

Should we retire statistical significance?

Devemos aposentar a significância estatística?

DOI 10.5935/2595-0118.20190037

Prezado editor,

Em publicação no periódico *Nature*¹, repercutiu o artigo intitulado “*Retire Statistical Significance*”, que traz uma reflexão crítica a respeito do dogmatismo estatístico, suscitando uma análise sobre os dois lados da mesma moeda. De um lado, o valor da reflexão trazida pelos autores. Do outro lado, as consequências não intencionais da aposentadoria do conceito de significância estatística. O primeiro ponto de vista relaciona-se com o viés da superestimativa do valor; o segundo ponto guarda relação com o viés do positivismo.

O conceito de significância estatística é dicotômico, isto é, categoriza a análise em “positiva” ou “negativa”. Categorizar agrega utilidade pragmática, porém toda categorização é um reducionismo arbitrário. Ao categorizar por questões pragmáticas, deveríamos entender categorias como algo de menor valor do que a visão do todo. O **paradoxo da categorização** ocorre quando passamos a valorizar mais a informação por esta ser categórica do que uma informação contínua²⁻⁴. A informação contínua aceita os tons de cinza, o intermediário, a dúvida, enquanto a categórica atribui um tom (pseudo) definitivo à afirmação.

Estatística é o exercício de reconhecer a incerteza, a dúvida, o acaso. A definição de significância estatística foi originalmente criada para dificultar afirmações decorrentes do acaso. O intervalo de confiança foi criado para descrever a imprecisão de nossas afirmações. Estatística é o exercício de integridade e humildade do cientista.

No entanto, o paradoxo da categorização fomenta um certo dogmatismo. Primeiro, os autores do artigo supracitado da *Nature* apontam a supervalorização de resultados negativos. Estudo negativo não é o que comprova inexistência, o que seria impossível; simplesmente, é um estudo que não comprovou existência. Portanto, a rigor, “ausência de evidência não é evidência de ausência”, como afirmou Carl Sagan. Isto é, “o estudo comprovou que não existe diferença” não constitui a melhor forma de descrever, sendo preferível pontuar “o estudo não comprovou diferença”.

Não se deve confundir tal colocação com a ideia de que um estudo negativo não quer dizer nada. Ele tem valor e tem impacto. O impacto de um estudo negativo ($p > 0,05$) encontra-se na redução da probabilidade de o fenômeno existir. Na medida em que bons estudos não conseguiram comprovar, a probabilidade do fenômeno cai progressivamente até o ponto em que se torna tão baixa que invalida o sentido de continuar tentando provar, tornando a hipótese nula o caminho de pensamento mais provável.

Um estudo negativo não é necessariamente contraditório em relação a um estudo positivo. Eventualmente, o resultado dos dois pode ser o mesmo, quando um não conseguiu rejeitar a hipótese nula, e outro estudo conseguiu rejeitar. Um não conseguiu ver e outro conseguiu ver. Na verdade, na maioria das vezes, apenas um dos dois estudos está correto.

Por fim, o paradoxo da categorização faz com que acreditemos em qualquer significância estatística, embora a maioria seja falso positivo (Ioannidis). Valor de $p < 0,05$ não é comprovação irrefutável. Estudos subdimensionados, multiplicidade de análises secundárias, vieses, podem fabricar falsa significância estatística.

Na verdade, o valor preditivo (negativo ou positivo) de estudos não reside apenas na significância estatística. Depende da qualidade do estudo e da análise, do ecossistema científico e da probabilidade pré-teste da ideia.

Portanto, os autores do artigo da *Nature* estão corretos em criticar a visão determinística da significância estatística.

MAS, PAIRA O QUESTIONAMENTO: SERÁ QUE DEVEMOS MESMO APOSENTAR A SIGNIFICÂNCIA ESTATÍSTICA?

Tal feito significaria aposentar um advento que historicamente foi responsável por uma grande evolução de integridade científica. Todavia, tudo que é bom tende a ser “sequestrado”. Artistas da falsa positividade de estudos “sequestraram” o advento do valor de p (feito para dificultar o erro tipo I) para provar coisas falsas.

Se por um lado a aposentadoria da significância estatística evitaria o paradoxo da categorização, por outro lado abriria espaço para o viés da positividade, nosso tropismo por criar ou absorver informações positivas.

A crítica à significância estatística, neste e em outros artigos de visibilidade⁵⁻⁷, não traz uma alternativa melhor. Por exemplo, o próprio autor do trabalho da *Nature* reconhece que outras abordagens estatísticas mais recentes (bayesiana, por exemplo) não abandonam o paradigma da categorização. Inclusive, em certas passagens, os autores mencionam que não propõem um total abandono da noção de significância estatística. Talvez o título que traduza o verdadeiro teor do artigo deveria conter uma interrogação: “*Retire Statistical Significance?*”

Atualmente, discute-se muito mais sobre integridade científica do que há duas décadas. Entretanto, ao abordar esse assunto com mais ênfase do que no passado, surge a impressão de que este é um problema pior nos dias de hoje. Não é o caso. Experimentamos clara evolução de integridade científica: conceitos de multiplicidade são mais discutidos hoje do que no passado, ensaios clínicos têm obrigatoriamente seus desenhos publicados *a priori*, normas CONSORT de publicação são exigidas por revistas, fala-se muito mais em transparência científica, *open science*, *slow science*. Estamos evoluindo. E o primeiro passo da integridade foi a criação da noção de significância estatística na primeira metade do século passado, por Ronald Fisher⁸.

Um estudo publicado na *PLoS One* (Bob Kaplan)⁹ analisou, durante um longo período de anos, os resultados de ensaios clínicos financiados pelo *National Institutes of Health* (NIH). Antes do ano 2000, quando não havia a necessidade de publicar previamente o protocolo, a frequência de estudos positivos era de 57%, caindo para apenas 7% de estudos positivos após a regra de publicação *a priori*. Antes, os autores positivavam seus estudos por análises múltiplas *a posteriori*. Hoje, isso tem se tornado menos frequente pela obrigatoriedade de publicação *a priori*.

A impressão que paira é que se tornou elegante criticar o valor de p , o que parece traição a um advento de grande importância histórica e que, até então, não encontrou um substituto melhor. Não é culpa de o p ter sido “sequestrado” por pesquisadores mal-intencionados. É culpa dos pesquisadores.

Portanto, propomos manter o valor de p e adotar as seguintes medidas:

- Descrever o valor de p apenas quando o estudo tiver uma dimensão adequada para o teste de hipótese. Do contrário, este ganharia um caráter mais descritivo, sem utilizar associações para testes de conceitos. Isso evitaria falso positivos decorrentes de “estudos pequenos”, a maioria dos artigos publicados. Para exemplificar, a mediana do poder estatístico de estudos em biomedicina é 20%;
- Não descrever o valor de p em análises de desfechos secundários;
- Em análises de subgrupo (exploratórias), utilizar apenas o P da interação (mais conservador e difícil de dar significativo), evitando o valor de p obtido pela comparação dentro de um subgrupo (estudos pequenos);
- Incluir no CONSORT a obrigatoriedade de os autores explicitarem no título de subestudos que aquela é uma análise exploratória e secundária de um estudo previamente publicado;
- Abandonar o termo “significância estatística”, substituindo-o por “**veracidade estatística**”. Estatística é utilizada para diferenciar associações causais verdadeiras de pseudocausalidades mediadas pelo acaso. Portanto, um valor de $p < 0,05$ conota veracidade. Se a associação é significativa (relevante), depende da descrição da diferença numérica ou das medidas de associação de desfechos categóricos. Destarte, utilizar “veracidade estatística” evita a confusão entre significância estatística e significância clínica.

Finalmente, propomos o advento do “**índice de integridade do pesquisador**”.

Esse índice será calculado pela razão entre: número de estudos negativos/número de estudos positivos. Um índice de integridade < 1 indica um pesquisador de integridade questionável. Esse índice se baseia na premissa de que a probabilidade de uma boa hipótese ser verdadeira é menor que 50%. Portanto, deveria haver mais estudos negativos do que estudos positivos. Isto não ocorre devidos às técnicas de positivação de estudos (pequenos trabalhos, multiplicidades, vieses, spin de conclusões) e pelo viés de publicação, que esconde os estudos negativos. Um autor íntegro seria aquele que não utiliza essas práticas, portanto teria vários estudos negativos e poucos positivos, resultando em índice de integridade bem superior a 1.

O artigo da *Nature* se faz útil para promover a reflexão sobre prós e contras da significância estatística. Não obstante, não chega a propor sua aposentadoria. Tal feito seria análogo a aposentar uma pessoa ainda muito produtiva. Em contrapartida, que a significância estatística continue atuante e evoluindo progressivamente na forma de utilização.

Que aprendamos a valorizar também um $p > 0,05$. Afinal, a imprevisibilidade da vida é representada por esta simbologia, boa parte do destino das pessoas é mediado pelo acaso.


Ou nada é por acaso?

Luis Claudio Lemos Correia


 <https://orcid.org/0000-0002-6910-1366>

E-mail: luisclcorreia@gmail.com

Gabriela Oliveira Bagano

 <https://orcid.org/0000-0002-6541-0372>

Milton Henrique Vitória de Melo

 <https://orcid.org/0000-0002-5130-1634>

Escola Bahiana de Medicina e Saúde Pública,
Hospital Aliança, Salvador, BA, Brasil.

REFERÊNCIAS

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-7.
2. Gigerenzer G. Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*. *Adv Meth Pract Psychol Sci*. 2018;1(2):198-218.
3. Greenland S. Invited Commentary: the need for cognitive science in methodology. *Am J Epidemiol*. 2017;186(6):639-45.
4. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. *Am Stat*. 2019;73(Suppl 1):235-45.
5. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. *Int J Cardiol*. 2014;177(3):1089-90 (2014).
6. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond “ $p < 0.05$ ”. *Am Stat*. 2019;73(Suppl)1-19.
7. Hurlbert SH, Levine RA, Utts J. Coup de Grâce for a tough old bull. “statistically significant” expires. *Am Stat*. 2019;73(Suppl):352-7. <https://doi.org/10.1080/00031305.2018.1543616>.
8. Fisher RA. The fiducial argument in statistical inference. *Ann Eug*. 1935;6:391-8.
9. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*. 2015;10(8):e0132382. <https://doi.org/10.1371/journal.pone.0132382>.

