PENSATA

Submetida 22.02.2019. Aprovada 19.09.2019
Avaliado pelo sistema *double blind review*. Editor Científico convidado: Marco Aurélio Carino Bouzada
Versão original
DOI: http://dx.doi.org/10.1590/S0034-759020200306

O P AINDA TEM VALOR?

INTRODUÇÃO

O paradigma positivista utilizado na pesquisa admite a existência de um mundo objetivo que pode ser medido e que admite relações claras entre as variáveis. Como evolução, o pós-positivismo incorpora a noção de que tais relações podem ser apenas probabilísticas (Gephart, 1999). Na tradição funcionalista, a possibilidade de replicação é fundamental, e são esperados resultados similares quando analisamos dados obtidos em situações semelhantes (Shah & Corley, 2006).

Várias áreas da Administração utilizam o paradigma pós-positivista e as técnicas estatísticas do Teste de Hipótese para suas conclusões obtidas a partir de observação ou experimentação. Em particular, livros-texto de Pesquisa em Marketing utilizam intensamente métodos estatísticos, apesar de conterem alertas sobre eventual valorização excessiva desses métodos. Por exemplo, Lehmann, Gupta e Steckel (1998) alertam contra o mito de estes serem "[...] uma forma elevada de lógica, pura e abstrata" (p. 8, tradução nossa).

Todavia, um importante periódico do campo da Psicologia recentemente baniu o Teste de Hipótese e o conceito do valor-p de seus artigos, recomendando outras técnicas para a generalização dos resultados amostrais (os editores defendem que o que foi sempre feito deveria ser interrompido imediatamente, para o bem da Psicologia): "De agora em diante, BASP [*Basic and Applied Social Psycology*] está banindo o NHSTP [*Null Hypothesis Significance Testing Procedure*]", afirmaram os editores Trafimow e Marks (2015, p. 1, tradução nossa).

De fato, existem questionamentos importantes: essa prática de obterem-se achados puramente por meio do valor-p contribui para a construção de teorias e conhecimento? Como toda discussão, há vais e voltas, até se obter um equilíbrio razoável. Mas essa importante discussão precisa chegar às áreas da Administração. Seguem alguns dos problemas relacionados com os artigos que utilizam o valor-p como base para suas conclusões:

a. má interpretação do valor-p obtido (uso de falácias lógicas), quando, a partir de algumas estrelinhas (***) ao lado dos valores-p calculados pelo *software*, pula-se diretamente para "fortes conclusões gerais";
b. excesso de importância dada ao valor-p, quando não se examina o tamanho do efeito;
c. realização de inúmeros experimentos até achar "conclusões importantes", desprezando os problemas causados pelo uso indiscriminado de p-hacking e HARKing, bem como pela falta de relato de toda a cadeia anterior de experimentos realizados; e
d. seleção adversa para publicação – experiências sem "bonitas conclusões" podem não ter o mesmo apelo de publicação, mesmo que sejam importantes para a Ciência.

NELSON LERNER BARTH[1]
nelson.barth@fgv.br
ORCID: 0000-0003-2546-4242

CARLOS EDUARDO LOURENÇO[1]
caerib@gmail.com
ORCID: 0000-0002-9278-8282

1Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo, São Paulo, SP, Brasil

POUCO ENTENDIMENTO DO SIGNIFICADO DO VALOR-P OBTIDO

Pesquisadores utilizam os conceitos do Teste de Hipótese, a saber, hipótese nula (H_o), hipótese alternativa (H_a) e valor-p. A probabilidade de H_o não ser verdadeira é uma formulação incorreta que aparece em vários artigos (Wasserstein & Lazar, 2016), nos quais, a partir de um valor-p ‹ 0,05 (ou ‹ 0,001, com várias estrelas na frente), parte-se diretamente para a uma conclusão na qual se menciona que Ho é altamente improvável.

Cohen (1994) compara dois silogismos utilizando uma argumentação de Pollard e Richardson (1987). O primeiro é correto: "Se uma pessoa é de Marte, então não é membro do congresso. A pessoa é membro do congresso. Logo, não é de Marte" (Cohen, 1994, p. 998, tradução nossa). O segundo é incorreto: "Se uma pessoa é americana, então a pessoa provavelmente não é membro do congresso. A pessoa é membro do congresso. Portanto, a pessoa provavelmente não é americana" (p. 998, tradução nossa). O segundo silogismo, que claramente é falso, equipara-se formalmente a: "Se H_o é verdadeira, então este resultado provavelmente não ocorre. Este resultado ocorreu. Então H_o provavelmente não é verdadeira e, portanto, formalmente inválida" (p. 998, tradução nossa).

Esse tipo de silogismo incorreto aparece também em alguns bons livros-texto de estatística aplicada, talvez por descuido ou má redação. Por exemplo: "[...] se obtivermos um valor experimental que caiu na região crítica, será pouco provável que a hipótese H_o seja verdadeira [...]" (Costa, 1977, p. 88); "[...] a comparação de afirmativas ou previsões com estatísticas amostrais permite decidir se as hipóteses estatísticas são aceitáveis ou não: a hipótese proposta é aceita sempre que provável; se improvável, aceita-se sua negação" (Milone, 2004, p. 235).

O problema é que a probabilidade de se obter um resultado amostral dada a hipótese H_o (que se faz corretamente no Teste de Hipótese) não é a mesma de H_o ser verdadeira dada o resultado amostral obtido. Ou seja, não podemos falar na probabilidade de H_o ser verdadeira, a menos que usemos o Teorema de Bayes, o que pode não ser muito simples porque normalmente não temos a probabilidade *a priori* de H_o ser verdadeira. Afinal, qual é a probabilidade *a priori* de uma teoria ser correta? Por exemplo, qual é a probabilidade *a priori* de a Teoria Geral da Relatividade ser correta? (Rozeboom, 1960).

Há várias décadas, Rozeboom (1960) já alertava sobre as diferentes interpretações da inferência, quando feitas por matemáticos (mais preocupados com o rigor formal), por filósofos (um embaraçoso mistério) e pelo cientista experimental (um necessário instrumento de pesquisa), levantando uma série de objeções ao método e fechando seu artigo com "[...] seu principal erro está em entender o objetivo da investigação científica como uma decisão, ao invés de uma avaliação cognitiva de proposições" (p. 426, tradução nossa).

CONCLUSÃO A PARTIR DO VALOR-P SEM EXAMIMAR O TAMANHO DO EFEITO

Ao usar o procedimento comum de Teste de Hipótese, pesquisadores estabelecem duas hipóteses mutuamente excludentes, H_o e H_a. Se o experimento tiver o sucesso desejado, será encontrada evidência para rejeitar H_o e, consequentemente, para aceitar H_a. Por exemplo, podemos ter H_o: $\theta = a$ e H_a: $\theta \neq a$. Todavia, se as hipóteses forem baseadas em uma variável contínua (assume valores de números reais), então nenhum experimento resultará exatamente em $\hat{\theta} =$ a com todas as casas imagináveis após a vírgula. Ou seja, para amostras suficientemente grandes, seriam obtidos valores-p suficientemente baixos e H_o seria sempre rejeitada e H_a seria sempre aceita. Qualquer nova teoria seria provada estatisticamente, independentemente de seu mérito real, se fosse possível realizar um experimento com um tamanho de amostra suficientemente grande (Kwan & Friendly, 2004).

Valores-p não servem para medir a importância ou a dimensão de um efeito, visto que dependem do tamanho da amostra (Wasserstein & Lazar, 2016). Um efeito desprezível pode ser estatisticamente significativamente se a amostra for suficientemente grande. Um impressionante efeito pode não ser estatisticamente significante se o tamanho da amostra for insuficiente. Mais ainda: na maioria das questões práticas das áreas da Administração, haver fortes evidências estatísticas de que o efeito gerado é diferente de zero corresponde a uma informação sem valor prático algum. Mesmo que seja utilizada uma interpretação correta do valor-p, a obtenção de um valor-p ‹ 0,05 não gerará conhecimento: há que se informar também o "tamanho do efeito" (*effect size*).

Um exemplo é citado com frequência (Sullivan & Feinn, 2012) sobre a questão do tamanho do efeito: em um estudo envolvendo mais de 22 mil pessoas, a aspirina foi associada à redução de infartos do miocárdio, com valor-p = 0,00001 (Bartolucci, Tendera, & Howard, 2011), após o que a droga passou a ser recomendada para prevenção geral. Todavia o tamanho do efeito (significância prática) era muito pequeno: uma diferença no risco de infartos no miocárdio de 0,77%, com $R^2 = 0,001$, o que causou uma revisão posterior nessa recomendação médica, em virtude dos efeitos colaterais da aspirina. Estudos que

 ISSN 0034-7590; eISSN 2178-938X

apresentam conclusões exclusivamente a partir do valor-p podem estar simplesmente evidenciando resultados absolutamente desprezíveis na prática. É fundamental, em paralelo, verificar o tamanho do efeito.

As medidas de tamanho do efeito permitem saber se determinado efeito de fato possui importância no seu campo de estudo. Por exemplo, em experimentos em lojas de varejo, se a conversão em vendas de determinado produto aumentou de 31% para 32% (valor-p = 0,001) ao adotar um estilo de música no ambiente, o resultado, apesar de estatisticamente provado, é absolutamente irrelevante para os gestores. Por outro lado, se a conversão em vendas aumentou de 31% para 52% (com o mesmo valor-p =0,001), temos um tamanho do efeito da música que agora é digno de nota e de decisões de gestão. Apesar de que o tamanho do efeito pode ser explicitado de inúmeras formas diferentes de acordo com a técnica estatística utilizada (diferenças entre médias em duas subpopulações, correlações, coeficientes de determinação R^2, coeficientes da regressão, *odds ratio* e vários outros), existem algumas medidas mais comuns, que facilitam posteriores estudos de meta-análise, por exemplo, D de Cohen (que nada mais é do que a diferença padronizada entre duas médias) e Correlação de Pearson (Borenstein, 2011).

CAÇADORES DE VALORES-P: *P-HACKING* E *HARKING*

Richard Bettis, professor da Kenan-Flager Business School, na University of North Carolina, perguntou, certa vez, a um aluno de doutorado de uma importante escola de negócios americana: "O que você está estudando?". E a resposta foi: "Eu estou procurando asteriscos" (Bettis, 2012, p. 108, tradução nossa). Valores-p menores que 0,05, 0,01 e 0,001 são assinalados pelo *software* estatístico R com asteriscos, a saber, * para 0,05, ** para 0,01 e *** para 0,001 (Navarro, 2017, p. 339).

Dois incentivos somam-se para que os pesquisadores às vezes se comportem como caçadores de valores-p ‹ 5%, cientes ou não de estarem sacrificando a boa ciência: a) publicação em bons periódicos é vital para pesquisadores terem emprego, promoções e verbas; b) resultados positivos e originais têm maior probabilidade de publicação em comparação com resultados negativos ou reprodução de experimentos (Nosek, Spies, & Motyl, 2012; Witteloostuijn, 2015). Colocado o objetivo de obter valor-p ‹ 5%, o pesquisador terá grande liberdade para conduzir sua análise, tornando-se fácil publicar resultados estatisticamente significantes: momento de cessar a coleta de dados, critérios para exclusão de observações, utilização de variáveis de controle, transformação e combinação de medidas, além de outros (Brodeur, Lé, Sangnier, & Zylberberg, 2016; Meyer, Witteloostuijn, & Beugelsdijk, 2017; Simmons, Nelson, & Simonsohn, 2011).

P-hacking (Starbuck, 2016), também conhecido como *data fishing* e, no Brasil, como "torturar os dados até confessarem", refere-se à busca de um valor-p ‹ 5% mediante variações no método de análise utilizado. *HARKing* significa *Hypothesizing After the Results are Known* (Kerr, 1998). Essas duas práticas não obrigatoriamente indicam má intenção do pesquisador, visto que decorrem do desejo de obter um resultado estatisticamente significante e da excessiva liberdade para a condução da análise, além de alguns fenômenos cognitivos (Munafò et al., 2017): a) apofenia (percepção de padrões ou conexões em dados puramente aleatórios); b) viés confirmatório (foco no que está em conformidade com nossas expectativas anteriores); c) viés de retrospectiva (tendência para enxergar um evento, que acaba de ocorrer, como tendo sido previsível).

P-hacking e HARKing possuem consequências graves para a boa ciência. Ao utilizar um nível de significância $\alpha = 5\%$ em cada análise de um experimento, declara-se aceitável a probabilidade de 5% de obter um falso positivo (ou Erro Tipo I) nessa análise. Mas a probabilidade de ao menos uma das várias análises realizadas produzir conclusão falsamente positiva pode ser muito maior que 5%. Se forem realizadas 100 análises diferentes, em busca de algum valor-p ‹ 5%, a probabilidade de se obter ao menos um falso positivo será de $1 - (1-0{,}05)^{100} = 99{,}4\%$, ou seja, quase uma certeza. Portanto, ao contrário, para fazer um conjunto de 100 análises, com uma probabilidade de 5% de ocorrência de falso positivo, dever-se-ia usar nível de significância $\alpha = 0{,}05\%$ em cada uma das análises individuais (Bettis, 2012; Benjamini & Braun, 2002).

Ao se analisarem coletâneas de estudos quantitativos publicados, pode-se analisar a distribuição dos valores-p presentes. Em virtude de *p-hacking* e *HARKing*, observa-se uma inesperada alta concentração de valores-p logo abaixo de 5% (e uma inesperada baixa concentração de valores-p logo acima de 5%). Esse fenômeno é relatado por Brodeur et al. (2016) em periódicos de Economia (*American Economic Review*, *Quaterly Journal of Economics* e *Journal of Poltical Economy*), por Masicampo e Ladance (2012) em periódicos de Psicologia (*Journal of Experimental Psychology*, *Journal of Personality and Social Psychology* e *Psychological Science*) e por Meyer et al. (2017) em uma análise dos artigos de 2015-2016 nas publicações *Journal of International Business Studies*, *Organizations Science* e *Strategic Management Journal*.

Um primeiro possível remédio para o *p-hacking* e para o *HARKing* é indicado pela *American Statistical Association* como

   ISSN 0034-7590; eISSN 2178-938X

fundamental para o uso de valores-p: "Pesquisadores precisam divulgar o número de hipóteses exploradas durante um estudo, todas as decisões tomadas durante a coleta de dados, todas as análises estatísticas conduzidas e todos os valores-p calculados" (Wasserstein & Lazar, 2016, p. 132, tradução nossa). Apesar de absolutamente precisa e correta, essa pesada recomendação não parece ser de fácil implementação e controle.

Um segundo remédio para o *p-hacking* e para o *HARKing* é a replicação das pesquisas, de modo a se poder de fato generalizar os resultados obtidos e gerar contribuição genuína para a Ciência. Todavia, Evanschitzky, Baumgarth, Hubbard e Armstrong (2007) relatam a baixíssima percentagem de publicação de artigos com replicações, na área de *Marketing*, nos *Journal of Marketing*, *Journal of Marketing Research* e *Journal of Consumer Reseach*: 1,2% no período 1990-2004 (contra 2,4% no período 1974-1989). Vários importantes periódicos (*Strategic Management Journal*, *Administrative Science Quaterly*, *Organization Science*), apesar de formalmente incentivarem trabalhos de replicação, têm raros exemplos publicados nos últimos anos (Witteloostuijn, 2015).

Do lado dos pesquisadores, estes devem ter como meta fazer real ciência. Portanto, há a necessidade de incentivos adequados, para que o pesquisador deva e possa: a) registrar todas as análises prévias realizadas, deixando de desprezar resultados negativos; b) publicar experimentos que geraram resultados negativos; c) publicar reproduções de experimentos. Um dos caminhos é por meio do pré-registro de pesquisa ("Promoting reproducibility", 2017), na qual o periódico aprova e garante a publicação após ter analisado o protocolo completo, independentemente do resultado a ser obtido, desde que o pesquisador siga tal protocolo completamente. A possibilidade de pré-registro tem se tornado comum em medicina clínica, evitando *p-hacking* e *HARKing*, mas ainda não está disseminado nas áreas de Ciências Sociais (Meyer et al., 2017; Munafò et al., 2017; Witteloostuijn, 2015).

Simmons, Nelson e Simonsohn (2013) sugerem aos pesquisadores que carimbem suas pesquisas como livres de *p-hacking*, escrevendo: "Nós relatamos como determinamos nosso tamanho de amostra, todas as eliminações de dados, todas as manipulações e todas as medidas realizadas no estudo" (p. 775, tradução nossa).

O EFEITO GAVETA E A POTENCIAL SOLUÇÃO POR META-ANÁLISE

Meyer et al. (2017), utilizando de bom humor, dizem que os acadêmicos devem ser capazes de prever o futuro, visto que obtêm evidências empíricas para a grande maioria de suas hipóteses. Segundo os autores, das 711 hipóteses testadas em artigos publicados pelos *Journal of International Business Studies*, *Strategic Management Journal* e *Organizational Science* em 2016, aproximadamente 89% obtiveram evidências a favor, com significância estatística. Na realidade, ocorreu um viés de publicação: grande parte dos estudos científicos que geraram resultados negativos ou inconclusivos simplesmente não foi publicada (porque os autores os descartaram ou porque os editores não os aceitaram). A esse viés de publicação, Rosenthal (1979) deu o nome de "efeito gaveta" (*file-drawer effect*), escrevendo de maneira provocativa: "A visão extrema do efeito gaveta é que os periódicos são preenchidos com os 5% dos estudos que mostram erros do Tipo I, enquanto as gavetas de arquivo são preenchidas com 95% dos estudos que mostram resultados não significativos [...]" (p. 638, tradução nossa).

A técnica denominada meta-análise reúne, de maneira quantitativa, os resultados de vários estudos anteriores, com o objetivo de estimar o tamanho de um efeito com melhor precisão. Esses estudos anteriores são considerados como uma amostra de todos aqueles que poderiam ser feitos e, assim, as conclusões da meta-análise costumam considerar a parte comum a todos os estudos individuais envolvidos (Card, 2012).

Sob o guarda-chuva da meta-análise, existem técnicas capazes de medir o efeito gaveta e até fazer correções para minimizar o seu impacto nas estimativas do tamanho do efeito. Uma delas, a técnica do Gráfico do Funil (Sterne, Becker, & Egger, 2005), tenta detectar o engavetamento de estudos a partir da distribuição dos tamanhos dos efeitos em estudos menores que deveria ser esperada a partir dos tamanhos dos efeitos encontrados nos estudos com amostras maiores.

Infelizmente, entretanto, as meta-análises, bem estabelecidas nos campos da Medicina e da Psicologia, são raramente publicadas nos periódicos de negócios e gestão, nomeadamente, *Academy of Management Journal*, *Administrative Science Quaterly* e *Journal of Management*, talvez em função da falta de similaridade entre os estudos na área (Witteloostuijn, 2015).

CAMINHOS

Dados os problemas inerentes ao uso impróprio do valor-p para, por si só, aferir a relevância científica de um estudo, alguns movimentos ocorreram. De modo mais radical, um periódico recentemente decidiu banir totalmente o uso do valor-p em seus artigos. Na ocasião, os editores de *Basic and Applied Social Psychology* (Trafimow & Marks, 2015) anteciparam algumas

   ISSN 0034-7590; eISSN 2178-938X

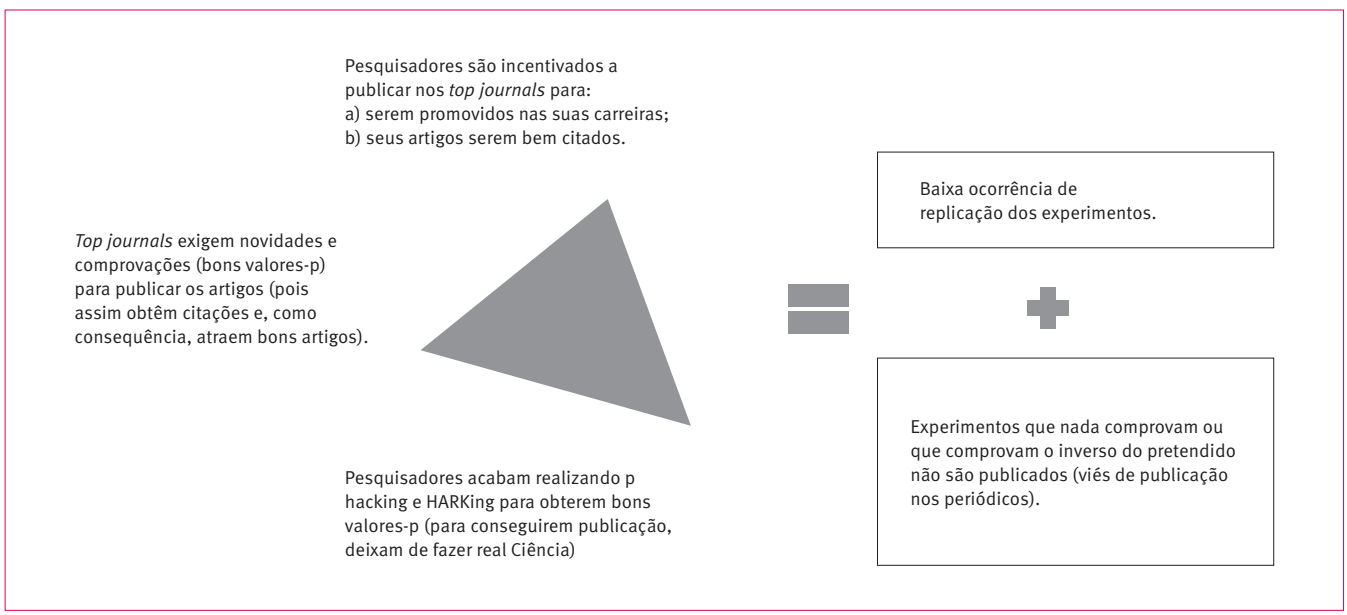
perguntas possíveis e, entre elas, a primeira era: "Os manuscritos que contenham valor-p serão rejeitados automaticamente?" (p. 1, tradução nossa). A resposta foi: "Não. Se os manuscritos passarem pela inspeção preliminar, eles serão enviados para a revisão. Mas, antes da publicação, os autores terão que remover todos os vestígios de NHSTP [*null hypothesis significance testing procedure*]." (p. 1, tradução nossa).

A esse movimento, veio o contra-ataque (García-Pérez, 2016): há de fato importantes críticas sobre o uso do Teste de Hipótese, mas estas aparecem devido a crenças falsas, interpretações incorretas e expectativas irreais de uma parcela dos pesquisadores. O problema não está no conceito do valor-p propriamente dito, mas no seu uso às vezes inadequado. O autor também alerta que o único caminho que poderia ser seguido a partir dessa decisão de banir o Teste de Hipótese seria o de restringir os estudos a amostras enormes, nos quais a Estatística Descritiva poderia ser usada de modo suficiente com tranquilidade.

A síntese desse movimento aparece claramente em 2019 em um longo editorial da *The American Statistician* (Wasserstein, Schirm, & Lazar, 2019), onde há alguns consensos estabelecidos sobre o que não fazer. Por exemplo, mesmo mantendo o uso do valor-p como uma grandeza contínua, não mais utilizar a expressão "estatisticamente significante" em nenhuma situação, para evitar crenças inadequadas sobre o seu significado. E há também uma grande coleção de recomendações sobre o que fazer, para autores, pareceristas e editores – todavia, como fica claro no texto, essas últimas estão longe de apresentarem unanimidade. Um comentário na *Nature* (Amrhein, Greenland, & McShane, 2019) recomenda também fortemente o banimento da expressão "estatisticamente significante", recomendando apresentar os valores-p com precisão adequada, sem comparações do tipo p ‹ 0,05 e sem estrelas.

A questão da reprodução das pesquisas científicas precisa urgentemente de uma política de incentivo. De alguma forma, o triângulo da Figura 1 precisa ser quebrado. Como pano de fundo para o estado atual das pesquisas quantitativas, há a dificuldade de se evitar o uso inadequado do valor-p como um árbitro superior da verdade. Goodman (2019) entende que o problema não é científico nem filosófico, mas sim sociológico, visto que o valor-p é usado devido ao seu valor para atestar conhecimento, permitir publicação, obter verbas para pesquisa e conseguir promoções acadêmicas. Ou seja, há necessidade de mudança em instituições acadêmicas, periódicos, agências provedoras de verbas para pesquisa e agências regulatórias. "Ao final, a única forma de resolver o problema da replicabilidade é fazer mais replicações e reduzir os incentivos que são impostos aos cientistas para produzirem trabalhos sem confiabilidade" (Colquhoun, 2019, p. 192, tradução nossa).

Figura 1. **O ciclo dos incentivos vai contra a Ciência**



Wasserstein et al. (2019) discutem o ritmo com o qual os periódicos devem implementar suas exigências para os artigos que utilizam inferência estatística. De fato, quebrar o ciclo de incentivos exposto na Figura 1 não é simples. Os autores desta Pensata sugerem que os periódicos brasileiros na área de Administração comecem a realizar passos pequenos e importantes, até

que a comunidade científica consiga chegar a novos padrões claros e unânimes. Esses cinco passos são: a) exigir que os autores declarem nos seus artigos, clara e formalmente, que não realizaram p-hacking ou HARKing; b) proibir que os autores classifiquem seus achados como estatisticamente significantes a partir do valor-p; c) exigir análise do tamanho do efeito; d) abrir espaço e incentivar estudos que utilizem o pré-registro (como forma de evitar o viés de publicação); e) incentivar ativamente a publicação de replicação de pesquisas (ou seja, não privilegiar apenas pesquisas inovadoras) e de meta-análises.

Foram examinadas as orientações para autores em todos os periódicos brasileiros, editados em território nacional, na área de Administração Pública e de Empresas, Ciências Contábeis e Turismo, com a classificação Qualis A2, no mínimo, conforme fotografia de fevereiro de 2019 (Brito, Luca, & Teixeira, 2017): *Advances in Scientific and Applied Accounting*, *Brazilian Administration Reviews*, *Brazilian Business Review*, *Cadernos EBAPE*, *Estudios y Perspectivas en Turismo*, *Contabilidade Vista & Revista*, *Organização & Sociedade*, *Revista Brasileira de Gestão de Negócios*, *Revista Brasileira de Pesquisa em Turismo*, *Revista Contabilidade & Finanças*, *Revista Contemporânea de Contabilidade*, *Revista de Administração Contemporânea*, *Revista de Administração da USP*, *Revista de Administração de Empresas*, *Revista de Administração Pública*, *Revista de Contabilidade e Organizações* e *Revista Universo Contábil*. As orientações para autores nesses 17 periódicos foram analisadas em agosto de 2019, e procuraram-se os cinco passos acima apontados. Verificou-se que os passos recomendados não foram encontrados de maneira explícita em qualquer um deles. Encontraram-se, todavia, em alguns desses periódicos, alertas genéricos sobre a fabricação e falsificação de dados e resultados (Byington & Felps, 2017), mas os autores desta Pensata consideram mais recomendável exigir que os próprios autores dos artigos científicos declarem que especificamente não realizaram p-hacking ou HARKing.

Vislumbra-se aqui uma oportunidade de inovação para esses periódicos brasileiros na área de Administração Pública e de Empresas, Ciências Contáveis e Turismo. Afinal, "Novidade e resultados positivos são vitais para a Publicabilidade, mas não para a Verdade" (Nosek et al., 2012, p. 617, tradução nossa).

REFERÊNCIAS

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305-307. doi: 10.1038/d41586-019-00857-9

Bartolucci, A. A., Tendera, M., & Howard, G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using Aspirin. *The American Journal of Cardiology*, *107*(12), 1796-1801. doi: 10.1016/j.amjcard.2011.02.325

Benjamini, Y., & Braun, H. (2002). John W. Tukeys contributions to multiple comparisons. *The Annals of Statistics*, *30*(6), 1576-1594. doi: 10.1214/aos/1043351247

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108-113. doi: 10.1002/smj.975

Borenstein, M. (2011). *Computing effect sizes for meta-analysis*. Oxford, Inglaterra: Wiley-Blackwell.

Brito, E. P. Z., Luca, M. M. M., & Teixeira, A J. C. (2017). *Considerações sobre Qualis Periódicos – Administração Pública e de Empresas, Ciências Contábeis e Turismo*. Recuperado de https://capes.gov.br/images/Qualis_periodicos_2017/Consideracoes_Qualis_Periodicos_Area_27_2017_-_final.pdf

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star Wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1-32. doi: 10.1257/app.20150044

Byington, E. K., & Felps, W. (2017). Solutions to the credibility crisis in management science. *Academy of Management Learning & Education*, *16*(1), 142-162. doi: 10.5465/amle.2015.0035

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, USA: The Guilford Press.

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

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-Values. *The American Statistician*, *73*(sup1), 192-201. doi: 10.1080/00031305.2018.1529622

Costa, P . L. O., Neto. (1977). *Estatística*. São Paulo, SP: Editora E. Blücher.

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication researchs disturbing trend. *Journal of Business Research*, *60*(4), 411-415. doi: 10.1016/j.jbusres.2006.12.003

García-Pérez, M. A. (2016). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*, *77*(4), 631-662. doi: 10.1177/0013164416668232

Gephart, R. (1999). Paradigms and research methods. *[Online] Research Methods Forum*, *4*(Summer). Recuperado de http://division.aomonline.org/rm/1999_RMD_Forum_Paradigms_and_Research_Methods.htm

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. doi: 10.1080/00031305.2018.1558111

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. doi: 10.1207/s15327957pspr0203_4

Kwan, E., & Friendly, M. (2004). Discussion and comments: Strong versus weak significance tests and the role of meta-analytic procedures. *Journal de la Societé Française de Statistique*, *145*(4), 47-53. Recuperado de http://www.numdam.org/item/JSFS_2004__145_4_47_0/

Lehmann, D. R., Gupta, S., & Steckel, J. H. (1998). *Marketing research*. Reading, USA: Addison-Wesley.

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279. doi: 10.1080/17470218.2012.711335

Meyer, K. E., Witteloostuijn, A. V., & Beugelsdijk, S. (2017). What's in a p? Reassessing best practices for conducting and reporting hypothesis-testing research. *Journal of International Business Studies*, *48*(5), 535-551. doi: 10.1057/s41267-017-0078-8

Milone, G. (2004). Estatística: *Geral e aplicada. São Paulo*, SP: Pioneira Thomson Learning.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. D., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021. doi: 10.1038/s41562-016-0021

Navarro, D. J. (2017). *Learning statistics with R*: *A tutorial for psychology students and other beginners (version 0.6)*. New South Wales, Australia: University of New South Wales. Recuperado de http://compcogscisydney.org/learning-statistics-with-r/

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

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

Promoting reproducibility with registered reports [Editorial]. (2017). *Nature Human Behaviour*, 1, 0034. doi: 10.1038/s41562-016-0034

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641. doi: 10.1037/0033-2909.86.3.638

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

Shah, S. K., & Corley, K. G. (2006). Building better theory by bridging the quantitative-qualitative divide. *Journal of Management Studies*, *43*(8), 1821-1835. doi: 10.1111/j.1467-6486.2006.00662.x

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. doi: 10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. In S. Botti & A. Labroo (Eds.), *NA: Advances in consumer research* (Vol. 41, p. 775). Duluth, USA: Association for Consumer Research. Recuperado de http://www.acrwebsite.org/volumes/1015833/volumes/v41/NA-41

Starbuck, W. H. (2016). 60th Anniversary essay: How journals could improve research practices in social science. *Administrative Science Quarterly*, *61*(2), 165-183. doi: 10.1177/0001839216629644

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Suton, & M. Borentein (Eds.), *Publicaton bias in meta-analysis* (pp. 75-98). West Sussex, Inglaterra: Wiley.

Sullivan, G. M., & Feinn, R. (2012). Using effect size: Or why the p-value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279-282. doi: 10.4300/jgme-d-12-00156.1

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. doi: 10.1080/01973533.2015.1012991

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

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p<0.05". *The American Statistician*, *73*(sup1), 1-19. doi: 10.1080/00031305.2019.1583913

Witteloostuijn, A. (2015). *What happened to Popperian falsification? A manifesto to create healthier business and management scholarship – towards a scientific Wikipedia*. Tilburg, Netherlands: Tilburg University. doi: 10.13140/rg.2.1.2455.6889

CONTRIBUIÇÃO DOS AUTORES

Os autores declaram que participaram de forma conjunta em todas as etapas do desenvolvimento do texto: conceitualização, abordagem teórico-metodológica, revisão teórica (levantamento de literatura), coleta de dados, redação e revisão final da Pensata.