A Biometria nos Estudos de Genetica

Eng. Agr. ALCIDES FRANCO
Chefe da Secção Technica do Serviço do Algodão
Ministerio da Agricultura

Num recente trabalho sob o titulo acima (Revista de Agricultura, vol. VI, n. 1-2, Janeiro-Fevereiro, 1931), o professor Carlos Teixeira Mendes procura demonstrar que a curva de frequencias de uma população, cuja amostra é estudada, não representa ou não define o seu gráo de puresa. Em outras palavras, a rigidez das formulas da analyse estatistica não traduz a sequencia dos phenomenos biologicos.

Para esse fim aquelle professor cita uns poucos de experimentos que realisou, utilisando caracteres quantitativos de diversas especies vegetaes, em numero de 1600 individuos para cada experimento.

Attendendo ao desejo manifestado pelo illustrado collega mas, de outro lado, sem pretender, de qualquer modo, melindrar-lhe o valor, me permitto fazer algumas considerações sobre o assumpto.

A Biometria nasceu, por assim dizer, da Genetica, pela necessidade de exprimir numericamente, os phenomenos desta ultima. Ambas essas sciencias são ainda muito jovens. Uma pleiade de scientistas, na Inglaterra, Estados Unidos, Allemanha, Dinamarca, Suecia, Russia, India, e na França, Belgica, etc., trazem quasi diariamente á publicidade novas contribuições para o esclarecimento de assumptos de que ambas se occupam. Parece, assim, de alguma utilidade que se desfaçam as duvidas levantadas pelo professor Teixeira Mendes, duvidas essas que podem traduzir-se em desconfiança no espirito daquelles que se iniciam nesses estudos.

Em preliminar, o preconceito de que a analyse estatistica não possa interpretar os phenomenos biologicos tem, hoje, graças especialmente aos estudos e pesquizas de R. A. Fisher uma significação muito restricta, não porque os phenomenos biolo-

gicos se passem, agora, de accordo com as formulas mathematicas, mas porque as investigações no dominio da analyse estatistica pura demonstraram *como e até que limite* aquellas formulas podem ser applicadas, isto é, satisfazem a propria condição de sequencia daquelles phenomenos. Está mesmo verificado que uma estimativa de erro de uma população, cuja amostra é estudada, somente pode ser feita quando as partes constituintes dessa amostra podem ser localisadas independentemente e ao acaso. Esta é a condição essencial (1).

Actualmente está se generalisando o conceito de que a probabilidade mathematica, tal como é definida nos livros, é *apenas* applicavel ao que pode ser chamada a "analyse deductiva provavel", como, por ex., a probabilidade que se tem de tirar uma bola branca em uma urna que contém um determinado numero de bolas brancas e pretas. A analyse *inductiva*, a estimativa da população partindo de um grupo ou amostra, é cousa differente.

Dr. R. A. Fisher, chefe do Departamento da Estatistica da Estação Experimental de Rothamsted, a quem se deve, em grande parte, a concepção e estabelecimento da escola estatistica moderna, e considerado como o seu *leader* intellectual, Fisher, diziamos, assim se exprime: (2)

> "Inferences respecting populations, from which known samples have been drawn, cannot be expressed in terms of probability, except in the trivial case when the population is itself a sample of a superpopulation the especification of which is known with accuracy. This is not to say that we cannot draw, from knowledge a sample, inferences respecting the corresponding population. Such a view would entirely deny validity to all experimental science. What is essential is that the mathematical concept of probability is inadequate to express our mental confidence in making such inferences, and that the mathematical quantity which appears to be appropriate for meassuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I

have used the term "likelihood" to designate this quantity ; since both the words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship".

A esse respeito, diz o professor H. Hotelling, chefe do Departamento de mathematicas da Stanford University (3)! "o gráo de confiança na analyse inductiva não pode ser medido na escala da probabilidade deductiva, mas a sua comprehensão requer um systema de medida inteiramente differente". E accrescenta aquelle professor que, si por ex., se toma uma moeda *não viciada* e a jogamos diversas vezes consecutivas, podemos dizer, *a priori*, que, em 100 vezes, a probabilidade é de cerca de 20 21 para que o numero de *caras* e *corôas* esteja comprehendido entre 40 e 60. Todavia, *si tudo quanto sabemos* é o numero de *caras* e *corôas* em um dado numero de observações, torna-se impossivel, de accordo *apenas* com a definição de probabilidade, derivar dahi um valor numerico da probabilidade para que, em outras series de observações, a proporção de *caras* esteja comprehendida entre 40 e 60 %. Isto significa que a definição mathematica de probabilidade não é sufficiente, por si só, para esclarecer o assumpto.

Exemplos da applicação da theoria do maximum likelihood a problemas de Genetica são dados no capitulo IX do livro de Fisher "Statistical Methods for Research Workers".

O gráo de confiança que se pode depositar nas conclusões estatisticas é medido, segundo Fisher, tomando-se a metade do logarithmo natural da *variance*. Fisher assim denomina o quadrado do erro standard. A differença entre dois quaesquer desses logarithmos é equivalente ao que Fisher denomina o *z test* e o seu erro standard depende somente do numero de *gráos de independencia* (degrees of freedom). Esta expressão é usada no sentido de comparações individuaes. Por exemplo, entre n quantidades cuja media é a unica, ha $n - 1$ comparações individuaes ou gráos de independencia.

O principio basico do methodo de Fisher reside em que, a variação total entre resultados individuaes, num conjuncto de observações, quando medida em funcção da differença entre a somma dos quadrados dos afastamentos e a sua media geral,

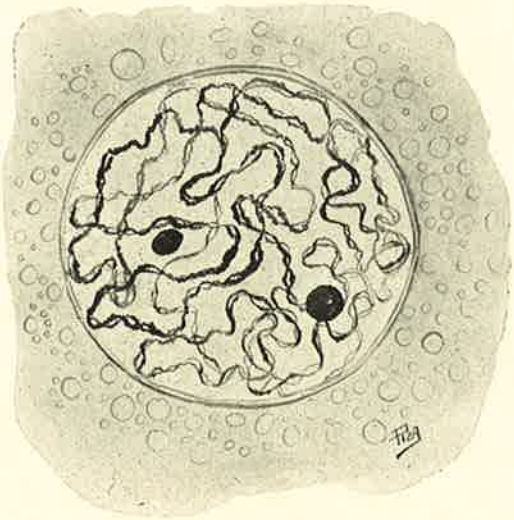

FIG. 1

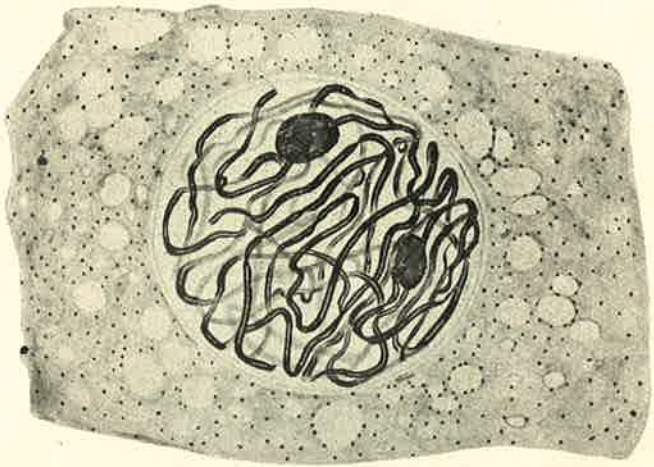

FIG. 2

FIG. 3

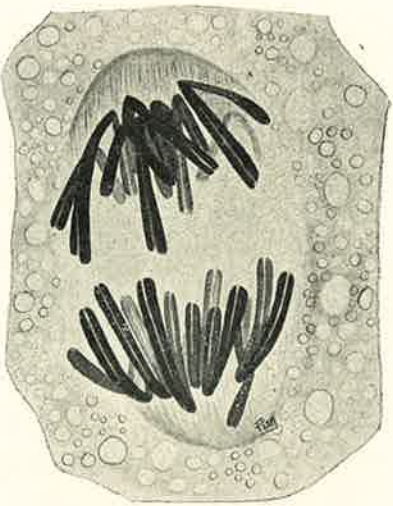

FIG. 4

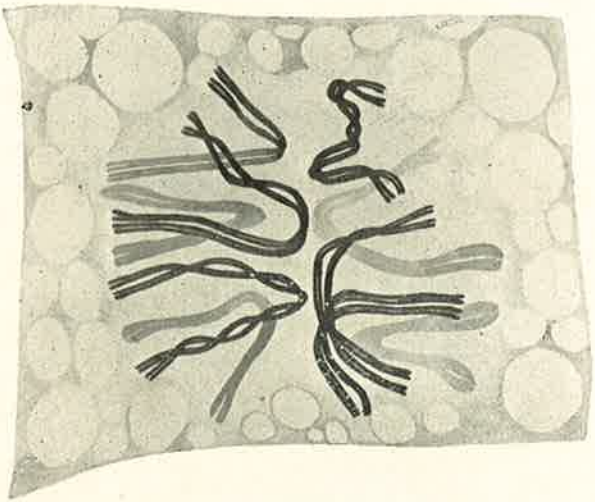

FIG. 5

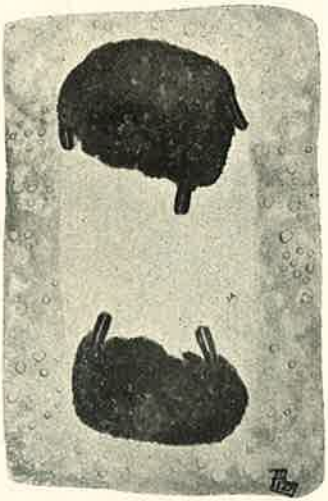

FIG. 6

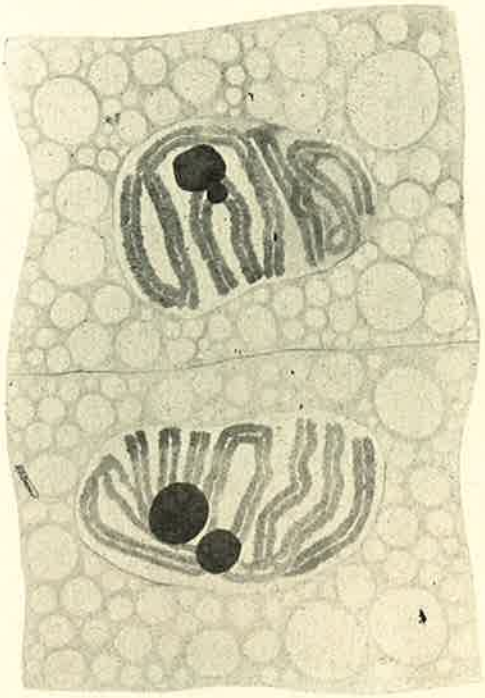

FIG. 7

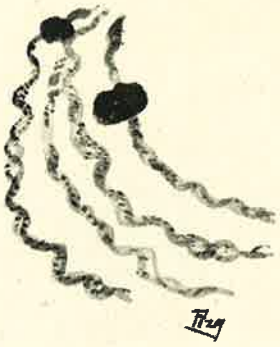

FIG. 8

pode ser analysada ou decomposta em um numero de partes ou fracções, pela applicação de uma identidade algebrica. Isto permitte decompôr se a somma total dos quadrados em varios factores conhecidos, ficando um residuo correspondente a factores desconhecidos ou que não podem ser controlados. Esta ultima fracção ou residuo é a base logica em que assenta uma estimativa do erro do experimento. Fisher demonstrou que o valor medio da fracção attribuida a qualquer factor, é obtida dividindo esta fracção pelo numero de gráos de independencia existentes no conjuncto de n quantidades.

O *z test* permitte a interpretação dos casos mais duvidosos, quando ha mistér determinar-se a significação das differenças entre dois quaesquer logarithmos da variance. No caso dos experimentos do professor Teixeira Mendes, a determinação do gráo de puresa das populações estudadas seria dada por esta medida.

O professor Teixeira Mendes não trabalhou com variedades puras ou hybridas, isto é, o gráo de puresa dessas variedades não foi, egualmente, a heterogeneidade do solo, cuja estimativa de erro não foi verificada. Tanto assim que elle diz: "... um lote que por tudo se nos affigurava homogeneo: terra, altura das plantas, sua apparencia, etc." e, mais adiante: "... Amparo, e que apresentava os caracteres de uma variedade mais ou menos pura". Com relação á variedade "Catteto" diz que: "... por todos os motivos nos parece ser mais pura que as anteriores..." e quanto á variedade "Hickory King": "... dá um doce a quem mostrar uma espiga que, NÃO SENDO HYBRIDA, possúa ou mais ou menos que oito fileiras de grãos". E conclúe que a curva de frequencias será, então, nesse caso, uma recta.

Com referencia a esse ultimo, é preciso considerar que o caracter *oito fileiras de grãos* está fixado e, assim, como poderá variar quando se não trata de individuos hybridos e, como, de outro lado, obter uma curva?

O nono experimento, relativo ao peso das bananas de um cacho, mostra que, augmentando o intervallo das classes, o typo de curva obtido é bem differente. Realmente, isto mostra a necessidade de grupar as observações em classes, pois, si ca-

da observação fosse considerada como constituindo uma classe, não teriamos uma curva mas uma linha quebrada, qualquer que fosse a naturesa da observação.

Não é muito facil, todavia, estabelecer promptamente qual o intervallo de classes conveniente a cada typo de observação. Estabelecido este, porém, pouco importa o typo de curva obtido, mas o essencial é que o conjuncto de observações tendo, embora, medias differentes, possúa não obstante, a mesma *variance*.

A analyse da variance determinará, pois, no caso, o gráo de puresa das populações estudadas dentro, é claro, de um limite de erro que, aliás, pode ser conhecido. E o gráo de puresa dentro da mesma variedade será conhecido em gerações successivas, pela comparação das populações.

A exposição feita pelo professor Teixeira Mendes não nos permitte conhecer qual a mais pura dentre as variedadee que estudou. De outro lado, o caracter *peso das espigas* não foi determinado em gerações succesivas. O caracter *duas espigas* está fixado, tanto assim que, dentre 300 individuos somente 11 não o possuiam (pag. 10).

Admittindo que o caracter *peso das espigas* possa representar o gráo de puresa da população estudada, somente com a analyse da variance, feita para cada geração, podemos saber si a variedade está ou não sendo melhorada, no sentido genetico do termo.

As pesquisas de Fisher foram feitas com os dados de observação, durante o longo periodo de 87 annos, de adubação de trigo em Rothamsted e, de outro lado, no seu trabalho intitulado "On the influence of rainfall on the yield of wheat at Rothamsted", talvez o mais importante de sua já grande bagagem scientifica, Fisher estabeleceu definitivamente novas bases para a escola estatistica. Realmedte, o julgamento da significação dos resultados, em que se procurava decidir sobre si algo havia sido descoberto ou si se tratava, apparentemente, de conclusões puramente accidentaes, exigia, de facto, um numero de novas descobertas no dominio da pura mathematica.

Com os diversos experimentos feitos pelo professor Teixeira Mendes, com populações da mesma variedade, podemos

comparal-as para determinar qual a mais pura, estabelecendo para isso, como vimos, o *z test*, dentro do criterio admittido do caracter peso das espigas. A esse respeito, aliás, o professor E. S. Beaven, do "National Institute of Agricultural Botany", de Cambridge (4) diz que: "the most valuabe characters of any cereal race are those wich affect its relative *productivity* in respect of grain as compared with other races of the same species. The measure of productivity is the weight of dry grain harvested and threshed from some unit of area".

O que não resta duvida, porém, é que se deve adoptar um criterio no modo de julgamento. Adoptando, segundo Fisher, os casos de probabilidade de 1:20 e 1:100 poder-se-ha distinguir, dentre os resultados, aquelles que devem constituir a base de conclusões e aquelles outros que devem ser ignorados.

O professor Teixeira Mendes serviu-se, apenas, de graphicos para illustrar os seus experimentos, sem se preoccupar com saber as estatisticas *media* e *erro standard* das populações que estudou. Tambem não fez referencia, como já accentuámos, á heterogeneidade do solo, factor da maior importancia em experimentação de campo, pois que o simples aspecto da vegetação regular e a apparente homogeneidade da área experimental não são condições essenciaes á determinação daquelle factor. Esta determinação só pode ser feita estimando-se o *erro* do experimento, do qual se deduz a sua validez, isto é, a confiança que inspira o resultado.

Os methodos de Fisher, reconhecipos recentemente como os mais perfeitos dentre quantos existem, methodos esses que já estão sendo generalisados em diversos paizes, notadamente nos Estados Unidos, são as formas mais simples que preenchem as condições de uma estimativa de erro do experimento, ao mesmo tempo que possuem a vantagem de eliminar uma grande fracção da heterogeneidade do solo.

A technica da experimentação originada por Fisher envolve dois pontos da maior relevancia: 1) a necessidade de uniformisar o processo, e 2) a necessidade de reduzir os erros experimentaes, ao mesmo tempo que se procura estimar esses erros por meio da analyse da variance. A condição essencial é, entretanto, que os lotes sejam distribuidos ao acaso.

Vamos considerar o methodo no seu caso mais simples, suppondo que se trate de comparar um numero *n* de variedades usando-se *m* replicações de cada uma. Teremos, assim *mn* pequenos lotes, e representando por *x* o caracter que se deseja conhecer, temos:

QUADRO I

1	2	3	. . .	n
1x1	1x2	1x3	. . .	1xn
2x1	2x2	2x2	. . .	2xn
			jxj	
mx1	mx2	mx3	. . .	mxn

em que *jxi* representa o *caracter* da fileira *j* na replicação *i*, emquanto que x_1, x_2, x_3... x_n representam o caracter medio nas replicações 1, 2, 3... n. Fazendo *x* egual ao caracter medio de todas as replicações, vem:

$$\bar{x} = \frac{\sum_{1}^{mn} (jxi) \sum_{i=1}^{n} (\bar{x}_i)}{mn} = \frac{}{n}$$

E a somma dos quadrados dos afastamentos em cada replicação e a media geral será, algebricamente:

$$S = \sum_{1}^{mn} (jxi - \bar{x})^2$$

Quando esta expressão é dividida por *mn-1* tem-se a estimativa da variance (o^2) da população, cujos *mn* lotes constituem a amostra.

S pode ser dividido em duas partes S_T e S_E em que S_T, é a variação entre as *n* medias dos grupos, e S_E a variação dentro desses grupos. Ou seja, algebricamente:

$$S_T = m \sum_{i=1}^{n} (jxi - \bar{xi} - \bar{x})^2$$

$$S_E = \sum_{i=1}^{n} \sum_{j=1}^{m} (jxi - \bar{xi})^2$$

em que a dupla somma mostra que, os quadrados dos afastamentos entre o caracter x e a media do grupo num lote considerado, devem ser addicionados juntos, pois que ha mn termos. Ha $mn-1$ gráos de independencia (total), isto é, para uma mesma media somente $mn-1$ lotes de outro experimento podem ser determinados ao acaso, emquanto que o ultimo será conhecido por differença.

A somma de mn quadrados da forma $(j^{xi}-\overline{x})$ é realmente equivalente á somma de $(mn-1)$ quadrados independentes. O numero de gráos de independencia para comparações entre as medias dos grupos é $n-1$, visto como ha n grupos, emquanto que dentro dos grupos ha $n(m-1)$ gráos de independencia. E' preciso notar, entretanto, que não se obtém uma identidade entre as varias sommas dos quadrados da forma S S_T S_E' mas sim temos uma identidade entre os gráos de independencia correspondentes á variação, entre e dentro, dos grupos:

$$mn-1=(n-1)+n(m-1)$$

Si se suppõe, por um instante, que não ha differenças entre o caracter observado nas diversas replicações, estão:

$$\frac{S_T}{n-1} \quad \text{e} \quad \frac{S_E}{n(m-1)}$$

representam as estimativas da variance na população cujos lotes experimentaes constituem a amostra.

Si fizermos $V_E=\frac{S_E}{n(m-1)}$ o erro standard de um lote será $S=\sqrt{V_E}$. O erro standard de n medias relativas a m valores será, cada um, equivalente a $\frac{S}{\sqrt{m}}$ e o erro standard da differença entre duas medias quaesquer de m será $S=\sqrt{\frac{2}{m}}$

Fazendo, agora, $V_T=\frac{S_T}{n-1}$ é preciso verificar si V_T é significantemente maior ou não que V_E.

A despeito da maior homogeneidade possivel, porventura existente no solo, na área experimental, existem sempre differenças no caracter observado. A significação ou validez dessas differenças é o meio por que se pode verificar si os dados fornecidos pelas observações podem ser considerados 1) ou for-

mando uma amostra homogenea proveniente da população; ou 2) si essas observações são exactamente consideradas como formando grupos de amostras de populações que, tendo embora medias differentes possuam, não obstante, a mesma variance. E' isto o que constitue o *z test* a que já nos referimos.

Si a relação $\frac{V_T}{V_E}$ fôr egual a e^{2z} a distribuição de z em uma amostra ou grupo, ao acaso, é constante na forma para um determinado numero de gráos de independencia. Para esse fim, torna-se preciso consultar as taboas organisadas por Fisher (2), taboas essas em que, para os valores de z correspondem differentes de P, n_1 e n_2. P é a probabilidade de exceder o valor de z, determinado ao acaso; n_1 e n_2 são, respectivamente, o numero de gráos de independencia correspondentes ao maior quadrado medio e ao menor quadrado medio. Por ex., si o valor de z excede o seu valor dado na taboa para 1 %, isto significa que um tal valor seria encontrado, ao acaso, menos de uma vez em 100, e isto exprime o gráo de confiança que se pode depositar no experimento.

Uma grande vantagem no adoptar o methodo de Fisher (randomised blocks), está em que o erro standard de uma differença entre duas medias pode ser calculado immediatamente, e para que todas as differenças, desde que o numero de replicações seja egual ao numero de variedades a experimentar. Este methodo tem ainda a vantagem de permittir a eliminação de fontes de erro causadas pelas differenças da composição do solo entre os lotes.

Partindo de *mn* pequenos lotes, referidos no quadro I, vejamos como se obtém as medias:

QUADRO II

	1	2	3	i	. . .	n	Medias
1	1x1	1x2	1x3	—	. . .	1xn	$1\bar{x}$
2	2x1	2x2	2x3	—	. . .	2xn	$2\bar{x}$
3	3x1	3x2	3x3	—	. . .	3xn	$3\bar{x}$
—	—	—	—	—	. . .	—	—
j	—	—	—	jxi	. . .	—	$j\bar{x}$
—	—	—	—	—	. . .	—	—
m	mx1	mx2	mx3	—	. . .	mxn	$m\bar{x}$
Medias	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_i$	. . .	$\bar{x}_n$	$\bar{x}$ = (m. geral)

Si se suppõe, agora, que a media geral $\overline{\overline{x}}$ pode ser dividida em duas partes, uma correspondente ás differenças entre as replicações (e egual á differença entre o valor numerico medio de $\overline{\overline{x}}i$ e $\overline{\overline{x}}$), e a outra parte correspondente ás differenças entre os lotes (e egual á differença entre o valor numerico medio de $j\overline{\overline{x}}$ e $\overline{\overline{x}}$), podemos dizer que o caracter quantitativo em estudo, j^xi, é da forma: $jXi - \overline{\overline{x}} (\overline{\overline{x}}i - \overline{\overline{x}}) + (j\overline{\overline{x}} - \overline{\overline{x}})$

ou seja: $$jXi = \overline{\overline{x}}i + j\overline{\overline{x}} - \overline{\overline{x}}$$

O afastamento do caracter j^xi calculado dos valores acima, representa um erro e, assim temos de calcular a somma dos quadrados de mn termos de forma $j^xi - jXi$, somma essa que pode ser deduzida da identidade:

$$\sum_{1}^{mn} (j\overline{\overline{x}}i - jXi)^2 = \sum_{1}^{mn} (j^xi - \overline{\overline{x}})^2 - m \sum_{1}^{n} (\overline{\overline{x}}i - \overline{\overline{x}})^2 - n \sum_{1}^{m} (j\overline{\overline{x}} - \overline{\overline{x}})^2$$

Si se faz S egual a $S - S_T'$ visto como no caso presente, S é reduzido por duas sommas de quadrados, uma dependendo das replicações e outra correspondente ás differenças observadas nos lotes, e desde que no calculo dos valores theoricos das medias de uma e outra ellas devem ser valores fixos, então segue-se que se perdem (mn-1) gráos de independencia, sendo (n-1) para as differenças entre as medias dos replicações e (m-1) para as differenças entre as medias dos lotes. Resta-nos, pois, (n-1) (m-1) que são correspondentes aos erros.

Temos, então: 1) para a somma dos quadrados:

$$S = S_T + S_B + S'_E$$

e 2) para os gráos de independencia:

$$mn-1 = (n-1) + (m-1) + (n-1)(m-1)$$

S'_E é a somma reduzida dos quadrados correspondentes aos erros emquanto que S_B só tem valor para verificar-se si a a sua eliminação resultou em apreciavel reducção do erro ou não.

A estimativa da variance correspondente ás replicações, isto é, $\frac{S_T}{n-1}$ deve, agora, ser comparada com a corresponden-

te ao erro $\frac{S'_E}{(n-1)\ (m-1)}$ por meio do *z test*, emquanto que o erro standard de um lote é dado por:

$$S = \sqrt{\frac{S'_E}{(n-1)\ (m-1)}}$$

A analyse da variance pode, pois, ser disposta da forma seguinte:

ANALYSE DA VARIANTE

Correspondente a:	Gráos de independencia	Somma dos quadrados	Quadrado medio
Replicacões	n-1	$m \sum_{1}^{n} (\bar{x}i - \bar{\bar{x}})^2 = S_T$	$V_T = \frac{S_T}{n-1}$
Lotes	m-1	$n \sum_{1}^{m} (\bar{x}j - \bar{\bar{x}})^2 = S_B$	$\frac{S_B}{m-1}$
Erros	(n-1) (m-1)	$\sum_{1}^{mn} (jxi - jXi)^2 = S'_E$	$V_E = \frac{S'_E}{(n-1)(m-1)}$

$$z = 1/2 \log_e \frac{V_T}{V_E} = \log_{10} \left[\frac{V_T}{V_E}\right] \times 1{,}15129$$

$$n_1 = n-1\ ;\ n_2 = (n-1)\ (m-1) \text{ quando } V_T > V_E$$

Emquanto que o caracter das replicações, tomado no seu conjuncto, é medido do modo como vemos, é obvio que se podem determinar as interacções existentes entre ellas.

Tal é, em suas linhas geraes, o novo methodo originado pelo Dr. R. A. Fisher,

Acreditando que com a divulgação desse methodo, entre nós, estamos prestando um serviço de interesse collectivo, especialmente aos nossos collegas, sem nenhuns intuitos, é claro de desmerecer o valor dos experimentos do professor Teixeira Mendes mas, antes de tudo, attendendo ao seu proprio appello, foi que nos resolvemos a escrever as linhas acima.

Rio, Junho, 1931.

BIBLIOGRAPHIA

1) CLAPHAM, A. R. — The estimation of yield in cereal crops

by sampling methods. The Journal of Agricultural Science, vol. XIX, Part II, 1923.
2) FISHER, Dr. R. A.—Statistical Methods for Research Workers, 3.ª ed., 1930.
3) HOTELLING, Dr. H. — Recent improvments in statistical inference. Proceedings of the American Statistical Association, vol. XXVI, n. 173 A, Março 1931.
4) BEAVEN, Prof. E. S. — Trials of new varietes of cereals. Jour. of the Ministry of Agriculture, vol. XXIX, ns. 4 e 5, 1922.
5) WISHART, Dr. J. — The analysis of variance illustrated in its application to a complex agricultural experiment ou sugar beet. Archiv fur Pflanzenbau, 1931.
6) FISHER, Dr. R. A. — On the influence of rainfall on the yield of wheat at Rothamsted, Philosophical Transactions of ihe Royal Society of London, 1925.
7) FISHER, Dr. R. A. — On the mathematical foundations of theorical statistics. Philosophical Transactions of the Royal Society, of London, 1922.
8) WISHART, Dr. J. e CLAPHAM, A. R. — A study in sampling technique: the effect of artificial fertilisers on the yield of potatoes. Jour. of Agricultural Science, vol. XIX, part IV, 1929.

VERDADES...

Exodo da terra, pauperismo, falta de trabalho, tudo isso tem origem na mesma causa: os privilegios de que goza a grande industria, privilegios que são pagos pela agricultura e pela pequena industria e em ultimo lugar pela massa. Para diminuir a desocupação o pauperismo e outros males, é preciso restabelecer uma proporção entre a produção e o consumo, entre a industria e a agricultura.

A agricultura é a unica actividade verdadeiramente produtora, a unica indispensavel á vida, verdadeiramente moral e moralizadora. Mais do que uma industria, é uma arte verdadeiramente completa, que desenvolve tanto os musculos como o espirito de organização, a atividade e a imaginação.

No Japão o pauperismo era desconhecido antes da implantação da grande industria.

GINA LOMBROSO