Uso dos testes estatísticos na experimentação agrícola

Armando Conagin
Divisão de Experimentação e Pesquisas (Instituto Agronômico)
Secretaria da Agricultura

I — INTRODUÇÃO

A origem e evolução de certas plantas cultivadas é ainda debatida e obscura.

Sabe-se, no entanto, que muitas das espécies hoje em cultivo, já existiam há centenas de anos; certas práticas agrícolas, são também, de origem muito antiga. Como exemplo, citaremos o fato dos índios americanos procurarem produzir espigas de milho muito heterozigotas, visando assim, obter maior produção (3).

Mas, ao lado dessas formas vegetais e dessas práticas agrícolas, cuja idade pode ser medida por séculos, outras formas vegetais e novos processos agrícolas entraram em uso.

Uma evolução gradual, operou-se na Agricultura, evolução essa que se tem intensificado muito, ultimamente.

Plantamos hoje variedades mais produtivas, mais precoces, menos suscetíveis às moléstias e de valor nutritivo superior às usadas no passado.

Procuramos dar ao solo um excelente preparo, utilizamos sementes selecionadas e colhemos tanto quanto possível, mecânicamente.

Visamos com isso o aumento e o barateamento da produção.

Isso só foi possível depois do desenvolvimento da Biologia e das outras ciencias em geral, que criaram condições para um posterior desenvolvimento da Cito-Genética e da Estatística, imprescindíveis ao fitotecnista.

Os melhoradores de plantas, usando os modernos métodos fitotécnicos, estão produzindo novas variedades e híbridos, superiores às variedades antigas.

Milhares de seleções são feitas anualmente e destas, são conservadas aquelas que demonstram maior produtividade, precocidade, resistências às moléstias, etc.

Antes dessas novas formas serem recomendadas, elas devem ser comparadas com as mais antigas, cultivadas na região.

Essa comparação deve ser feita, de modo a permitir o julgamento das diferenças, se estas existem; ensaios de competição são então instalados de acôrdo com planos experimentais os mais eficientes e os dados obtidos são depois analisados estatisticamente.

Se as formas selecionadas conseguem sobrepujar as mais antigas, elas são então recomendadas aos lavradores.

Em outro artigo (1), já salientamos as dificuldades do contrôle experimental em Biologia e o inestimável auxílio que a Estatística vem prestando ùltimamente, a todos os ramos da experimentação biológica.

Quando nós estamos experimentando um grupo de variedades de trigo, amendoim ou milho, pretendemos chegar ao fim do experimento, com uma série de informações, obtidas a partir dos dados de produção, que nos permitam conhecer qual foi o comportamento dessas variedades e o que razoàvelmente poderemos esperar com relação a futuras experimentações da mesma natureza.

O que desejamos é, a partir dos dados obtidos, fazer inferências e tirar conclusões.

A estatística nos permite, a partir de resultados parciais concluir sôbre bases mais gerais.

Para isso, é preciso aplicar testes estatísticos.

Precisamos, então, conhecer como são efetuados os testes estatísticos e com que segurança neles podemos confiar.

Para tanto, é necessário saber quais são os valores possíveis como variam e com que frequência êsses valores são encontrados. Precisamos, em síntese, conhecer a distribuição dêsses valores.

II — A DISTRIBUIÇAO NORMAL

A curva normal é conhecida há mais de duzentos anos. Os matemáticos acharam que ela representava muito bem os êrros de observação nas ciências físicas.

A expressão matemática da distribuição normal foi descoberta por De Moire (4).

A experiência tem mostrado que ela explica satisfatòriamente a distribuição de diversos fenômenos biológicos, sociológicos, educacionais, etc.

Se temos uma variável que pode ter muitos valores possíveis, e se dizemos que essa variável é normalmente distribuida, o que queremos dizer é que os resultados possíveis dessa variável estão de acôrdo com a distribuição normal.

Essa distribuição pode ser analiticamente representada por:

$$y = f(x) = \frac{1}{\sigma \sqrt{2\pi}} \, e^{-\frac{1}{2} \frac{(X - \mu)^2}{\sigma^2}}$$

onde:

e = 2,7183 (base dos logarítmos naturais).

π = 3,1416

σ = desvio padrão da população

μ = média da população

X = valor da variável

Os valores possíveis de X na distribuição normal estão compreendidos entre $-\infty$ e $+\infty$.

A equação (1) representa analiticamente uma curva simétrica em tôrno do valor μ em forma de sino e asintótica em relação ao eixo dos X.

Essa equação é bastante simplificada, se fizermos a transformação:

$$p = \frac{X - \mu}{\sigma}$$

Então :

$$f(p) = \frac{1}{\sqrt{2\pi}} e^{-\frac{p^2}{2}}$$

Essa curva é simétrica ao redor do ponto p igual a zero, com variância um : ($\sigma^2 = 1$).

Em Cálculo, aprendemos que a área compreendida entre uma curva f(p) e o eixo das abcissas, é dada pela integral.

$$\int f(p).dp$$

Como no nosso caso os valores de p variam de $-\infty$ a $+\infty$, a área total, será entre êsses limites, abrangendo todos os valores possíveis da variável.

Sendo :

$$\int_{-\infty}^{+\infty} f(p).dp = 1$$

O valor da integral entre os limites a e b,

$$\int_a^b f(p).dp$$

representará a probabilidade de obtermos os valores de X maiores que a e menores que b.

Como a curva é simétrica ao redor do ponto p = 0 (zero), temos :

$$\int_{-\infty}^{0} f(p).dp = \int_{0}^{\infty} f(p).dp = 1/2 = 0{,}5000$$

Essa integral foi tabelada e poderemos encontrar nas tabelas, que :

$$\int_{0}^{1{,}96} f(p).dp = 0{,}4750 \quad \text{ou :} \quad \int^{2{,}57} f(p).dp = 0{,}4949$$

Sabemos que :

$$\int_{0}^{\infty} f(p).dp = \int_{0}^{1.96} f(p).dp + \int_{1{,}96}^{\infty} f(p).dp = 0{,}5000$$

Portanto, se quizermos calcular só uma parte dêsses valores, temos :

$$\int_{1{,}96}^{\infty} f(p).dp = 0{,}5000 - \int_{0}^{1{,}96} f(p).dp = 0{,}0250$$

Considerando só a metade positiva da curva de distribuição, temos que, os valores que se afastam da origem no eixo das abcissas mais que 1,96, são encontrados na proporção de 2,5%.

Se tomarmos os valores maiores que 1,96 e os menores que — 1,96 (portanto valores maiores que 1,96 em valor absoluto), êsses valores serão encontrados na proporção de 5% da população, sendo 2,5% na extremidade negativa e 2,5% na extremidade positiva (figura 1).

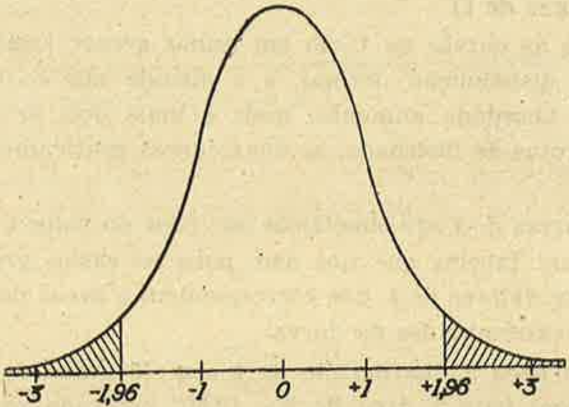

Fig.1 Curva normal em unidades de desvio padrão

III — A DISTRIBUIÇÃO DE t

A diferença entre duas médias de amostra, medida em unidades da estimativa do desvio padrão da diferença, entre as médias, distribui-se de acôrdo com a distribuição de **t** de Student

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{diferença}}$$

onde Sdif. é estimativa do desvio padrão da população de diferenças.

Essa distribuição de **t** depende do número de graus de liberdade utilizado na obtenção da estimativa do desvio padrão da diferença. Assim, para cada grau de liberdade teremos uma curva representativa de **t**.

Essas distribuições têm tôdas, certos característicos comuns e, para graus de liberdade maiores de 14, encontramc mais de 99% da área da curva, compreendidos entre — 3 e + 3 (valores de **t**).

Tôdas as curvas de **t** são em linhas gerais, semelhantes à curva da distribuição normal, e à medida que o número de graus de liberdade aumenta, mais e mais dela se aproxima. Para 50 graus de liberdade, as duas curvas pràticamente se superpõem.

As curvas de **t** são simétricas ao redor do valor $t = 0$.

Existem tabelas que nos dão, para os vários graus de liberdade, os valores de **t**, que correspondem a áreas de 5% e 1% nas duas extremidades da curva.

Assim para a distribuição de **t** com 15 graus de liberdade encontramos para a área de 5% (2,5% em cada extremidade da curva), o valor $t = 2{,}13$.

Os valores de **t** são dados em valor absoluto e nos indicam que devem ser considerados valores menores que — 2,13 e maiores que 2,13 como no nosso exemplo, para representar a área de 5% da curva.

IV — A ANÁLISE DA VARIÂNCIA

Decomposição Entre — Dentro

Já vimos em outro artigo, que as estatísticas, como por exemplo a média aritmética, o desvio padrão, etc., são valores determinados a partir de amostras; isto é, são estimativas dos parâmetros. Parâmetros são valores que caracterizam a população; na distribuição normal temos dois parâmetros que são : μ e σ.

Também já foi feito o estudo comparativo de duas amostras.

Vejamos como efetuar um estudo comparativo, para um número maior de amostras.

Suponhamos três amostras — A, B e C, cada uma, composta de 4 valores, Sejam os valores que caracterizam essas amostras, respectivamente

Amostra A	Amostra B	Amostra C			
(1)	(2)	(3)	(4)	(5)	(6)
$X_{11} = 3$	$X_{21} = 8$	$X_{31} = 9$	9	64	81
$X_{12} = 5$	$X_{22} = 5$	$X_{32} = 7$	25	25	49
$X_{13} = 3$	$X_{23} = 11$	$X_{33} = 7$	9	121	49
$X_{14} = 1$	$X_{24} = 8$	$X_{34} = 5$	1	64	25
$\Sigma X_{1j} = 12$	$\Sigma X_{2j} = 32$	$\Sigma X_{3j} = 28$	44	274	204

$\bar{X}_1 = 3{,}00$ $\qquad$ $\bar{X}_2 = 8{,}00$ $\qquad$ $\bar{X}_3 = 7{,}00$

$$\text{média geral } \bar{\bar{X}} = \frac{\sum\limits_{i}^{3} \sum\limits_{j}^{4} X_{ij}}{12} = \frac{12+32+28}{12} = 6{,}00$$

É fácil provar que o desvio dos valores individuais em relação à média geral, pode ser decomposto em duas partes :

a) desvio dos valores individuais em relação às médias parciais de amostra, dando origem à variação dentro das amostras;

b) desvio das médias parciais de amostra, em relação à média geral, dando origem ao desvio entre amostras.

Assim, para um valor qualquer **X12**, podemos ter :

$$(X_{12} - \bar{\bar{X}})^2 = [(X_{12} - \bar{X}^1) + (\bar{X}_1 - \bar{\bar{X}})]^2$$
$$= (X_{12} - \bar{X}_1)^2 + (\bar{X}_1 - \bar{\bar{X}})^2 + 2\,(X_{12} - \bar{X}_1)\,(\bar{X}_1 - \bar{\bar{X}})$$

Fazendo a soma para os 12 valores individuais, teremos :

$$\sum_{i=1}^{3}\sum_{j=1}^{4}(\bar{X}_{ij} - \bar{X})^2 =$$

$$\sum_{i=1}^{3}\sum_{j=i}^{4}(X_{ij} - \bar{X}_i)^2 + 4\sum_{i=1}^{3}(\bar{X}_i - \bar{\bar{X}})^2,$$

pois a soma dos duplos produtos é nula A variação total foi decomposta em duas partes : variação dentro das amostras e variação entre amostras.

Nas somatórias acima, os valores de **i** variam de um até três, isto é, podem tomar os valores um, dois e três e correspondem ao primeiro índice dos valores de **X**; os valores de j podem ser um, dois, três e quatro.

(7)	(8)	(9)	(10)	(11)	(12)
— 3	2	3	0	0	+ 2
— 1	— 1	1	+ 2	— 3	0
— 3	5	1	0	+ 3	0
— 5	2	— 1	— 2	0	— 2
			0	0	0

Para calcular a variação total, nós somamos o quadrado, dos valores das colunas (7), (8) e (9), onde os valores da coluna (7), por exemplo, são as diferenças entre os valores da coluna (1) e a **média geral.**

Temos :

$$9 + 1 + 9 + 25 + 4 + 1 + 25 + 4 + 9 + 1 + 1 + 1 = 90$$

A variação total pode ser calculada de um modo mais simples, pois já foi visto que :

$$\sum_{i} \sum_{j} (X_{ij} - \bar{X})^2 = \sum_{i} \sum_{j} X_{ij}^2 - \frac{(\sum_{i} \sum_{j} X_{ij})^2}{12}$$

$$\text{Variação total} = 522 - \frac{72)^2}{12} = 522 - 432 = 90$$

O valor 522 é a soma das colunas (4), (5) e (6), que são os quadrados dos valores das colunas (1), (2) e (3).

a) A variação dentro de amostras é calculada, somando os quadrados das colunas (10), (11) e (12).

Os valores da coluna (10) são os desvios dos valores individuais da coluna (1) em relação à média parcial, $\bar{X}_1 = 3{,}00$

Efetuando os cálculos, temos :

$$\sum_{i} \sum_{j} (X_{ij} - \bar{X}_i)^2 = 34$$

b) A variação — entre amostras pode ser calculada assim : Os desvios das médias parciais em relação à média geral, são elevados ao quadrado, somados e o resultado multiplicado por quatro:

$$4\ [3^2 + 2^2 + 1^2] = 4 \times 14 = 56$$

mostrando-nos que,

Variação total (9) = Variação dentro (34) + Variação entre (56)

De maneira semelhante são decompostos os graus de liberdade.

Como temos 12 valores, à variação total corresponderão onze graus de liberdade.

Sendo três o número de médias parciais, ficamos com dois graus de liberdade para a Variância entre amostras

Os nove gráus de liberdade, restantes, serão os gráus de liberdade da Variância Dentro das Amostras.

Vamos representar a análise da Variância efetuada, em um quadro :

Fonte de Variação	Variação	Gráus de liberdade	Variância	Valor de F
Total	90,00	11	8,18	—
Entre amostras	56,00	2	28,00	**7,41***
Dentro das amostras	34,00	9	3,78	—

Quando nós efetuámos a análise da Variância, admitimos que as três médias de amostras faziam parte da mesma população de médias.

Para que a hipótese acima não fosse rejeitada, as médias parciais não deviam diferir estatìsticamente.

Para saber se elas diferem estatìsticamente ou não nós efetuamos o teste de F, que é igual ao quociente de duas variâncias. No numerador temos sempre a variância entre amostras e no denominador a variância dentro das amostras.

O resultado a ser esperado para os valores de 5% e 1%, estão tabelados, sendo que entramos na tabela para o nosso caso, com os graus de liberdade respectivamente, dois e nove.

Resultados de F maiores que os esperados para 5%, são considerados como significativos.

O resultado que obtivemos, que é 7,41 para os gráus de liberdade dois e nove, está compreendido entre 4,26 (esperado para 5%) e 8,02 (esperado para 1%).

O resultado, tendo sido significativo, rejeitamos a nossa hipótese inicial, de que as amostras pertencem a uma mesma população de médias. Elas possuem, portanto, médias diferentes.

Se as três amostras, A, B e C tivessem sido respectivamente três variedades de milho, Amparo, Cateto e Armour e os dados representassem número de espigas por canteiro, nós poderíamos efetuar um julgamento, quanto ao número de espigas produzidas pelas variedades.

Vimos que cada uma das amostras pode ter seu desvio dentro, respectivo, por exemplo :

$$S_A{}^2 = \frac{\Sigma (X_A - \bar{X}_A)^2}{3} = 8/3 = 2{,}67$$

$$S_B{}^2 = \frac{\Sigma (X_B - \bar{X})_B{}^2}{3} = 18/3 = 6{,}00 \qquad \text{e, portanto :}$$

$$S_A = \sqrt{2{,}67} = 1{,}29 \; S_B = \sqrt{6{,}00} = 2{,}45 \text{ e } S_C = 1{,}29$$

Mas, desde que êsses três desvios possam ser considerados como estimativas de um único desvio padrão, a melhor estimativa é aquela dada por :

$$S_D{}^2 = \frac{\Sigma (X_A - \bar{X}_A)^2 + \Sigma (X_B \quad \bar{X}_B)^2 + (X_C - \bar{X}_C)^2}{9}$$

$$\text{ou} \qquad S_D = \sqrt{34{,}00/9} = 1{,}94$$

pois essa estimativa é determinada com maior precisão.

Se quizermos efetuar um estudo das diferenças entre médias de amostras, precisamos calcular :

$$t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{2 S_D{}^2/_4}}$$

onde i ou j podem ser A, B ou C.

Pàra calcular, tomamos t com nove graus de liberdade, que é o número de graus de liberdade correspondente ao desvio dentro das amostras SD; para o limite de 5%, o nosso t vale 2,01.

Então se $\bar{X}_i - \bar{X}_j > 2{,}01 \sqrt{2\,(3{,}78)/4}$ ou :

$\bar{X}_i - \bar{X}_j > 2{,}01 \;.\; (1{,}38)$ isto é :

se : $\bar{X}_i - \bar{X}_j > 2{,}77$ a diferença será significativa.

Duas daquelas médias que diferirem por mais de 2,77, serão estatìsticamente diferentes.

Então $\bar{X}_B$ e $\bar{X}_C$ não diferem; ambas são superiores a $\bar{X}_A$

V — INTERPRETAÇÃO DO RESULTADO DOS EXPERIMENTOS E SUA BASE RACIONAL

Por conhecer a distribuição dos valores, nós temos a possibilidade de avaliar a probabilidade com que podemos esperar, para um certo número de graus de liberdade, digamos nove, um valor de t maior ou igual a 2,01. Vimos anteriormente que essa probabilidade era de 5%. Do mesmo modo, podemos [illegible]abilidade de obter valores da variável, maiores que 1,96 por exemplo, na distribuição normal. Êsses valores nós encontramos na tabelas e no caso acima, a probabilidade é também de 5%.

A probabilidade de valores quociente de duas variâncias também pode ser determinada..

Temos, portanto, meios de avaliar os resultados possíveis dos testes estatísticos, efetuados com os dados obtidos no nosso experimento.

VI — TESTE DE SIGNIFICANCIA

No exemplo anterior comparamos três amostras A, B e C, cada uma com 4 repetições.

Dissemos que as três amostras poderiam ser lotes de três variedades de milho, Amparo, Armour e Cateto.

Quando efetuamos um estudo comparativo da produção das variedades, admitimos inicialmente que, estas são equivalentes quanto à produção.

Essa hipótese que nós fazemos e que vamos pôr à prova, é chamada "hipótese de nulidade".

Para pôr em prova essa hipótese efetuamos um teste como o de F

$$F = \frac{\text{variância entre}}{\text{variância experimental}}$$ e si F fôr significativo calculamos os valores de $t = \frac{\text{diferenças das médias}}{\text{desvio da diferença das médias}}$

e adotamos para t, o limite de 5%, considerando como significativos os valores de t maiores que os esperados pela tabela.

Qual a razão de procedermos assim ? Vimos que a curva de t é simétrica ao redor do valor t igual a zero e que os valores de t, pequenos, próximos ao centro da curva, são mais frequentes que os grandes.

Procuramos escolher regiões nas caudas da curva, pois essas regiões correspondem a diferenças maiores, pouco prováveis de acontecer, simplesmente por circunstâncias acidentais.

Quando calculamos as diferenças mínimas, tomamos o valor de $t = 2{,}01$, correspondendo ao limite de 5%, para nove graus de liberdade.

Calculamos as diferenças mínimas e obtivemos : 2,77.

Diferenças maiores que 2,77 serão consideradas significativas, pois essas diferenças serão obtidas por circunstâncias exclusivamente acidentais, com uma probabilidade de 5% (que é a probabilidade quando a hipotese de nulidade é verdadeira).

Tôda vez que julgamos um experimento, estamos sujeitos a dar uma interpretação errada.

Podemos ou rejeitar uma hipótese, quando ela é verdadeira, ou aceitar uma outra, quando esta é falsa. (5).

A Estatística nos permite escolher a frequência com que estamos dispostos a dar uma interpretação errada (2).

Se adotarmos o critério de 5%, é porque estamos dispostos a rejeitar a hipótese de nulidade, por exemplo, com ela sendo verdadeira, com uma frequência de cinco vezes em cem.

É facultado ao experimentador, ser mais ou menos rigoroso, quanto à probabilidade escolhida para a rejeição da hipótese formulada.

Se julgarmos 5% um limite pouco rigoroso, poderemos escolher outros limites como 1% ou 1 por mil.

Em Agricultura, bons resultados têm sido obtidos com a adoção de limite de 5% e êste é o limite geralmente aceito.

Êsse critério (5% ou 1%), deve ser fixado, antes da realização do experimento.

Ao efetuarmos um experimento, temos duas classes de resultados possíveis :

1) Aqueles resultados que se mostram bastante diferentes dos que deveríamos esperar se nossa hipótese fosse correta, e que determinarão a rejeição da hipótese formulada. Êstes, são os resultados, que têm pequena probabilidade de serem devidos, exclusivamente, às circunstâncias acidentais.

2) Os resultados obtidos têm grande probabilidade de serem devidos ùnicamente às circunstâncias acidentais.

Neste caso "os dados do experimento não permitem a rejeição da hipótese formulada".

VII — RESUMO DO JULGAMENTO DE UM EXPERIMENTO E CONCLUSÕES

Se estamos efetuando um experimento em que comparamos variedades de milho, por exemplo, começamos admitindo que as variedades são equivalentes quanto à produção.

Essa é a nossa hipótese de nulidade, isto é, é a hipótese que vai ser posta à prova no nosso experimento. Para tanto começamos investigando as variâncias dentro das amostras. Aplicamos um teste de homogeneidade das variâncias (Teste de Bartlett). Se foram homogêneas, efetuamos a análise da Variância (vêr quadro).

Aplicamos em seguida o teste de F

$$F = \frac{\text{Variância entre variedades}}{\text{Variância experimental}}$$

e se F fôr significativo como sucedeu no nosso exemplo, onde $F = 7{,}41$ a hipótese de equivalência das variedades será rejeitada.

Determinamos em seguida, as diferenças mínimas significativas entre as variedades, para classificá-las pela produtividade. Para isso aplicamos o t teste. Êsse teste é perfeitamente válido, mesmo quando as variáveis individuais não tenham uma distribuição perfeitamente normal; pois, mesmo assim, as médias de amostras tendem a uma distribuição normal, a medida que o tamanho da amostra cresce.

Como resultado da determinação das diferenças mínimas significativas, tivemos que $\overline{X}_B$ e $\overline{X}_C$ não diferiram estatisticamente, tendo sido, ambas superiores a $\overline{X}_A$.

Aqui termina a análise estatística.

É conveniente acentuar que os resultados obtidos em um único ano de experimentação são inseguros, pois sabemos da grande variabilidade das condições ecológicas. Só depois de alguns anos de experimentação, podemos ter idéias fidedignas com relação ao comportamento das variedades.

VIII — LITERATURA CONSULTADA

(1) CONAGIN, A — Algumas Noções de Estatística. Revista de Agricultura (Piracicaba) 22 págs. 119-134 (1947)

(2) FISHER, R. A. —Em The Design of Experiment, 4.a edição, pag. 1-240. Hafner — Publishing C°., New York, 1947.

(3) HAYES, H. K. e R. J. GARBER — Em Breeding Crop Plants, 2.a edição, pag. 1-438. Mc-Graw Hill C°., 1927.

(4) KENNEY, J. F. — Em Mathematics of Statistics, Vol. 1, pag. 1-248, 4.a impressão. D. Van Nostrand C°., New York, 1943.

(5) TIPETT, L. H. C. — Em The Methods of Statistics, 2.a edição, pag. 1-280, Williams and Norgate Ltda., Londres, 1937.

AGRADECIMENTO

Ao Dr. Constantino G. Fraga Jor., que nos deu preciosas sugestões, e à Srta. Elza S. Berquó, os nossos agradecimentos.