ALGUMAS NOÇÕES DE ESTATÍSTICA

Sua utilização na Experimentação Agrícola

A. Conagin

Divisão de Experimentação e Pesquisas

(Instituto Agronômico) — Secretaria da Agricultura

1 — INTRODUÇÃO

Quando efetuamos um experimento, chegamos ao seu final, com uma série de números que exprimem quantitativamente o fenômeno que estamos estudando.

Se nós pretendemos conhecer o efeito de um tratamento A, que devemos fazer ?

Para analisar a ação do tratamento A, sôbre um conjunto de indivíduos, precisamos comparar uma parte dêsse conjunto que sofreu a ação do tratamento A, com a outra parte qual se subtrairam as condições peculiares ao tratamento A

Aparece-nos agora uma pergunta :

Quantos indivíduos serão necessários, para obtermos uma boa informação acêrca da possível influência do tratamento

Será suficiente tomarmos um único indivíduo e submetê-lo à ação do tratamento A e depois, comparar o resultado d do com o de outro indivíduo ao qual aplicamos o tratamento comum ?

Evidentemente, não.

Para que tenhamos segurança em acreditar que o tratamento A exerce influência, é preciso que essa influência tenha sido constatada como um resultado positivo, para um certo número de indivíduos : isso porque alguns dos indivíduos que foram submetidos à ação do tratamento A, poderiam apresentar-se com um resultado superior ao do tratamento contrôle

e êsse resultado ser, não consequência do tratamento A, mas, exclusivamente, de circunstâncias acidentais.

Uma das grandes vantagens da Estatística é de, justamente poder calcular as várias probabilidades da superioridade do tratamento A sôbre o tratamento contrôle, como consequência, exclusivamente de circunstâncias acidentais e poder adotar para julgamento, certos limites que reduzem essas probabilidades de causas acidentais, a valores muito pequenos.

Suponhamos que em uma experiência de Física, dispomos de 2 cilindros idênticos aos quais estão adaptados pistões.

Enchemos êsses cilindros de um mesmo gás e os colocamos a uma mesma temperatura. Em seguida submetemos o primeiro cilindro a uma pressão dupla do segundo.

No fim de certo tempo verificamos que o volume do primeiro se reduz à metade do volume do segundo, portanto, que o volume dos gases está na razão inversa das pressões.

Não poderíamos ter a mesma confiança no resultado, se ignorássemos a natureza e a temperatura dêsses gases.

No nosso experimento tornamos constantes todos os fatôres que poderiam ter influência sôbre o volume dos gases e fizemos variar exclusivamente um dos fatores, cuja influência desejávamos conhecer, que foi a pressão.

O que nós fizemos, foi submeter o nosso experimento a um contrôle que é chamado contrôle experimental.

Êsse tipo de contrôle pode ser usado nos experimentos de Física Química e de outras ciências em que os fatores extranhos aos que queremos estudar podem ser mantidos constantes.

Entretanto em experiências biológicas, cada indivíduo, além da sua constituição hereditária própria e portanto, variável de indivíduo para indivíduo, está sujeito à ação das condições ecológicas — clima — solo, etc., cuja perfeita homogeneidade é muito difícil obter.

Para eliminar essas dificuldades, tornou-se necessário desenvolver métodos que permitam efetuar um contrôle, quando o contrôle experimental não é possível.

Uma nova modalidade de contrôle foi obtida com o auxílio da Estatística e é chamado contrôle estatístico.

Admitamos que queremos comparar duas variedades de milho A e B que podem ser, por exemplo, as variedades Cateto e Armour. Estamos interessados em saber se, com relação à produção em pêso dos grãos, as variedades são equivalentes, ou se uma delas é melhor.

Vamos então plantar um certo número de grãos de milho das duas variedades. Para julgar as variedades, vamos lançar mão de um característico "produção", que é o que mais nos interessa no momento.

Depois de obtidas as espigas de cada pé, estas são debulhadas e pesadas e seus pesos anotados.

Chegamos assim, ao fim do nosso experimento com um conjunto de dados que representam a produção por pé de milho, das duas variedades.

Se a produção total de 200 pés de milho da variedade A foi, digamos, de 20,0 kg de milho em grão e a da variedade B de 18,0 kg, poderemos considerar a variedade A como mai produtiva que a B ? Será confirmado em outras experiências o resultado que obtivemos ? As sementes que plantamos teriam sido uma boa amostra das respectivas variedades ? Teriam as condições de clima e solo sido idênticas para as duas variedades, ou teria uma delas sido beneficiada ?

A Estatística desenvolveu métodos experimentais e de análise que nos permitem um contrôle dos fatores acidentais, oferecendo-nos ainda soluções para as questões atrás mencionadas. Por permitir ainda, um julgamento dos resultados dos experimentos com uma certa segurança, dela se vêm valendo cada vez mais, as ciências experimentais.

HISTÓRICO

O têrmo Estatística, em inglês "Statistic", tem atualmente um conceito diferente daquele que possuiu no passado.

As palavras "Statistics" aparecem tôdas como derivadas, direta ou indiretamente, do vocábulo latino "Status" no sentido de "Estado Político".

No século XVIII os escritores alemães usavam o têrmo "Inquerito Estatistico", para designar os levantamentos que se fa-

ziam no interêsse do Estado como os referentes à população, nascimentos e mortes.

Nessa época, o modo de exposição dos dados obtidos era de natureza, preponderantemente descritiva.

Pouco a pouco, a brevidade das descrições e o caráter definido dos dados numéricos foram sendo mais apreciados, a condensação dos dados preferida e as descrições verbais, abandonadas.

Com o desenvolvimento da teoria das probabilidades, ela passou pouco a pouco a exercer uma função importante na interpretação do comportamento das estimantivas.

É difícil dizer em que época a palavra "Estatística" passou a ser empregada em seu conceito atual.

3 — DEFINIÇÕES

A palavra Estatística é usada em pelo menos dois diferentes sentidos.

a) Em um dêles, refere-se a uma apresentação sistemática dos dados quantitativos, sendo sua exposição, principalmente descritiva. O resultado dos censos de população de um país, pertencem a êste tipo.

b) O outro, refere-se ao conjunto de métodos que têm por objetivo a classificação e análise dos dados quantitativos. Existem os que preferem para êste último, a expressão "Métodos Estatísticos".

Métodos Estatísticos, são todos aquêles processos usados na obtenção e análise dos dados.

A Teoria Estatística é a Exposição dos Métodos Estatísticos. Nela encontramos o porque da utilização das várias fórmulas, a derivação dessas fórmulas, etc.

Essa Teoria Estatística é de natureza matemática.

4 — FÓRMULAS ELEMENTARES

A Estatística trata em geral, de um conjunto de números

Êsses números representam uma série de medidas, qualitativas ou quantitativas tomadas sôbre um número determinado de **indivíduos variáveis.**

Suponhamos o caso de um melhorador, que esteja a estudar pela primeira vez, duas variedades de milho A e B.

Para ficar conhecendo o seu material, êle fará ensaios de germinação, medições da altura da planta e da espiga, anotará o aparecimento das inflorescências e por fim, pesará as espigas, etc.

O experimentador, por mais minucioso que queira ser, estará cingido às medições de um número parcial das plantas que êle possui, pois, sua capacidade de trabalho é limitada.

Como êle não pode medir tôdas as plantas por serem estas muito numerosas, êle escolherá, ao acaso, um número menor de plantas, que constituirão uma amostra do material.

Por uma escolha ao acaso, devemos entender uma escolha feita de tal modo, que tôdas as plantas de que êle dispõe, tenham a mesma probabilidade de serem escolhidas para compôr a amostra em questão.

Vamos supôr, para facilitar, que êle tomou ao acaso, duas amostras, uma de A e outra de B, de 10 plantas cada uma (normalmente, o número de plantas de cada amostra é bem maior).

Suponhamos ainda, que os pesos obtidos para as espigas sejam os seguintes :

Amostra da var. A : 135, 140, 155, 140, 145, 140, 120, 130, 145 e 150.
Amostra da var. B : 110, 150, 145, 130, 155, 175, 105, 160, 130 e 140.

Em Estatística, nós chamamos de **População** a totalidade

dos indivíduos que representam certos caraterísticos fundamentais comuns a todos êles. No nosso caso, a população seria constituída por todos os pés de milho que apresentam as características que definem as variedades A e B, representadas na nossa **amostra.**

Como as medidas são tomadas dentro de um grupo de indivíduos semelhantes, porém, não idênticos, o pêso da espiga varia de um indivíduo para outro.

Se o número de variáveis medidas é muito grande, é impossível termos uma noção do significado que êsses dados encerram, antes dos mesmos serem relacionados, condensados, enfim, de um modo **compreensível.**

Para efetuarmos essa condensação, utilizamos-nos de uns tantos valores, geralmente chamados **estatísticos,** entre êles, a média e o desvio padrão. A média geralmente utilizada é a média aritmética, resultado do quociente da soma de todos os valores, pelo número **dêsses valores.**

$$\bar{x}_A = \frac{135 + 140 + 155 + \ldots + 145 + 150}{10} = \frac{1.400}{10} = 140{,}0$$

De um modo geral, podemos representar assim :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{1}^{n} X$$

Observando-se as amostras A e B, vemos que elas possuem a mesma média. Entretanto, se calcularmos os desvios de cada variável em relação à média, notaremos : (quadro abaixo, colunas 2 e 6).

AMOSTRA A

(1)	(2)	(3)	(4)
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x^2
135	— 5	25	18.225
140	0	0	19.600
155	+ 15	225	24.025
140	0	0	19.600
145	+ 5	25	21.025
140	0	0	19.600
120	— 20	400	14.400
130	— 10	100	16.900
145	+ 5	25	21.025
150	+ 10	100	22.500
Σ = + 1.400	= 0	= 900	= 196.900

AMOSTRA B

(5)	(6)	(7)	(8)
x	$(x-\bar{x})$	$(x-\bar{x})^2$	x^2
110	— 30	900	12.100
150	+ 10	100	22.500
145	+ 5	25	21.025
130	— 10	100	16.900
155	+ 15	225	24.025
175	+ 35	1.225	30.625
105	— 35	1.225	11.025
160	+ 20	400	25.600
130	— 10	100	16.900
140	0	0	19.600
	= 0	= 4.300	= 200.300

1) Os valores obtidos em B, diferem mais da média que os encontrados em A.

$$A \begin{cases} 155 - 140 = +15 \\ 120 - 140 = -20 \end{cases} \qquad B \begin{cases} 175 - 140 = +35 \\ 105 - 140 = -35 \end{cases}$$

2) A diferença entre o valor maior e o menor de cada amostra, é ainda maior na amostra B.

$$A \text{ ——— } 155 - 120 = +35$$
$$B \text{ ——— } 175 - 105 = +70.$$

Essa diferença, chamada **amplitude de dispersão** (em inglês, range), dá-nos uma idéia da variabilidade encontrada na amostra; é portanto, uma medida de dispersão; outras existem, ainda.

A primeira idéia que nos poderia ocorrer, seria a de determinarmos a soma dos desvios de todos os valores em relação à média. Entretanto, a soma algébrica dêsses valores é nula, isto é

$$\sum_{1}^{n} (x-\bar{x}) = \sum_{1}^{n} d = 0 \text{ (colunas 2 e 6).}$$

Poderíamos determinar a média dos desvios, não levando em consideração o sinal de cada desvio. Êsse valor é chamado **média dos desvios absolutos**, e poderia ser calculado assim :

$$\left| \bar{d} \right| = \frac{\sum_{1}^{n} |d|}{n}$$

Poderíamos obter êsses valores fàcilmente a partir das colunas (2) e (6).

$$\left| \bar{d}_A \right| = \frac{5+0+15+0+5+0+20+10+5+10}{10} = \frac{70}{10} = 7$$

$$\left| \bar{d}_B \right| = \frac{170}{10} = 17$$

Entretanto os matemáticos acharam que, a fórmula que melhor caraterizava a variabilidade seria a rais quadrada dos desvios médios quadrados.

$$\sigma = \pm \sqrt{\frac{\Sigma (x - \mu)^2}{N}} \quad (1)$$

onde μ e σ são parâmetros da população. Não podendo obter-se $\frac{\Sigma (x - \mu)}{N}$ pois, que $\Sigma (x - \mu) = 0$ (onde μ é a média da população), fazemos $\frac{\Sigma (x - \mu)^2}{N}$

Entretanto, como elevamos os desvios ao quadrado, para voltarmos aos valores simples, precisamos extrair a rais.

O nosso valor σ é chamado afastamento padrão ou desvio padrão. Vamos dar um exemplo do que seja um parâmetro : e relacão 1 : 1 de nascimento macho-fêmea nos mamíferos, cuja causa a Genética satisfatòriamente explicou, ao descobrir que os machos eram portadores dos cromossômios sexuais do tipo X Y e as fêmeas do tipo X X, é um parâmetro da população, pois é um valor que carateriza essa população.

O desvio padrão para uma amostra é calculado pela fórmula :

$$s = \pm \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} \quad (2)$$

Nessa fórmula, temos :

X = valor das variáveis.

$\bar{X}$ = média aritmética da amostra.

n = número de indivíduos da amostra.

n-1 = número de graus de liberdade.

Vamos ver porque dividimos pelo número de graus de liberdade :

Vimos que $x_1 + x_2 + \ldots + x_{10} = 10.\bar{x}$

Dêsses 10 valores, 9 dêles poderiam oscilar à vontade; o

décimo, teria que ser tal, que a soma dos 9 anteriores, mais o décimo, valeria o total de 1.400.

Há uma limitação para o décimo valor que é a de ter que totalizar o valor 1.400, quando somado aos outros nove.

Há na nossa amostra, ùnicamente uma limitação. Então, êsses nove valores que podem oscilar à vontade, constituem os nove valores livres de nossa amostra. Temos portanto, **nove graus de liberdade.**

"O número de graus de liberdade, é dado pelo número de variáveis, menos o número de estatísticas calculadas a partir dessas variáveis (1 média no nosso caso)".

Então:

$$s_A = \pm \sqrt{900/9} = \pm \sqrt{100} = \pm 10{,}0 \text{ gramas.}$$

$$x_E = \pm \sqrt{4300/9} = \pm \sqrt{477{,}7} = \pm 21{,}8 \text{ gramas.}$$

Ao calcularmos os desvios padrão das amostras, dividimos por n 1, quando originalmente, dividíamos por N.

Student, provou que dividindo por n-1, a estimativa **s** é uma estimativa mais segura de σ (sigma) da **população; di**-divindo por n-1, fica elimnado o pergo de s ser sitstemàticamente menor que σ

Além do mais, n-1 representa o número de graus de liberdade com que poderemos contar: pode-se provar fàcilmente, que a esperança matemática de s^2 é σ^2.

Vemos pelos valores sA e sB que o desvio padrão dá uma bôa noção da variabilidade das amostras. Pelo resultado obtido a vriabilidade da amostra B foi maior que a de A.

O desvio padrão que calculamos, é uma estimativa do desvio padrão da população, estimativa essa, que é tanto mais perfeita, quanto maior fôr o número de indivíduos existentes na amostra.

5 — FÓRMULA SIMPLIFICADA PARA O CÁLCULO DO DESVIO PADRÃO

O processo de calcular o desvio padrão, usando a fórmula

(2), é um tanto trabalhoso. Por uma simples transformação algébrica, poderemos provar que :

$$\sum_{1}^{n} (x-\bar{x})^2 = \sum_{1}^{n} x^2 = \frac{\sum_{1}^{n} X)^2}{n}$$

Prova : Vamos desenvolver os vários quadrados dos desvios e depois, somá-los :

$$(x_1 - \bar{x})^2 = x_1^{\,2} - 2\,x_1\,\bar{x} + \bar{x}^2$$
$$(x^2 - \bar{x})^2 = x_2^{\,2} - 2\,x_2\,\bar{x} + \bar{x}^2$$
$$(xn - \bar{x})^2 = xn^{\,2} - 2\,xn\,\bar{x} + \bar{x}^2$$

Somando-se teremos :

$$\sum_{1}^{n} (x-\bar{x})^2 = \sum_{1}^{n} x^2 - 2\,\bar{x}.\sum_{1}^{n} x + n\bar{x}^2$$

Mas,

$$\bar{x} = \frac{\sum_{1}^{n} X}{n}$$

donde :

$$\bar{x}^2 = \frac{(\sum_{1}^{n} x)^2}{n^2}$$

Substituindo, temos :

$$\sum_{1}^{n} (x - \bar{x})^2 = \sum_{1}^{n} x^2 - 2\,\frac{(\sum_{1}^{n} x)^2}{n} + \frac{(\sum_{1}^{n} x)^2}{n} \text{ ou :}$$

$$\sum_{1}^{n} (x - \bar{x})^2 = \sum_{1}^{n} x^2 - \frac{(\sum_{1}^{n} x)^2}{n}$$

Então, $S = \pm \sqrt{\dfrac{\sum_{1}^{n} (x - \bar{x})^2}{n-1}} = \pm \sqrt{\dfrac{\sum_{1}^{n} x^2 - (\sum_{1}^{n} x)^2/n}{n-1}}$

Calculamos já a média e o desvio padrão. A média é uma medida do tipo, ao passo que o desvio padrão é uma medida da dispersão dos dados, ao redor da média.

Nos nossos trabalho experimentais, a nossa média é baseada em um número limitado de observações.

Ela não é, portanto, o valor exato da média da população, e sim uma estimativa dessa média.

Se fizermos a determinação de várias amostras, retiradas ao acaso daquela população, nós obteremos as médias de amostras x1, x2. ..., xk, tôdas oscilando em tôrno da média da população. algumas dessas médias sendo maiores que // outras menores.

Dessas médias de amostras vão ser mais frequentes aquelas que diferirem pouco da média da população e menos frequentes as que apresentam desvios maiores.

Foi provado matemàticamente, que uma boa estimativa do desvio de nossa média de amostras, é obtida pela fórmula

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

onde s = desvio padrão da amostra

n = número de indivíduos da amostra.

Porisso, caraterizamos a produção de uma variedade pela sua média, mais ou menos um desvio.

$\bar{x}_A = 140{,}0 \pm 10{,}0/\sqrt{10{,}0} = 140{,}0 \pm 3{,}16$ gramas.

$\bar{x}_B = 140{,}0 \pm 21{,}8/\sqrt{10{,}0} = 140{,}0 \pm 6{,}89$ gramas.

6 — ESTUDO COMPARATIVO DE DUAS AMOSTRAS

Se admitirmos que essas duas amostras pertencem a uma mesma população, poderemos aplicar o "t teste de Student",

Sabemos que :

$$t = \frac{\text{Diferença}}{\text{Desvio da diferença}} = \pm \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_{\bar{x}_A}^2 + s_{\bar{x}_B}^2}}$$

Entretanto, como são poucos os graus de liberdade de cada amostra, o s da diferença é calculado por uma estimativa ponderada do êrro padrão.

Se uma das amostras, por exemplo, a amostra A, tem n1 valores e a B, n2 sendo n1 ± n2, o desvio ponderado é calculado assim :

$$sm = \pm \sqrt{\frac{\Sigma (x_A - \bar{x}_A)^2 + \Sigma (x_B - \bar{x}_B)^2}{n_1 + n_2 - 2}}$$

No nosso caso n1 = n2 = 10, a fórmula se simplificando :

$$t = \pm \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{2\ sm^2}{10}}} = \frac{\bar{x}_A - \bar{x}_B}{sm \sqrt{\frac{2}{10}}}$$

onde :

$$sm = \pm \sqrt{\frac{900 + 4.300}{18}} \sim 17{,}90$$

Para 18 graus de liberdade o t vale no limite de 5%, 2,10.

Portanto, se :

$$\bar{x}_A - \bar{x}_B \geq 2{,}10 \times 17{,}00 \sqrt{\frac{2}{10}} \text{ ou se}$$

$\bar{x}_A - \bar{x}_B \geq 15.96$, a diferença será significativa.

Ora, no nosso oaso, $\bar{x}_A = 140{,}0$ e $\bar{x}_B = 140{,}0$ também.

Não houve diferença entre as médias das amostras. Elas pertencem portanto, a uma mesma população. A conclusão estatística seria :

Os resultados encontrados não se opõem à admissão de que a hipótese de equivalência das variedades A e B seja verdadeira.

Admitamos agora, que as amostras A e B sejam caraterizadas :

$$\bar{x}_A = 140{,}0 \pm 3{,}16 \text{ gramas.}$$
$$\bar{x}_B = 160{,}0 \pm 6{,}89 \text{ gramas.}$$

A diferença será significativa como anteriormente, se ela fôr maior que 15,96. Fazendo a diferença das médias, teremos agora, 20,0 gramas.

A diferença das médias sendo significativa, rejeitams a hipótese de serem aquelas amostras pertencentes a uma mesma população.

Um teste estatístico, como o t teste, baseia-se em uma hipótese matemática, por meio da qual, a probabilidade de che-

garmos a uma conclusão falsa proveniente de uma série de observações, pode ser limitada a um determinado valor em probalididades.

No caso acima citado, a diferença ($\bar{x}B - \bar{x}A = 20$) foi maior do que a esperada pela tabela t para 18 graus de liberdade e para o limite de 5 %.

Só em menos de 5% dos casos, poderíamos considerar A e B como pertencente a uma mesma população.

Então, se considerarmos B diferente e superior a A, por ter maior produção, teremos mais de 95% de probabilidade de estarmos acertando, a nossa probabilidade de êrro sendo de menos de 5%.

Como vimos, a Estatística nos permite fazer um estudo comparativo entre amostras e julgar a natureza das diferenças encontradas.

Devido aos trabalhos de grandes matemáticos, como Fisher, Yates, Cochran, Neyman e outros, foi possíivel a obtenção de planos experimentais, desenvolvidos inicialmente para a Agricultura e mais tarde extendidos a outras ciências, com os quais um contrôle estatístico muito eficiente foi obtido.

A Estatística é portato, uma parte da Matemática aplicada, que deve ser cultivada com carinho pelos engenheiros-agrônomos e biologistas, principalmente experimentadores, e seu valor já é de há muito, reconhecido em todos os países que cuidam de uma experimentação racional.

Encerrando uma de suas conferências em Rothamsted em 1931, disse Sir A. D. Hall :

"Eu irei mais longe, e insisto em que tôda a investigação biológica, compreende uma consderação estatística dos resultados".

BIBLIOGRAFIA

G. UDNY YULE e M. G. KENDALL — An Introduction to the Theory of Statistics. 1940.

PAUL RIDER — An Introduction to modern Statistical Methods. 1939.

R. A. FISHER — Statistical Methods for Research Workers. 1944

D. D. PATERSON — Statistical Technique in Agricultural Research. 1939.

G. W. SNEDECOR — Statistical Methods. 1940.

AGRADECIMENTOS

Quero deixar aqui os meus sinceros agradecimentos ao Dr. Constantino G. Fraga Júnior e á Srta. Elza S. Berqué, pelas preciosas sugestões e revisão do texto.