Algumas Considerações Sôbre os Testes Estatísticos

J. M. POMPEU MEMORIA (*)

Em certos trabalhos experimentais, o objetivo é algumas vezes estimar um parâmetro; por exemplo, um agrônomo deseja saber qual é a produção média de uma linhagem de milho híbrido. Suponhamos que êle obteve, em quatro talhões de um are, o seguinte resultado em Kg: 48, 53, 57, 42. Êstes dados constituem uma amostra da população hipotética, constituida de um número infinito de produções de talhões de um are plantados da linhagem do milho híbrido. A estatística $\bar{X} = 50$ Kg/a, média da amostra é a estimativa do parâmetro μ, média da população. O interêsse do investigador é generalizar o resultado obtido, isto é, passar da amostra à população. E', entretanto, impossível dar êsse salto indutivo com perfeição. Tôda indução envolve um certo grau de incerteza; esta é medida em termos de probabilidade e é uma das finalidades dos métodos estatísticos fornecer a técnica apropriada para medí-la. Portanto, o valor $\bar{X} = 50$ Kg/a é de pequena significação se não for acompanhado de uma medida do êrro cometido na sua estimativa.

E' sabido que os dados do exemplo citado obedecem à distribuição normal de frequência. Suponhamos, a título de simplicidade, que a variabilidade do material é conhecida, seja $\sigma = 10$ Kg/a, o desvio padrão. Sendo σ o desvio padrão dos itens de uma amostra de tamanho n, o desvio padrão da média será $\sigma / \sqrt{n}$, no nosso caso 5 Kg/a. Por outro lado, a quantidade $y = (\bar{x} - \mu)\sqrt{n}/\sigma$ distribui-se normalmente com média nula e desvio padrão unitário. Se quisermos, pois, estimar a média não com um único valor, mas com um intervalo com 95% de confiança de incluir o verdadeiro parâmetro μ, devemos ter:

$$\int_{-1,96}^{1,96} e^{-y^2/2} \, \frac{1}{\sqrt{2\pi}} \, dy = 0,95 \quad (1),$$

isto é, $P\,(-1,96 < (\bar{x} - \mu)/5 < 1,96) = 0,95$ (2), onde P indica a "probabilidade de"

(*) Eng. Agrônomo, M. S., Prof. do Depto. de Engenharia Rural da ESAV.

Donde: $\mu > \bar{X} - 9{,}8$ e $\mu < \bar{X} + 9{,}8$.

Substituindo-se $\bar{X}$ por 50, obtém-se:

$$P\ (\ 40{,}2\ <\ \mu\ <\ 59{,}8\ ) = 0{,}95\quad (3)$$

Foram assim obtidos o limite inferior 40,2 e o limite superior 59,8. O significado de (3) deve, entretanto, ser examinado com mais cautela. Parece, à primeira vista, que μ é variável e que isto implica que a probabilidade de μ estar entre 40,2 e 59,8 é de 95%. Naturalmente que esta conclusão não tem sentido, pois, μ é um valor fixo — a média da população, e está ou não entre 40,2 e 59,8. A probabilidade refere-se ao intervalo aleatório $\bar{X} - 9{,}8$ a $\bar{X} + 9{,}8$, cuja probabilidade de incluir μ é de 95%. Se fossem obtidas várias amostras de 4 ítens e se êsse intervalo fôsse construido em cada uma delas, então, seria de esperar que, em 95% dos casos, os intervalos obtidos contivessem a verdadeira média, isto é, erraríamos apenas 1 vez em 20. Isto é ilustrado gràficamente na fig. 1.

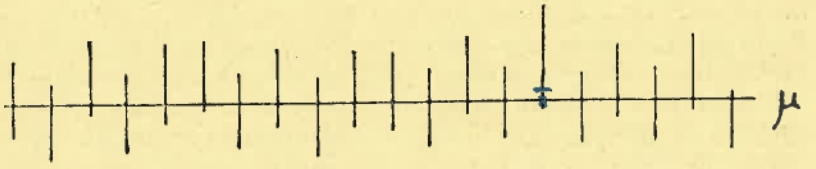

Fig. 1

Temos, portanto, uma confiança considerável de que o intervalo 40,2 a 59,8 contenha μ, e a medida da nossa confiança é 95% de probabilidade, antes da amostra ter sido tirada. Em (2) temos uma verdadeira probabilidade, mas em (3) temos uma medida da nossa confiança. Por isso, Fisher propôs denominá-la "probabilidade fiducial" P_F:

$$P_F\ (\ 40{,}2 < \mu < 59{,}8\ ) = 0{,}95\ .$$

Os limites 40,2 e 59,8 são chamados limites fiduciais. Poderíamos obter limites fiduciais com qualquer grau de confiança — 75%, 90%, 99%, etc.

Com a mesma probabilidade fiducial de 95% outros intervalos poderiam ser obtidos, como o indicado na fig. 2:

$$\int_a^b f(y)\,dy = 0{,}95$$

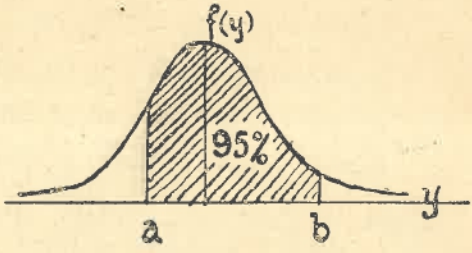

Fig. 2

Ordinàriamente deseja-se que o intervalo seja o menor possível. Para uma determinada área, a distância $b - a$ é mínima quando $f(b) = f(a)$ e como $f(y)$ é simétrica em relação a $y = 0$, o menor valor de $b - a$ ocorre quando $b = -a$ que, no nosso exemplo, é $b = 1{,}96$ e $a = -1{,}96$.

Convém, entretanto, salientar que o exemplo citado tem certa artificialidade. O caso natural é ser desconhecida a variabilidade do material. Nesta situação o desvio padrão deverá ser estimado com os dados da amostra. Quando σ é conhecido, ou a amostra é suficientemente grande de modo que êle possa ser estimado com êrro desprezível, pode-se utilizar a curva normal para se determinar os limites fiduciais de μ. Mas, no caso de amostras pequenas, isto pode acarretar grande êrro. Deve-se ao estaticista inglês Gosset, que escreveu sob o pseudônimo de Student, o engenho empregado para se determinar os limites fiduciais de μ para pequenas amostras. Student mostrou que a quantidade $(\bar{X} - \mu)\sqrt{n}/s$, distribui-se segundo a distribuição de t com $n - 1$ graus de liberdade, sendo $s = \sqrt{S\,x^2/(n-1)}$ a estimativa de σ, $x = X - \bar{X}$ e S o símbolo de somatório. Temos então que, no nosso exemplo: $s = 6{,}5$ e $(\bar{X} - \mu)/3{,}25$ obedece à distribuição de t com 3 graus de liberdade.

$$\int_{-t_{0,05}}^{t_{0,05}} f(t)dt = 0{,}95 \quad (4)$$

O valor $t_{0,05}$ é o valor tal que:

$$\int_{t_{0,05}}^{\infty} f(t)\,dt = 0{,}025 \quad (5),\ t_{0,05} \text{ com 3 graus de liberdade} = 3{,}18$$

e desde que $f(t)$ é simétrica em relação a

$t = 0$, (6) dá o menor intervalo com 95% de confiança: $P(-3{,}18 < (\bar{X} - \mu) / 3{,}25 < 3{,}18) = 0{,}95$ (6).

A distribuição de $\underline{t}$ aproxima-se da normal à medida que $\underline{n}$ cresce; daí o uso desta distribuição quando a amostra é grande. Com (6) obtem-se: $P_F(39{,}7 < \mu < 60{,}3) = 0{,}95$. Usando-se consistentemente 95% de probabilidade como medida de nossa confiança para estimar parâmetros, e declarando-se que o intervalo obtido contém o verdadeiro parâmetro, pode-se esperar estar errado em 5% das vezes ou seja 1 em cada 20.

Consideremos agora uma outra situação, aliás mais comum em experimentação do que a mera estimativa de parâmetro. Suponhamos que uma companhia produtora de sementes de milho híbrido deseja aconselhar o plantio de uma certa linhagem, sòmente se a produção média for igual ou maior que 50 Kg/a. Em outras palavras, a companhia de sementes deseja verificar a hipótese de que a produção média da linhagem de milho híbrido é proveniente de uma população de média igual a 50 Kg/a; isto é, se qualquer diferença existente entre a média encontrada e 50 deve ter sido ocasionada pelo acaso. Fisher denomina isso de verificação da "hipótese nula". Na verdade a hipótese nula nunca é finalmente provada, mas possívelmente refutada pela experimentação. O teste estatístico de significância é o meio pelo qual se obtém a probabilidade de que qualquer diferença igual ou maior que a observada tenha sido ocasionada pelo acaso, se realmente não houver diferença. No nosso exemplo, a companhia está apenas interessada em saber se a amostra sugere que os dados são provenientes de uma população de média inferior a 50, pois, se for superior, ela nada tem a perder, na sua reputação. Se a hipótese é rejeitada porque uma amostra ($n = 4$), tirada ao acaso, sugere uma produção menor que 50 Kg/a, a companhia não recomendará o plantio da linhagem, mas o aconselhará, se for 50 ou maior.

Na verificação de uma hipótese, dois tipos de erros podem ocorrer:

Êrro I — Rejeitar uma hipótese verdadeira.

Êrro II — Aceitar uma hipótese falsa.

No nosso exemplo, o êrro I consiste em rejeitar a hipótese, na base da amostra obtida, sendo esta realmente proveniente de uma população de média 50. O êrro II consiste em aceitar a hipótese quando, na verdade, a amostra é proveniente de uma população de média diferente de 50. O caso ideal seria tornar mínimo ambos os erros. Isto, entretanto,

não é possível. O êrro I pode ser feito tão pequeno quanto se queira. A escolha do seu valor mínimo é puramente arbitrária, mas já se tornou convencional os limites de 5% e 1%, consagrados pelo uso, respectivamente conhecidos por nível de "significância" e de "alta significância". A escolha dêsse nível depende em grande parte da natureza do problema. Por exemplo, se é feita uma competição de variedades de milho para verificar qual a que deve ser aconselhada, no caso de nenhuma delas acarretar mudanças nos tratos culturais ou de qualquer outra natureza, pode-se usar um nível de 10% ou maior. Já em pesquisas de certos tratamentos médicos perigosos podem-se usar valores tão baixos quanto 1‰. Convém salientar que a diminuição do êrro I: rejeitar uma hipótese verdadeira, aumenta a probabilidade do êrro II: aceitar uma hipótese falsa.

Suponhamos que no nosso exemplo foi escolhido 5% como margem do êrro I — rejeitar uma hipótese verdadeira. Admitamos também que a variabilidade do material é conhecida, seja $\sigma = 10$ Kg/a, donde $\sigma/\sqrt{n} = 5$ Kg/a para $n = 4$. Com a rejeição de 5% podemos construir várias regiões críticas, mas consideraremos apenas as três da Fig. 3.

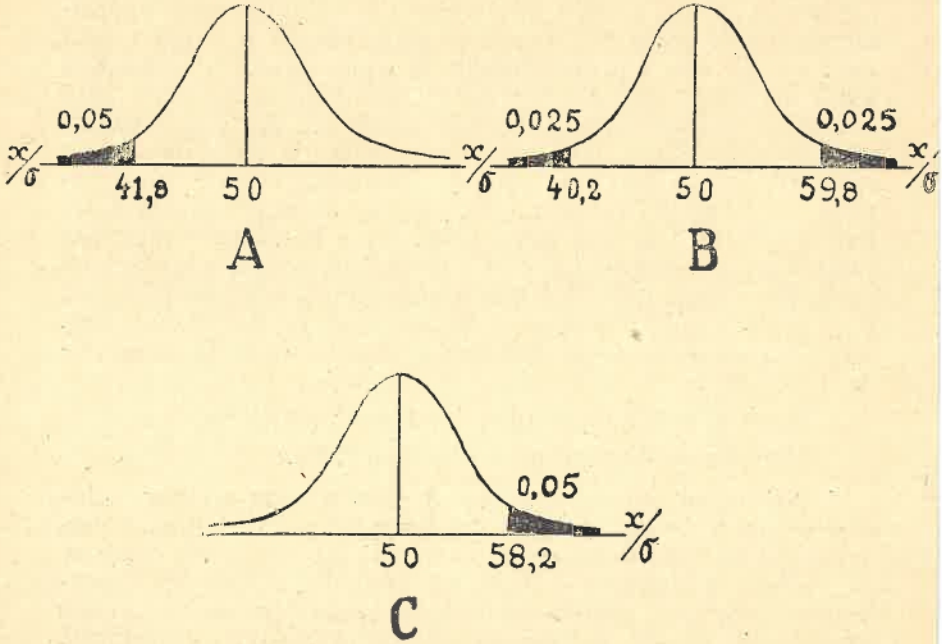

Fig. 3

Essas três regiões apresentam o mesmo êrro I. A região de rejeição mais conveniente é aquela em que, para um determinado êrro I, acarreta o menor êrro II — aceitar uma hipótese falsa.

Em *A* qualquer valor da amostra abaixo de 41,8 rejeita $\mu = 50$ mas aceita para valores maiores que 41,8. O limite mínimo 41,8 é igual a 50 — 1,64x5. A razão disso reside no fato de 1,64 vezes o desvio padrão a infinito incluir 5% da área sob a curva normal padrão. Em *B* a hipótese é aceita para os valores entre 40,2 a 59,8, e rejeitada para valores menores e maiores, respectivamente. Note-se que de 1,96 vezes o desvio padrão até o infinito inclui 2,5% da área da curva. Finalmente, em *C* a rejeição da hipótese se dá para valores acima de 58,2, sendo aceita para qualquer valor inferior a 58,2. Ordinàriamente uma região crítica será a melhor se a população diferir da hipótese num certo sentido, enquanto que outra região será preferível se a divergência for no sentido oposto. Se o pesquisador está interessado na diferença em qualquer sentido, a região crítica deve ser bilateral.

Analisemos as três regiões da Fig. 3 em relação ao problema que temos em vista. A melhor região será aquela em que para o êrro I de 5% tornará mínimo o êrro II. A companhia de sementes deseja evitar a aceitação da hipótese $\mu = 50$ quando, na verdade, μ for menor, porém não importa aceitar esta mesma hipótese, quando μ for maior. As Figs. 4, 5 e 6 ilustram gráfica e intuitivamente o êrro de aceitar $\mu = 50$ quando for na verdade 45, usando-se respectivamente as três regiões críticas da Fig. 3. A área hachuriada mostra a probabilidade de se aceitar a hipótese falsa, isto é, de que μ é igual a 50, quando na verdade, é 45. Vê-se facilmente que êste risco é menor na Fig. 4, quando se utiliza a região *A* da, Fig. 3. Esta é a região crítica a ser usada, para se verificar a hipótese de μ ser igual a um certo valor μ_1 relativo a alternativa de $\mu < \mu_1$.

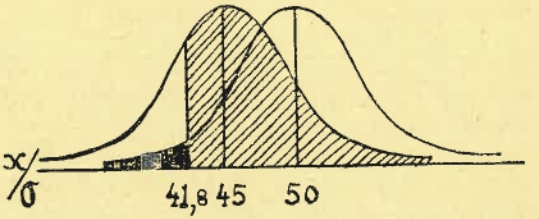

Fig. 4.

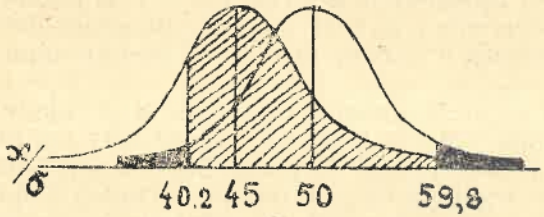

Fig. 5

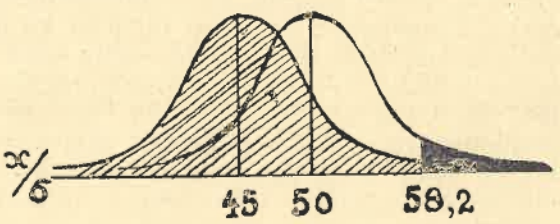

Fig. 6

Êste tipo de problema é comum na indústria, quando um fabricante deseja aconselhar um novo processo de manufatura, sòmente se êle é superior ao processo comumente usado. A região *C* seria a aconselhada se fôsse verificada a hipótese μ de ser igual a um certo valor μ' relativo à alternativa de $\mu > \mu'$. Finalmente *B* seria a região mais eficiente se a companhia de sementes fôsse indiferente a superestimar ou subestimar a produção média da linhagem de milho híbrido.

O tamanho da amostra é de grande importância, pois, quanto maior a amostra menor é o êrro padrão, e, portanto, mais próximos da média estão os limites da região crítica. Disto resulta uma menor probabilidade de ser aceita uma certa hipótese quando esta não é verdadeira. Se σ não é conhecido, ao invés da curva normal seria usada a distribuição de t que goza de propriedades análogas relativamente à escolha das regiões críticas.

A distribuição de t é também usada para verificar se duas médias podem ser consideradas como oriundas da mesma população. Nesta situação, um valor significante de t pode ser atribuído à diferença de médias ou à diferença de variâncias (quadrado do desvio padrão) ou a ambas. Convém

salientar aqui que, o teste de *t* só é válido sob a condição de ser verdadeira a igualdade das variâncias.

Um resultado não significante não exprime, necessàriamente, que a hipótese nula seja verdadeira. E' conveniente sempre construir os limites fiduciais para um certo grau de confiança, por exemplo, 95%. Muitas pessoas assumem uma atitude tirânica em relação aos testes de significância, aceitando sem restrições a hipótese, se o resultado não é significante, e rejeitando, no caso contrário. Isto revela uma apreciação inadequada sôbre amostragem (sampling). Não é finalidade dos testes estatísticos provar ou refutar hipóteses. Sòmente o pesquisador através de uma crítica lógica e sensata dos seus dados experimentais e dos de outros investigadores poderá fazer a decisão final. Os métodos estatísticos fornecem apenas os meios de medir o grau de incerteza dessas conclusões.