

Uso de Estatística em Periódicos Brasileiros de Análise do Comportamento

(The use of Statistics in Brazilian Journals of Behavior Analysis)

Bruna Rodrigues Lins¹ e Bruno Angelo Strapasson

Universidade Federal do Paraná

(Brasil)

Resumo

O uso de testes estatísticos de hipótese nula (NHST) tem sido criticado na ciência em geral e na Análise do Comportamento (AC), mas tem crescido em jornais internacionais da área. Entretanto, avaliações da literatura brasileira não estavam disponíveis até agora. Neste estudo analisamos os relatos de estatísticas inferenciais em todos os artigos empíricos publicados na Revista Brasileira de Análise do Comportamento, na Revista Brasileira de Terapia Comportamental e Cognitiva e na Perspectivas em Análise do Comportamento publicados até abril de 2021. Cerca de um quarto (25,2%) dos artigos publicados nos jornais avaliados utilizaram o NHST para fazer inferências estatísticas com tendência crescente ao longo dos anos. Desses, apenas 3,2% incluíram estatísticas inferenciais que utilizam medidas por estimativa. Considerando apenas os estudos que utilizaram estatísticas inferenciais, 69,5% apresentaram seus resultados de modo incorreto ou incompleto. Os periódicos brasileiros de AC apresentam tendências similares aos periódicos internacionais no uso de NHST e o relato incompleto das estatísticas associadas é pernicioso para a área, indicando necessidade de maior atenção ao relato de estatísticas inferenciais nas políticas editoriais dos periódicos da área.

Palavras-chave: estatística inferencial, teste de significância de hipótese nula (NHST), Análise do Comportamento, análise estatística

¹ Endereço para correspondência: Bruna Rodrigues Lins (UFPR). Endereço: Rua Rafael Andrade Duarte, nº580, apto 11 - CEP: 13100011. Campinas, São Paulo. E-mail: rodrigueslinsb@gmail.com

Abstract

The use of Null Hypothesis Significance Tests (NHST) has been a topic of debate in science, psychology, and behavior analysis due to difficulties in interpreting them correctly, their tendency towards dichotomous thinking, and the frequent associated questionable practices. The behavior analysis literature has a tradition of criticizing the use of NHST for studying individual behavior, given its preference for single-case research designs (SCRD). Despite this, there has been an increase in the use of NHST in international behavior analysis journals over the last six decades. However, there has been a lack of evaluations of Brazilian behavior analysis literature. In this study, we examined the reports of inferential statistics in all empirical articles published in the *Revista Brasileira de Análise do Comportamento* (Brazilian Journal of Behavior Analysis), the *Revista Brasileira de Terapia Comportamental e Cognitiva* (Brazilian Journal of Behavioral and Cognitive Therapy), and *Perspectivas em Análise do Comportamento* (Perspectives in Behavior Analysis) until April 2021. Our findings reveal that almost half of the papers published in the mentioned journals were empirical. Among these, 76.6% were experimental studies, and approximately a quarter (25.2%) used NHST for statistical inferences, with a noticeable upward trend in its use. However, only 3.2% included inferential statistics with estimation metrics such as confidence intervals and effect sizes. Importantly, we found that 69.5% of the studies using inferential statistics presented their results incorrectly or incompletely according to the APA standards in force at the time of their publication, hindering the verification and reevaluation of the statistical calculations presented. Our study reveals that Brazilian behavior analysis journals show similar trends to international journals in the use of NHST, and the incomplete reporting of associated statistics is damaging to the field. This emphasizes the pressing need for improvement in the editorial policies of behavior analysis journals regarding the reporting of inferential statistics.

Keywords: inferential statistics, null hypothesis significance test (NHST), Behavior Analysis, statistical analysis

A tomada de decisões adequadas na formulação de um método de pesquisa e na definição da estratégia de análise de dados a ser empregada é uma etapa essencial de qualquer planejamento de pesquisa. Na psicologia, é muito comum o recurso a uma análise estatística que mistura duas abordagens de testes de hipótese desenvolvidas por Fisher e Newman e Pearson, respectivamente (Perezgonzalez, 2015). Tal abordagem ficou conhecida como teste de significância da hipótese nula (Null Hypothesis Significance Testing - NHST) e, ainda que careça de uma definição formal, se refere a aplicação de um modelo estatístico para calcular a probabilidade de que um resultado encontrado em um estudo específico (ou um resultado mais extremo que esse) seja obtido tomando-se como pressuposto a verdade de uma hipótese que se pretende rejeitar (conhecida como hipótese nula - em geral expressa como a ausência de diferenças entre dois grupos ou uma correlação de valor zero entre duas variáveis - Perezgonzalez, 2015). Tal probabilidade (o p -valor) é julgada em relação a um critério pré-definido de aceitabilidade. Caso o p -valor seja menor

que esse critério, rejeita-se a hipótese nula e considera-se o resultado obtido como estatisticamente significativo (Perezgonzalez, 2015). Ainda que amplamente disseminado, o uso do NHST tende a produzir um raciocínio dicotômico (ou um resultado é ou não é estatisticamente significativo) e tem sido amplamente criticado desde seu desenvolvimento e disseminação (Kline, 2013).

No âmbito da Análise do Comportamento (AC), críticas ao uso de NHST, comum em delineamentos de pesquisa que comparam grupos de sujeitos, estão presentes também nos primeiros e nos principais manuais de metodologia de pesquisa na área (e.g., Iversen, 2013; Johnston et al., 2020; Sidman, 1960). Paradoxalmente, o uso do NHST tem aumentado na literatura internacional da AC (e.g., Acuña, 2010; Foster et al., 1999; Kyonka et al., 2019), mas até então não estava claro se esse padrão se repetia na AC brasileira. Neste artigo descrevemos os padrões de uso do NHST na literatura nacional de AC.

Problemas Relacionados ao uso do NHST

Desde seu surgimento na década de 1930, o uso do NHST tem sido controverso. Um de seus principais proponentes, R. A. Fisher (1890-1962), concordava que essa estratégia de análise de dados não era a ideal e que constituía apenas uma aproximação estatística que poderia ser útil se outras formas mais adequadas não estivessem disponíveis (Fisher, 1956). No entanto seu uso se espalhou tanto na ciência a ponto de ser tomado como sinônimo de análise científica (Cowles, 2001).

A Psicologia parece ter sido uma das primeiras disciplinas na qual o uso do NHST foi popularizado (Gigerenzer et al., 2004). A ampla expansão da experimentação com uso de grupos na Psicologia coincidiu aproximadamente com a introdução das estatísticas fisherianas nessa disciplina, e os dois rapidamente se tornaram inseparáveis (Cowles, 2001). Desde então, o uso do NHST parece estar se ampliando nos diferentes âmbitos da psicologia científica. Em uma análise do *Journal of Applied Psychology*, por exemplo, o NHST subiu de 17% nos seus anos iniciais (entre 1917 e 1929) para 75% nos anos 1950 e para 94% durante os anos 1990 (Hubbard et al., 1997). A mesma tendência foi identificada em 12 jornais proeminentes da APA entre 1911 e 1998 (Hubbard & Ryan, 2000) e expandida para novos 10 jornais de psicologia entre 1998 e 2006, período no qual mais de 95% dos artigos empíricos em língua inglesa incluíam alguma forma de NHST (Cumming et al., 2007).

Críticas ao uso do NHST na análise de dados em psicologia são recorrentes na literatura. Um exemplo é a constatação de que a significância estatística é produto de múltiplas variáveis (por exemplo, do poder do teste adotado, do tamanho da amostra etc.) e que, portanto, a depender dos arranjos metodológicos utilizados, ela pode ser alcançada mesmo que um efeito relevante não exista na população (e.g., Bakan, 1966; Branch, 2014; Carver, 1978; Cohen, 1994; Lambdin, 2012; Loftus & Loftus, 1996; Meehl, 1967, 1978; Simmons et al., 2011). Outra crítica pondera que o *p*-valor é uma probabilidade condicional que toma a verdade da hipótese nula como certa (como condição). Desse modo, a lógica do NHST assume o oposto ao que se quer testar e, portanto, é no mínimo de difícil interpretação na análise de resultados

de pesquisa (e.g., Cohen, 1994; Falk, 1998; Loftus & Loftus, 1996; Meehl, 1978; Thompson, 1999). Ademais, a tomada de decisão sobre a “significância estatística” baseada exclusivamente no p -valor produz inevitavelmente um raciocínio dicotômico (de rejeitar ou não a hipótese nula), o que limita o raciocínio científico (e.g., Cumming, 2012).

Acrescente-se a esses problemas as práticas questionáveis na aplicação do NHST que incluem a não observância dos pré-requisitos necessários para o uso de testes específicos, o uso sucessivo e não justificado de diversos testes estatísticos até que um resultado positivo seja encontrado (p -hacking - Head et al., 2015), a formulação de hipóteses após os dados terem sido analisados (HARKing - Kerr, 1998), o relato apenas de resultados estatisticamente significativos (cherry-picking), a busca extensiva de relações entre variáveis não relacionadas às hipóteses inicial do estudo e, principalmente, a diversas falácias na interpretação do significado do p -valor (para uma compilação dessas falácias ver Kline, 2013). Por fim, a valorização do raciocínio dicotômico típico do NHST tem levado também a vieses no processo editorial que favorecem a publicação de resultados estatisticamente significativos em detrimento dos não significativos (Krawczyk, 2015; Leggett et al., 2013).

Ainda que alguns dos problemas mencionados possam ser resultado de má fé (ver Simmons et al., 2011), a maior parte dos problemas parece estar relacionada ao fato de que psicólogos não compreendem corretamente o NHST (e.g., Badenes-Ribera et al., 2015, 2016), são mal ensinados a respeito dele por seus professores e pelos livros didáticos disponíveis (e.g., Cassidy et al., 2019; Friedrich et al., 2018; Huberty, 1993) e mantêm foco excessivo nos resultados do NHST para as conclusões de suas pesquisas.

Na tentativa de mitigar os problemas do uso incorreto do NHST a American Psychological Association (American Psychological Association, 1996, 2019) e a American Statistical Association (Wasserstein & Lazar, 2016) publicaram orientações para o uso apropriado desse recurso metodológico. As orientações descritas nesses documentos assumem ao menos duas frentes diferentes: a de orientação para melhores práticas ao usar o NHST, apresentando as informações corretamente e de forma completa, de modo a permitir que os pesquisadores possam refazer os cálculos e analisar os dados, mas também o incentivo a utilização de outras práticas estatísticas complementares ou alternativas ao NHST. Tais práticas incluem o uso de estatísticas inferenciais que contemplam estimativas (não dicotômicas), o uso de estatística bayesiana e a ampliação do uso de estatísticas descritivas (ver também Young, 2018).

A despeito disso, o NHST tem sido amplamente utilizado na literatura internacional de Psicologia (Hubbard et al., 1997; Hubbard & Ryan, 2000) e a tendência parece ser de ampliação desse uso (Cumming et al., 2007; ver também Imam & Frate, 2019). Adicione-se a isso o fato de que as alternativas disponíveis no campo da estatística inferencial, como, por exemplo, a valorização de medidas que fazem uso de estimativa (e.g., intervalos de confiança – IC, tamanhos de efeito – ES – etc.) e o uso de outras estatísticas inferenciais alternativas (ver número suplementar do *The American Statistician*, vol. 73, 2019), são ainda pouco utilizadas, criando um cenário preocupante (Cumming et al., 2007). Uma dimensão adicional

do problema é que as estatísticas utilizadas na psicologia, além de questionáveis e eventualmente mal aplicadas, frequentemente são descritas de forma insuficiente nos estudos, impedindo a revisão por terceiros (Bakker & Wicherts, 2011, fig. 1; ver também Berle & Starcevic, 2007).

Críticas ao uso do NHST na AC

Os problemas identificados no uso do NHST na Psicologia têm sido amplamente discutidos também pelos analistas do comportamento (e.g., Baron, 1990; Branch & Pennypacker, 2013; Chase & Tucker, 1976; Imam, 2021; Iversen, 2013; Johnston et al., 2020; Killeen, 2019; Michael, 1974. Para um contraponto, ver Young, 2018) que acrescentam ao debate certa suspeição sobre o delineamento de grupos como forma de responder questões importantes sobre as relações funcionais relevantes, dado que compreensão de processos comportamentais são fenômenos individuais (e.g., Branch & Pennypacker, 2013; Imam, 2021). Além disso, há indícios de que quando o NHST é usado na AC, isso frequentemente ocorre por razões incorretas, como a pressão social de colegas e editores que eventualmente valorizam essa prática mais por tradição do que por alguma razão científica (Ator, 1999).

Por outro lado, diferentes manifestações a favor do uso de delineamentos de grupo – e da estatística inferencial comumente associada a esses delineamentos – podem ser encontradas na literatura da área. Dentre os argumentos a favor do uso do NHST na AC estão o apelo para que métodos formais sejam empregados no julgamento da variabilidade (Davison, 1999; Reese, 1998, 1999), a sugestão de que algumas perguntas de pesquisa relevantes para os analistas do comportamento só podem ser respondidas por meio de delineamentos de grupo (Reese, 1998), a indicação de que a maioria dos problemas do NHST vêm do seu mau uso e não de suas características intrínsecas (Crosbie, 1999; Reese, 1998, 1999; Shull, 1999) e a constatação de que mesmo que os analistas do comportamento não realizem experimentos de grupo, seria imprudente ignorar os resultados produzidos com essa metodologia e, para aproveitar esses dados, seria importante compreender bem o uso do NHST (Acuña, 2010; Crosbie, 1999; Johnston et al., 2020).

Independente de quão adequados são esses argumentos, parece relevante notar que o uso de delineamentos de grupo e do NHST tem aumentado também na AC. Já na década de 1970, a proporção de artigos em jornais aplicados com orientação comportamental (Behavior Therapy [BT], Behavior Research and Therapy [BRT], Journal of Applied Behavior Analysis [JABA] e Journal of Behavior Therapy e Experimental Psychiatry [JBTEP]) vinha se mostrando alta (Kratochwill & Brody, 1978), representando entre 89% e 97% dos estudos empíricos. Dados mais específicos obtidos nos jornais de AC demonstram que o uso de delineamentos de grupo e de NHST também vem aumentando ao longo dos anos. No Journal of the Experimental Analysis of Behavior (JEAB), por exemplo, a porcentagem de estudos que incluíam estatísticas inferenciais parece ter aumentado consistentemente, passando de 5% nos anos 1960 para 50% na década de 1990 (Foster et al., 1999 ver tb. Baron, 1990) o que se manteve estável pelo menos até 2008 (Acuña, 2010). Replicações parciais dessas avaliações entre os anos de 2000 e 2013 (Zimmermann

et al., 2015) e entre 1997 e 2017 (Kyonka et al., 2019) sugerem que o aumento progressivo do uso de estatísticas inferenciais, em especial do NHST, se manteve vigente no JEAB (ver tb. Imam & Frate, 2019). Padrão similar foi encontrado também na Revista Mexicana de Análisis de la Conducta (Acuña, 2010). No Journal of Applied Behavior Analysis (JABA), entretanto, essa tendência não foi encontrada. Apesar de os delineamentos de grupo e mistos aumentarem o recurso a estatísticas inferenciais na década passada (especialmente entre 2002 e 2018) tais delineamentos constituem a minoria dos estudos publicados nesse jornal e não houve aumento do uso desse recurso em delineamentos de caso único (Rader et al., 2021).

AAC, historicamente, tem privilegiado métodos alternativos de análise de dados (Sidman, 1960), mas no cenário internacional tem cada vez mais seguido tendências já presentes na psicologia como um todo (e.g., Acuña, 2010; Foster et al., 1999; Kyonka et al., 2019). Apesar do Brasil ser um dos maiores centros da AC no mundo, até o momento, nenhum estudo parece ter avaliado como a estatística inferencial tem sido utilizada nos periódicos nacionais. O presente estudo tem como objetivo analisar relatos de estatísticas inferenciais nos periódicos brasileiros de AC.

Método

Foram identificados e avaliados por meio de leitura de resumo, método e descrição dos resultados todos os artigos que relatavam pesquisas empíricas publicados nas três revistas brasileiras de AC vigentes na época da pesquisa (Revista Brasileira de Análise do Comportamento [REBAC], na Revista Brasileira de Terapia Comportamental e Cognitiva [RBTC] e na Perspectivas em Análise do Comportamento [Perspectivas]), desde sua criação até abril de 2021. As pesquisas foram divididas de acordo com o delineamento utilizado sendo classificadas como experimentais (1a) de caso único, (1b) de grupos, (1c) misto; ou (2) estudo de caso, (3) caracterização ou (4) correlação. Essas categorias foram operacionalmente definidas do seguinte modo: (1a) delineamento no qual se tem medidas repetidas do mesmo sujeito, participante ou caso e a comparação das diferentes condições da variável independente são feitas sempre do sujeito, participante ou caso com ele mesmo em uma perspectiva individual. Em situações nas quais duplas/trios/grupos são tratados como se fossem um mesmo sujeito, por exemplo, casos em que a interação de uma dupla produziu um dado único de interação (quantidade de distribuições igualitárias de pontos vs. distribuições desiguais) esse estudos foram também considerados como estudos de caso único, dado que não foram utilizadas medidas agregadas, mas sim medidas únicas de desempenho; (1b) delineamento que centra sua análise de dados na comparação de dois ou mais conjuntos de sujeitos ou participantes. Usualmente, são estudos que comparam o desempenho de diferentes grupos submetidos a condições experimentais diferentes e utilizam medidas agregadas de desempenho; (1c) delineamento que inclui tanto análises de medidas repetidas em uma perspectiva individual quanto análises com medidas agregadas. Delineamentos que utilizam apenas um grupo submetido a condições diferentes e faz isso utilizando de medidas agregadas de desempenho, comparando

o grupo com ele mesmo, também foram considerados estudos com delineamento misto; (2) um estudo de caso fornece uma descrição de uma pessoa, instituição ou ambiente. O estudo de caso envolve a análise de um problema, relação social ou estrutura específica, podendo, por exemplo, focar nas características de um problema de saúde, nos arranjos únicos encontrados na vida de um paciente ou no desenvolvimento histórico de uma instituição específica, mas sem o registro de medidas repetidas das variáveis de interesse; (3) estudos que visam caracterizar as variáveis que compõem um determinado contexto, ambiente, prática ou pessoa (e.g., reforçadores mantenedores do comportamento de prostituir-se), ou que buscam identificar a distribuição de uma variável (e.g., prevalência de depressão) em um determinado contexto ou população; (4) estudos correlacionais, que visam identificar a co-ocorrência de duas ou mais variáveis, mas não manipulam diretamente nenhuma dessas variáveis. A co-ocorrência é normalmente descrita numericamente em sua versão normalizada (correlação), nesses casos, usualmente, a ausência de manipulação sistemática de uma variável independente impede asserções sobre relações de dependência entre as variáveis investigadas.

Em cada estudo empírico foi identificado se havia o uso de estatísticas descritivas, estatísticas inferenciais dicotômicas e/ou estatísticas inferenciais que faziam uso de estimativas em uma categorização não-excludente. Foram considerados como estatísticas descritivas: uso de gráficos; tabelas; medidas de tendência central como média, mediana e moda; medidas de posição como percentil; e medidas de dispersão, como amplitude total, amplitude interquartilica, desvio padrão e variância (com ou sem o uso de barras de erro). As estatísticas inferenciais foram separadas em duas: sendo o NHST representante do uso de estatísticas inferenciais que apresentam informações dicotômicas, e medidas inferenciais que fazem análise por estimativa, representadas por intervalo de confiança (IC) e tamanho de efeito (Effect Size – ES). Foi avaliado o percentual do uso de cada estatística no geral, por periódico e por ano em cada periódico. Foram avaliados e categorizados os tipos de procedimentos/testes estatísticos usados em experimentos de caso único, misto e de grupos, como por exemplo, teste do Qui-quadrado, teste t ou ANOVA. Artigos cujo o teste estatístico não foi especificado, por exemplo, que relataram apenas um *p*-valor ou que incluíam a afirmação que os resultados foram “estatisticamente significativos” sem especificar o teste, foram classificados como incluindo “teste não especificado”. O relato completo das estatísticas apropriadas segundo as normas da APA também foi avaliado.

Para todas as categorias estipuladas um segundo avaliador analisou independentemente 56,2% dos artigos, selecionados aleatoriamente, destes artigos a concordância geral entre os categorizadores na avaliação foi de 94,6%. Em avaliação individual das categorias analisadas, apenas o tipo de delineamento teve concordância abaixo de 90%, com 87,8% de acordo entre observadores. Discordâncias foram avaliadas por um terceiro categorizador independente.

Resultados

Na Tabela 1 estão dispostas a quantidade de artigos e o percentual de estudos empíricos em cada revista no período avaliado. Embora a RBTCC seja a revista mais antiga de AC no Brasil e tenha também a maior quantidade de estudos empíricos publicados, o percentual de estudos empíricos publicados nessa revista é inferior ao da REBAC. Na revista Perspectivas a proporção de artigos empíricos publicados é a menor dentre as revistas analisadas, constituindo 33,8% dos estudos realizados. De modo geral, 47,4% dos estudos publicados em periódicos brasileiros de AC são empíricos e fazem uso de algum tipo de representação estatística de dados, sejam eles experimentais ou não experimentais.

Tabela 1

Quantidade e Percentual de Estudos Empíricos por Revista

Revista	Período	Nº de artigos	Nº estudos empíricos	% de estudos empíricos
RBTCC	1999 - 2021	489	230	47,0
REBAC	2005 - 2021	305	173	56,7
Perspectivas	2010 - 2021	195	66	33,8
Total	-	989	469	47,4

Conforme pode ser deduzido da Tabela 2, a maior parte dos estudos empíricos analisados é constituída de estudos experimentais (75,4% a soma do total de estudos de grupo, caso único e mistos). Na REBAC o percentual de estudos experimentais foi de 96,2% dentre os estudos empíricos avaliados e o menor percentual de estudos experimentais foi publicado na revista Perspectivas, representando 60,5% dos estudos empíricos. Dentre os estudos experimentais, o estudo de caso único foi o mais utilizado, representando 42,3% do total de estudos empíricos. Comparações entre sujeitos (delineamentos de grupo 13,4% e mistos 19,8% somados) estavam presentes em 33,2% dos estudos empíricos.

Tabela 2

Percentual de Estudos de Caracterização, Estudos de Caso, de Correlação e Experimentais por Revista

Revistas	Caracterização	Estudo de Caso	Correlação	Estudos experimentais			
				Grupos	Caso único	Misto	Total
Perspectivas	22,7%	7,6%	7,6%	6,0%	40,9%	13,6%	60,5%
RBTC	14,8%	13,9%	7,8%	16,5%	34,3%	12,6%	63,4%
REBAC	2,3%	0,0%	0,6%	12,0%	53,2%	31,1%	96,2%
Total Geral	11,3%	7,9%	5,3%	13,4%	42,3%	19,8%	75,4%

Nota. As categorias apresentadas na tabela não são excludentes.

Na Tabela 3 constam a quantidade e percentuais de estatísticas encontradas. O uso de NHST esteve presente em 25,2% dos artigos empíricos, sendo a menor porcentagem na Perspectivas (21,2%) e a maior na RBTC (26,5%). Análises inferenciais que incluíam estimativas foram usadas em apenas 3,2% dos estudos empíricos, sendo que todos os estudos que incluíam esses também utilizaram NHST.

Tabela 3

Quantidade e Percentuais de Estudos que Utilizaram Estatísticas Descritivas e Inferenciais por Revista

Revistas	Estatísticas Descritivas		Estatísticas Inferenciais			
			NHST		Medidas de estimativa	
	Nº	%	Nº	%	Nº	%
Perspectivas	64	97,0	14	21,2	1	1,5
RBTC	213	92,6	61	26,5	10	4,3
REBAC	173	100,0	43	24,9	4	2,3
Total Geral	450	95,9	118	25,2	15	3,2

Nota. As categorias apresentadas na tabela não são excludentes.

Quando analisado o uso de estatísticas em função do tipo de delineamento utilizado, tal como apresentado na Tabela 4, percebe-se que estatísticas inferenciais

são mais comuns em estudos de correlação (88%) e estudos experimentais que incluem comparações entre grupos (com delineamento de grupo – 73% - ou delineamento misto – 37,6%).

Tabela 4

Percentuais de Estudos que Utilizaram Estatísticas por Delineamento

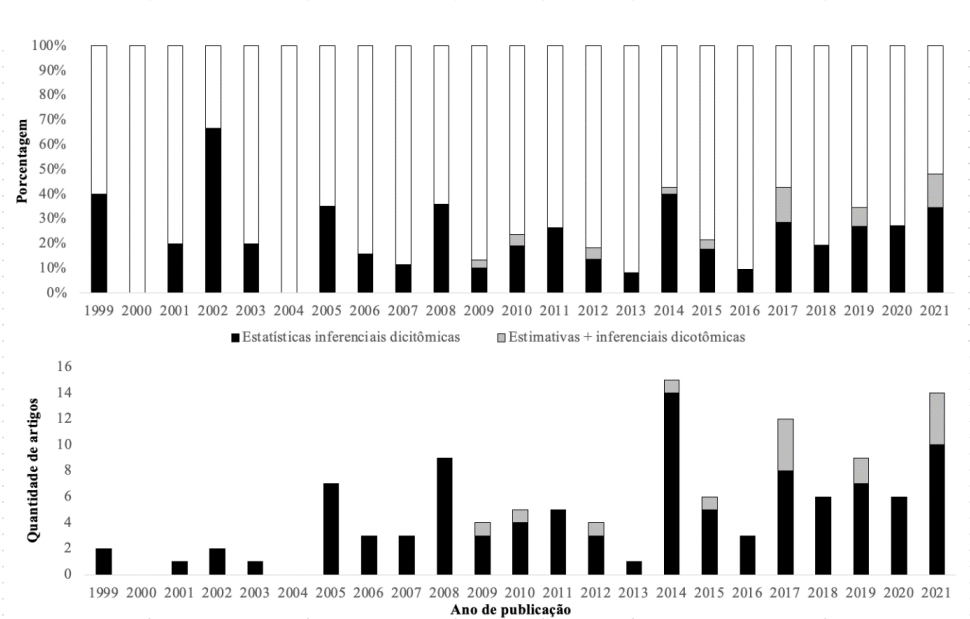
Delineamento	Uso de Estatística	Descritiva	Estatísticas Inferenciais	
			NHST	Medidas de Estimativa
Caracterização	86,8%	86,8%	5,7%	0,0%
Estudo de Caso	67,6%	67,6%	0,0%	0,0%
Correlação	100,0%	100,0%	88,0%	8,0%
Experimental	Grupos	100,0%	73,0%	16,0%
	Caso Único	100,0%	6,0%	0,5%
	Misto	100,0%	37,6%	2,2%
Total	95,9%	95,9%	25,2%	3,2%

Nota. As categorias apresentadas na tabela não são excludentes.

Embora os analistas do comportamento tenham dado preferência ao uso de casos individuais na realização de estudos experimentais e, portanto, privilegiem o uso de estatísticas descritivas, ao observar a Figura 1 é possível notar um leve crescimento na utilização de estatísticas inferenciais ao longo do tempo nas revistas avaliadas, em especial nos últimos anos. Entre 2014 e 2021 foram publicados quase o dobro de artigos com estatísticas inferenciais (59, o que corresponde a porcentagem média de 33,59%) do que nos oito anos anteriores (31, o que corresponde a uma porcentagem média de 20,38%). Para os anos iniciais, os dados de porcentagem podem distorcer a análise dada a pequena quantidade de artigos publicados nesses anos. O pico percentual de 67% de artigos empíricos com estatística inferencial em 2002, por exemplo, parece ser apenas um artefato da representação gráfica, pois apenas três artigos empíricos foram publicados nesse ano.

Estatísticas inferenciais que incluem estimativas são ainda pouco frequentes e bastante recentes. Foi apenas a partir de 2009 que encontramos esse tipo de estatística e a moda ocorre em 2017 e em 2021, com apenas 4 artigos com esse tipo de métrica.

Figura 1
Quantidade e Percentual de Estudos Empíricos com NHST e que Fazem uso de Medidas por Estimativa ao Longo dos Anos



Avaliando o uso de estatística inferencial por delineamento utilizado identifica-se maior percentual do uso de estatística inferencial em pesquisas de grupo na RBTCC e na REBAC enquanto esse uso é mais comum em pesquisas de Correlação na revista Perspectivas conforme demonstrado na Tabela 5.

Tabela 5
Percentual do uso de Estatísticas Inferências em Estudos de Caracterização, de Caso, de Correlação, de Grupos, de Caso Único e Misto por Delineamento em cada Revista

Revistas	Caracterização	Estudos		Estudos Experimentais		
		de caso	Correlação	Grupos	Suj. Único	Misto
Perspectivas	0,0%	0,0%	83,3%	75,0%	3,8%	71,4%
RBTCC	8,8%	0,0%	94,1%	68,4%	4,3%	48,1%
REBAC	0,0%	0,0%	100,0%	89,4%	10,1%	39,5%
Total Geral	5,7%	0,0%	91,7%	75,4%	6,8%	45,4%

Nota. As categorias apresentadas na tabela não são excludentes.

Os testes de NHST mais utilizados na literatura avaliada foram: Teste *t* (41 estudos), Anova (34 estudos), Mann-Whitney (24 estudos), Qui-quadrado (16 estudos) e Wilcoxon (16 estudos). Tamanhos de efeito padronizados (η^2 [5 estudos], d' Cohen [4 estudos], R Rosenthal [2 estudos], e g' Hedges [1 estudo]) ou não padronizados (e.g., diferenças entre médias [2 estudos]) foram pouco frequentes. Intervalos de confiança foram apresentados em apenas 3 estudos. Nove dos estudos de caso único incluíram estatísticas inferenciais na análise de dados, os testes utilizados nesses estudos incluíram o Teste *t*, Wilcoxon, Correlação de Spearman, ANOVA, Tukey, Correlação de Pearson e Coeficiente de Kendall. Nenhum estudo analisado incluiu estatística bayesiana.

Os 118 estudos que utilizaram estatísticas inferenciais foram avaliados quanto a seguirem apropriadamente as recomendações do manual da APA vigente na data de publicação de cada artigo para relato de estatísticas inferenciais. Essa análise só não foi realizada para os dois estudos empíricos mais antigos analisados (Falcone, 1999; Amaral & Magna, 1999) porque eles foram publicados quando estava vigente a quarta edição do manual da APA e não tivemos acesso a essa edição do manual. Entre 2001 e 2009, quando estava vigente a quinta edição do manual da APA, a principal recomendação quanto a apresentação de informações estatísticas inferenciais era de que houvesse informações suficientes para que o leitor entendesse as análises que foram conduzidas e possíveis explicações alternativas para os resultados dessas análises. Tais como resultados do cálculo completo do teste (e.g., $t(60) = 1,99$, $p = 0,03$) mais a informação de que o teste era bilateral/bicaudal ou unilateral/unicaudal. Dentro desse período, 30 estudos que incluíram estatísticas inferenciais foram publicados. Desses, 24 não apresentaram algum dos critérios das recomendações, seja por não apresentar o cálculo completo, não apresentarem o *p*-valor exato ou porque indicaram apenas se houve ou não significância sem essa informação vir acompanhada de um *p*-valor. Apenas um estudo (Fernandes & Santos, 2009) relatou medida de ES. Outros 72 estudos, que utilizaram estatísticas inferenciais dicotômicas, foram publicados enquanto estava vigente a sexta edição do manual da APA. Nele, além da recomendação da inclusão de informações estatísticas suficientes para a compreensão das análises que foram conduzidas, também consta orientação para a inclusão de ES e IC associados. Dos estudos publicados nesse período, 20 apresentaram os dados resultantes do NHST de forma completa, 2 estudos apresentaram dados de IC e 9 estudos apresentaram algum tipo de ES. No entanto, nenhum estudo seguiu fielmente as normas da APA no que diz respeito à apresentação conjunta dos dados do NHST completas, ES e IC. Os dois estudos que apresentaram IC nesse período o relataram em tabelas, indicando o intervalo e o valor de confiança definido e os interpretaram na descrição dos dados, seguindo as orientações do manual. No ano de 2021 (até Abril), 14 estudos que continham estatísticas inferenciais foram publicados, na vigência da sétima edição do manual da APA, este manual contém orientações semelhantes quanto a apresentação das informações estatísticas da sexta edição. Desses estudos, 8 continham descrições adequadas de NHST, mas apenas 3 deles apresentaram ES e nenhum deles incluiu IC.

Dos 82 estudos que não apresentaram informações do NHST completas (69,5% do total de estudos que utilizou estatística inferencial), 57 não apresentavam todas as informações recomendadas, 46 não apresentavam *p*-valores exatos e 14 estudos apenas indicavam a presença ou não de significância estatística sem apresentar *p*-valor. Vinte e dois dos 82 estudos cometeram mais de um erro.

Discussão

Os delineamentos utilizados em estudos empíricos nos periódicos de AC no Brasil são variados. Embora a AC tenha historicamente privilegiado o estudo de caso único como delineamento de preferência, a utilização de delineamentos com grupos é ampla e o uso de estatísticas inferenciais tem aumentado em frequência (ver Figura 1).

Este movimento (tendência levemente crescente em anos recentes), presente nas revistas nacionais, tem paralelo no que vem acontecendo em revistas de AC fora do Brasil em periódicos como o JEAB, JABA e RMAC (Acuña, 2010; Imam & Frate, 2019; Kratochwill & Brody, 1978; Kyonka et al., 2019; Zimmermann et al., 2015), ainda que ocorra em menor escala.

O crescimento no uso de estatísticas inferenciais nos periódicos de AC no Brasil acompanha outra tendência identificada em periódicos de Psicologia em geral, principalmente fora do Brasil: o uso frequente de NHST, não acompanhado de análises estatísticas inferenciais que utilizam medidas estimativas como ES e IC. Nos periódicos de AC no Brasil estatísticas inferenciais que utilizam este tipo de medida só começaram a ser incluídas em 2009, neste ano esse tipo de estatística esteve presente em apenas um (3,23%) dos estudos empíricos publicados. O uso de estatísticas que fazem análises que contemplam estimativas permaneceu infrequente ao longo dos anos com seu ápice pontual dentre os estudos empíricos em 2017 (12,50%) e 2021 (12,12%). Estes valores são considerados mais expressivos se comparados ao início das publicações, mas são marcadamente inferiores ao uso de estatísticas inferenciais dicotômicas. Em periódicos da AC internacionais padrão similar foi identificado. Imam e Frate (2019) encontraram quantidades muito pequenas de artigos com relato de IC no JEAB - apenas em um estudo em 2011 e três estudos em 2015, com apenas o estudo de 2011 tendo seu resultado de IC interpretado. No BP, nesse mesmo período, cinco estudos relataram IC em 2011 e 25 em 2015, sendo um deles interpretado em 2011 e três em 2015 (25% e 17%, dos descritos respectivamente). Kyonka et al. (2019) encontrou um aumento tímido (de apenas 2,5% a cada 5 anos) do uso de ICs no JEAB entre 1992 até 2017. Os números de IC relatados nas revistas internacionais, embora ainda pequenos, são maiores do que os identificados nos periódicos de AC no Brasil nos quais, contabilizando todo o período de publicação das revistas, apenas dois (1,69%) dos estudos dentre os que apresentaram informações estatísticas inferenciais incluíram IC em suas análises. Já sobre ES nos periódicos de AC, não foram identificadas análises completas do quanto esse tipo de estatística vem sendo utilizada, apenas dois periódicos avaliados nos anos de 2011 e 2015 – BP e JEAB – foram analisados enquanto tipo de ES que utilizavam (padronizado ou não) sem calcular a proporção da utilização desse tipo de estatística em comparação com o NHST nas revistas

(Imam & Frate, 2019). No Brasil, 12 (11%) dos estudos que utilizaram estatísticas inferenciais fizeram uso de algum tipo de ES (padronizado ou não).

Ao olhar mais amplamente para a Psicologia Internacional, os números em torno do NHST podem ser tão preocupantes quanto aos identificados na AC. Cumming et al. (2007) avaliou 10 periódicos de Psicologia que publicam especialmente pesquisas empíricas em três períodos de tempo (1998, 2003-2004 e 2005-2006) e identificou que o NHST foi usado em quase todos os artigos (97,8%, 97,7% e 96,9% nos três períodos, respectivamente). ICs raramente foram relatados, mas seu uso aumentou (3,7%, 9,2% e 10,6%).

O uso de estatísticas inferenciais não é restrito a grupos. Ainda que muitos dos testes utilizados na análise de dados obtidos em experimentos com grupos não seja adequado para análises de dados provenientes de experimentos de caso único (ver Hartmann, 1974), há estatísticas inferenciais relacionadas a análises sequenciais (e.g., McComas et al., 2009), séries temporais (e.g., Borckardt et al., 2013) e outras estratégias bayesianas que são especialmente adaptadas para o tipo de informação produzidas nesse último tipo de delineamento (Natesan, 2019). Além disso, é comum que correlações entre as principais variáveis dos estudos experimentais e outras variáveis registradas sejam também informações de interesse do pesquisador. Os demais estudos mais recentes que fizeram revisões semelhantes dos periódicos de AC fora do Brasil não avaliaram este aspecto. Em periódicos de AC no Brasil, o uso de estatísticas inferenciais foi identificado em doze (6,06%) dos estudos de caso único e os testes utilizados foram variados, incluindo tanto testes de diferenças de médias, como Teste *t*, ANOVA, Wilcoxon, Mann-Whitney, Kruskal-Wallis; quanto testes de correlação, como Pearson, Spearman, e Coeficiente de Kendall. Análises de séries temporais, análises sequenciais ou estatísticas bayesianas não foram encontradas. Não avaliamos, neste estudo, se o uso de testes com NHST foi adequado para o delineamento metodológico utilizado, mas a existência desses testes em delineamentos de caso único mostra que as preocupações relacionadas ao NHST não se restringem aos delineamentos de grupo.

A apresentação inadequada dos dados provenientes de testes estatísticos inferenciais é também um problema importante na literatura em psicologia. Nos periódicos de AC no Brasil 82 estudos (69,4%) dos que utilizaram estatísticas inferenciais dicotômicas representadas pelo NHST não apresentaram as informações de forma completa e adequada segundo as recomendações da APA, impedindo que outros pesquisadores possam refazer as análises conforme orientação do manual da APA desde sua 5ª edição. Além disso, a apresentação de estimativas complementares ao *p*-valor (IC e ES) foram infrequentes, simplificando demasiadamente a interpretação dos testes estatísticos empregados e impondo um raciocínio dicotômico problemático típico da análise isolada do *p*-valor.

Em síntese, os problemas do uso do NHST na ciência parecem se repetir na análise do comportamento brasileira. Contrariando eventuais expectativas de que a análise do comportamento, com sua preferência por delineamentos de caso único, seria pouco afetada pelas dificuldades do uso do NHST e as decorrências desses problemas para a conhecida crise da replicabilidade na psicologia (Branch, 2019), os resultados do presente estudo sugerem que a AC brasileira também é diretamente

afetada por esses problemas, uma vez que o uso do NHST em seus periódicos é frequente e vem aumentando. O problema é agravado quando se considera que parte expressiva do relato de estatísticas relacionadas ao NHST é relatada fora dos padrões preconizados pela APA, dificultando expressivamente a possibilidade de a comunidade conferir se o uso do NHST foi adequado e de fazer interpretações mais sofisticadas do resultado obtido. Exemplo de como esse tipo de verificação pode ser feita é encontrado em Bakker e Wicherts (2011) e em Francis e Thunell (2022) que utilizaram softwares para recalcular as estatísticas apresentadas em diversos estudos da literatura. Esse tipo de análise, entretanto, só é possível se as estatísticas apropriadas forem adequadamente relatadas.

O uso de estatística inferencial, em especial na forma do NHST, tem sido amplamente criticado na ciência em geral (e.g., Cumming, 2012; Goodman, 2019), na Psicologia (e.g., Bakan, 1966) e na Análise do Comportamento (e.g., Sidman, 1960; Branch, 2019). Sua adoção como método preferencial na análise de dados tem gerado vieses na publicação científica e frequentes deturpações do raciocínio científico. Independente desses problemas serem resultados do mau uso do NHST, da dificuldade de interpretação dos seus resultados ou da limitada contribuição que ele tem sozinho para a consolidação do raciocínio científico (e.g., Killeen, 2019), parece claro que a comunidade brasileira de analistas do comportamento precisa olhar mais atentamente para o uso desse recurso metodológico nas suas práticas de produção de conhecimento, afinal, seu uso é frequente e seu relato incompleto.

Ações que podem colaborar para a atenuação deste problema incluem maior investimento no ensino apropriado do uso e interpretação de estatísticas inferenciais, bem como a adoção de políticas editoriais explícitas em relação ao relato de estudos com NHST. O caso da revista *Memory & Cognition* é ilustrativo da efetividade da segunda ação mencionada. Em 1993, o então editor da revista publicou um editorial sugerindo a ampliação de estatísticas descritivas (como barras de erro) e o cuidado para que as conclusões não sejam exclusivamente baseadas em raciocínios dicotômicos derivados do NHST (Loftus, 1993), tal editorial foi seguido de incentivos nessa direção durante o processo editorial no período de 1994 e 1997. Uma análise demonstrou que o uso de NHST diminuiu e a apresentação de barras de erro aumentou durante o período de gestão desse editor, mas a frequência dessas estatísticas voltou a padrões anteriores após a saída de Loftus da editoria (Finch et al., 2004). Mais recentemente, a comparação da implementação de boas práticas no relato de estatística na *Psychological Science*, tendo o *Journal of Experimental Psychology: General* como controle, demonstrou que o incentivo ao uso de ES, IC e meta-análises bem como o uso de práticas de ciência aberta pode ser substancialmente aprimorado com a adoção de políticas editoriais nessa direção (Giofrè et al., 2017).

Portanto, considerando o potencial pernicioso da adoção desmedida do NHST, a constatação de que esse uso é frequente e tem aumentado nos periódicos brasileiros de AC e que políticas editoriais tem se mostrado formas efetivas de intervenção sobre esse problema, sugere-se que os periódicos da área revisem suas políticas editoriais e implementem práticas que garantam o uso adequado desse tipo de estatística e incentivem o uso de formas alternativas de representação e análise dos dados.

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. (1996). *Initial Report Task Force on Statistical Inference*. 1-4.
- American Psychological Association. (2019). *Publication manual of the American Psychological Association*. American Psychological Association.
- Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *The Behavior Analyst*, 22(2), 93-97. <https://doi.org/10.1007/BF03391985>
- 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, 1-9 <https://doi.org/10.3389/fpsyg.2016.01247>
- Badenes-Ribera, L., Frias-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>
- Baron, A. (1990). Experimental designs. *The Behavior Analyst*, 13(2), 167-171. <https://doi.org/10.1007/BF03392533>
- Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research*, 16(4), 202-207. <https://doi.org/10.1002/mpr.225>
- Borckardt, J. J., Nash, M. R., Balliet, W., Galloway, S., & Madan, A. (2013). Time-series statistical analysis of single-case data. En G. J. Madden, W. V. Dube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), *APA handbook of behavior analysis, Vol. 1: Methods and principles*. (pp. 251-266). <https://doi.org/10.1037/13937-011>
- Branch, M. N., & Pennypacker, H. S. (2013). Generality and generalization of research findings. *APA handbook of behavior analysis, Vol. 1: Methods and principles*. (pp. 151-175). <https://doi.org/10.1037/13937-007>
- 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>

- 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>
- Chase, L. J., & Tucker, R. A. Y. K. (1976). Statistical Power: Derivation, development, and data-analytic implications. *The Psychological Record*, 26(4), 473-486. <https://doi.org/10.1007/BF03394413>
- 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>
- Cowles, M. (2001). *Statistics in psychology: An historical perspective*. Psychology Press <https://doi.org/10.4324/9781410612380>
- Crosbie, J. (1999). Statistical inference in behavior analysis: Useful friend. *The 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*. 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>
- Davison, M. (1999). Statistical inference in behavior analysis: Having my cake and eating it? *The Behavior Analyst*, 22(2), 99-103. <https://doi.org/10.1007/BF03391986>
- 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>
- Finch, S., Cumming, G., Williams, J., Palmer, L., 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. <https://doi.org/10.3758/BF03195577>
- Fisher, R. A. (1956). *Statistical methods and statistical inference*. Oliver and Boyd.
- Foster, T. M., Jarema, K., & Poling, A. (1999). Inferential statistics: Criticized 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>
- Francis, G., & Thunell, E. (2022). Data detective methods for revealing questionable research practices. En W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 123-145). Springer International Publishing. https://doi.org/10.1007/978-3-031-04968-2_6
- 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. En D. Kaplan (Ed.). *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>
- Goodman, S. N. (2019). Why is getting rid of *p*-values so hard? Musings on science and statistics. *The American Statistician*, 73(sup1), 26-30. <https://doi.org/10.1080/00031305.2018.1558111>
- 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), 635-638. <https://doi.org/10.1901/jaba.1974.7-635>
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLOS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- 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>
- 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. (2021). Historically recontextualizing Sidman's *Tactics*: How behavior analysis avoided psychology's methodological Ouroboros. *Journal of the Experimental Analysis of Behavior*, 115(1), 115-128. <https://doi.org/10.1002/jeab.661>
- 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. (2013). Single-case research methods: An overview. En G. J. Madden (Ed.), *APA Handbook of Behavior Analysis: Methods and Principles* (pp. 3-32). American Psychological Association.
- Johnston, J. M., Pennypacker, H. S., & Green, G. (2020). *Strategies and tactics of behavioral research and practice*. Routledge. <https://doi.org/10.4324/9781315537085>
- 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

- 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*. American Psychological Association. <https://doi.org/10.1037/14136-000>
- Kratochwill, T. R., & Brody, G. H. (1978). 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. <https://doi.org/10.1177/014544557823001>
- Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of p-values in Experimental psychology literature. *PLoS ONE*, 10(6), e0127872. <https://doi.org/10.1371/journal.pone.0127872>
- 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*, 111(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>
- Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *Quarterly Journal of Experimental Psychology*, 66(12), 2303-2309. <https://doi.org/10.1080/17470218.2013.863371>
- 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>
- McComas, J. J., Moore, T., Dahl, N., Hartman, E., Hoch, J., & Symons, F. (2009). Calculating contingencies in natural environments: Issues in the application of sequential analysis. *Journal of Applied Behavior Analysis*, 42(2), 413-423. <https://doi.org/10.1901/jaba.2009.42-413>
- 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>
- 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. <https://doi.org/10.1037/0022-006X.46.4.806>
- Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7(4), 647-653. <https://doi.org/10.1901/jaba.1974.7-647>
- Natesan, P. (2019). Fitting bayesian models for single-case experimental designs. *Methodology*, 15(4), 147-156. <https://doi.org/10.1027/1614-2241/a000180>
- Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 6(2), 134-138. <https://doi.org/10.3389/fpsyg.2015.00223>

- Rader, A. E., Rader, K. A., Katz, J. S., & Leaf, J. B. (2021). The progression of experimental design and data analysis in applied behavior analysis. *European Journal of Behavior Analysis*, 22(2), 152-172. <https://doi.org/10.1080/15021149.2021.1932199>
- 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. <https://doi.org/10.5514/rmac.v1.i2.27099>
- Reese, H. W. (1999). Problems of statistical inference. *Mexican Journal of Behavior Analysis*, 25(1), 39-68.
- 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>
- 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>
- 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>
- Young, M. E. (2018). A place for statistics in behavior analysis. *Behavior Analysis: Research and Practice*, 18(2), 193-202. <https://doi.org/10.1037/bar0000099>
- Zimmermann, Z. J., Watkins, E. E., & Poling, A. (2015). JEAB research over time: Species used, experimental designs, statistical analyses, and sex of subjects. *The Behavior Analyst*, 38(2), 203-218. <https://doi.org/10.1007/s40614-015-0034-5>

Notas

Agradecemos a Amanda Carolina Alexandre Petinati e Glória Maria Oliveira Nogueira, alunas do curso de Psicologia da Universidade Federal do Paraná, pelo auxílio na coleta de dados.

(Received: September 09, 2024; Accepted: February 05, 2025)