Estado da publicação: Não informado pelo autor submissor

Teste de Significância de Hipótese Nula na Análise do Comportamento: Problemas e recomendações

Bruna Rodrigues Lins, Bruno Strapasson

https://doi.org/10.1590/SciELOPreprints.7933

Submetido em: 2024-04-20
Postado em: 2024-10-08 (versão 2)
(AAAA-MM-DD)

A moderação deste preprint recebeu o endosso de:
Helder Gusso (ORCID: https://orcid.org/0000-0002-8688-2010)

Justificativa da versão: In this second version, the discussion related to behavior analysis was deleted from the text, and the pertinent adjustments were made. This modification was necessary to shorten the text and make it suitable for publication in an academic journal.

Teste de Significância de Hipótese Nula na Psicologia: Problemas e recomendações

Autores:

Bruna Rodrigues Lins - Universidade Federal do Paraná – Curitiba, Paraná, Brasil

https://orcid.org/0009-0000-9670-2581

Bruno Angelo Strapasson - Universidade Federal do Paraná – Curitiba, Paraná, Brasil

https://orcid.org/0000-0002-1720-6182

Contribuição dos autores: Bruna Rodrigues Lins foi responsável pela redação original do manuscrito e por sua revisão final; Bruno Angelo Strapasson foi responsável pela conceitualização e revisão final do manuscrito.

Declarações sobre conflito de interesse: Os Autores Bruna Rodrigues Lins e Bruno Angelo Strapasson declaram não ter conflitos de interesse neste manuscrito.

Este estudo não recebeu financiamentos.

Teste de Significância de Hipótese Nula na Psicologia: Problemas e recomendações

Resumo

Amplamente adotada na Psicologia em geral, a estatística inferencial tem sido alvo de longos debates na ciência. O aumento do uso de pesquisas com grupos que utilizam testes de significância de hipótese nula (Null Hypothesis Significance Testing - NHST) na análise de dados tem crescido nessa área e traz consigo problemas relacionados. Tais problemas muitas vezes passam despercebidos no atual sistema de revisão por pares, comprometendo a confiabilidade de algumas conclusões disponíveis na literatura científica. Neste artigo, explicamos os problemas relacionados ao uso indevido, à má interpretação do NHST e ao uso de práticas questionáveis de pesquisa envolvendo NHST, bem como compilamos orientações para editores, revisores e autores que podem ser adotadas para minimizar os problemas mencionados.

Palavras-Chave: Estatística inferencial, Análise do Comportamento, Teste de significância de hipótese nula

Null Hypothesis Significance Testing in Psychology: Problems and Recommendations

Abstract

Widely supported in general psychology, inferential statistics have been subject to long scientific debates. The increase in research use with groups that use null hypothesis significance testing (Null Hypothesis Significance Testing - NHST) in data analysis has grown in this area. It raises related problems. These problems often go unnoticed in the current peer review system, compromising the reliability of some conclusions available in the scientific literature. In this article, we explain the problems related to misuse, to further interpret NHST and to the use of research question practices involving NHST, and how we

compile guidance for editors, reviewers, and authors that can be used to minimize the problems mentioned above.

Keywords: Inferential statistics, Behavior Analysis, null hypothesis significance testing

Pruebas de significación de hipótesis nula en psicología: Problemas y recomendaciones

Resumen

La estadística inferencial, ampliamente aceptada en psicología general, ha sido objeto de largos debates científicos. El uso de la estadística inferencial en la investigación con grupos que utilizan pruebas de significación de hipótesis nulas (Null Hypothesis Significance Testing - NHST) en el análisis de datos ha aumentado en esta área, lo que plantea problemas relacionados. Estos problemas a menudo pasan desapercibidos en el sistema actual de revisión por pares, lo que compromete la fiabilidad de algunas conclusiones disponibles en la literatura científica. En este artículo, explicamos los problemas relacionados con el uso indebido, la interpretación más detallada de la NHST y el uso de prácticas questionables de investigación que involucran la NHST. También recopilamos orientación para editores, revisores y autores que se puede utilizar para minimizar los problemas antes mencionados.

Palabras clave: Estadística inferencial; Análisis de comportamiento, pruebas de significación de hipótesis nula

Teste de Significância de Hipótese Nula na Psicologia: Problemas e recomendações

O uso de estatística inferencial, em especial por meio da aplicação de testes de hipótese nula (null hypothesis significance testing – NHST), é pervasivo na psicologia e fundamenta parte considerável da tomada de decisões sobre achados empíricos nesta ciência. Se, de um lado, o uso de estatísticas inferenciais tem aumentado nos periódicos de psicologia (e.g., Cumming et al., 2007; Hubbard & Ryan, 2000; Imam & Frate, 2019), por outro, há uma

suspeição importante de metodólogos e psicólogos para com essa forma de análise de dados (e.g., Bakan, 1966; Ioannidis, 2005; Wasserstein et al., 2019), de modo que o debate é um tanto complexo e merece esclarecimento. As críticas à aplicação do NHST se relacionam à (a) interpretação incorreta dos seus resultados, à (b) decorrências típicas do seu uso na tomada de decisões sobre resultados empíricos e à (c) práticas questionáveis de pesquisa. Neste artigo explicitaremos quais são os problemas relacionados a aplicação do NHST e apresentaremos orientações para pesquisadores, editores e revisores a respeito de boas práticas no sentido de aumentar a probabilidade de que o uso de estatística inferencial seja adequado.

O teste de significância de hipótese nula (NHST)

Os testes inferenciais comumente utilizados podem ser explicados considerando tanto as proposições de R. A. Fisher (1890-1962), conhecido como teste de significância, quanto às proposições de J. Neyman (1894-1981) e E. S. Pearson (1895-1980), conhecido como teste de hipóteses, ou por uma alternativa híbrida comum na literatura conhecida como teste de significância de hipótese nula (NHST) que embora tenham algumas diferenças importantes, conduzem a um raciocínio semelhante. Fisher foi quem criou o teste de significância em que o *p*-valor é o guia para definir conclusões sobre uma hipótese. Nesse caso escolhe-se o nível de significância a ser aceito, geralmente definido em 0,05, posteriormente deve ser aplicado um teste estatístico apropriado, como por exemplo o teste t de Student, e calculado o *p*-valor, sendo o *p* a probabilidade de um resultado pelo menos tão extremo quanto aquele encontrado ser obtido caso a hipótese nula (H0) seja verdadeira (probabilidade condicional). Se o resultado for um *p*-valor menor que o valor de significância escolhido ele é frequentemente descrito como estatisticamente significativo. Para Fisher o *p*-valor é utilizado como uma força de evidência de hipótese, portanto ainda que todo *p*-valor menor que o nível de significância (por exemplo, e.g., $p = 0,048$) possa ser considerado estatisticamente significativo, *p*-valores menores (e.g., $p < 0,001$), aumentam a confiança na significância estatística do resultado.

Discordando de Fisher, Neyman e Pearson criaram o teste de hipóteses e introduziram a noção de hipótese alternativa (HA) a ser testada. Nessa abordagem o pesquisador faz uma escolha entre duas hipóteses (H0 vs. HA). Além da introdução do conceito de HA os autores instituíram também os conceitos de erro de tipo I e erro de tipo II. Nesta teoria, o erro tipo I ocorre quando a H0 é verdadeira e o pesquisador a rejeita. Neste caso a probabilidade de cometer um erro tipo I é indicada por ⟨ (um ⟨ de 0,05 indica que se "aceita" uma chance de erro de 5% ao rejeitar a H0). O erro tipo II acontece quando a H0 é falsa e o pesquisador não a rejeita, a probabilidade de cometer um erro tipo II é β e seu valor depende do poder do teste escolhido. Para explicar a proposição dos dois autores pode-se utilizar um exemplo simples proposto por Cumming (2012) que avaliou a aplicação do raciocínio considerando teste de um novo medicamento para insônia. Neste caso, propunha que o novo medicamento era melhor do que o antigo. Primeiro, deve-se escolher qual a H0 – neste caso o autor opta pela H0: μ = 0, sendo μ (letra grega mu) a média das diferenças nos dois medicamentos em toda a população de interesse, ou seja, a H0 diz que não há diferenças entre os medicamentos. Quando o raciocínio proposto por Neyman e Pearson é utilizado para analisar os dados, uma HA também é especificada. Nesse caso ela sugere que existe alguma (qualquer) diferença entre o remédio em teste e o remédio antigo, HA: μ ≠ 0. Após aplicação do teste estatístico apropriado avalia-se se foi atingido um nível de significância (nesse contexto chamado de ⟨ - alfa) menor que o nível pré-definido (0,05 por exemplo), se sim rejeita-se a H0, o que é interpretado como fortalecendo a HÁ: apoia-se a hipótese de que o novo medicamento tem efeitos diferentes (talvez melhores) quando comparado ao anterior. Na lógica proposta por Neyman e Pearson diferentes *p*-valores não representam diferentes graus de confiabilidade, trata-se de uma decisão dicotômica: ou a H0 é

rejeitada ou ela não é. Portanto, o NHST como é conhecido hoje possui duas vertentes que, embora semelhantes, implicam formas diferentes de analisar e tirar conclusões sobre os dados.

A despeito das diferenças nos dois modelos, em ambos tomamos decisões sobre a confiabilidade estatística de um resultado a partir de uma comparação entre o *p*-valor obtido em um teste apropriado e o limiar de significância (ou ⟨) previamente estabelecido. É importante notar, entretanto, que o *p*-valor tem um significado muito específico. Trata-se da probabilidade de que os resultados (que foram de fato obtidos no estudo, ou valores mais extremos que esse) sejam encontrados se não houver efeito ou associação entre as variáveis estudadas (ou seja, se a H0 for verdadeira) e se todos os demais pressupostos envolvidos na análise (e.g., que o registro das variáveis foi adequado e preciso, que os procedimentos atendem às exigências dos testes utilizados – como a aleatoriedade da amostra e aspectos distribucionais do dados – e que os testes utilizados têm poder adequado para a análise pretendida – ver Greenland et al., 2016) tiverem sido atendidos.

Note-se que a não existência de um efeito ou associação (a verdade da H0) é um pressuposto nesses testes, não é algo que está sendo testado: o *p*-valor diz respeito a uma probabilidade condicional. Nesse contexto, dizer que um resultado é estatisticamente significativo é o mesmo que dizer que seria improvável alcançá-lo se a variáveis investigadas não forem associadas ou não estiverem funcionalmente relacionadas. O *p*-valor, portanto, não informa (a) sobre a probabilidade do resultado obtido ser verdadeiro ou falso, (b) sobre a significância prática do resultado obtido, (c) sobre a magnitude de uma diferença ou associação entre variáveis (o tamanho de efeito – ES – do inglês Effect Size), (d) sobre a objetividade do estudo, (e) sobre o sucesso do estudo, (f) sobre a qualidade do delineamento metodológico utilizado, (g) etc. e constitui uma medida imprecisa e difícil de interpretar a

respeito da confiabilidade de que o resultado obtido no estudo (baseado em uma amostra) de fato representa o que ocorre com a população de interesse.

As ressalvas apresentadas no parágrafo anterior são importantes porque o entendimento incorreto do significado do *p*-valor na estatística inferencial é uma fonte muito comum de erros na disseminação e divulgação do conhecimento científico. Nas próximas seções apresentaremos exemplos dos problemas mais comuns, faremos um panorama do quão disseminados são esses problemas e daremos indicações de como diminuir a probabilidade de que tais problemas persistam.

Iniciaremos a discussão apresentando a interpretação padrão do NHST para depois discutir seus problemas relacionados a sua má interpretação, suas decorrências na avaliação de dados empíricos e às práticas questionáveis de pesquisa a ele associadas. Por fim, apresentaremos sugestões para autores, revisores e editores de psicologia para minimizar esses problemas.

A interpretação dos resultados de NHST

Uma das principais e mais comuns interpretações incorretas dos dados resultantes do NHST é a chamada ilusão de prova probabilística por contradição. Nesta, acredita-se que ao obter significância estatística ($p < 0,05$ por exemplo) a H0 é improvável (Cohen, 1994; Falk, 1998; Gigerenzer et al., 2004; Haller & Krauss, 2002; Kalinowski et al., 2008; Kline, 2013; Nickerson, 2000; Pollard & Richardson, 1987). A interpretação incorreta resulta do seguinte raciocínio: se assumirmos que a H0 é verdadeira, então a significância estatística provavelmente não será alcançada, logo, se a significância estatística for alcançada, então a H0 é improvável. Ainda que os cientistas frequentemente rejeitem a H0 se o *p*-valor no teste estatístico apropriado for menor que o ⟨, é incorreto dizer que o *p*-valor representa a probabilidade de H0 ser falsa. A veracidade da H0 (bem como dos demais pressupostos

estatísticos envolvidos) é o elemento que condiciona o cálculo do *p*-valor (é um pressuposto), não é algo que esteja sendo avaliado nos NHSTs.

Se é incorreto interpretar que o *p*-valor é a probabilidade da H0 ser falsa, também é incorreto, ainda que também seja comum, interpretar que a probabilidade complementar ao *p*-valor (1-*p*) é a probabilidade da HA ser verdadeira (Carver, 1978; Cohen, 1994; Gigerenzer et al., 2004; Kline, 2013; Lambdin, 2012; Nickerson, 2000). Se assim fosse, seria plausível reivindicar que se obtido um *p*-valor menor do que 0,05 haveria ao menos 95% de chance de a hipótese alternativa ser verdadeira. Mas como dito anteriormente, a verdade da H0 é um pressuposto nos NHSTs, ela não está sendo testada.

O entendimento de que o complemento de que o *p*-valor é a probabilidade de um resultado ser replicado também representa uma interpretação incorreta das informações produzidas pelo NHST (Carver, 1978; Falk & Greenbaum, 1995; Gigerenzer et al., 2004; Kline, 2013; Nickerson, 2000; Sohn, 1998). A interpretação realizada por pesquisadores que defendem essa conjectura é a de que se a significância estatística for alcançada em 0,05 isso significa que o pesquisador pode afirmar que a cada 100 experimentos a diferença observada se manterá em 95 deles (Carver, 1978). Para Carver (1978) nada na lógica das estatísticas com NHST fornece base para concluir que um resultado estatisticamente significativo seja interpretado como a probabilidade de replicação de um resultado. É uma fantasia afirmar que significância estatística pode estabelecer confiança na replicabilidade de um estudo. Segundo Sohn (1998) é uma prática comum em Psicologia tratar significância estatística como replicabilidade, embora não seja possível fazer esse tipo de afirmação, visto que um resultado estatisticamente significativo não é base para fazer nenhum tipo de inferência sobre probabilidade de nenhuma das hipóteses avaliadas.

Um engano semelhante ao descrito na falácia da ilusão de prova probabilística por contradição é a falácia da significância, nesta acredita-se que ao rejeitar a H0 confirma-se a

HA e a hipótese de pesquisa por trás dela. Segundo Kline (2013), ao cometer essa falácia o pesquisador comete dois erros conceituais, o primeiro é que ao rejeitar a H0 em um único estudo não significa que a HA esteja comprovada e segundo, mesmo que estatisticamente a HA ganhe apoio, isso não significa que a hipótese apontada pelo pesquisador também esteja correta. Para exemplificar esse segundo aspecto o autor cita um estudo realizado por um pesquisador que estudou nascimentos de bebês em Londres por 82 anos. Nesta pesquisa foi identificado que mais meninos do que meninas nasceram nesse período, com base em seus dados o pesquisador rejeitou a H0 de que o número de meninos e meninas nascido era igual, no entanto a HA do pesquisador era de que por providência divina, mais meninos nasciam para compensar o maior número de mortes de homens em guerras, acidentes e coisas do gênero, o que claramente é uma hipótese questionável.

O entendimento de que α determina a probabilidade de cometer um erro tipo I, também é um erro de interpretação do NHST. Para Pollard e Richardson (1987). O valor de α apenas indica qual o percentual de erro de tipo I é aceitável naquele experimento, assim se ao realizar um experimento e definir previamente o α como 0,05, ou 5%, isso quer dizer que 5% ou menos das ocasiões em que a H0 for a verdadeira é aceito que ela seja rejeitada e um erro tipo I seja cometido. Deste modo, o nível α apenas fornece a probabilidade de cometer um erro tipo I quando a H0 é verdadeira, mas ao realizar um experimento não é possível saber se de fato ela é.

O entendimento de que a significância estatística é equivalente à significância teórica ou prática também parece ser um erro comum na interpretação dos resultados do NHST. Segundo Kühberger et al. (2015) as duas interpretações equivocadas mais comuns relacionadas à significância são: (1) a de que a falha em rejeitar H0 significa que o efeito ou associação na população é zero e (2) a que o tamanho do *p*-valor indica a força do tratamento ou ainda magnitude de efeito ou associação (Kühberger et al., 2015). A falha em rejeitar a H0

pode significar duas coisas: (A) a H0 é verdadeira o que, na maioria dos casos, significa que não há efeito ou associação entre as variáveis investigadas, ou (B) que existe na população de interesse um efeito ou associação entre as variáveis investigadas, mas o teste aplicado não teve poder suficiente para detectar esse efeito ou associação (Erro do tipo II, aceitar a H0 quando ela é falsa) (Kline, 2013). O *p*-valor não diz absolutamente nada sobre a força do tratamento ou ainda sobre magnitude de efeito ou associação.

Ao avaliar dados resultantes de NHST, alguns pesquisadores cometem a falácia da probabilidade contra o acaso. Grande parte dos experimentos que testam H0 atribuem o seu valor como de nenhum efeito ou associação, ou seja, se o *p*-valor resultado do experimento for maior do que ⟨ é porque não há efeito ou associação entre as variáveis investigadas e os resultados ocorreram devido ao acaso. Considerando que *p*-valor é a probabilidade de obter os resultados da pesquisa quando se assume que a H0 é verdadeira, Carver (1978) sugere que:

> Portanto, é impossível que o *p*-valor seja a probabilidade de que o acaso causou a diferença média entre dois grupos de pesquisa, uma vez que (a) o *p*-valor foi calculado assumindo que a probabilidade era 1,00 de que o acaso causou a diferença média, e (b) o *p*-valor é usado para decidir se aceita ou rejeita a ideia de que a probabilidade é 1,00 de que o acaso causou a diferença média (p. 382).

Tanto sugerir que os resultados foram obtidos como produto do acaso quando a H0 não é rejeitada, quanto sugerir que foi encontrado um efeito ou associação real quando a H0 é rejeitada são proposições falaciosas.

Kline (2013) apresenta ainda a falácia da objetividade, segundo a qual o teste de significância seria um método objetivo de teste de hipótese(s), enquanto os outros métodos para inferência, incluindo a inspeção visual de gráficos, são subjetivos. O NHST é objetivo apenas na aparência, pois o uso de testes de hipótese nula exige decisões subjetivas, como qual será o valor de ⟨ adotado, se a análise será unicaudal ou bicaudal etc.

Um par adicional de falácias relevantes diz respeito à falácia da qualidade e sucesso assim como o seu complemento, a falácia da falha (Kline, 2013). Na falácia da qualidade e sucesso acredita-se que ao obter significância estatística está confirmada a qualidade do desenho experimental, porém um desenho experimental ruim ou um erro de amostragem podem levar a uma rejeição incorreta da H0 e, logo, a um erro de Tipo I, assim como a não rejeição da H0 pode ser produto de um estudo bem desenhado. Na falácia da falha, não alcançar a significância estatística classificaria o estudo como um fracasso, o que também não é verdade.

Outras três falácias menos conhecidas foram descritas por Kline (2013): a da santificação, da reificação e da robustez. A falácia da santificação se refere ao pensamento dicotômico. Se o critério de avaliação do *p*-valor for definido em 0,05 e estudos resultarem em $p = 0,049$ ou $p = 0,051$, é discutível considerar essa diferença como suficiente para rejeitar ou não a H0, tal como sugere a abordagem de Neyman-Pearson. Esse critério deve ser especialmente considerado para o teste de significância postulado por Fisher, que apesar de considerar o *p*-valor como força de evidência contra a H0, também implica um raciocínio dicotômico em relação ao limiar adotado. Na falácia da reificação, a falha em tomar a mesma decisão de estudos anteriores a respeito da H0 significaria falha em replicar um resultado. Neste pensamento, um resultado não é considerado replicado se H0 for rejeitada em um estudo e em outro não, ou seja, quando se falha em tomar a mesma decisão sobre a H0. Tal raciocínio é bem demonstrada no exemplo de Dixon (1999):

> suponha que um pesquisador compare memória para uma lista de palavras concretas com aquela para uma lista de palavras abstratas; suponha ainda que a precisão durante lembrar para o primeiro é de 60% e 40% para o último, e que a H0 de nenhuma diferença é rejeitada com o *p*-valor de 0,02. É possível para um segundo pesquisador ao fazer um estudo com menos poder, obtenha a mesma amostra média de 60% e 40%,

> mas obtenha um *p*-valor de 0,10. Como uma consequência, o segundo pesquisador não conseguiria rejeitar a hipótese nula. A falácia da reificação seria descrever este segundo experimento como uma falha de replicação; na verdade, tal descrição reifica o resultado do processo de decisão como uma descrição dos resultados. Na verdade, neste exemplo, o padrão de resultados é exatamente o mesmo nos dois experimentos, e eles são inteiramente consistentes um com o outro. (p. 137)

Segundo o autor, a falha em replicar deve ser considerada apenas quando existirem evidências claras de que um padrão diferente de resultados foi obtido por meios diferentes, não apenas ao analisar resultados que levaram a decisões diferentes sobre a rejeição da H0.

A escolha inapropriada do teste estatístico também é uma má prática relativamente comum na análise de dados. Tal escolha deve considerar sempre o teste (a) com maior poder e (b) que tenham respeitados os pressupostos nos quais foram fundados seus cálculos. Por exemplo, testes estatísticos paramétricos têm, em geral, mais poder que testes não paramétricos, mas eles frequentemente exigem que a distribuição dos dados obtidos siga o padrão da chamada distribuição normal. Testes não paramétricos não exigem um padrão específico de distribuição dos dados, mas têm menos poder. Além da distribuição, diversas outras características dos dados (e.g., aleatoriedade da amostra, independência, homogeneidade, homoscedasticidade, esfericidade etc.) são necessárias a depender dos testes que serão utilizados. Ainda que mandatório, nem sempre o atendimento dessas características é garantido ou verificado antes da realização dos testes, mesmo havendo formas estatísticas padronizadas e populares para identificar tais características. Kline (2013) chama uma variação específica desse problema de falácia da robustez, que se refere ao fato de que os testes estatísticos clássicos paramétricos não são robustos contra outliers (dados que se diferenciam drasticamente de todos os outros) ou problemas com as premissas distribucionais

(e.g., ausência de distribuição normal nos dados) e, portanto, não seriam apropriadamente aplicados quando os resultados têm essas características.

O uso acrítico de teste de significância e o desconhecimento sobre o que de fato seus dados representam podem levar a erros de interpretação de resultados de pesquisa. Ao interpretar dados de forma incorreta, segundo raciocínios falaciosos, os pesquisadores tendem a, mesmo sem querer, distorcer o conhecimento divulgado e, portanto, distorcer o conhecimento científico da área. É relevante que os pesquisadores tenham conhecimento a respeito da interpretação dos resultados de NHST ao produzir seus estudos e também ao acessar o conhecimento científico disponível em sua área de atuação, seja na ciência em geral, na Psicologia ou na AC.

Decorrências típicas do uso de NHST sobre a tomada de decisões a respeito de resultados empíricos

Parte dos problemas relacionados à utilização do NHST podem ser resolvidos ao melhorar a forma com que estes são ensinados e aplicados, pois assim se resolveriam as falácias decorrentes de má interpretação e os problemas da descrição imprecisa ou incorreta dos seus resultados. Porém existem questões mais essenciais que são relacionadas ao próprio raciocínio lógico que embasa o NHST.

A primeira crítica expressa por analistas do comportamento (Branch & Pennypacker, 2013; Branch, 2014), psicólogos (Bakan, 1966; Cohen, 1994; Lambdin, 2012; Loftus & Loftus, 1996; Lykken, 1968; Meehl, 1967; Nunnally, 1960; Simmons et al., 2011; Thompson, 1999) e cientistas em geral (e.g., Carver, 1978) é a de que a significância estatística é um produto de um conjunto amplo de variáveis, e não um produto exclusivo do padrão de dados obtido, nesse sentido ela sempre pode ser alcançada. Bakan (1966) explicitou cinco fatores que levam à rejeição da H0 independente da real relação entre variável independente (VI) e dependente (VD) na população, sendo elas:

> se o teste é unicaudal ou bicaudal, o nível de significância, o desvio padrão, a amplitude de desvio da hipótese nula e a quantidade de observações. A escolha de um teste unilateral ou bicaudal é do investigador; o nível de significância também é baseado na escolha do investigador; o desvio padrão é um dado da situação e é caracteristicamente razoavelmente bem estimado; o desvio da hipótese nula é o que é desconhecido; e a escolha da quantidade de casos avaliados (tamanho da amostra) é caracteristicamente arbitrária ou apressada (p.426).

Portanto independente da existência real de um efeito de uma VI sobre uma VD, diversos outros fatores irão determinar o *p*-valor obtido em um teste estatístico. Além disso, é comum que a H0 seja definida como a ausência completa de diferenças entre os grupos ou condições experimentais e é pouco razoável supor que em condições naturais os resultados sejam rigorosamente iguais. Se houver qualquer desvio da H0 na população, ou seja, algum efeito ou associação entre as variáveis investigadas, mesmo que ele seja ínfimo ou irrelevante na prática, um número suficientemente grande de observações precisas levará a rejeição da H0.

Outro aspecto criticado é que a lógica do NHST assume o oposto ao que se quer testar. O teste de significância não fornece informações diretas sobre a hipótese de interesse, a HA (Cohen, 1994; Falk, 1998; Falk & Greenbaum, 1995; Loftus & Loftus, 1996; Meehl, 1967; Thompson, 1999). Esta crítica se refere ao fato de que o resultado do teste de significância não avalia a probabilidade de a H0 ser verdadeira considerando os dados (D) obtidos (P(H0|D)), que é o que se gostaria de saber, mas sim qual a probabilidade de os dados encontrados serem obtidos tendo a verdade da H0 como condição de partida (P(D|H0)). A ordem dos fatores é um aspecto fundamental no estabelecimento de probabilidades condicionais. Por exemplo, a probabilidade de que você encontre uvas na salada de frutas do seu restaurante favorito (P(uva na salada de frutas | restaurante favorito) é muito diferente da probabilidade de você estar no seu restaurante favorito se soubermos que você está comendo

salada de frutas com uvas (P(restaurante favorito | uva na salada de frutas). É por isso que é importante notar que a probabilidade dos dados obtidos sendo a H0 verdadeira, diz pouco ou nada sobre a probabilidade da a H0 ser ou não verdadeira. Considere o seguinte paralelo que Branch e Pennypacker (2013) apresentam para ilustrar o problema lógico da inversão da probabilidade condicional frequentemente empregado na aplicação do NHST:

> O fato importante é que a significância estatística, via *p*-valores pequenos, não implica que a hipótese nula é improvável. A lógica incorreta que subjaz essa conclusão equivocada (cf. Falk & Greenbaum, 1995) aparentemente segue do seguinte modo: se a hipótese nula é verdadeira, dados com certas características são improváveis. Se os dados obtidos têm essas características, então a hipótese nula é improvável. Essa (pseudo) lógica tem como paralelo preciso o seguinte: Se a próxima pessoa que eu encontrar for um estadunidense, é improvável que ele ou ela seja o presidente [estadunidense]. Portanto, se eu encontrar o presidente na rua é improvável que ele seja um estadunidense. (p.152).

Um problema adicional é que o *p*-valor permite apenas uma decisão dicotômica sobre os resultados de uma pesquisa: ou os dados encontrados são estatisticamente significativos ou não são. O foco excessivo no resultado do NHST na análise de dados promove o que Branch (2014) chamou de "ciência sem tamanho", pois os resultados do NHST não nos informam sobre a magnitude da diferença entre a H0 e o que se esperava encontrar – seja esse resultado esperado a dependência funcional entre uma VI e uma VD ou uma associação mais ou menos forte entre duas variáveis. É possível, por exemplo, que uma VI afete sim uma VD, mas em uma magnitude tão pequena que seus efeitos podem ser negligenciáveis em qualquer situação prática. Tanto casos como esses como casos em que uma VI tem um efeito muito expressivo sobre uma VD resultariam, sem distinção, em resultados estatisticamente significativos. Quando os pesquisadores se atêm apenas à significância estatística eles deixam de lado o que

seria o mais relevante para os pesquisadores: identificar a magnitude das relações entre as variáveis estudadas. É verdade que há formas estatísticas de se calcular o tamanho de efeito em uma análise de dados, mas além de ainda infrequente, a análise do tamanho de efeito é frequentemente condicionada à significância estatística, ou seja, só é analisada nos casos em que o *p*-valor se mostrou menor que o limiar estabelecido.

Por fim, se por um lado há diversas críticas em relação ao uso incorreto, exclusivo e limitado do uso de NHST, uma vez que assim ele se expressa em uma porção considerável dos estudos científicos; por outro lado há também uma descrença de que o resultado do NHST controle o comportamento dos pesquisadores tal como se esperaria que o fizesse se ele constituir como uma boa prática científica. Rozeboom (1960), por exemplo, apresenta a seguinte reflexão:

> Quem já desistiu de uma hipótese só porque um experimento gerou uma estatística de teste na região de rejeição? E qual cientista em sã consciência em algum momento ignoraria que há uma diferença apreciável entre o significado interpretativo dos dados, digamos, para os quais p=0,04 unilateral e os dados para os quais p=0,06, mesmo que o ponto de "significância" fora definido em p=0,05? Na verdade, o leitor pode não se sentir perturbado pelas acusações levantadas aqui contra o procedimento tradicional de NHST precisamente porque, talvez sem perceber, ele nunca levou o método a sério de qualquer maneira. (p. 424)

A despeito dos problemas levantados até aqui, a confiança de que o NHST seria uma medida objetiva e confiável se tornou tão disseminada na ciência psicológica que a exigência de apresentação de *p*-valores significativos acabou se tornando um critério não explícito para a publicação de artigos científicos que incluam análise de dados empíricos (e.g., Ferguson & Heene, 2012). Tal padrão criou vieses de publicação (van Aert & Niemeyer, 2022) que, combinados com políticas de financiamento e de carreira científica baseados na quantidade de

artigos publicados (e.g., Lilienfeld, 2017; Nosek et al., 2012) tornaram a inclusão de *p*-valores significativos quase que mandatória para o sucesso acadêmico. Tais contingências promovem a disseminação de práticas questionáveis de pesquisa que já têm sido amplamente descritos na literatura e, infelizmente, parecem mais prevalentes do que se gostaria de admitir (John et al., 2012; Swift et al., 2022).

Práticas questionáveis de pesquisa e o NHST

Na tentativa de produzir resultados estatisticamente significativos, alguns pesquisadores flexibilizam indevidamente o NHST de modo a produzir resultados que aumentem a probabilidade de publicarem seus artigos, tais práticas são popularmente conhecidas como *p-hacking*. Práticas que seguem nessa direção incluem parar a coleta de dados antes do planejado ou continuar a coleta de dados até que um resultado significativo seja obtido, transformar a distribuição obtida ou excluir seletivamente outliers da amostra sem justificativa plausível, tentar múltiplos métodos analíticos (e.g., aplicar diferentes testes estatísticos ou controlar seletivamente as covariáveis ou moderadores) até quem um resultado estatisticamente significante seja obtido, descrever apenas os resultados positivos obtidos (*cherry-picking*) ou arredondar o *p*-valor (Reis & Friese, 2022)[1]. Além disso, também não é incomum que autores formulem hipóteses após os resultados serem conhecidos (*HARKing*, Kerr, 1998). Práticas como essa, infelizmente, não têm sido tão incomum quanto gostaríamos (John et al., 2012; Moran et al., 2023) e têm resultado sistematicamente em alta frequência de falsos resultados positivos e tamanhos de efeito superestimados (van Aert & Niemeyer, 2022).

Ainda que em parte substancial dos casos essas práticas sejam deliberadas (Simmons et al., 2011), elas não são necessariamente cometidas por má fé e podem ser produto do desconhecimentos de boas práticas e da correta interpretação dos produtos da estatística

[1] Para uma lista mais completa de práticas questionáveis relacionadas ao uso de testes de hipótese nula ver Wicherts et al. (2016)

inferencial (e.g., Badenes-Ribera et al., 2015, 2016) que, por sua vez, está relacionada ao ensino deficitário da estatística inferencial (e.g., Cassidy et al., 2019; Friedrich et al., 2018; Huberty, 1993).

Muitas saídas têm sido propostas para esses problemas, mas as mais populares são o investimento no melhor ensino da estatística e o uso de pré-registros e *registered reports*. A recomendação para o uso de pré-registros públicos se refere a criação de um documento aberto e datado no qual o pesquisador descreve suas predições, hipóteses e plano de análise de dados antes de começar a coleta de dados (Krypotos et al., 2022). Bases de dados específicos para esse fim foram disponibilizados por diferentes instituições científicas sendo a Opens Science Framework e o site AsPredicted os mais conhecidos. *Registered Reports* são casos específicos de pré-registros nos quais os autores submetem a introdução justificativa e plano de análise de dados para a revista na qual almejam publicar antes da coleta de dados e tal revista pode, a partir de uma análise prévia por pares, concordar com um pré-aceite do artigo independente de serem obtidos resultados estatisticamente significativos ou não. Infelizmente, ainda são poucos os jornais que aceitam *Registered Reports* (Montoya et al., 2021), mas tanto essa quanto outras formas de pré-registro são bastante recomendadas como formas de minimizar práticas questionáveis de pesquisa, em especial para estudos que incluem NHST (Krypotos et al., 2022).

Orientações para os pesquisadores, editores e revisores sobre melhores práticas de estatística inferencial na Psicologia

As críticas em relação ao uso do NHST dividem os cientistas e estatísticos. Enquanto alguns entendem que seus problemas são graves demais para que se mantenha o uso adequado na ciência, outros sugerem que seu uso ponderado pode ainda ser mantido desde que tomados alguns cuidados. O que parece consensual, contudo, é que o uso apropriado do NHST deveria ser mais limitado do que tem ocorrido e que, quando implementado, seus resultados devem

ser analisados judiciosamente. Tais preocupações se concretizam em uma série de prescrições feitas para autores, revisores e editores de revistas científicas que serão apresentadas a seguir. Tal como na reflexão a respeito das críticas ao NHST, aqui não serão apresentadas propostas originais ou inovadoras, mas sim adaptações de diferentes prescrições já disponíveis na literatura.

Orientações para Autores

Os autores de estudos empíricos são os principais responsáveis pelas análises dos dados apresentados no estudo e pelas conclusões derivadas dessas análises. Mesmo que serviços de estatísticos sejam contratados pontualmente, é dos autores a responsabilidade pela apresentação e julgamento adequado dos resultados da pesquisa. Os autores são, portanto, os principais agentes que podem impedir a perpetuação dos erros que infelizmente se tornaram comuns na produção científica. A seguir sintetizamos sugestões para autores, revisores e editores retiradas, do manual de publicação da American Psychological Association (2019), das prescrições da American Statistical Association (Wasserstein & Lazar, 2016), de comentadores da área de estatística (Bakker & Wicherts, 2011; Counsell & Harlow, 2017; Hand, 2022; Wasserstein et al., 2019; Wicherts et al., 2016) e de editoriais a esse respeito publicados em diferentes periódicos científicos (Harrison et al., 2020; McCarren et al., 2017; Schreiber, 2020):

- Faça o pré-registro do seu estudo antes da coleta de dados, preferencialmente em uma base conhecida como a Open Science Framework (https://osf.io) ou a AsPredicted (aspredicted.org), ou considere submeter seu projeto de pesquisa a um jornal que aceite Registered Reports.
- Não embase suas conclusões apenas no achado de que uma associação ou efeito está relacionado a um *p*-valor que ultrapassou um limiar pré-definido (e.g., $p < 0,05$).

- Não afirme que uma associação ou efeito existe apenas porque ele foi considerado "estatisticamente significativo". Mesmo que seu resultado sugira que há uma associação ou efeito, ele está baseado apenas em uma de um conjunto infinito de amostras.
- Não afirme que uma associação ou efeito não existe apenas porque não se mostrou "estatisticamente significativo".
- Não sugira que o *p*-valor obtido indica (1) a probabilidade de que seu resultado foi produzido pelo acaso ou (2) a probabilidade de que sua hipótese sob teste é verdadeira.
- Não conclua nada sobre a significância científica ou prática dos seus resultados baseado na significância estatística (ou na falta dela).
- Evite, sempre que possível, o uso da expressão estatisticamente significativo (bem como suas variantes, e.g., "estatisticamente significante"). Essa expressão induz ao raciocínio dicotômico que carrega boa parte dos problemas do *p*-valor.
- Desde as fases iniciais de planejamento, programe-se de modo a facilitar a replicação.
- Amplie a apresentação de estatísticas descritivas e o uso de recursos visuais (tabelas e gráficos).
- Descreva detalhadamente seus métodos, se não no próprio texto do artigo, ao menos em material suplementar. Seja específico na menção aos testes utilizados. Por exemplo, indique que utilizou o "Teste T para amostras pareadas" ao invés de dizer apenas que usou o "Teste T" ou que aplicou um "teste de diferença de médias".
- Descreva o tratamento dado aos *outliers* e a dados faltantes, justifique o tratamento dado a essas informações.
- Dê preferência para o uso de estatísticas não dicotômicas como intervalos de confiança e tamanhos de efeito, mas não julgue a significância estatística a partir da constatação de que a hipótese nula cai ou não dentro do intervalo de confiança, isso seria cair novamente em um raciocínio dicotômico que compartilha alguns dos problemas do *p*-valor.

- Se optar por manter o uso de NHST, apresente o *p*-valor de modo apenas descritivo (e não como critério único de decisão sobre a validade dos dados) e o acompanhe de outras medidas de avaliação (e.g., fatores bayesianos, valores s, tamanhos de efeito, intervalos de confiança)
- Siga as normas da APA para relatos de estatísticas inferenciais.
- Apresente todas as estatísticas relevantes, não apenas aquelas que se mostraram estatisticamente significativas. Vieses de seleção dos resultados são altamente prejudiciais para o avanço científico.
- Explicite textualmente todas hipóteses a serem testadas no estudo;
- Certifique-se de que os testes utilizados são apropriados (avaliam corretamente a relação que se quer investigar);
- Justifique o tamanho da amostra em função do poder estatístico que ela pode produzir.
- Certifique-se de que a análise estatística implementada é robusta e que todos os pré-requisitos (e.g., distribuição, homoscedasticidade, variância) foram atendidos.
- Utilize controles apropriados para múltiplas comparações (e.g., correções de Bonferroni, Benjamini & Hochberg, Holm) e justifique a estratégia utilizada.
- Apresente os *p*-valores exatos e com pelo menos 3 casas decimais.
- Não interprete o *p*-valor isoladamente.
- Apresente os tamanhos de efeito apropriados e os intervalos de confiança para os tamanhos de efeito.
- Defina e justifique os níveis de tamanhos de efeito (e.g., baixo, médio, alto) em relação a significância prática.
- Interprete os tamanhos de efeito no contexto para informar sobre a significância prática.
- Certifique-se que você não está cometendo nenhuma interpretação falaciosa dos resultados de testes estatísticos.

- Sempre que possível, adote práticas de ciência aberta: realize o pré-registro do seu estudo antes de realizar a análise dos dados e disponibilize protocolos detalhados da análise dos dados bem como os dados brutos da pesquisa como materiais suplementares do seu estudo.

Orientações para Revisores

- Insista que os autores sigam as recomendações da APA no relato do método e dos resultados de estatísticas inferenciais.
- Não use o *p*-valor como um critério para avaliar se um resultado é importante ou não.
- Exija métodos detalhados e dedique mais tempo na análise de tais métodos, especialmente se eles incluírem dados derivados de NHST.
- Certifique-se que os autores descreveram o método com suficiente detalhe para permitir a replicação do estudo; que descreveram os resultados de modo a permitir análises alternativas e que interpretaram corretamente os resultados de análises em NHST caso existam.
- Sugira que os autores disponibilizem dados abertos de sua pesquisa em material suplementar.
- Consulte as orientações para editores a seguir, muitas delas cabem para o revisor.

Orientações para Editores

O editor que tramita um artigo também é responsável por detectar interpretações incorretas de estatísticas inferenciais e avaliar se os autores declararam e avaliaram os pré-requisitos dos modelos estatísticos utilizados.

- Evite tomar decisões editoriais baseadas na obtenção de significância estatística.
- Sugira que os autores evitem o uso de expressões como “estatisticamente significante” ou “estatisticamente significativo”.
- Avalie como expandir a seção de método dos artigos ou crie condições para que informações detalhadas sejam apresentadas em material suplementar.

- Exija dos autores descrições detalhadas do método e dos procedimentos de análises de dados.
- Exija dos autores relatem quantos testes estatísticos foram realizados, quais dados foram descartados, quais pré-requisitos foram testados ou garantidos e quais testes estatísticos foram considerados.
- Exija dos autores justifiquem a escolha do nível de significância adotado. Usar $p<0,05$ simplesmente porque é tradicional na área não é uma boa prática científica. O $\langle$ deve ser definido a partir de uma avaliação racional sobre as probabilidades de erros do Tipo I ou do Tipo II em cada caso.
- Insista na descrição de quantos testes estatísticos foram desenvolvidos (quantas hipóteses foram testadas) e como esses testes foram usados e interpretados. Cada hipótese testada representa uma chance de falso positivo ou falso negativo. A quantidade total de hipóteses testadas é importante para interpretar as estatísticas inferenciais isoladas.
- Exija dos autores que interpretem seus resultados baseados não apenas no *p*-valor obtido, mas na adequação do design metodológico, na possibilidade de vieses e no contexto dos resultados já disponíveis na literatura.
- Incorpore práticas da ciência aberta para permitir análises independentes dos resultados descritos nos estudos. Disponibilizar não apenas os resultados resumidos no corpo do texto, mas os resultados brutos e uma descrição detalhada dos métodos empregados em materiais suplementares permite análises por outros cientistas, a conferência da adequação das análises empregadas e a reanálise dos dados a luz de outras estratégias científicas. Criar condições para disponibilizar essas informações, bem como políticas de incentivo (ou até exigência) de que os dados brutos e métodos detalhados sejam apresentados, promoverá o avanço consistente das análises científicas.

- Sensibilize a equipe editorial a respeito da importância do cuidado com os aspectos aqui discutidos.
- Inclua nas políticas editoriais menções aos problemas aqui discutidos.
- Considere a possibilidade de incluir um estatístico que revise e editore o uso de estatísticas nos artigos publicados no periódico: um editor de estatística.

É de fundamental importância que todos os envolvidos na produção do conhecimento científico estejam atentos aos possíveis erros e vieses presentes em suas pesquisas. O cuidado na definição, aplicação, descrição e interpretação das informações estatísticas, como meio de chegar a conclusões dos estudos deve ser considerado em todas as etapas da produção do conhecimento. Portanto a responsabilidade pela produção de resultados cientificamente significativos deve ser compartilhada por todos os envolvidos desde seu processo de ensino, até suas aplicações e na disseminação do conhecimento produzido.

Referências

American Psychological Association. (2019). *Publication manual of the American Psychological Association*. American Psychological Association.

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frías-Navarro, D., Monterde-I-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p-value: A national survey study in academic psychologists from Spain. *Psicothema*, *27*(3), 290–295. https://doi.org/10.7334/psicothema2014.283

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-

011-0089-5

Branch, M. N., & Pennypacker, H. S. (2013). Generality and generalization of research findings. *APA handbook of behavior analysis, Vol. 1: Methods and principles.*, *1*, 151–175. https://doi.org/10.1037/13937-007

Branch, Marc. N. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology*, *24*(2), 256–277. https://doi.org/10.1177/0959354314525282

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399. https://doi.org/10.17763/haer.48.3.t490261645281841

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233–239. https://doi.org/10.1177/2515245919858072

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Counsell, A., & Harlow, Lisa. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology*, *58*(2), 140–147. https://doi.org/10.1037/cap0000074

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology is anything changing? *Psychological Science*, *18*(3), 230–232. https://doi.org/10.1111/j.1467-9280.2007.01881.x

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology*, *53*(2), 133–149. https://doi.org/10.1037/h0087305

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist*, *53*(7), 798–799. https://doi.org/10.1037/0003-066X.53.7.798

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98. https://doi.org/10.1177/0959354395051004

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Friedrich, J., Childress, J., & Cheng, D. (2018). Replicating a national survey on statistical training in undergraduate psychology programs: Are there "new statistics" in the new millennium? *Teaching of Psychology*, *45*(4), 312–323. https://doi.org/10.1177/0098628318796414

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. Em *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409). SAGE. https://doi.org/10.4135/9781412986311.n21

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20.

Hand, D. J. (2022). Trustworthiness of statistical inference. *Journal of the Royal Statistical Society*, *185*(1), 329–347. https://doi.org/10.1111/rssa.12752

Harrison, A. J., McErlain-Naylor, S. A., Bradshaw, E. J., Dai, B., Nunome, H., Hughes, G. T. G., Kong, P. W., Vanwanseele, B., Vilas-Boas, J. P., & Fong, D. T. P. (2020). Recommendations for statistical analysis involving null hypothesis significance testing. *Sports Biomechanics*, *19*(5), 561–568. https://doi.org/10.1080/14763141.2020.1782555

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. *Educational and Psychological Measurement*, *60*(5), 661–681. https://doi.org/10.1177/00131640021970808

Huberty, C. J. (1993). Historical origins of statistical testing practices. *The Journal of Experimental Education*, *61*(4), 317–333. https://doi.org/10.1080/00220973.1993.10806593

Imam, A. A., & Frate, M. (2019). A snapshot look at replication and statistical reporting practices in psychology journals. *European Journal of Behavior Analysis*, *20*(2), 204–229. https://doi.org/10.1080/15021149.2019.1680179

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy. *Methodology*, *4*(4), 152–158. https://doi.org/10.1027/1614-2241.4.4.152

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and*

Social Psychology Review, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*. American Psychological Association. https://doi.org/10.1037/14136-000

Krypotos, A.-M., Mertens, G., Klugkist, I., & Engelhard, I. M. (2022). Preregistration: Definition, advantages, disadvantages, and how it can help against questionable research practices. Em W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 343–357). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_15

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes*, *8*(84), 1–9. https://doi.org/10.1186/s13104-015-1020-4

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—Significance tests are not. *Theory & Psychology*, *22*(1), 67–90. https://doi.org/10.1177/0959354311429854

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664. https://doi.org/10.1177/1745691616687745

Loftus, G. R., & Loftus, R. G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*(6), 161–171. https://doi.org/10.1111/1467-8721.ep11512376

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159.

McCarren, M., Hampp, C., Gerhard, T., & Mehta, S. (2017). Recommendations on the use and nonuse of the p-value in biomedical research. *American Journal of Health-System*

Pharmacy, *74*(16), 1262–1266. https://doi.org/10.2146/ajhp160443

Meehl, P. E. (1967). Theory-testing in Psychology and Physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Montoya, A. K., Krenzer, W. L. D., & Fossum, J. L. (2021). Opening the door to Registered Reports: Census of journals publishing Registered Reports (2013–2020). *Collabra: Psychology*, *7*(1), 24404. https://doi.org/10.1525/collabra.24404

Moran, C., Richard, A., Wilson, K., Twomey, R., & Coroiu, A. (2023). I know it's bad, but I have been pressured into it: Questionable research practices among psychology students in Canada. *Canadian Psychology*, *64*(1), 12–24. https://doi.org/10.1037/cap0000326

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Nunnally, J. (1960). The palce of statistical in psychology. *Educational and Psychological Measurement 1960*, *20*(4), 641–650. https://doi.org/10.1177/001316446002000401

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*(1), 159–163. https://doi.org/10.1037/0033-2909.102.1.159

Reis, D., & Friese, M. (2022). The myriad forms of p-hacking. Em W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 101–121). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_5

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

Schreiber, J. B. (2020). New paradigms for considering statistical significance: A way forward for health services research journals, their authors, and their readership. *Research in Social and Administrative Pharmacy*, *16*(4), 591–594. https://doi.org/10.1016/j.sapharm.2019.05.023

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sohn, D. (1998). Statistical significance and replicability. *Theory & Psychology*, *8*(3), 291–311. https://doi.org/10.1177/0959354398083001

Swift, J. K., Christopherson, C. D., Bird, M. O., Zöld, A., & Goode, J. (2022). Questionable research practices among faculty and students in APA-accredited clinical and counseling psychology doctoral programs. *Training and Education in Professional Psychology*, *16*(3), 299–305. https://doi.org/10.1037/tep0000322

Thompson, B. (1999). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*(2), 191–196. https://doi.org/10.1177/095935439992007

van Aert, R. C. M., & Niemeyer, H. (2022). Publication bias. Em W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding Questionable Research Practices in Applied Psychology* (pp. 213–242). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_10

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p <

0.05". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.