Entendendo o Valor p

Texto extraído e adaptado de David Salsburg (2009)[1]

Ronald Aylmer Fisher (1890-1962) desenvolveu a maioria dos métodos de teste de significância que hoje tem uso geral e referiu-se à probabilidade que permite declarar a significância como o "**valor de p**". R. A. Fisher mostra como calcular os valor de p em seu livro *Statistical Methods for Research Workers* (1ª edição publicada em 1925).

R. A. Fisher considerava que valores de p altos (um fracasso em encontrar significância) indicavam a inadequação dos dados para se chegar a uma decisão.

O que representa ter um resultado não significativo em um teste de significância? Egon S. Pearson (1895-1980) e Jerzy Neyman (1894-1981) investigaram essa ampla questão. Pearson e Neyman exploraram vários paradoxos que surgiram dos testes de significância, casos em que a utilização impensada de um teste de significância levava à rejeição de uma hipótese obviamente verdadeira. Fisher considerava que os testes de significância estavam sendo aplicados incorretamente. Assim, Neyman perguntou que critérios vinham sendo usados para decidir quando um teste de significância era aplicado corretamente. De modo gradual, Pearson e Neyman desenvolveram as ideias básicas dos testes de hipóteses.

Uma versão simplificada do teste de hipótese pode ser encontrada atualmente em todos os livros didáticos elementares de estatística. Quando a formulação de Neyman-Pearson é ensinada nessa versão simplificada do que Neyman de fato desenvolveu, ela distorce suas descobertas ao concentrar-se nos aspectos errados da formulação. Sua maior descoberta foi a de que os testes de significância não faziam sentido a não ser que houvesse pelo menos duas hipóteses possíveis. Não se pode, portanto, testar se os dados se ajustam a uma distribuição a não ser que exista alguma outra distribuição ou conjunto de distribuições a que se acredita que eles se ajustarão. A escolha dessas hipóteses alternativas dita a forma como é feito o teste de significância. A probabilidade de detectar aquela hipótese alternativa, se for verdadeira, é o "poder" do teste. Para distinguir entre a hipótese que está sendo usada para computar do valor de p de Fisher e a outra possível hipótese ou hipóteses, Neyman e Pearson chamaram a hipótese testada de "hipótese nula" (o títere[2]) e as outras de "hipóteses alternativas". Em sua formulação, o valor de p é calculado para testar a hipótese nula, mas o poder se refere a como esse valor de p se comportará se a alternativa for de fato verdadeira.

Isso levou Neyman a duas conclusões:

- A primeira é a de que o poder de um teste é uma medida de quão bom ele é; o mais poderoso de dois testes é o melhor a ser usado.
- A segunda conclusão é a de que o conjunto de alternativas não pode ser demasiado grande.

[1] Fonte: Salsburg, D. (2009). *Uma senhora toma chá: como a estatística revolucionou a ciência no século XX*. Tradução: José Maurício Gradel; revisão técnica: Suzana Herculano-Houzel. Rio de Janeiro: Zahar, 2009, p. 93; 99-103.

[2] "boneco que se move por meio de cordéis e articulações; marionete" (fonte: http://www.infopedia.pt/).

O valor p é uma probabilidade calculada, uma probabilidade associada com os dados observados sob a suposição de que a hipótese nula seja verdadeira.

Por exemplo, suponhamos que queremos testar uma nova droga para a prevenção de recorrência de câncer de mama em pacientes que sofreram mastectomias, comparando-as com um placebo[3]. A hipótese nula é que a droga não é melhor do que o placebo. Suponhamos que, depois de cinco anos, 50% das mulheres tratadas com o placebo tenham tido recorrência, contra nenhuma mulher tratada com a nova droga. Isso prova que a nova droga "funciona"? A resposta, claro, depende de quantas pacientes esses 50% representam.

Se o estudo incluísse apenas quatro mulheres em cada grupo, isso significa que teríamos oito pacientes, duas das quais tiveram recorrência. Suponhamos que tomemos um grupo qualquer de oito mulheres, marquemos duas delas (**mulher vermelha** e **mulher verde**) e dividamos as oito aleatoriamente em dois grupos de quatro (**grupo A** e **grupo B)**. Existiriam quatro combinações, todas com as mesmas probabilidades desde que o grupo seja formado aleatoriamente:

- mulher vermelha no grupo A e mulher verde no grupo B (25%).
- mulher vermelha no grupo B e mulher verde no grupo A (25%).
- ambas as mulheres no grupo A (25%).
- ambas as mulheres no grupo B (25%).

Portanto, a probabilidade de que ambas as pessoas marcadas caiam em um mesmo grupo (por exemplo, B) é de 25%. Se houvesse apenas quatro mulheres em cada grupo, o fato de que todas as recorrências caíram no grupo placebo não é significante. Se o estudo incluísse 500 mulheres em cada grupo, seria altamente improvável que todas as 250 mulheres com recorrência ocorressem no grupo placebo, a não ser que a droga estivesse funcionando. A probabilidade de que todas as 250 caíssem em um único grupo, se a droga não fosse melhor do que o placebo, é o valor de p, que nesse caso é inferior a 0,0001.

O valor de p é uma probabilidade, e assim é computado. Nesse exemplo, o valor de $p < 0{,}0001$ significa que a probabilidade de todas as 250 mulheres com recorrência serem do grupo placebo por mero acaso, sem que a droga tenha efeito real, é menor do que apenas 1 em 10.000. Como é usado para mostrar que a hipótese sob a qual é calculado é falsa, o que ele realmente significa? É uma probabilidade teórica associada às observações sob condições que muito provavelmente são falsas. Nada tem a ver com a realidade. É uma medição indireta de plausibilidade. Não é a probabilidade de que estivéssemos errados ao dizer que a droga funciona. Não é a probabilidade de qualquer tipo de erro. Não é a probabilidade de que uma paciente ficará igualmente tratada com o placebo ou com a droga.

[3] "medicamento inerte ministrado com fins sugestivos ou psicológicos..." (fonte: http://www.infopedia.pt/).