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: 2024-05-02 (versão 1)
(AAAA-MM-DD)

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

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

Resumo

Amplamente adotada na Psicologia em geral, a estatística inferencial também é frequente na Análise do Comportamento (AB), abordagem que historicamente favoreceu estudos experimentais de caso único. 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 (intrínsecos e por mau uso). 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 e à má interpretação do NHST e 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 Behavior Analysis: Problems and Recommendations

Abstract

Widely adopted in Psychology in general, inferential statistics is also prevalent in Behavior Analysis (BA), an approach that historically favored single-case experimental studies. The dissemination of research with groups using null hypothesis significance testing (NHST) in data analysis has grown in this area. It brings with it related problems (intrinsic and due to

misuse). Such 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 the misuse and misinterpretation of the NHST and compile guidance for editors, reviewers, and authors that can be adopted to minimize the problems mentioned.

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

Prueba de significancia de hipótesis nuls en la Análisis de la Conducta: Problemas y recomendaciones

Resumen

Ampliamente adoptada en la Psicología en general, la estadística inferencial también prevalece en la Análisis de la Conducta, una perspectiva que históricamente favoreció los estudios experimentales de caso único. En este ámbito ha crecido la difusión de investigaciones con grupos que utilizan pruebas de significancia de hipótesis nula (Null Hypothesis Significance Testing - NHST) en el análisis de datos y eso trae problemas relacionados (intrínsecos y por mal uso). Estos problemas a menudo pasan desapercibidos en el actual sistema de revisión por pares, comprometiendo la confiabilidad de algunas conclusiones disponibles en la literatura científica. En este artículo, explicamos los problemas relacionados con el mal uso y la mala interpretación del NHST y recopilamos pautas para editores, revisores y autores que pueden adoptarse para minimizar los problemas mencionados.

Palabras clave: Estadística inferencial; Análisis de la Conducta, prueba de significancia de hipótesis nula

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

O uso de métodos quantitativos e a busca por maneiras de avaliar os dados obtidos nas pesquisas empíricas não é recente na Psicologia. De acordo com Cowles (2001), Gustav Theodor Fechner (1801-1887) foi quem começou a estudar processos mentais por meio de métodos quantitativos, publicando em 1860 o "Elemente der Psychophysik", no qual apresenta uma lei psicofísica que descreve, no seu entendimento, a relação entre mente e corpo. Após esse período, foram adotadas na Psicologia Experimental duas estratégias amplas para lidar com a variabilidade dos dados, a abordagem analítica experimental e a inferência estatística (Cowles, 2001). Atualmente ainda predominante na Psicologia, o uso de estatística inferencial passou também a ser adotado na Análise do Comportamento (AC), abordagem que tem como método preferível estudos de caso único nos quais o uso da estatística inferencial tradicional é limitado. Portanto, torna-se oportuna uma avaliação quanto a utilização de estatística inferencial nesta abordagem.

O uso de estatística para a análise de dados e análises de probabilidades se fizeram presente em todas as áreas da ciência desde seu surgimento. Segundo Fienberg (1992), a estatística próxima àquela utilizada hoje, denominada estatística moderna, surgiu no século XX. O uso da estatística aplicada a pesquisas experimentais com o tratamento de pequenas amostras e o objetivo inferencial ganhou destaque com Ronald Aylmer Fisher (1890-1962). Fisher foi quem desenvolveu as noções de modelo estatístico, os conceitos de suficiência, probabilidade e aleatorização (Fienberg, 1992). Apesar de seu impacto na

estatística moderna e o grande valor atribuído ao seu trabalho, algumas de suas proposições geraram e ainda geram controvérsias, como o teste de significância. Por meio desse teste Fisher "forneceu um método de inverter declarações de probabilidade sobre observações de valores de parâmetros em declarações de probabilidade sobre parâmetros de observações sem usar o teorema de Bayes, que fornece um mecanismo para tal inversão" (Fienberg, 1992, p. 219).

Adicionalmente ao trabalho de Fisher, outros dois autores se destacaram no desenvolvimento de uma nova estatística inferencial: Jerzy Neyman (1894-1981) e Egon Sharpe Pearson (1895-1980). A proposta de Neyman e Pearson consistia em substituir o teste com hipótese nula única de Fisher pelo seu próprio teste de hipóteses, no qual duas hipóteses rivais são propostas e o resultado da pesquisa deve ser avaliado por meio destas alternativas (Gigerenzer et al., 1989).

A utilização da estatística inferencial nas ciências, seja pela abordagem de Fisher, pela abordagem de Neyman-Pearson ou pela "mistura" delas, logo fora disseminada para a Psicologia. Porém muitas críticas ao seu (mau) uso foram desenvolvidas. Além dos problemas já identificados e relatados na Psicologia, problemas relacionados ao mau uso da estatística inferencial e teste de significância de hipótese nula (*null hypothesis significance testing* - NHST) e problemas intrínsecos que são evidentes na Psicologia em geral podem também ser identificados na AC, quando os pesquisadores dessa área realizam estudos com grupos e aplicam alguma forma de NHST como estratégia de avaliação de dados. Este artigo pretende explicitar quais são os problemas relacionados ao mau uso ou má interpretação decorrente do uso do NHST e quais são os problemas intrínsecos que podem ser identificados na Psicologia em geral e na AC. Além disso serão apresentadas

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 Fisher (teste de significância), quanto a proposições de Neyman e Pearson (teste de hipóteses), ou por uma alternativa hibrida comum na literatura conhecida como teste de significância de hipótese nula (*null hypothesis significance testing* – NHST) que embora tenham algumas diferenças importantes, conduzem a um raciocínio semelhante. Fisher foi quem criou o teste de significância em que o valor-p 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 valor-p*,* 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 valor-p menor que o valor de significância escolhido, a H0 e o resultado é frequentemente descrito como estatisticamente significativo. Para Fisher o valor-p é utilizado como uma força de evidência de hipótese, portanto ainda que todo valor-p menor que o nível de significância (por exemplo, e.g., $p = 0{,}048$) possa ser considerado estatisticamente significativo, valores-p menores (e.g., $p < 0{,}001$), aumentam a confiança de que o resultado é estatisticamente significativo.

Discordando de Fisher, Neyman e Pearson criaram o teste de hipóteses e introduziram a noção de hipótese alternativa (HA) a ser testada, então esta 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, então tomando como H0 a proposição de que não haveria diferença entre os dois tratamentos. Primeiro, deve-se escolher qual a H0 – neste caso o autor opta pela H0: $\mu = 0$, sendo μ (letra grega mu) a média das diferenças nos dois medicamentos em toda a população de interesse. Quando o raciocínio proposto por Neyman e Pearson é utilizado para analisar os dados, uma HA também é especificada. Nesse caso é que existe alguma (qualquer) diferença entre o remédio em teste e o remédio antigo, HA: $\mu \neq 0$. 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 e a HA ganha força, apoiando a hipótese de pesquisa de que o novo medicamento é melhor que o anterior. Na lógica proposta por Neyman e Pearson diferentes valores-p 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 valor-p obtido em um teste apropriado e o limiar de significância (ou alfa) previamente estabelecido. É importante notar, entretanto, que o valor-p 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: *p* é 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 valor-p, portanto, não informa absolutamente nada (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 valor-p na estatística inferencial é uma fonte muito comum de erros na disseminação e divulgação do conhecimento científico, incluindo a Psicologia em geral e a AC em específico. 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.

Mau uso ou à má intepretação dos resultados de NHST

As críticas ao uso dos testes de significância na Psicologia podem ser divididas entre aspectos relacionados ao mau uso ou à má intepretação dos resultados desses testes e a aspectos intrínsecos, ou seja, que necessariamente fazem parte da lógica de funcionamento dos testes. As críticas relacionadas ao mau uso ou a má interpretação dos resultados provenientes do NHST podem ser divididas em algumas falácias identificadas na literatura que serão descritas a seguir.

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., 2012; 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 valor-p no teste estatístico apropriado for menor que o ⟨, é incorreto dizer que o valor-p 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 valor-p (é um pressuposto), não é algo que esteja sendo avaliado nos NHSTs.

Se é incorreto interpretar que o valor-p é a probabilidade da H0 ser falsa, também é incorreto, ainda que também seja comum, interpretar que a probabilidade complementar ao valor-p (1-p) é a probabilidade da HA ser verdadeira (Carver, 1978; Cohen, 1990; Gigerenzer et al., 2012; Kline, 2013; Lambdin, 2012; Nickerson, 2000). A crença apresentada por muitos pesquisadores é de que o valor-p é a probabilidade da H0, portanto para estes a probabilidade aceitável, dados os resultados, de que a H0 é verdadeira é de 0,05 ou menos e que seu complemento, 0,95, é a probabilidade de que a HA é verdadeira.

O entendimento de que o complemento de que o valor-p é a probabilidade de um resultado ser replicado também representa uma interpretação incorreta das informações produzidas pelo NHST (Badenes-ribera et al., 2015; Carver, 1978; Falk & Greenbaum, 1995; Gigerenzer et al., 2012; 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 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 o pesquisador identificou 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) a significância prática, definida como a força da relação entre duas variáveis, foi descrita por volta do século XVIII, porém, mais recentemente, tem-se confundido a significância prática com o valor-p. Para o autor, as duas interpretações equivocadas mais comuns relacionadas a significância são: (1) a que a falha em rejeitar H0 significa que o efeito ou associação na população é zero e (2) a que o tamanho do valor-p 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).

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 valor-p resultado do experimento for acima do valor de 0,05 (normalmente adotado) é porque não há efeito ou associação entre as variáveis investigadas e os resultados ocorreram devido ao acaso. Considerando que valor-p 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 valor-p seja a probabilidade de que o acaso causou a diferença média entre dois grupos de pesquisa, uma vez que (a) o valor de p foi calculado assumindo que a probabilidade era 1,00 de que o acaso causou a diferença

> média, e (b) o valor de p 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 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 ao definir que se usará testes de hipótese nula, outras decisões subjetivas devem ser tomadas, como o tamanho da amostra, valor de ⟨, a decisão se será realizada uma análise unicaudal ou bicaudal etc. Além disso, o NHST não é a única estrutura estatística para testar hipóteses, sendo a estimativa Bayesiana um exemplo de alternativa disponível.

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 a falácia da robustez. A falácia da santificação se refere ao

pensamento dicotômico. Se o critério de avaliação do valor-p 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 considera o valor-p como força de evidência contra a H0. Também deve-se considerar que ao avaliar os efeitos ou associações entre variáveis, mudanças grandes no valor-p podem significar mudanças pequenas na variável avaliada (no tamanho do efeito em estudos experimentais ou no grau de associação entre variáveis em estudos correlacionais). Na falácia da reificação, a falha em tomar a mesma decisão de estudos anteriores sobre a H0 significa 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. Ao adotar essa crença se ignora tamanho da amostra, tamanho do efeito e poder do teste em diferentes estudos (Kline, 2013). Um exemplo para melhor compreensão do porquê falhar em tomar a mesma decisão sobre a H0 não pode ser considerada falha na replicação foi expresso por 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 valor-p de 0,02. É possível para um segundo pesquisador para fazer um estudo com menos poder, obtenha a mesma amostra média de 60% e 40%, mas obtenha um valor de p 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.

O uso inapropriado dos testes estatísticos também é uma má prática relativamente comum na análise de dados. A escolha do teste estatístico apropriado 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., 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).

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 própria AC.

Problemas intrínsecos do NHST na Psicologia e na Ciência em geral

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 ainda mais essenciais que são relacionadas ao próprio raciocínio lógico que embasa o NHST. Esses problemas são chamados aqui de problemas intrínsecos ao NHST.

A primeira crítica expressa por analistas do comportamento (Branch, 2014; Branch & Pennypacker, 2013) psicólogos (Bakan, 1966; Cohen, 1990, 1994; Lambdin, 2012; G. R. Loftus & Loftus, 1996; Lykken, 1968; Meehl, 1967, 1978; Nunnally, 1960; Simmons et al., 2011; Thompson, 1998) 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 o número 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 do número de casos no trabalho psicológico é caracteristicamente arbitrária ou apressada (p.426).

Portanto se houver algum desvio da H0 na população, ou seja, algum efeito ou associação entre as variáveis investigadas, por menor que ele seja, 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, 1990, 1994b; Falk, 1998; Falk & Greenbaum, 1995; G. R. Loftus & Loftus, 1996; Meehl, 1978; 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). A probabilidade de que alguém morra se for contaminado com o COVID-19 no Brasil, é consideravelmente diferente da probabilidade de os mortos (em geral) no Brasil terem sido contaminados por COVID-19. 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 (Cohen, 1990; Falk & Greenbaum, 1995):

> O fato importante é que a significância estatística, via valores-p 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. Portanto, é improvável que ele ou ela seja um estadunidense. (p.152).

Um problema adicional é que o NHST permite apenas uma decisão dicotômica sobre os resultados de uma pesquisa: ou os dados encontrados são estatisticamente significativos ou não são. Branch (2014) indicou também que a utilização exclusiva do NHST na análise de dados promove o que ele chamou de “ciência sem tamanho”, pois se o valor-p representa apenas a probabilidade de os resultados serem obtidos dada a H0 verdadeira ou alfa a probabilidade da ocorrência de um erro tipo I (considerando as proposições de Fisher e de Neyman e Pearson, respectivamente). Neste caso 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 um 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 ou a grandeza das relações entre as variáveis estudadas.

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)

As críticas apresentadas a partir de problemas intrínsecos do uso do NHST são bastante comuns na Psicologia. Conforme descrito, as principais críticas indicadas por pesquisadores da área são: o fato de que a significância sempre pode ser alcançada; de que a lógica do NHST assume algo diferente ao que se quer testar; NHST promove “ciência sem tamanho” não fornecendo informações sobre a grandeza das relações de modo a ser muito difícil aplicar verdadeiramente comportamento inferencial na pesquisa científica.

Embora muito difundido não se pode negar que a análise de dados baseadas no NHST são bastante discutíveis e polêmicas na ciência em geral e na Psicologia. Na AC devido ao seu modo de fazer ciência utilizando caso único o uso de NHST e estatística inferencial apresentam ainda alguns problemas particulares.

Problemas intrínsecos do NHST na AC

Em periódicos voltados a divulgação de pesquisas em AC, as críticas sobre o uso de NHST também se fazem presentes. Alguns argumentos apresentados nestes periódicos são semelhantes aos fornecidos por autores da Psicologia em geral como: a defesa de que a significância estatística sempre pode ser alcançada (Branch 1999, 2019; Chase & Tucker 1976; Hopkins et al., 1998; Perone, 1999) a lógica de que ao utilizar NHST assume-se o oposto ao que se quer provar, logo o teste de significância não fornece informações diretas sobre a hipótese de interesse (Branch & Pennypacker, 2013; Killeen, 2019); testes de significância estatística não fornecem estimativas quantitativas de confiabilidade de um resultado (Branch 1999, 2019), ou seja, os testes de significância não fornecem informações de tamanhos de efeito ou quantidades práticas, mas apenas informações estatísticas baseadas em uma amostra através da teoria das probabilidades; ao fazer uso de testes de significância estatística os pesquisadores projetam seus experimentos para obter esse tipo de dados, criando uma dependência do teste para a análise dos resultados e deixando de realizar análises experimentais completas (Baron, 1990, 2000; Michael, 1974; Perone, 1999), ou seja aplicando método limitado a produzir um resultado dicotômico entre aceitar ou não uma hipótese, deixando de explorar outros aspectos que poderiam ser analisados.

Na AC, o método preferível para se fazer pesquisa se diferencia das demais abordagens utilizadas na Psicologia e nas demais Ciências Sociais de várias maneiras, a

começar pela definição de seu objeto de estudo. Diferentemente das demais abordagens psicológicas, na AC compreende-se que o comportamento é o objeto de estudo e não apenas um indicador de alguma outra coisa. A metodologia preferencialmente utilizada foi desenvolvida para estudar esse objeto. Na AC, para entender porque certos comportamentos são emitidos, a avaliação é feita identificando as variáveis das quais o comportamento é função. Portanto, experimentos são realizados promovendo alterações em VIs e avaliando se houve um efeito ou uma alteração na VD ou comportamento avaliado. Logo, ao analisar e identificar variáveis das quais o comportamento é função (VIs) é possível prever o comportamento e, em alguns casos, controlar o comportamento na medida que essas variáveis podem ser manipuladas (Skinner, 1953).

Diferentemente da lógica hipotético-dedutiva típica do NHST, na AC o meio preferível de fazer experimentação é indutivo, ou seja, a partir da identificação na regularidade dos dados é que se baseia a teoria. Segundo Sidman (1960), nesse cenário o experimentador escolhe uma área de pesquisa para investigá-la cuidadosamente, ao fazer isso este avaliará as inter-relações existentes entre os fenômenos, a forma e semelhança entre as variáveis que são relevantes para esses fenômenos e poderá desenvolver técnicas de controle experimental do comportamento. Para o autor, as técnicas empregadas para o controle experimental são caracterizadas em termos de variáveis que são manipuladas e das consequências comportamentais resultantes de tal manipulação. Logo para se indicar que uma técnica é adequada ao experimento deve ser estabelecida a precisão e confiabilidade do controle que realiza.

A metodologia preferencial adotada na AC é denominada delineamento de caso único (também conhecido como delineamento de N=1, de sujeito único ou delineamento intrasujeito) e sua característica principal é tratar os sujeitos individualmente, tanto nas

decisões relativas ao próprio delineamento quanto na análise dos dados (Sampaio et al., 2008). Estudos que utilizam delineamento de caso único costumam realizar medição dos comportamentos de um indivíduo antes, durante e, em alguns casos, após uma intervenção (Iversen, 2012). Nesta abordagem substitui-se a aferição do comportamento de muitas pessoas submetidas a diferentes situações (típicas do delineamento de grupos) por muitas aferições do comportamento do mesmo indivíduo submetido a todas as situações planejadas (Sampaio et al., 2008). Essas avaliações ocorrem a partir da identificação, registro e manipulação da(s) VI's, e da verificação de seu efeito sobre a(s) VD's.

O delineamento de caso único é um meio de aplicar a ciência do comportamento ao nível do sujeito individual. É importante salientar que embora a aplicação do experimento, bem como a análise de dados seja feita com sujeitos individuais, isso não quer dizer que apenas um sujeito é avaliado, ou seja, em uma pesquisa diversos sujeitos podem ser expostos às mesmas condições experimentais e variações dessas condições, mas seus dados devem sempre ser tratados individualmente (Velasco et al., 2010). Nesse sentido, cada sujeito que participa de um experimento constitui uma replicação do experimento, dado que seus resultados são analisados individualmente. É por meio da replicação de dados de diferentes sujeitos e condições que é possível obter representatividade e generalidade sobre os resultados do estudo, assim, alcançando validade externa.

O uso de grupos em pesquisas na AC, portanto, não costuma ser a preferência dos pesquisadores. A avaliação feita com medidas agregadas do desempenho das pessoas que compõem um grupo não é equivalente a avaliar o comportamento individual, logo, ao medir uma curva de desempenho de um grupo, mesmo que esta apresente forma semelhante à de um indivíduo, ambas não fornecem a mesma informação. Portanto, segundo Sidman (1960) um experimento de grupos não pode ser um substituto mais controlado ou

generalizável que os dados individuais pois eles produzem dados diferentes. Como critério de fidedignidade e generalidade a replicação feita por meio da avaliação com casos únicos é um instrumento mais poderoso do que a replicação entre grupos (Sidman, 1960). Por meio deste tipo de replicação, cada novo experimento aumenta a representatividade dos resultados, estabelecendo a generalidade de dados entre os indivíduos de uma população. Na replicação entre grupos a fidedignidade é demonstrada por meio de mudanças na tendência central que podem ser repetidas.

De modo resumido, Sampaio et al. (2008) explicitam os principais motivos pelos quais analistas do comportamento preferem utilizar avaliação com caso único. Os autores afirmam que ao avaliar o comportamento se observa um fenômeno que é característico de sujeitos individuais que interagem de maneira única com o mundo, ou seja, dois indivíduos não se comportam da mesma maneira mesmo que diante dos mesmos fenômenos. Portanto, para alguns tipos de estudo em que é importante compreender se determinadas variáveis estão funcionalmente relacionadas ao comportamento do sujeito, cálculos que agregam resultados de desempenho como por exemplo médias de grupos na maioria dos casos não representarão corretamente o desempenho de nenhum dos membros avaliados. Entendendo que raramente um sujeito se comportará exatamente como a média, logo a utilização de estudos com caso único é mais indicada.

Para alguns pesquisadores estudiosos de estatística e AC (e.g., Baron, 1990; Baron & Derenne, 2000; Branch, 1999, 2019; Perone,1999), o uso do NHST nesta abordagem não é adequado, visto que testes de significância estatística enfatizam os parâmetros populacionais e o comportamento é um fenômeno individual, portanto ele não pode ser representado pela média de um grupo. Branch (1999) cita como exemplo “uma taxa de 2% de gravidez em uma população pode ter um significado importante para formuladores de

políticas (seguros e público), mas não tem significado para um indivíduo mulher, que nunca ficará 2% grávida" e também não tem 2% de chance de ficar grávida, assim embora estudos com grupos e o uso de estatísticas possam ser muito úteis para alguns tipos de objetivos, quando se busca avaliar comportamentos a análise com caso único é mais recomendada. Baron e Derenne (2000) ainda acrescentam que mesmo que as análises sejam conduzidas avaliando grupos que foram expostos a tratamentos diferentes, as relações identificadas nas análises dos grupos não têm origem em nenhum indivíduo em particular, ou seja, a análise de grupos pode levar a um resultado completamente diferente do que ocorreria ao analisar um único indivíduo, portanto pode ser um problema generalizar esse resultado e dizer que os indivíduos da população avaliada se comportarão daquela forma. Outro argumento é de que o NHST promove confusão do comportamento atuarial (uma ciência de parâmetros da população) com a ciência comportamental: "as médias amostrais de um grupo de indivíduos permitem inferências sobre a média da população, mas essas médias não permitem inferências para os indivíduos, a menos que seja demonstrado que a média é, de fato, representativa dos indivíduos." (Branch, 2014, p. 264).

O uso do NHST na Psicologia é duramente criticado por evidenciar diversos problemas intrínsecos e extrínsecos, na AC conforme indicado existem ainda argumentos adicionais contra o uso do NHST e estatísticas inferenciais tradicionais. Na AC, além das críticas já apontadas pela Psicologia, os problemas que são pertinentes ao seu modo de fazer ciência podem ser resumidos em: a problemática de que testes de significância estatística enfatizam os parâmetros populacionais e o comportamento é um fenômeno individual, o uso de testes de significância não abrange informações de sujeitos individuais podendo estes resultados ser muito diferentes de dados produzidos por grupos. Considerando os argumentos da área somados aos já levantados pela Psicologia de modo

geral, autores da AC expressam que o uso de estatísticas inferenciais também não é uma prática objetiva para examinar efeitos de variáveis independentes sobre dependentes, portanto, podem acrescentar pouco ou nada às informações que são produzidas pela metodologia de caso único (Hopkins et al., 1998). Autores mais radicais como Branch (2014) sugerem que o uso do NHST deve ser abandonado relatando que o que o NHST comunica é a desinformação sobre uma suposta confiabilidade dos dados obtidos e que nenhuma linguagem comum produzida a partir do NHST pode fornecer benefícios para a Ciência. No entanto mesmo não sendo o estudo com grupos e o uso de estatísticas inferenciais o meio preferível na AC, conhecer e aplicar alguns tipos de estatísticas inferenciais e estudos com grupos ainda podem ser úteis nesta abordagem.

Motivos para utilizar estatísticas Inferenciais na AC

Apesar de todas as críticas relacionadas ao NHST na Psicologia e a sugestão de abandono do uso do teste e da própria estatística inferencial, na AC muitos autores fazem uso desse tipo de análise estatística. Em periódicos relevantes para a área como Journal of Applied Behavior Analysis (JABA), Journal of the Experimental Analysis of Behavior (JEAB) e Revista Mexicana de Análisis de la Conducta (RMAC) foi identificado crescimento sistemático no uso de NHST em suas publicações ao longo dos anos (ver Acuña, 2010; Foster et al., 1999; Imam & Frate, 2019; Kratochwill & Brody, 1978b; Kyonka et al., 2019; Zimmermann et al., 2015), paralelo a esse crescimento é possível observar pesquisadores que defendem o uso de estatística inferencial na área (ver Crosbie, 1999; Reese, 1998, 1999; Shull, 1999), ainda que essa defesa não seja restrita ao uso do NHST.

Pressões externas relacionadas a publicações das pesquisas parece ser um dos motivos que está levando os analistas do comportamento a adotar o uso do NHST e outras

estatísticas inferenciais. Sabe-se que é prática comum que os analistas do comportamento atuem em instituições que os impulsionam a desenvolver pesquisas, no entanto muitas vezes para a publicação dessas pesquisas é comum a exigência ou favorecimento de publicação de pesquisas que contenham análises estatísticas nas práticas editoriais de alguns periódicos (Ator, 1999). Outro agravante para esse uso pode ser o favorecimento para a concessão de bolsas e financiamentos, visto que em muitos casos, ao avaliar propostas de bolsas e financiamentos, revisores favorecem estudos que contenham análises estatísticas e os analistas do comportamento precisam fazer suas propostas de pesquisa considerando a eventual "preferência" dos avaliadores para conseguir financiamento para realizar seus estudos (Ator, 1999; Johnston et al., 2020)

Adicional a aspectos extrínsecos de possíveis favorecimentos ou exigências, analistas do comportamento têm defendido alguns motivos pelos quais é importante que os estudiosos da área entendam os métodos utilizados em estudos com grupos e o uso de análises estatísticas. Um dos motivos indicados por Johnston et al. (2020) é de que os analistas do comportamento devem se tornar consumidores informados, visto que seu próprio público possivelmente entrará em contato com muitos estudos que contemplem análises com grupos. Logo é importante que os analistas do comportamento possam auxiliar seus consumidores a tomar decisões informadas a partir dessa literatura. Para atuar desse modo os analistas do comportamento devem saber avaliar:

> (a) quais tipos de questões de pesquisa podem ou devem ser abordadas em estudos que usam projetos de pesquisa de grupo e análises estatísticas; (b) como interpretar os resultados relatados de tais estudos; (c) as limitações de tais estudos, especialmente o que pode e não pode ser inferido sobre os efeitos das intervenções a partir de análises estatísticas de médias de grupo e outras abstrações matemáticas; e (d) como

> esses métodos de pesquisa diferem dos métodos de pesquisa em análise do comportamento (Johnston et al., 2020, p. 409).

Essas questões explicitam a necessidade dos analistas do comportamento em se comunicar com membros de outras áreas ou profissões, buscando também advogar por políticas que apoiem a AC em cumprir suas responsabilidades éticas de manter seus consumidores informados.

Existem questões que requerem comparações de grupos e, portanto, análises estatísticas. Existem algumas questões que são melhor abordadas em estudos que enfocam medidas de grupo em comparação com caso único, como por como exemplo “Qual texto de ciências o distrito escolar deve adotar para os alunos da sexta série?” (Johnston et al., 2020, p. 409). Perguntas como essa se referem a grupos de pessoas e não ao comportamento individual. Projetos de pesquisa de grupos e análises estatísticas podem ser utilizados para comparar efeitos de duas intervenções médicas ou comportamentais que por quaisquer razões que sejam, não podem ser aplicadas em um mesmo sujeito. Deste modo, pesquisadores na AC devem adquirir habilidades para identificar que tipos de questões podem ser abordadas de forma adequada por meio de estudos com grupos e, ao optar por realizar esse tipo de estudo, devem saber como projetá-los e conduzi-los de modo a produzir resultados adequados.

O terceiro motivo indicado por Johnston et al. (2020) para a utilização do uso de estatísticas inferenciais na AC é a possibilidade de que existem circunstâncias em que mesmo em pesquisas realizadas com caso único o uso de análises estatísticas de dados podem ser informativos para complementar análises visuais de dados obtidos por meio de análises de caso único. Para tal os autores sugerem que podem incluir a análise de dados agregados em vários estudos de caso único fazendo o uso de meta-análise, desde que os

analistas do comportamento saibam identificar quando esse tipo de análise é apropriado, bem como quais ferramentas estatísticas podem ser apropriadas para interpretar corretamente os resultados. Os autores citam como exemplo de situação em que a estatística inferencial pode auxiliar em análises de dados de comportamentos individuais casos em que há alta variabilidade de uma série de dados, ou ocorrências inesperadas em que podem ser difíceis de analisar simplesmente por observação. Nestes casos os autores sugerem que uma análise estatística que indique que os dados parecem estáveis, ou se existe uma tendência, podem ser úteis para interpretar a variabilidade registrada. Ainda que esses casos não exijam necessariamente o uso de NHST, um conhecimento de estatística inferencial ainda é necessário.

São diversos os motivos que fazem com que os analistas do comportamento utilizem de estatísticas inferenciais e/ou estudos com grupos de sujeitos. Alguns dos principais motivos identificados são: possível favorecimento para publicação ou financiamentos; tornar-se consumidores bem informados; existem questões que são melhor abordadas por meio de estudos com grupos; estudos que por algum motivo a intervenção não possa ser reaplicada em um mesmo sujeito; pela possibilidade de aplicar estatística inferencial em estudos com caso único. Independente de qual seja o motivo ou justificativa pelos quais analistas do comportamento utilizem estatística inferencial e/ou estudos com grupos, o importante é que este uso seja feito de forma a produzir resultados adequados e que a interpretação desses dados seja feita corretamente.

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

De modo a evitar o mau uso, má interpretação e aplicação de testes incoerentes com a pesquisa, e assim prejudicar os achados científicos, em 1996 o Conselho de Assuntos

Científicos (Board of Scientific Affairs - BSA) da APA se reuniu para discutir sobre duas questões: o papel do NHST na Psicologia e a modificação da prática adotada no tratamento quantitativo de dados na Psicologia. A reunião resultou na elaboração de quatro tópicos de orientação sobre o tratamento quantitativo de dados de pesquisa nos quais foram enfatizadas outras maneiras já existentes além do NHST, porém pouco utilizadas de fazer inferência estatística para analisar os dados, classificados como parte da chamada abordagem por estimativa. Baseados nesses tópicos e com algumas explicações adicionais pretende-se fornecer orientações aos os pesquisadores, editores e revisores sobre melhores práticas de estatística inferencial na Psicologia e AC.

O primeiro tópico de orientação sobre o tratamento quantitativo de dados fornecido pela APA está relacionado, às abordagens para melhorar a qualidade do uso de dados e para diminuir o risco de interpretações incorretas dos resultados quantitativos. A APA propõe a extensão de descrições dos dados incluindo informações como médias, desvios padrão, tamanhos de amostra, resumos de cinco pontos (*five-point sumaries*), gráficos como o de caixa-e-bigode (*box-plot*), descrições relacionadas aos dados ausentes, a melhor caracterização dos resultados e recomendação de não utilizar apenas valor-p*,* mas também incluir estatísticas por estimativa.

A utilização de estatísticas por estimativa sugeridas pela APA são: tamanho de efeito (*effect size* – ES), intervalos de confiança (IC) e barras de erro. Um ES é uma quantidade de qualquer coisa que seja de interesse, podendo ser uma média, diferença entre médias, percentual, mediana ou correlação, pode também ser um valor-padronizado como o *d* (Cohen) ou um coeficiente de regressão (Cumming, 2012). O IC, por sua vez, fornece estimativas sobre eventuais replicações do estudo calculando-se a faixa de valores na qual se espera encontrar outras médias caso a amostragem fosse repetida. Quando se calcula o

IC de uma amostra é importante considerar que esta é apenas uma amostra dentro de um grande número de possibilidades, assim ao definir um IC de 95%, esse valor se refere a que possivelmente 95% das amostras coletadas conteriam o parâmetro populacional enquanto 5% não conteriam esse parâmetro (Cumming, 2012). O autor descreve também as barras de erro dentre as possibilidades para representar os dados obtidos. Segundo Cumming (2012), as barras de erro representam medidas de variabilidade ou incerteza, ou seja, podem definir uma série de valores em torno de uma estimativa pontual como uma média, mediana, frequência, entre outros. As barras de erro podem também ser utilizadas para representar vários tipos de dados como: um IC, desvio padrão ou erro padrão.

O segundo tópico proposto pela APA (1996), enfatiza a necessidade de estudos da teoria, ou seja, de que o pesquisador busque realizar estudos mais exploratórios em que não haja a necessidade de formulações de hipóteses definidas previamente e faz uma crítica ao modelo hipotético-dedutivo adotado no NHST. O terceiro tópico se refere a orientação do uso de estratégias analíticas minimamente suficientes. A recomendação da força tarefa foi de que fosse aplicado o princípio da parcimônia na seleção de projetos e análises, sugerindo aos pesquisadores que evitem se deixar levar pela exigência de revisores de bolsas e artigos, editores de periódicos e orientadores de dissertação que obrigam os pesquisadores a selecionar cada vez estratégias analíticas e designes mais complexos que não necessariamente são necessários para análise dos dados e utilizem de design e análise minimamente suficientes. O quarto tópico descrito se refere aos problemas com análises computadorizadas de dados. Segundo a força tarefa, com o avanço da tecnologia também aumentaram as probabilidades de usos indevidos na análise computadorizada de dados, como apresentação de relatórios estatísticos sem compreender os cálculos e o que eles significam e assim, o relato de resultados indicados como contendo maior precisão do que

de fato são suportados pelos dados. A força tarefa deixou explícito o incentivo para evitar a santificação da análise informatizada de dados.

O presente artigo buscou evidenciar características relacionadas ao mau uso ou má interpretação interpretação das informações, aspectos intrínsecos relacionados ao uso de estatística inferencial e NHST e também a propor alternativas para evitar que os problemas aqui descritos continuem sendo difundidos. Como pesquisador ou profissional responsável em alguma medida por avaliar pesquisas em processo de publicação, o conhecimento sobre a adequação do uso de estatística inferencial, assim como de outros tipos de estatísticas que podem ser utilizados na análise de dados de pesquisas é de fundamental importância para manutenção da confiabilidade dos estudos que são publicados na área. As orientações da American Psychological Association - APA (1996), assim como as orientações publicadas nos demais manuais posteriores voltados ao uso de estatística por estimativa e de outros meios para garantir a confiança sobre a análise e interpretação de resultados das pesquisas são importantes para que essa adequação seja realizada com sucesso.

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 como sugestões também para esses atores na AC. 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 apresentadas 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 (2020), das prescrições da American Statistical Association (Wasserstein & Lazar, 2016), de comentadores da área de estatística (Hand, 2022; Wasserstein et al., 2019) e de editoriais a esse respeito publicados em diferentes periódicos científicos (Harrison et al., 2020; McCarren et al., 2017; Schreiber, 2020):

- Não embase suas conclusões apenas no achado de que uma associação ou efeito está relacionado a um valor-p 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 valor-p 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 valor-p.
- Desde as fases iniciais de planejamento, programe 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.
- 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 valor-p.
- Se optar por manter o uso de NHST, apresente o valor-p 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 valores-p exatos e com pelo menos 3 casas decimais.
- Não interprete o valor-p 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: 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 valor-p 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.
- É preciso relatar quantos testes estatísticos foram realizados, quais dados foram descartados, quais pré-requisitos foram testados ou garantidos e quais testes estatísticos foram considerados.

- É preciso justificar a escolha do nível de significância adotado. Usar $p<0,05$ simplesmente porque é tradicional na área não é uma boa prática científica. O valor-p 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 valor-p 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 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 e 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

Acuña, L. (2010). El uso de estadística en análisis de la conducta: ¿Cuándo usarla y cuándo no? *Revista Mexicana de Análisis de La Conducta*, *36*(1), 133–145. https://doi.org/10.5514/rmac.v36.i1.18020

American Psychological Association. (2020). Publication Manual of the American Psychological Association (Vol. 7). American Psychological Association. https://doi.org/10.2753/CES1097-1475050304177

American Psychological Association. (1996). *Initial Report Task Force on Statistical Inference*. 1–4.

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental Determinants? *Behavior Analyst*, *22*(2), 93–97. https://doi.org/10.1007/BF03391987

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.

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

Baron, A. (1990). Experimental designs. *The Behavior Analyst*, *13*(2), 167–171.

Baron, A., & Derenne, A. (2000). Quantitative summaries of single-subject studies: What do group comparisons tell us about individual performances? *The Behavior Analyst*, *1*(1), 101–106.

Betensky, R. A. (2019). The p -Value requires context, not a threshold. *The American Statistician*, *73*(sup1), 115–117. https://doi.org/10.1080/00031305.2018.1529624

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *Behavior Analyst*, *22*(2), 87–92. https://doi.org/10.1007/BF03391984

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

Branch, M. N. (2019). The “reproducibility crisis”: Might the methods used frequently in behavior-analysis research help? *Perspectives on Behavior Science*, *42*(1), 77–89. https://doi.org/10.1007/s40614-018-0158-5

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

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399.

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

Chase, L. J., & Tucker, R. A. Y. K. (1976). Statistical power: Derivation, development, and data-analytic implications. *The Psychological Record*, *26*, 473–486.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Cowles, M. (2001). Statistics in psychology: An Historical Perspective. In *Making Sense of Data and Statistics in Psychology* (2nd ed.). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1007/978-0-230-35799-0_1

Crosbie, J. (1999). Statistical inference in behavior analysis: Useful friend. *Behavior Analyst*, *22*(2), 105–108. https://doi.org/10.1007/BF03391987

Cumming, G. (2012). *Understanding The New Statistics: Effect sizes, confidence intervals, and meta-analysis*. Taylor & Francis Group.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138–146. https://doi.org/10.1111/j.1742-9536.2011.00037.x

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

Davison, M. (1999). Statistical inference in behavior analysis: Having my cake and eating it? *Behavior Analyst*, *22*(2), 99–103. https://doi.org/10.1007/BF03391986

Dixon, M. R., & Hayes, L. J. (1999). A behavioral analysis of dreaming. *The Psychological Record*, *49*(4), 613–627. https://doi.org/10.1007/BF03395331

Falk, R. (1998). In criticism of the Null Hypothesis Statistical Test. *American Psychologist*, *7*, 798–799.

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

Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science*, *7*(2), 208–225. https://projecteuclid.org/download/pdf_1/euclid.ss/1177013437

Finch, S., Cumming, G., Williams, J., Palmer, L. E. E., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 312–324.

Fisher, R. A. (1956). *Statistical Methods and Statistical Inference*. Oliver and Boyd.

Foster, T. M., Jarema, K., & Poling, A. (1999). Inferential statistics: Criticised by Sidman (1960), but popular in the Journal of the Experimental Analysis of Behavior. *Behaviour Change*, *16*(3), 203–204. https://doi.org/10.1375/bech.16.3.203

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Kruger, L. (1989). *The Empire of Chance How Probability Changed Science and Everyday Life* (1st ed.). Cambridge Universoty Press.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2012). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In *The SAGE Handbook of*

Quantitative Methodology for the Social Sciences (pp. 392–409). SAGE. https://doi.org/10.4135/9781412986311.n21

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *Plos One*, *12*(4), 1–16. https://doi.org/10.1371/journal.pone.0175583

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 Online 2002, 7*(1), 1–20.

Hand, D. J. (2022). Trustworthiness of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in 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

Hartmann, D. P. (1974). Forcing square pegs into round holes: Some comments on "an analysis-of-variance model for the intrasubject replication design." *Journal of Applied Behavior Analysis*, *7*(4), 1311680. https://doi.org/10.1901/jaba.1974.7-635

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *Behavior Analyst*, *21*(1), 125–137. https://doi.org/10.1007/BF03392787

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917-1994. *Theory & Psychology*, *7*(4), 545–554. https://doi.org/10.1177/0959354397074006

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

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

Iversen, I. H. (2012). Single-case research methods: An overview. In *APA handbook of behavior analysis, Vol. 1: Methods and principles.* (Vol. 1, pp. 3–32). https://doi.org/10.1037/13937-001

Johnston, J. M., Pennypacker, H. S., Green, G. (2020). *Strategies and Tactics of Behavioral Research and Practice* (4th ed.). Taylor & Francis Group.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Experimental Psychology, 4*(44), 152–158. https://doi.org/10.1027/1614-2241.4.4.152

Killeen, P. R. (2019). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science*, *42*(1), 109–132. https://doi.org/10.1007/s40614-018-0171-8

Kline, R. B. (2013). *Beyond Significance Testing Statistics Reform in the Behavioral Sciences* (2nd ed.). American Psychological Association.

Kratochwill, T. R., & Brody, G. H. (1978a). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification*, *2*(3), 291–307.

Kratochwill, T. R., & Brody, G. H. (1978b). Single subject designs. *Behavior Modification*, *2*(3), 291–307. https://doi.org/10.1177/014544557823001

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

Kyonka, E. G. E., Mitchell, S. H., & Bizo, L. A. (2019). Beyond inference by eye: Statistical and graphing practices in JEAB, 1992-2017. *Journal of the Experimental Analysis of Behavior*, *2*, 155–165. https://doi.org/10.1002/jeab.509

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

Loftus, G. R. (1993). Editorial comment. *Memory & Cognition*, *21*(1), 1–3. https://doi.org/10.3758/BF03211158

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

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.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, *7*(4), 647–653.

Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs. *Methodology*, *15*(4), 147–156. https://doi.org/10.1027/1614-2241/a000180

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

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

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*(2), 109–116. https://doi.org/10.1007/BF03391988

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making type I errors. *Psychological Bulletin*, *102*(1), 159–163.

Reese, H. W. (1998). Utility of group methodology in behavior analysis and developmental psychology. *Revista Mexicana de Análisis de La Conducta*, *24*(2), 137–151.

Reese, H. W. (1999). Problems of statistical inference. *Revista Mexicana de Análisis de La Conducta*, *25*(1), 39–68.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428.

Sampaio, A. A. S., De Azevedo, F. H. B., Cardoso, L. R. D., De Lima, C., Pereira, M. B. R., & Andery, M. A. P. A. (2008). Uma introdução aos delineamentos experimentais de sujeito único. *Interação em Psicologia*, *12*(1), 151–164. https://doi.org/10.5380/psi.v12i1.9537

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

Shull, R. L. (1999). Statistical inference in behavior analysis: Discussant's remarks. *The Behavior Analyst*, *22*(2), 117–121. https://doi.org/10.1007/BF03391989

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Basic Books.

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

Skinner, B. F. (1953). *Science and human behavior.* MacMillan.

Skinner, B. F. (2013). *Contingencies of reinforcement: A theoretical analysis*. B. F. Skinner Foundation.

Sohn, D. (1998). Statistical significance and replicability. *Theory & Psychology*, *8*(3), 291–311.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, *53*(7), 1997–1998. https://doi.org/10.1037/0003-066X.53.7.799

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

Velasco, S. M., Garcia-mijares, M., & Tomanari, G. Y. (2010). Fundamentos metodológicos da pesquisa em análise experimental do comportamento. *Psicologia em Pesquisa*, *4*(02), 150–155.

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*(sup. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Zimmermann, Z. J., Watkins, E. E., & Poling, A. (2015). JEAB research over time: Species used, experimental designs, statistical analyses, and sex of subjects. *Behavior Analyst*, *38*(2), 203–218. https://doi.org/10.1007/s40614-015-0034-5

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.