Sobre o Significado da Significância Estatística

On the Meaning of Statistical Significance

Pedro Miguel TEIXEIRA✉[1]
Acta Med Port 2018 May;31(5):238-240 ▪ https://doi.org/10.20344/amp.9375

Palavras-chave: Estatística como Assunto; Interpretação Estatística de Dados
Keywords: Data Interpretation, Statistical; Statistics as Topic

A obtenção de resultados estatisticamente significativos (i.e. valor *p* < 0,05) é muitas vezes motivo de apreço e de celebração entre investigadores, mas o seu significado nem sempre é compreendido e utilizado de forma adequada.[1-5] O uso do valor *p* como principal, senão único, elemento de suporte estatístico às conclusões em investigação científica tem sido bastante criticado e merece atenção.[6-8]

Como interpretar a significância estatística

A afirmação de que algo é significativo pode ser muito subjectiva. Contudo, em estatística o termo 'resultado estatisticamente significativo' traduz um consenso objectivo cujo significado é amplamente aceite. Porém, o que significa que o resultado é estatisticamente significativo? Considera-se um resultado significativo caso o valor *p* obtido seja inferior a 0,05. E o que significa que o valor *p* seja inferior a 0,05? O uso de valor *p* ganhou notoriedade através do trabalho de Ronald Fisher que o definiu em 1925 como: '*the probability of the observed result, plus more extreme results, if the null hypothesis were true*'.[9] Ou seja, o valor *p* é a probabilidade de o resultado observado, ou um resultado ainda mais extremo, ocorrer se a hipótese nula fosse verdadeira. Convencionou-se, posteriormente, que um valor *p* < 0,05 traduziria um resultado estatisticamente significativo (i.e. critério *alpha*). Assim, um resultado observado numa dada estatística (e.g. um valor do coeficiente de correlação) com menos de 5% de probabilidade de ocorrer, por mero acaso, caso a hipótese nula se verificasse na população (i.e. H_0: $r = 0$ na população) sugere que a hipótese nula poderá ser rejeitada e que se poderá encarar a hipótese alternativa (i.e. H_1: $r \neq 0$ na população). Assim, por exemplo, a hipótese de que na população (de onde veio a amostra) *r* será diferente de zero tem alguma sustentação e, de acordo com esta abordagem estatística, a melhor estimativa que temos para esse *r* é o valor estimado na amostra. A incerteza associada a essa estimativa pode ainda ser calculada (e.g. intervalo de confiança a 95%). A conclusão do estudo estatístico suporta assim a hipótese alternativa e aponta uma estimativa ou conjunto de estimativas (i.e. intervalo de confiança) para o valor da população em estudo (i.e. parâmetro). Probabilidade e hipótese são termos chave do uso apropriado do valor *p* e da noção de significância estatística.

Problemas com o uso do valor *p*

Contudo, a aplicação e a interpretação do valor *p* são muitas vezes inadequadas. O valor *p* é o resultado do cálculo de uma probabilidade obtido sobre determinados pressupostos. O uso do valor *p* requer o uso de amostragem representativa e aleatorizada, a ausência de erro sistemático nos dados observados e de uma correta interpretação do seu significado. A probabilidade de se obterem resultados idênticos, ou ainda mais extremos, numa outra amostra da mesma dimensão e recolhida da mesma população, caso a hipótese nula se verificasse na população. Este cálculo de probabilidade é, desde logo, influenciado pelo número de observações contidas na análise. Quanto maior for a amostra maior a possibilidade de o valor *p* ser um valor mais baixo, logo mais significativo. Por outro lado, a inferência estatística sobre o efeito em estudo com base no valor *p* é válida para a população de onde a amostra foi recolhida. A definição de critérios de inclusão/exclusão de um estudo condiciona a sua abrangência de generalização das conclusões. Daí que o que os estudos experimentais controlados ganham em validade interna mas perdem em validade externa. E, pelo mesmo motivo, todos os estudos desenvolvidos com amostras de estudantes universitários não podem ser generalizados para a população geral. Ademais, o uso do valor *p* para a inferência estatística pressupõe a ausência de erro sistemático (e.g. viés de confundimento, viés de seleção, erros de codificação). A aleatorização da amostra de uma população, bem definida, é um instrumento de controlo do erro sistemático. Contudo, muita da investigação realizada utiliza amostras de conveniência o que fragiliza o uso do valor *p*. Novamente, probabilidade e hipótese são termos chave em significância estatística.

O valor *p* é apenas uma parte da informação estatística que um estudo pode conter. É importante compreender o que o valor *p* diz e não diz. Não diz, por exemplo, qual a probabilidade de encontrar resultados semelhantes ou mais extremos caso a hipótese alternativa se verifique na população. O valor *p* também não indica a probabilidade de

1. Instituto de Investigação em Ciências da Vida e Saúde - ICVS/3b's Laboratório Associado. Escola de Medicina. Universidade do Minho. Braga. Portugal.
✉ Autor correspondente: Pedro Miguel Teixeira. teixeira.pms@gmail.com
Recebido: 03 de julho de 2017 - Aceite: 06 de fevereiro de 2018

a hipótese nula ser verdadeira. Não se refere à probabilidade da hipótese estar certa, refere-se à probabilidade do efeito observado ocorrer numa amostra com determinada dimensão, caso a hipótese nula fosse verdadeira na população. Por outro lado, um resultado com um valor $p > 0,05$ não indica que a hipótese nula seja verdadeira ou que se deva assumir a ausência de efeito entre as variáveis em estudo. Uma probabilidade de o efeito observado ocorrer numa amostra (com determinada dimensão) superior a 5%, caso a hipótese nula fosse verdadeira na população, pode refletir inúmeros aspectos, nomeadamente a ausência dos pressupostos na base dos quais o valor *p* foi calculado (e.g. erro sistemático). De igual forma, se vários estudos forem realizados de forma independente e não encontrarem resultados estatisticamente significativos (i.e. valor $p > 0,05$) isso não constitui uma evidência de que não existe efeito entre as variáveis em estudo. A síntese de resultados não significativos, obtidos em estudos independentes, pode gerar evidência de efeito relevante e estatisticamente significativo.[10]

Vários autores como Goodman[1] e Greeland[2] sintetizam estas e outras falácias comuns no mau uso e interpretação do valor *p*. A Associação Norte-Americana de Estatística emitiu, em 2016, um comunicado sobre o problema do mau uso do valor *p* em ciência.[5] E, como é referido nesse comunicado, o raciocínio científico não pode ser reduzido nem substituído por um único indicador estatístico.

Aliás, em algumas circunstâncias o valor *p* será o indicador menos relevante na óptica restrita da estatística. Por exemplo, com a análise de grandes quantidades de informação (e.g. *Big Data*) é possível encontrar efeitos de baixa ou até elevada magnitude que apesar de serem estatisticamente significativos (*viz*. pelo tamanho amostral) podem ser espúrios e sem significado prático, clínico ou epidemiológico. É fundamental examinar a estimativa do efeito em estudo, a sua magnitude, a sua direcção e o seu significado. Todavia, sem análise do contexto, do significado e de um raciocínio científico que enquadre os indicadores estatísticos a possibilidade de erro é imensa porque ainda que tenhamos uma resposta assente numa grande quantidade de dados, com resultados estatisticamente significativos e com muito poder estatístico, corremos o risco de não acertar na pergunta. Resultados estatisticamente significativos por si só não significam relevância prática (*viz*. clínica).

Valerá a pena abandonar ou ocultar o valor *p*?

A forma como o conhecimento científico é atualmente desenvolvido, validado e disseminado apresenta várias falhas de aplicação e de interpretação, mas sobretudo de acesso transparente aos resultados obtidos e isso gera desconforto entre os investigadores. O valor *p* surge associado a vários desses problemas (e.g. viés de publicação). Contudo, não será pela exclusão de um protagonista que o filme passará a ter melhor qualidade. As alternativas sugeridas de abandono do valor *p* em detrimento de estatísticas mais informativas (e.g. uso do tamanho do efeito e do seu intervalo de confiança) ou de soluções mais sofisticadas (e.g. estatística Bayesiana) apresenta os mesmos problemas e não se constituem como verdadeiras alternativas, antes devem ser encaradas como abordagens complementares. Se o valor *p* tende a ser interpretado numa visão redutora de que $p < 0,05$ é sinónimo de estatisticamente significativo, também os intervalos de confiança das estimativas de tamanho de efeito podem ser interpretados na mesma visão redutora pelo facto de o intervalo conter ou não conter o valor que indica ausência de efeito (i.e. 0 para diferenças ou 1 para rácios). Por outro lado, os erros de compreensão e de interpretação de estatísticas mais complexas como a estatística Bayesiana poderão ser ainda mais problemáticos do que os atuais erros com o valor *p*.

O valor *p* continuará a fazer parte dos estudos estatísticos e dificilmente será abandonado na investigação. Importa que o seu uso seja adequado na forma como é aplicado e interpretado. Tal desafio implica uma aposta mais determinada numa sólida formação estatística e metodológica de forma continuada dos investigadores que produzem conhecimento, dos editores e dos revisores que o disseminam e dos clínicos que o consomem. Por outro lado, o acesso aos resultados dos estudos efectuados não pode ser condicionado pelo facto de os resultados serem ou não serem estatisticamente significativos.

Os estudos cujos resultados não são estatisticamente significativos são igualmente significativos para a síntese do conhecimento sobre o tema a que se dedicam. Com o recurso a estudos secundários de síntese estatística, como a meta-análise, torna-se possível obter estimativas mais robustas dos efeitos observados em estudos independentes e ainda estimar a percentagem da variação total entre estudos que é devida à heterogeneidade e não apenas ao acaso (e.g. I^2).[10] Esta abordagem requer o acesso aos resultados de todos os estudos efectuados, independentemente do seu valor *p*. Assim, o problema não é de significância estatística, ou da falta dela, é do acesso significativo aos resultados de todos os estudos desenvolvidos, cuja qualidade tem de ser aferida pela coerência e pelo rigor metodológicos e não apenas pelo resultado do valor *p*. A possibilidade de identificar de forma sistemática os estudos desenvolvidos sobre um determinado tópico e de agregar os seus resultados estatísticos permitirá desenvolver uma evidência mais completa das estimativas existentes para uma apreciação mais informada da relevância clínica dos resultados estatísticos e da sua significância para a prática.

No processo editorial, a avaliação por pares tem sido assegurada de forma anónima pelos revisores. Existe presentemente uma tendência para que este processo se torne mais participado, amplo e transparente. Por exemplo, o registo dos protocolos dos estudos antes da sua realização limita a flexibilidade analítica e o fenómeno de *p-hacking* (i.e. a exploração exaustiva de dados com manipulação de diferentes modelos analíticos até a obtenção de resultados estatisticamente significativos) e poderá reduzir a médio prazo o efeito de viés de publicação. Ademais, a avaliação pós-publicação com plataformas como o PubMed Commons ou o Pubpeer começa também a ganhar alguma

relevância. Contudo, a decisão de rever artigos de forma cega, no que concerne os resultados, talvez seja a maior mudança a implementar a curto prazo.[11] Assim, não se trata de eliminar o valor *p* dos estudos, antes de estes serem avaliados pela pertinência, coerência e pelo rigor metodológico com que foram desenhados, desenvolvidos e reportados e de serem publicados independentemente desse valor *p* e de outras estatísticas.

REFERÊNCIAS

1. Goodman S. A dirty dozen: twelve p-value misconceptions. Seminars in Hematol. 2008;45:135-40.
2. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337-50.
3. Chavalarias D, Wallach J, Li A, Ioannidis J. Evolution of reporting p values in the biomedical literature, 1990-2015. JAMA. 2016;315:1141.
4. Altman N, Krzywinski M. Points of significance: p values and the search for significance. Nat Methods. 2016;14:3-4.
5. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70:129-33.
6. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychol. 2015;37:1-2.
7. Ioannidis J, Fanelli D, Dunne D, Goodman S. Meta-research: evaluation and improvement of research methods and practices. PLOS Biol. 2015;13:e1002264.
8. Munafò M, Nosek B, Bishop D, Button K, Chambers C, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:0021.
9. Fisher R. Statistical methods for research workers. Edinburgh: Oliver and Boyd;1925.
10. Chalmers T, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. Stat Med. 1996;15:1263-8.
11. Button K, Bal L, Clark A, Shipley T. Preventing the ends from justifying the means: withholding results to address publication bias in peer-review. BMC Psychol. 2016;4:59.