamento casualizado (*randomised design*) condena com demasiada facilidade o delineamento sistemático. Esperamos mostrar neste artigo que, empregando os métodos de análise apropriados, as conclusões tiradas de um experimento sistemático são capazes de atingir, senão o rigor lógico das conclusões baseadas num delineamento moderno, pelo menos uma objetividade suficiente para satisfazer o homem prático.

2—EXPERIMENTO E OS DADOS

Mendes (3) apresenta uma descrição completa do experimento. Para a compreensão da análise estatística, é suficiente notar aqui que estamos comparando seis variedades designadas por A, B, C, D, E, F, de acôrdo com a relação abaixo

A	Nacional	D	Bourbon Amarelo
B	Amarelo Botucatu	E	Sumatra
C	Bourbon	F	Maragogipe

Estas variedades foram colocadas em trinta fileiras de 50 plantas cada uma, pela ordem sistemática :

A B C D E F A B C D E F A B C D E F A B C D E F A B C D E F.

Cada fileira é colhida independentemente e constitui, portanto, um canteiro ou talhão (*plot*), na terminologia genérica dos ensaios de campo. Há cinco repetições de cada variedade. Os resultados nos doze anos, de 1935-1946, são apresentados no quadro 1. Notamos que do terceiro ano em diante (1937) os resultados são notàvelmente regulares (fig. 1). Sendo pequenas as colheitas dos primeiros dois anos, não hesitamos em pôr de lado os dados referentes a êsses anos, baseando as nossas conclusões inteiramente nos resultados dos dez anos, de 1937-1946. Neste período, quase tôdas as fileiras manifestam os máximos de produção nos anos pares. A fileira 4 é, contudo, uma exceção evidente. Desconhecemos as colheitas das plantas individuais, mas é evidente que a maioria das plantas, embora não tôdas, segue o mesmo ciclo de produção.

3—ANÁLISE ESTATÍSTICA

3.1—CONSIDERAÇÕES GERAIS

A quantidade primária que nos interessa é, naturalmente, a colheita total ou a colheita média. Vemos, logo, que êsse total tem que ser baseado num número *par* de anos. Se fôsse obtido a partir de um número ímpar de anos, começando e terminando, digamos, com um ano ímpar, as plantas

QUADRO 1.—Produção, em quilos, de café em côco. Colheita total de fileiras de 50 plantas dos diversos anos do ensaio.

Variedade	Número da fileira	Ano 1935	Ano 1936	Ano 1937	Ano 1938	Ano 1939	Ano 1940	Ano 1941	Ano 1942	Ano 1943	Ano 1944	Ano 1945	Ano 1946	Produção total
A Nacional	1	37,4	33,8	63,3	85,2	43,9	110,8	13,2	18,3	80,8	53,2	24,9	108,1	672,9
	7	40,2	35,4	75,6	83,7	56,1	101,2	25,2	46,9	51,2	109,6	69,1	92,9	787,1
	13	32,5	50,2	68,3	141,5	32,9	130,1	24,9	80,0	54,1	155,4	68,8	108,1	946,8
	19	34,6	43,0	75,7	177,0	36,3	181,8	21,0	98,7	63,7	167,8	82,1	154,3	1136,0
	25	38,4	51,9	69,0	164,4	25,0	167,0	13,2	88,7	57,7	166,8	71,9	148,6	1062,6
	Total	183,1	214,3	351,9	651,8	194,2	690,9	97,5	332,6	307,5	652,8	316,8	612,0	4605,4
B Amarelo Botucatu	2	38,6	42,6	69,8	95,1	71,5	85,8	27,0	27,5	87,5	76,2	33,5	117,3	772,4
	8	36,2	44,4	74,3	95,3	62,6	92,2	34,1	69,3	38,7	142,6	58,0	114,3	862,0
	14	30,8	44,5	69,9	139,9	44,2	128,3	29,8	87,8	39,9	149,5	55,7	109,6	929,9
	20	26,6	42,8	61,6	161,8	20,7	146,4	12,4	102,5	21,9	170,3	53,4	125,8	946,2
	26	32,8	48,2	63,3	186,0	18,7	171,2	7,3	104,0	38,2	172,4	62,0	128,6	1032,7
	Total	165,0	222,5	338,9	678,1	217,7	623,9	110,6	391,1	226,2	711,0	262,6	595,6	4543,2
C Bourbon	3	45,0	51,3	110,0	96,1	121,7	97,4	42,4	47,8	97,1	116,6	79,2	152,8	1057,4
	9	46,4	56,4	114,0	130,2	128,0	132,2	67,9	104,1	72,0	203,4	123,1	168,0	1345,7
	15	33,9	64,9	124,4	170,3	129,6	185,1	68,2	125,1	89,5	195,0	130,2	166,3	1482,5
	21	35,6	56,4	109,1	203,6	76,4	203,6	28,7	146,2	52,5	209,1	99,9	184,8	1405,9
	27	35,0	55,1	108,8	241,5	57,0	233,7	29,8	126,9	89,8	216,1	103,9	216,3	1513,9
	Total	195,9	284,1	566,3	841,7	512,7	852,0	237,0	550,1	400,9	940,2	536,3	888,2	6805,4
D Bourbon Amarelo	4	46,7	58,2	106,6	104,2	113,9	116,2	66,2	47,9	137,8	99,8	137,0	181,8	1216,3
	10	43,4	59,4	98,1	109,5	117,2	125,8	81,2	91,0	97,1	201,5	123,2	191,0	1338,4
	16	34,2	66,8	106,5	182,8	70,1	201,9	48,6	142,5	95,1	224,1	125,9	239,0	1537,5
	22	29,5	69,5	99,7	218,2	56,7	213,9	33,8	144,8	58,4	257,8	83,5	251,0	1516,8
	28	42,1	76,4	112,8	252,6	45,3	245,0	30,1	148,4	99,5	218,7	103,1	268,2	1642,2
	Total	195,9	330,3	523,7	867,3	403,2	902,8	259,9	574,6	487,9	1001,9	572,7	1131,0	7251,2
E Sumatra	5	42,4	46,9	82,3	116,2	64,8	126,8	29,4	56,4	81,1	125,5	93,6	148,9	1014,3
	11	47,1	52,3	86,9	140,7	70,0	145,9	42,6	97,9	84,5	177,8	110,9	162,3	1218,9
	17	31,8	53,5	83,5	164,3	51,5	176,5	25,2	114,1	60,0	197,2	92,6	166,8	1217,0
	23	35,3	49,8	75,0	182,5	24,6	170,0	14,3	107,1	55,6	164,2	87,1	161,3	1126,8
	29	37,8	64,3	83,9	211,1	28,3	208,4	13,0	106,1	106,4	134,7	95,6	165,1	1254,7
	Total	194,4	266,8	411,6	814,8	239,2	827,6	124,5	481,6	387,6	799,4	479,8	804,4	5831,7
F Maragogipe	6	6,1	18,1	28,3	82,9	64,1	111,4	30,7	44,1	53,9	67,3	41,2	151,8	699,9
	12	6,0	19,5	36,9	97,8	63,3	121,9	37,1	70,4	53,6	105,0	52,3	183,0	846,8
	18	5,5	18,5	31,9	141,0	64,4	156,1	37,8	73,2	71,1	93,7	57,5	208,2	958,9
	24	4,9	17,6	22,1	146,5	55,2	155,2	29,2	62,7	79,9	79,2	42,0	195,0	889,5
	30	4,8	17,0	21,5	148,5	56,4	155,5	20,4	67,6	108,0	72,5	34,3	202,0	908,5
	Total	27,3	90,7	140,7	616,7	303,4	700,1	155,2	318,0	366,5	417,7	227,3	940,0	4303,6
Total geral		961,6	1408,7	2333,1	4470,4	1870,4	4597,3	984,7	2648,0	2176,6	4523,0	2395,5	4971,2	33340,5

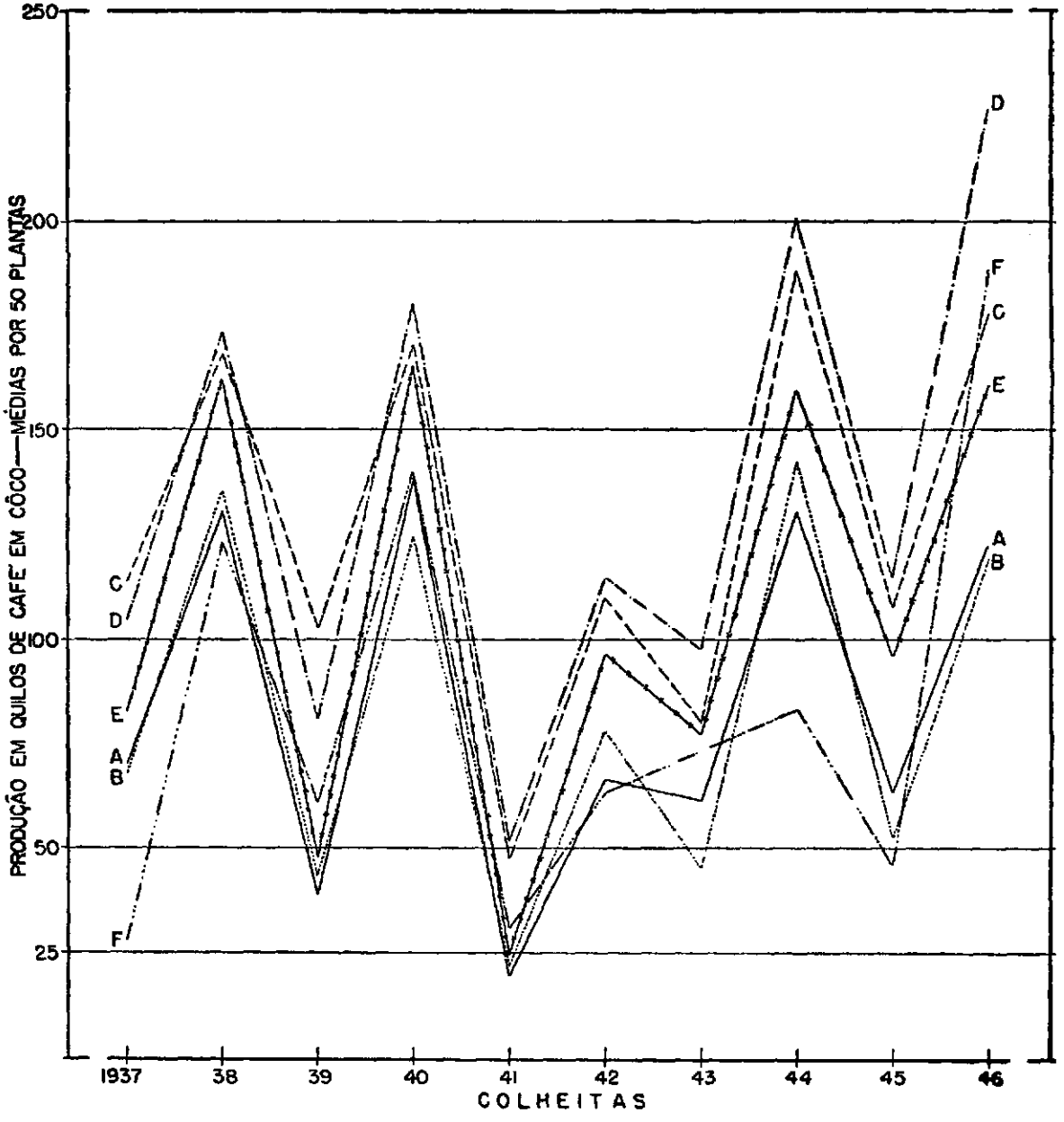

FIGURA 1.—Produção, em quilos, de café em côco, médias de 50 plantas das variedades do ensaio, nas colheitas de 1937 a 1946. Vairedades: *A* — Nacional, *B* — Amarelo Botucatu, *C* — Bourbon, *D* — Bourbon amarelo, *E* — Sumatra, *F* — Maragogipe.

que produzem bem nos anos ímpares seriam favorecidas injustamente em comparação com as que produzem bem nos anos pares. A escolha de um período de dez anos é, portanto, correta. (Daí por diante a análise deveria ser refeita de dois em dois anos).

A magnitude de oscilação é medida convenientemente por :

Total de anos pares menos total de anos ímpares.

Pretendemos, a seguir, construir uma medida de "tendência" (*trend*), apropriada para distinguir as variedades cujas taxas de incremento de produção são superiores às das outras. A medida mais evidente é a função linear das colheitas dos dez anos com coeficientes :

$$-9 \quad -7 \quad -5 \quad -3 \quad -1 \quad +1 \quad +3 \quad +5 \quad +7 \quad +9.$$

(Esta função pode ser dividida por 330, para obtermos o coeficiente de regressão linear).

Uma tal função não é, contudo, independente do fenômeno oscilação. As plantas que produzem bem nos anos pares seriam favorecidas em comparação com as que produzem bem nos anos ímpares. Compreende-se isso fàcilmente, calculando-se os valores da função para as duas séries de colheitas:

a) 2 1 2 1 2 1 2 1 2 1

e

b) 1 2 1 2 1 2 1 2 1 2

Os valores da função proposta são, respectivamente, —5 e +5.

Mais formalmente devemos dizer que esta função não é *ortogonal* com a função escolhida para medir a oscilação de produção. Não há, contudo, nenhuma dificuldade em construir uma função que, sendo uma medida de tendência, seja, ao mesmo tempo, ortogonal tanto com o total como com a medida da oscilação. Os coeficientes desta função são :

$$-2 \quad -2 \quad -1 \quad -1 \quad 0 \quad 0 \quad +1 \quad +1 \quad +2 \quad +2.$$

Podem ser construídas outras funções lineares ortogonais até se obter um grupo completo de dez. A seguinte, por exemplo, é apropriada para revelar a falta de linearidade do gráfico de colheita de ano em ano :

$$+2 \quad +2 \quad -1 \quad -1 \quad +1 \quad +1 \quad -1 \quad -1 \quad +2 \quad +2.$$

(Verificamos que esta é ortogonal com as três funções lineares já construídas).

Um exame da figura 1 sugere-nos, contudo, ser improvável que funções do segundo grau em diante nos forneçam informações de valor prático. Restringir-nos-emos, porisso, às três funções lineares ortogonais dadas no quadro 2.

QUADRO 2.—Funções Lineares Ortogonais

Símbolo	Descrição	Coeficientes									
T	Total	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
S	Oscilação	—1	+1	—1	+1	—1	+1	—1	+1	—1	+1
R	Regressão Linear	—2	—2	—1	—1	0	0	+1	+1	+2	+2

O problema de eliminar, quanto possível, os efeitos da heterogeneidade do terreno tem, formalmente, uma solução relativamente simples. Tudo que temos a fazer é considerar o número da fileira (1 . . . . 30) como variável independente e eliminar a contribuição desta variável por meio de uma análise de covariância. Na prática, contudo, o emprêgo dos polimônios

ortogonais reduz muito o trabalho de cálculo. É verdade que a necessidade de eliminar as diferenças entre variedades, antes de calcular os coeficientes de regressão destrói, em parte, as relações de ortogonalidade, mas algumas dessas relações subsistem e permitem alguma simplificação da análise.

3.2—COMPARAÇÃO DOS TOTAIS

O esquema da análise dos totais é apresentado no quadro 3. Calculamos as diferenças e as somas de pares de fileiras que são equidistantes do centro :

fileira (16) ± fileira (15) = 1436 ± 1384 = 52 (diferença) e 2820 (soma).

fileira (17) ± fileira (14) = etc.

Os coeficientes dos polinômios ortogonais, obtidos de Fisher e Yates, Tabela XXIII, (2), são associados : os de grau ímpar com as diferenças e os de grau par com as somas. É conveniente dividir os coeficientes de 3.°, 4.° e 5.° graus, respectivamente, por 10, 100 e 1000. Designando o total de uma fileira por y, temos :

$S(y\xi_1) = (52)\,(1) + (278)\,(2) + \ldots\ldots\ldots\ldots + (284)\,(29) \ldots\ldots = 58\,571$

$S(y\xi_3) = (2\,820)\,(-112) + (1\,986)\,(-109) + \ldots\ldots + (1\,488)\,(203) = -159\,077$

$S(y\xi_5) =$ etc.

As somas dos quadrados de ξ_1, etc., são dadas em Fisher e Yates (2). Temos, agora, que subtrair as somas de quadrados e produtos *entre* variedades. Os subtotais por variedade já são dados na primeira tabela do quadro 3. Designamos subtotais de ξ_1, etc., por Ξ_1, etc. Temos, por exemplo :

$\Xi_1 = -29 - 17 - 5 + 7 + 19 \qquad = -25 \quad = -(5 \times 5)$ etc.

$\Xi_3 = -182{,}7 + 88{,}4 + 53{,}5 - 71{,}4 - 70{,}3 = -182{,}5 = -(5 \times 36{,}5)$ etc.

Convém dividir êstes subtotais por cinco. Sendo simétricos (com mudança de sinal no caso dos graus ímpares) podemos trabalhar com as diferenças e as somas dos subtotais das variedades, como no caso das observações individuais. Assim :

$$6725 \pm 6326 = 399 \text{ (diferença) e } 13051 \text{ (soma).}$$

Então, fàcilmente se obtêm as somas dos produtos entre variedades. Por exemplo :

$$S(Y\Xi_3)/5 = (399)\,(6{,}8) + (1\,217)\,(20{,}9) + (-23)\,(36{,}5) = 27\,309{,}0$$

Subtraindo êsse valor de $S(y\xi_3)$, obtemos a soma de produtos *dentro de variedades*,

$$29\,383{,}3 - 27\,309{,}0 = 2\,074{,}3 \text{ etc.}$$

Calculamos, semelhantemente, as somas dos quadrados e produtos dos subtotais dos coeficientes, por exemplo :

$$S(\Xi_3^2)/5 = 2 \times 5[(6,8)^2 + (20,9)^2 + (36,5)^2] = 18\,153,00$$

Subtraindo de $S(\xi_3{}^2)$, que é dado em Fisher e Yates (2), obtemos a soma de quadrados *dentro* de variedades :

$$213\,602,40 - 18\,153,00 = 195\,449,40.$$

Observe-se que uma soma dos produtos dos subtotais de coeficientes de ordem par e ímpar, respectivamente, é nula. Por conseguinte, qualquer polinômio par é ortogonal com qualquer polinômio ímpar, mesmo depois da eliminação das diferenças *entre* variedades. É por essa razão que as equações normais se dividem em dois grupos, fornecendo o primeiro grupo os coeficientes de regressão de ordem ímpar e o segundo os coeficientes de regressão de ordem par. Dentro de cada grupo, a construção da análise de variância segue o método usual para regressão com duas ou mais variáveis independentes (4).

Quadro 3

TOTAIS DOS DEZ ANOS (*TOTALS OF THE TEN YEARS*), 1937–1946

Variedade	Fileiras (*Rows*) 1–6	7–12	13–18	19–24	25–30	Total	Média
A	602	712	864	1058	972	4208	841,6
B	691	781	854	877	952	4155	831,0
C	961	1243	1384	1314	1424	6326	1265,2
D	1111	1236	1436	1418	1524	6725	1345,0
E	925	1120	1132	1042	1153	5372	1074,4
F	676	821	935	867	886	4185	837,0
Total	4966	5913	6605	6576	6911	30971	1032,4

	Graus de liberdade (*degrees of freedom*)	Somas dos quadrados (*sums of squares*)
Entre variedades (*between varieties*)	5	1 344 132
Dentro de variedades (*within varieties*)	24	498 983
Total	29	1 843 115

(QUADRO 3—continuação)

Cálculo dos polinômios ortogonais (*Computation of the Orthogonal Polynomials*).

Diferenças	ξ_1	ξ_3	ξ_5	Somas	ξ_2	ξ_4
52	1	—11,2	+ 17,68	2820	—112	+123,76
278	3	—33,1	+ 50,83	1986	—109	+112,7
71	5	—53,5	+ 77,53	1799	—103	+ 91,31
237	7	— 71,4	+ 94,08	1879	— 94	+ 60,96
—243	9	— 85,8	+ 97,68	1997	— 82	+ 23,76
78	11	— 95,7	+ 86,79	2550	— 67	— 17,49
175	13	—100,1	+ 61,49	2661	— 49	— 59,29
261	15	— 98,0	+ 23,84	1823	— 28	— 97,44
155	17	— 88,4	— 21,76	1579	— 4	—127,04
296	19	— 70,3	— 68,21	1648	+ 23	—142,49
27	21	— 42,7	—105,35	1877	+ 53	—137,49
313	23	— 4,6	—119,60	2535	+ 86	—105,04
563	25	+ 45,0	— 93,60	2485	+122	— 37,44
462	27	+107,1	— 5,85	1844	+161	+ 73,71
284	29	+182,7	+169,65	1488	+203	+237,51

$S(y\ \xi_1) = +58\ 571$

$S(y\ \xi_3) = +29\ 383{,}3$

$S(y\ \xi_5) = -28\ 193{,}05$

$S(y\ \xi_2) = -159\ 077$

$S(y\ \xi_4) = -\ 44\ 365{,}85$

Subtotais por variedade (*Varietal subtotals*).

Diferenças	$\frac{\Xi_1}{5}$	$\frac{\Xi_3}{5}$	$\frac{\Xi_5}{5}$	Somas	$\frac{\Xi_2}{5}$	$\frac{\Xi_4}{5}$
+ 399	1	+ 6,8	+ 3,676	13051	—4	—19,10
+1217	3	+20,9	+15,298	9527	—1	— 4,95
— 23	5	+36,5	+39,910	8393	+5	+24,05

(Quadro 3—continuação)

$S(y\,\Xi_1)/5 = +3\,935$

$S(y\,\Xi_2)/5 = -19\,766$

$S(y\,\Xi_3)/5 = +27\,309{,}0$

$S(y\,\Xi_4)/5 = -94\,581{,}10$

$S(y\,\Xi_5)/5 = +19\,166{,}46$

$S(\Xi_1^2)/5 = 350$ $\quad S(\Xi_1\Xi_3)/5 = 2\,520{,}0$ $\quad S(\Xi_1\Xi_5)/5 = 2\,491{,}20$

$S(\Xi_3^2)/5 = 18\,153{,}00$ $\quad S(\Xi_3\Xi_5)/5 = 18\,014{,}400$

$S(\Xi_5^2)/5 = 18\,403{,}4988$

$S(\Xi_2^2)/5 = 420$ $\quad S(\Xi_2\Xi_4)/5 = 2\,016{,}00$

$S(\Xi_4^2)/5 = 9\,677{,}1500$

Coeficientes de regressão ímpares (*Odd Coefficients of Regression*).

Regressão quíntica (*quintic regression*) :

Matriz de quadrados e produtos dos $\xi\xi$ .....

8 640,0	—2 520,0	—2 491,2
	195 449,4	—18 014,4
		196 169,8

	Matriz de Cofactores			Produtos	Coeficientes	
$10^9 \times$	38,016 751	0,539 225	0,532 300	+54 636,00	+6,269 921	$= b_1$
		1,688 701	0,161 922	+ 2 074,30	+0,077 646	$= b_3$
			1,682 332	—47 359,51	—0,154 263	$= b_5$

Determinante $= 10^9 \times 325\,779{,}8$

Soma dos quadrados (*sum of squares*) = 350 030

Regressão cúbica (*cubic regression*) :—

Cofactores	195 449,4	2 520,0	+54 636,0	+6,350 590	$= b_1$
		8 640,0	+ 2 074,3	+0,092 493	$= b_3$

Determinante $= 10^9 \times 1{,}682\,332$

Soma dos quadrados (*sum of squares*) = 347 163

Regressão linear (*linear regression*) :—

Quadrado (*square*) $= (54\,636)^2/8\,640 = 345\,497$

Coeficientes de regressão pares (*Even Coefficients of Regression*)

Regressão quártica (*quartic regression*) :—

Matriz =	301 644,0	—2 016,0
		357 481,6

Cofactores	357 481,6	2 016,0	—139 311,00	—0,460 918	$= b_2$
		301 644,0	+ 50 215,21	+0,137 870	$= b_4$

Determinante $= 10^9 \times 107{,}8281$

Soma dos quadrados (*sum of squares*) = 71 134

Regressão quadrática (*quadratic regression*) :—

Quadrado (*square*) $= (-139\,311)^2/301\,644 = 64\,339$

(QUADRO 3—continuação)

Análise de regressão (*Regression Analysis*)

	Graus de liberdade	Somas dos quadrados	Quadrado médio
Regressão :			
do 1.º grau (*1st. degree*)	1	345 497	
diferença (*difference*)	1	1 666	
dos 1.º e 3.º graus (*1st. & 3rd. degrees*)	2	347 163	
diferença (*difference*)	1	2 867	
dos 1.º, 3.º e 5.º graus (*1st., 3rd. & 5th*)	3	350 030	
do 2.º grau (*2nd. degree*)	1	64 339	
diferença (*difference*)	1	6 795	
dos 2.º e 4.º graus (*2nd. & 4th. degrees*)	2	71 134	
Resto (*remainder*)	19	77 819	4 096
Total, dentro de variedades (*Total, within varieties*)	24	498 983	

Equação de regressão (*Regression Equation*) :—

$$y = 1032{,}4 + 6{,}351\ \xi_1 - 0{,}462\ \xi_2 + 0{,}092\ \xi_3$$

Análise de regressão modificada (*Revised Regression Analysis*)

	Graus de liberdade	Somas dos quadrados	Quadrado médio
Regressão dos 1.º e 3.º graus (*1st. & 3rd.*)	2	347 163	
Regressão do 2.º grau (*2nd. degree*)	1	64 339	
	3	411 502	
Resto (*remainder*)	21	87 481	4 166
Total, dentro de variedades	24	498 983	

Colheita corrigida por fileira por ano

(*Corrected Yield per Row per Year*) *kg*

Variedade	Média	Limites fiduciais 10%	
A	87,9	84,1	91,7
B	85,2	81,4	89,0
C	127,0	123,2	130,8
D	133,6	129,8	137,4
E	105,3	101,5	109,1
F	80,4	76,6	84,2

Regressão total (*Total Regression*)

1.º grau	(*1st. degree*)	$(58\ 571)^2/(8\ 990)$	= 381 597,6
2.º grau	(*2nd. degree*)	$(159\ 077)^2/(302\ 064)$	= 83 775,3
3.º grau	(*3rd. degree*)	$(29\ 383{,}3)^2/(213\ 602)$	= 4 042,0
			469 415

Análise de covariância (*Analysis of Covariance*)

Análise original			Regressão		Análise final		Quadrado médio
Variedades ...	(5)	1343 132			(5)	1286 219	257 244
Resto.....	(24)	498 983	(3)	411 502	(21)	87 481	4 166
Total	(29)	1843 115	(3)	469 415	(26)	1373 700	

Os números entre parêntesis são os graus de liberdade
(*Numbers in brackets are the degrees of freedom*).

QUADRO 4

COMPONENTE DE OSCILAÇÃO (*OSCILLATION COMPONENT*)

Variedade	Fileiras (*Rows*) 1–6	7–12	13–18	19–24	25–30	Total	Média
A	150	157	366	501	499	1673	334,6
B	113	246	376	537	573	1845	369,0
C	60	233	300	581	645	1819	363,8
D	—12	202	544	754	742	2230	446,0
E	223	330	506	528	498	2085	417,0
F	239	335	410	410	406	1800	360,0
Total	773	1503	2502	3311	3363	11452	381,7

Análise de regressão (*Regression Analysis*)

	Graus de liberdade	Somas dos quadrados	Quadrado médio	F
Regressão do 1.º grau (*1st. degree*)	1	813 869		
Diferença (*difference*)	1	35 351	35 351	4,35
1.º e 3.º graus (*1st. & 3rd. degrees*)	2	849 220		
Regressão do 2.º grau (*2nd. degree*)	1	47 414	47 414	5,83
	3	896 634		
Resto (*remainder*)	21	170 738	8 130	
Total, dentro de variedades	24	1067 372		

Com P = 5% e graus de liberdade = 1 ; 21, F = 4,32

Equação de regressão :–

$$y = 381{,}7 \quad +9{,}581\ \xi_1 \quad -0{,}396\ \xi_2 \quad -0{,}426\ \xi_3$$

Análise de covariância (*Analysis of Covariance*)

Análise original			Regressão		Análise final		Quadrado médio
Variedades	(5)	42 758			(5)	32 125	6 425
Resto.....	(24)	1067 372	(3)	896 634	(21)	170 738	8 130
Total	(29)	1110 130	(3)	907 267	(26)	202 863	

QUADRO 5

COMPONENTE DE TENDÊNCIA LINEAR (*COMPONENT OF LINEAR TREND*)

Variedades	Fileiras (*Rows*) 1–6	7–12	13–18	19–24	25–30	Total	Média
A	— 52	+ 9	— 19	—19	+ 7	— 74	— 14,8
B	— 22	+ 32	— 72	—63	— 97	—222	— 44,4
C	+ 46	+109	— 27	—74	— 45	+ 9	+ 1,8
D	+224	+269	+198	+79	+ 40	+810	+162,0
E	+103	+138	+ 52	+ 7	— 65	+235	+ 47,0
F	+109	+175	+130	+86	+101	+601	+120,2
Total	+408	+732	+262	+16	— 59	+1359	+45,3

Análise de regressão (*Regression Analysis*)

	Graus de liberdade	Somas dos quadrados	Quadrado médio	F
Regressão do 1.º grau (*1st. degree*)	1	45 375		
Diferença (*difference*)	1	9 315	9 315	4,72
1.º e 3.º graus (*1st. & 3rd. degrees*)	2	54 690		
Regressão do 2.º grau (*2nd. degree*)	1	7 480	7 480	3,79
	3	62 170		
Resto (*remainder*)	21	41 426	1 973	
Total, dentro de variedades	24	103 596		

(Quadro 5—continuação)

Equação de regressão :-

$$y = 45{,}3 \quad -2{,}228\ \xi_1 \quad -0{,}157\ \xi_2 \quad +0{,}219\ \xi_3$$

Análise de covariância (*Analysis of Covariance*)

Análise original			Regressão		Análise final		Quadrado médio
Variedades	(5)	163 910			(5)	159 553	31 911
Resto.....	(24)	103 596	(3)	62 170	(21)	41 426	1 973
Total	(29)	267 506	(3)	66 527	(26)	200 979	

Aumento de colheita por fileira por ano
(*Increase in Yield per Row per Year*) *kg*

Variedade	Aumento médio	Limites fiduciais 10%
A	—0,43	—1,07 .....+0,21
B	—1,17	—1,80 .....—0,53
C	+0,01	—0,63 .....+0,65
D	+4,05	+3,42 .....+4,69
E	+1,22	+0,59 .....+1,86
F	+3,10	+2,47 .....+3,74

3.3—ANÁLISE DE REGRESSÃO

Um breve exame da análise de regressão revela que uma regressão do segundo grau é necessária e suficiente para representar o efeito da heterogeneidade do terreno. Verificaremos, mais tarde, contudo, que, para as outras funções, é preferível uma regressão do terceiro grau. É, então, por motivo de uniformidade que usaremos aqui uma regressão cúbica. A análise de regressão é devidamente alterada, passando para o "resto" os dois graus de liberdade que correspondem às funções do quarto e quinto graus.

3.4—EXAME DOS RESÍDUOS

Se subtrairmos do total de qualquer fileira a média da variedade correspondente, ficará um resíduo que é puramente uma medida da heterogeneidade do solo.

$$602 - 841{,}6 = 239{,}6 \text{ etc.}$$

Os trinta resíduos são indicados na figura 2*a*. Ora, a curva de regressão dada por

$$y = \bar{y} + b_1\xi_1 + b_2\xi_2 + b_3\xi_3$$

devia representar o componente sistemático dêstes trinta resíduos. O desvio padrão dos desvios, em relação à curva de regressão, é:

$$s = \sqrt{4166} = 64{,}5.$$

Os limites 2,5% ficam, então, em

$$\pm\,(1{,}96)\,(64{,}5) = \pm\,126{,}4$$

acima e abaixo da curva de regressão (ver fig. 2*a*).

Temos, assim, uma espécie de "gráfico de contrôle" (*control chart*) que devia ser interpretado como o são os "gráficos de contrôle" usados na indústria. Parece-nos ser razoável supor que os desvios em relação à curva de regressão sejam aleatórios. Notamos, especialmente, que não há nada excepcional na primeira e última fileiras, embora não fôssem guardadas por fileiras marginais. Frisamos que a análise subsequente é rigorosamente lógica, sòmente na hipótese de os desvios residuais serem aleatórios.

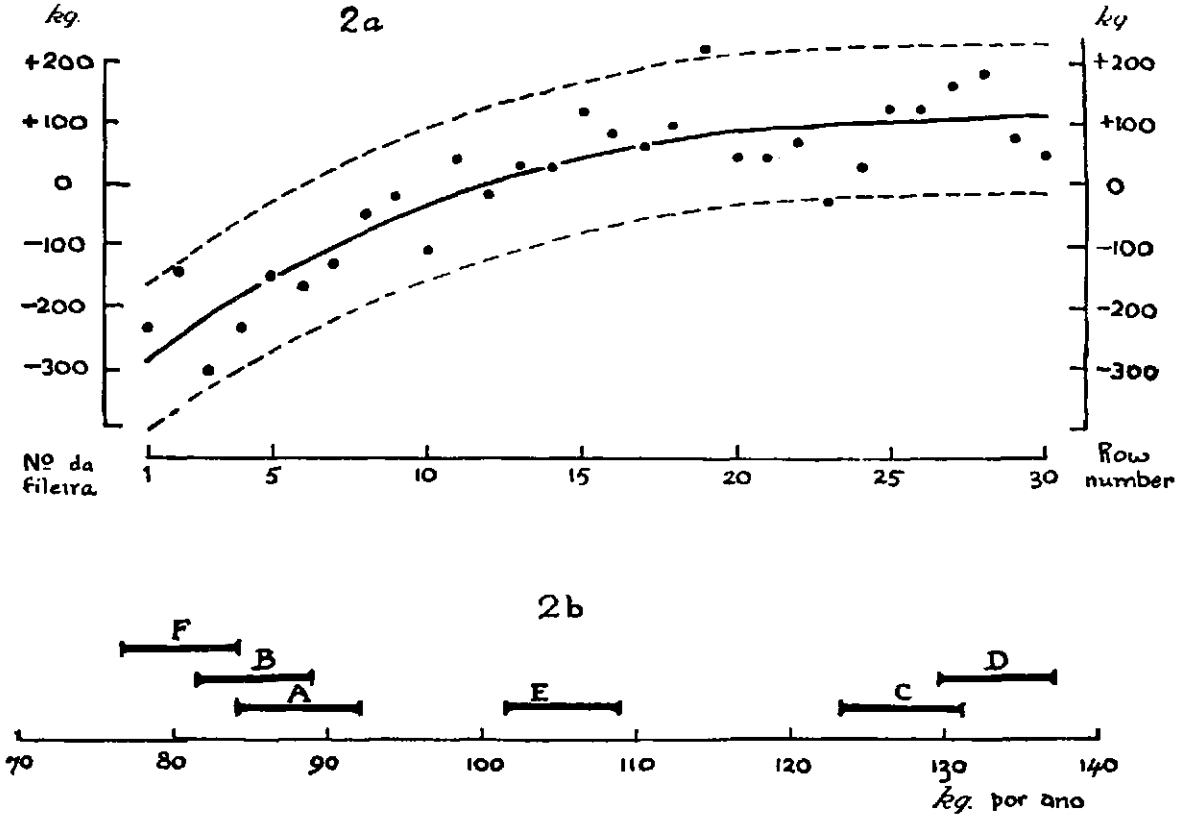

FIGURA 2.—2*a* — Resíduos dos totais e regressão sôbre posição da fileira. 2*b* — Colheitas médias por ano.

3.5—CORREÇÃO DAS MÉDIAS DAS VARIEDADES

Eliminamos, quanto possível, o efeito da heterogeneidade do terreno, *subtraindo*, de cada fileira, a quantidade

$$b_1\xi_1 + b_2\xi_2 + b_3\xi_3.$$

Para corrigir a média de uma variedade temos, então, que subtrair

$$\frac{b_1\Xi_1 + b_2\Xi_2 + b_3\Xi_3}{5}$$

Por exemplo, a média corrigida da variedade *A* será:

$$841{,}6 + 5(6{,}350) - 5\,(-0{,}462) + 8{,}25(0{,}092) = 879.$$

Não tomando em conta os erros da estimação dos coeficientes de regressão, obtemos para a estimativa da variância da média de uma variedade

$$\frac{4166}{5} = 833{,}2$$

$$\text{desvio padrão} = \sqrt{833{,}2} = 28{,}86.$$

Para calcular os limites da média, multiplicamos o desvio padrão pelo valor de *t* (de "Student"), que corresponde a P = 20% e graus de liberdade = 21.

$$(28{,}86)\ (1{,}323) = 38.$$

Somamos e subtraímos esta quantidade da média corrigida

$$879 \pm 38 = 841 \text{ e } 917.$$

Finalmente podemos dividir por dez, para apresentar os resultados em têrmos de colheita por fileira *por ano*. (Ver quadro 3). Os intervalos são representados gràficamente na figura 2*b*. Aproximadamente, podemos dizer que a probabilidade é de 80% de que o valor verdadeiro esteja dentro do intervalo indicado e que, quando dois intervalos se sobrepõem, as duas variedades correspondentes não são significativamente diferentes.

Os intervalos deviam ser, de fato, um pouco mais largos para tomar em conta os erros dos coeficientes de regressão. O teste de significância exato é, contudo, fornecido pela *análise de covariância*.

3.6—ANÁLISE DE COVARIÂNCIA

Para construir uma análise de covariância, temos que calcular a soma dos quadrados atribuível a uma regressão do terceiro grau, *sem a eliminação das variedades*. As funções lineares, agora, são ortogonais, permitindo-nos calcular, independentemente, cada quadrado, como se vê na secção "Regressão total" no quadro 3.

Temos, agora, todos os elementos necessários para a análise de covariância. Note-se que a soma final dos quadrados de "entre variedades" se obtém por diferença.

$$1\,373\,700 - 87\,481 = 1\,286\,219.$$

É claro que, neste exemplo, as diferenças entre as variedades são tão grandes que a análise de covariância é desnecessária. Apresentamo-la sòmente para exemplificar o método.

3.7—OSCILAÇÃO E TENDÊNCIA

Calculamos para cada fileira o valor da função S que dá uma medida da amplitude de oscilação de ano em ano.

$$-63{,}3+85{,}2-43{,}9+110{,}8-13{,}2+18{,}3-80{,}8+53{,}2-24{,}9+108{,}1=150, \text{ etc.}$$

Daqui por diante o cálculo segue exatamente o mesmo esquema observado no caso da análise dos totais. Foi, por isso, abreviado antes da sua apresentação no quadro 4.

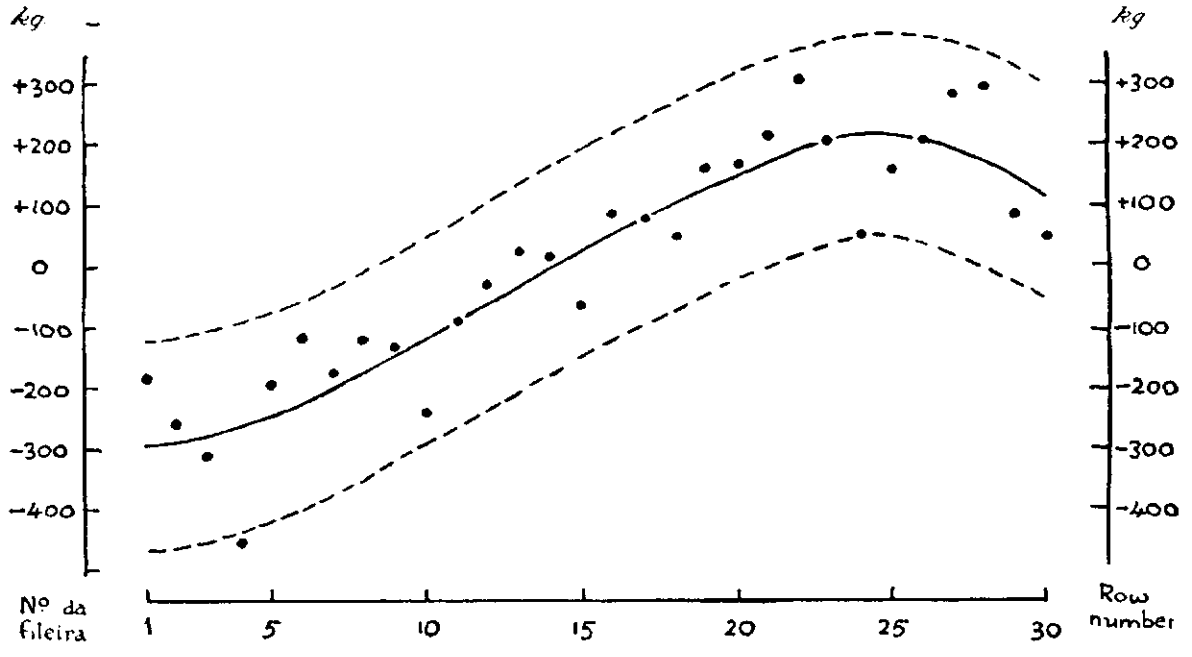

FIGURA 3.—Resíduos da função de oscilação e a regressão dêstes resíduos sôbre posição da fileira.

Notamos aqui que vale a pena incluir o têrmo do terceiro grau na equação de regressão. Os resíduos são apresentados, gràficamente, na figura 3. A análise de covariância revela que não há nenhuma indicação de diferenças entre variedades, no que diz respeito a esta função. A diferença média entre um ano par e um ano ímpar é de 76 kg por fileira.

A análise da terceira função R, que mede o componente linear de variação secular, apresenta-se no quadro 5.

A figura 4*a* mostra os resíduos e o "gráfico de contrôle". Se dividirmos o valor de R por 40, ficaremos com uma estimativa da taxa de aumento de produção em quilos por ano. As médias corrigidas e os intervalos são, assim, convertidos e apresentados no quadro 5 e na figura 4*b*.

4—DISCUSSÃO

4.1—HETEROGENEIDADE DO SOLO

Notamos que a regressão leva em conta uma grande parte da variação atribuível à falta de uniformidade do terreno (82% no caso de T, 84% no de S e 60% no de R). Em nenhum caso, contudo, é adequada uma regressão do primeiro grau. Disso resulta que uma simples análise de covariância é insuficiente para eliminar o efeito da posição da fileira.

De modo geral, a fertilidade vai aumentando à medida que passamos da fileira 1 até a fileira 30. Por causa disso, o arranjo sistemático falha na sua intenção de eliminar o efeito da heterogeneidade do campo. As colheitas

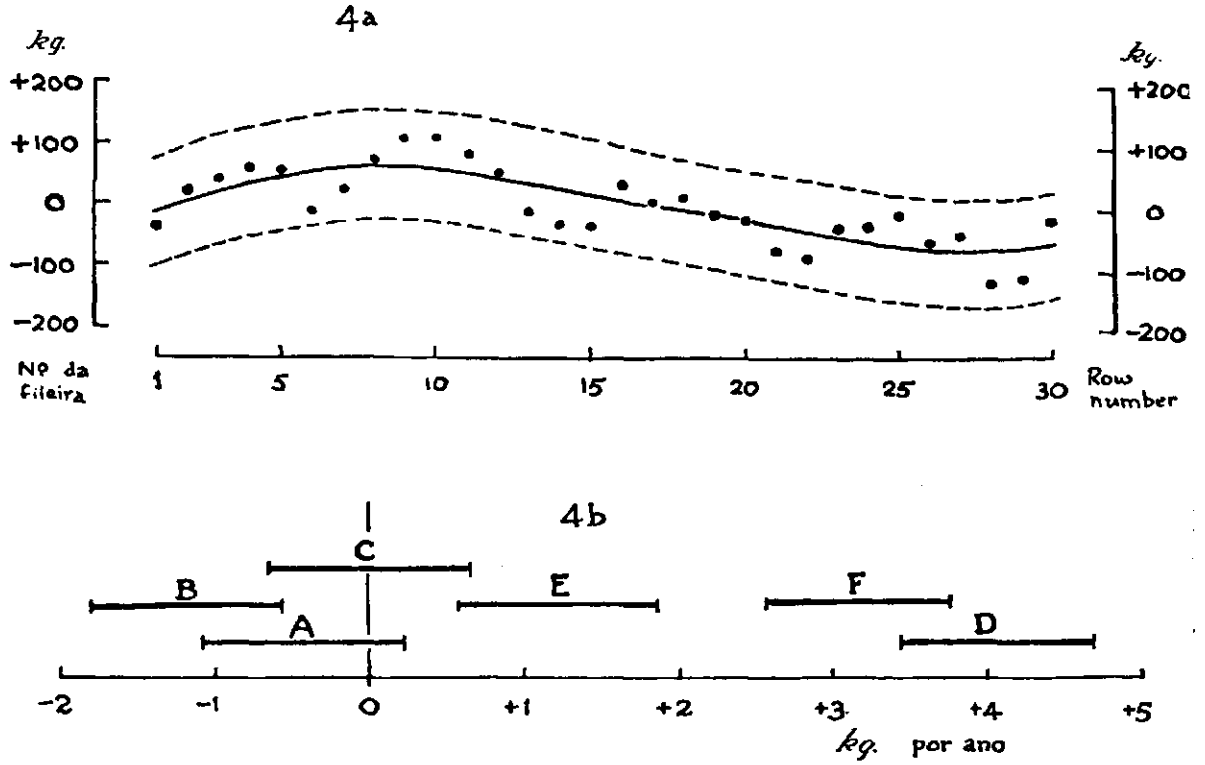

FIGURA 4.—4*a* — Resíduos da função de tendência linear e a regressão dêstes resíduos sôbre posição da fileira. 4*b* — Taxas de aumento de colheita.

de A e B são diminuídas e as de F aumentadas. Consideremos, por exemplo, as colheitas médias por fileira por ano, antes e depois da correção :

Variedades	Originais	Corrigidas
A	84,1	87,9 kg
F	83,7	80,4 kg

Comparando a figura 2*a* com a figura 3, notamos que a variação atribuível à heterogeneidade do solo é maior no caso da função S do que no do total T. Verificamos que

$$T + S = 10 \text{ (total dos anos pares)}$$
$$T - S = 10 \text{ (total dos anos ímpares).}$$

Obtêm-se os coeficientes das equações de regressão de T + S e T — S, somando-se e subtraindo-se, respectivamente, os coeficientes das equações de regressão de T e S.

Coeficientes	T	S	T+S	T—S
b_1	+6,351	+9,581	+15,932	—3,230
b_2	—0,462	—0,396	—0,858	—0,066
b_3	+0,092	—0,426	—0,334	+0,518

Vemos, assim, que a curva de regressão de T+S se apresenta ao contrário da curva de regressão de T—S. Isto indica que uma fileira que produz melhor do que as outras nos anos pares produzirá pior do que as outras nos anos ímpares. Por outras palavras, num canteiro relativamente *mais* fértil, a produção será menor nos anos de pequena produção, embora no total de um período maior de anos a produção seja maior. Isso sugere que, se tentássemos aumentar a produção por meio de melhoramentos da terra, adubação, etc., o resultado seria uma diminuição da colheita nos anos de produção mínima.

Olhando para a figura 4*a*, notamos que a curva de regressão desce da esquerda para a direita. Isto quer dizer que as diferenças de fertilidade entre as fileiras vão diminuindo no curso do tempo.

4.2—DIFERENÇAS ENTRE VARIEDADES

As diferenças principais entre as médias de produção das seis variedades já eram bastante evidentes, mesmo antes da análise estatística. Essa análise corrigiu, contudo, um "vício" (*bias*) nas estimativas das médias e, especialmente, na comparação de A com F. Mostra, também, que as diferenças entre C e D, e entre A, B e F não são estatìsticamente significativas.

Um dos fatos mais interessantes revelados pela análise é a falta completa de qualquer sinal de heterogeneidade no valor da função S, o critério da oscilação. Ora, o fenômeno da oscilação é desvantajoso do ponto de vista comercial. Mas o experimento mostrou que, selecionando-se uma variedade altamente produtiva, não se aumentará a amplitude da oscilação. Por outras palavras, o incremento de produção das melhores variedades é obtido igualmente nos anos pares e ímpares.

Não obstante a grande alternância de ano em ano, a análise conseguiu revelar diferenças entre as tendências seculares das variedades : As colheitas de D, E e F vão crescendo ; a de B vai diminuindo, ao passo que não po-

demos demonstrar uma tendência positiva ou negativa no caso de A e C. Notamos que, embora C e D não sejam distinguíveis nas suas produções médias (figura 2*b*), a consideração das taxas de aumento (figura 4*b*) mostra a superioridade de D. Será interessante observar se a taxa de incremento de D, relativamente a C, irá manter-se no futuro.

4.3—OUTRAS APLICAÇÕES DO MÉTODO

Num artigo dedicado principalmente a uma descrição de *método*, devemos chamar a atenção para o fato de que o mesmo tipo de análise estatística pode ser aplicado em investigações de natureza inteiramente diferente. Vamos tomar um exemplo no campo da economia. Temos, digamos, dados referentes à produção ou venda de qualquer mercadoria, mês por mês, durante três anos e queremos comparar os meses. Além das diferenças gerais entre os meses e a variação aleatória, haverá uma tendência secular (*secular trend*) que pode ser representada por uma regressão de grau apropriado. Identificando formalmente os doze meses de um ano com doze variedades, e os trinta e seis meses com trinta e seis fileiras, reconhecemos que o problema será resolvido exatamente pelo mesmo tipo de análise que foi aplicado ao problema atual da comparação de variedades de café.

4.4—CONCLUSÕES

Pondo de lado as questões que interessam ao produtor de café (com os quais não nos preocupamos neste trabalho), podemos tirar três conclusões importantes :

a) Um experimento sistemático não atinge o seu próprio objetivo — a eliminação do efeito de heterogeneidade do campo — a não ser que seja sujeito a uma análise estatística muito rigorosa. Essa análise, geralmente, dá mais trabalho que a dos delineamentos modernos que contêm um elemento de casualização (blocos ao acaso, quadrados latinos, etc.).

b) O método das funções ortogonais pode ser explorado com grande vantagem : Na análise dêste experimento foi empregado duas vêzes — na construção das funções independentes das colheitas anuais e no cálculo da regressão sôbre a posição da fileira.

c) Uma inspeção dos gráficos dos resíduos sugere-nos que um delineamento em blocos ao acaso (cinco blocos de seis fileiras cada um) daria comparações, aproximadamente com a mesma precisão, mas com menos trabalho de cálculo e mais objetividade lógica.

SUMMARY

This paper describes the statistical analysis of a varietal trial with two unusual characteristics :

(i) The plant (coffee) is one of those which show strong maxima and minima of production in alternate years. This phenomenon must be prevented from masking or biasing the other varietal comparisons in which we are interested.

(ii) The design of the experiment is systematic. It was laid down in Campinas, Brazil, in 1933 at a time when the principles of randomisation were not so widely known as they are today.

THE EXPERIMENT AND DATA.

Six varieties are compared, denoted by A B C D E and F (see page 104). They are planted in thirty rows, each with 50 plants, according to the systematic design :

A B C D E F A B C D E F A B C D E F A B C D E F A B C D E F

Data for twelve years are available in *quadro* 1 but those of the small and irregular yields in the first two years were discarded. The mean yields of the remaining ten years (1935-1946) appear by figure 1 to be fairly regular and consistent in their behaviour. Most of the plants, but by no means all, showed their maxima in the even years.

STATISTICAL ANALYSIS.

The quantity of primary interest is the mean yield over the whole period. It is essential that these means should be based (as here) on an even number of years in order to eliminate, from their comparisons, the effect of the alternations of maxima and minima.

The magnitude of the oscillation is conveniently measured by total of even years *minus* total of odd years.

Finally we need a linear function of the annual yields for measuring secular trend in order to discriminate varieties which are slowly gaining on the others. The usual linear orthogonal polynomial (with coefficients —9, —7, —5, etc.) is unsuitable because it is not independent of the component of oscillation. A suitable function is obtained instead by using the coefficients

$$-2 \quad -2 \quad -1 \quad -1 \quad 0 \quad 0 \quad +1 \quad +1 \quad +2 \quad +2.$$

The coefficients of the three linear functions thus defined are set out in *quadro* 2 (page 107), where it will be verified that they are mutually orthogonal.

The effect of the heterogeneity of the soil is as far as possible eliminated (separately for the three functions) by an analysis of covariance, using the number of the row (1-30) as the concommitant observation. A simple linear regression formula is however inadequate. The regressions were taken to the fifth degree by means of orthogonal polynomials. Since the "between varieties" contribution must be removed from the sums of squares and products, the regression coefficients are no longer independently obtainable. It is found however that the normal equations fall into two sets, one yielding the regression coefficients of odd degree and the other those of even degree. Consequently the use of orthogonal polinomials still effects a considerable saving of work. The computations are set out in full in *quadro* 3 and in abbreviated form in *quadro* 4 and 5 for the total, the oscillation and the trend respectively. (Note that the comma indicates the decimal point.)

We find that a quadratic regression is adequate for the first and cubic regressions for the others. For the sake of uniformity, a cubic regression was used in every case. The residuals found by subtracting the varietal means from the rows are plotted in figures 2a, 3 and 4a. respectively, together with the regression curves and the 2.5% control limits. These control charts suggest that it is not unreasonable to suppose that the remaining variation is random.

Next we use the regression formulae to correct the varietal means. The approximate 80% fiducial intervals of the mean annual yields (kg per row) and the rate of increase of yield (kg per row per year per year) are shown in figures 2b and 4b respectively. In the case of the component of oscillation, the analysis of covariance failed to show the slightest suggestion of differences between varieties.

DISCUSSION.

An examination of the regressions on number of row reveals the interesting fact that the more fertile portions of the field produce *lower* yields in the odd years than the less fertile portions. The reason is presumably that the heavier yields in the even years, by exhausting the plant, depress the yields in the following years.

The major differences between varietal means over the ten years were sufficiently clear even before the analysis though some of the adjustments are appreciable. A striking fact is that, although there are big general differences between varieties, there are no significant differences between them in respect of the amplitude of oscillation. In other words, the increment of yield in the better varieties is obtained equally in odd and even years. In spite of the large component of oscillation, it is possible to discriminate varieties in respect of their rate of increase of yield (figure 4b).

CONCLUSIONS.

(i) The extra difficulty introduced by the strong alternations of yield from year to year can be solved by the choice of suitable orthogonal functions of yearly yields.

(ii) Once again a systematic design is found wanting — it fails to eliminate the effect of soil heterogeneity from varietal comparisons. This defect can however be removed, for practical purposes, by an adequate analysis of covariance on row number.

LITERATURA CITADA

1. **Brieger, F. G.** Melhoramento de *Coffea arabica* var. *bourbon* — Análise estatística da experiência de café bourbon e seleção de café por métodos modernos. Bragantia **1** : 26-119. 1941.

2. **Fisher, R. A.** and **F. Yates.** *Em* Statistical Tables, pág. i—viii+1-90, Oliver and Boyd, 98 Great Russel Street, W. C., London. 1938.

3. **Mendes, J. E. T.** Ensaio de variedades de cafeeiros. Bol. Tec. do Instituto Agronômico de Campinas 65 : 1-36. 1939.

4. **Rodrigues de Carvalho, M. J.** *Em* A estatística na experimentação agrícola, pág. I—XV+1—174, secção 11.2, Sá Costa, Lisboa. 1946.