ESSAY

Submitted 02.22.2019. Approved 09.19.2019
Evaluated by double blind review system. Invited Scientific Editor: Marco Aurélio Carino Bouzada
Translated version
DOI: http://dx.doi.org/10.1590/S0034-759020200306

DOES P STILL HAVE VALUE?

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

[1]Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo, São Paulo, SP, Brazil

INTRODUCTION

The positivist research approach acknowledges the presence of a predictable outcome that can be measured and that admits clear relationships between the variables. As an evolution, post-positivism incorporates the notion that such relationships can only be probabilistic (Gephart, 1999). In the functionalist tradition, the possibility of replication is indispensable and similar results are expected when we analyze data obtained in similar situations (Shah & Corley, 2006).

Several management fields use the post-positivist paradigm and the statistical techniques of null hypothesis significance testing for their conclusions obtained from observation or experimentation. In particular, marketing research textbooks make heavy use of statistical methods, despite perpetual warnings about possible overvaluation of these methods. For example, Lehmann, Gupta, and Steckel (1998) warn against the myth that they are "[...] a high form of logic, pure and absolute" (p. 8).

However, an important academic journal of psychology recently banned null hypothesis significance testing and the concept of the p-value, recommending other techniques for the generalization of sample results (the editors argue that what was always done should be halted immediately for the sake of psychology). Editors Trafimow and Marks (2015, p. 1) stated, "From now on, BASP [Basic and Applied Social Psychology] is banning the NHSTP [Null Hypothesis Significance Testing Procedure]."

Indeed, there are important questions: Does this practice of obtaining findings purely based on p-values contribute to building theories and knowledge? Like any other discussion, there are vigorous debates until a reasonable balance is reached. However, this important discussion must reach several management fields. Here are some of the problems related to papers that use the p-value as a basis for the conclusions:

a. misinterpretation of the obtained p-value (use of logical fallacies), when researchers forejudge "strong general conclusions" based on those asterisks (***), next to the p-values calculated by statistical software;
b. excessive importance given to the p-value by itself, when the effect size is completely ignored;
c. carrying out numerous experiments until "important conclusions" are found, disregarding the problems caused by the indiscriminate use of p-hacking and HARKing, as well as the underreporting of the total number of experiments; and
d. biased selection for publication—experiments without "pretty conclusions" may not have the same appeal for publication, even if they are important for science.

   ISSN 0034-7590; eISSN 2178-938X

LITTLE UNDERSTANDING REGARDING THE MEANING OF THE P-VALUE

Researchers extensively use the concepts of the NHST (Null Hypothesis Significance Testing), namely, null hypothesis (H_o), alternative hypothesis (H_a), and p-value. The probability that H_o is not true is an incorrect formulation that appears in several papers (Wasserstein & Lazar, 2016), in which p-value ‹ 0.05 (or ‹ 0.001, with several asterisks) allows researchers to reach the conclusion that H_o is highly unlikely.

Cohen (1994) compares two syllogisms using an argument posed by Pollard and Richardson (1987). The first claim is correct: "If a person is a Martian, then he is not a member of Congress. This person is a member of Congress. Therefore, he is not a Martian" (Cohen, 1994, p. 998). The second claim is incorrect: "If a person is an American, then he is probably not a member of Congress. This person is a member of Congress. Therefore, he is probably not an American" (p. 998). The second syllogism, which is clearly false, formally equates the statement, "If H_o is true, then this result (statistical significance) would probably not occur. This result has occurred. Then H_o is probably not true and therefore formally invalid" (p. 998).

This type of incorrect syllogism also appears in some eminent textbooks on applied statistics, perhaps due to carelessness or bad writing. For example: "[...] if we get an experimental in the critical region, the H_o hypothesis is unlikely to be true [...]" (Costa, 1977, p. 88); "[...] the comparison of statements or forecasts with sample statistics allow to decide whether the statistical hypotheses are acceptable or not: the proposed hypothesis is accepted whenever probable; if unlikely, its denial is accepted" (Milone, 2004, p. 235).

The problem is that the probability of obtaining a sample result, given the H_o hypothesis (which is done correctly in the NHST), is not the same for H_o to be true, given the sample result obtained. This means that we cannot discuss the probability that H_o is true unless we use Bayes' Theorem, which may not be simple since we normally do not have the *a priori* probability that H_o is true. After all, what is the *a priori* probability for a theory to be correct? For example, what is the *a priori* probability that the general theory of relativity is correct? (Rozeboom, 1960).

Several decades ago, Rozeboom (1960) had already warned about the different interpretations of statistical inference, such as when made by mathematicians (more concerned with formal rigor), by philosophers (an embarrassing mystery), and by experimental scientists (a necessary research instrument), raising a series of disagreements with the method. He closed his paper with "[...] its most basic error lies in mistaking the aim of a scientific investigation to be a decision, rather than a cognitive evaluation of propositions" (p. 426).

DRAWING CONCLUSIONS FROM THE P-VALUE WITHOUT EXAMINING THE EFFECT SIZE

Using the standard NHST procedure, researchers establish two mutually exclusive hypotheses: H_o and H_a. If the experiment has the desired success, evidence will be found to reject H_o and, consequently, to accept H_a. For example, we can have H_o: $\theta = a$ e H_a: $\theta \neq a$. However, if the hypotheses are based on a continuous variable (real numbers), then no experiment will result in exactly $\hat{\theta} = a$, with extreme precision to include innumerous decimal places. This means that, for sufficiently large samples, sufficiently low p-values would be obtained, H_o would always be rejected, and H_a would always be accepted. Any new theory would be proved statistically, regardless of its real merit, if it were possible to carry out an experiment with a sufficiently large sample size (Kwan & Friendly, 2004).

P-values are not used to measure the importance or the dimension of an effect as they depend on the size of the sample (Wasserstein & Lazar, 2016). A negligible effect can be statistically significant if the sample is large enough. An impressive effect may not be statistically significant if the sample size is insufficient. Furthermore, in most practical matters in management, having strong statistical evidence that the effect generated is different from zero corresponds to information with no practical value. Even if a correct interpretation of the p-value is used, obtaining a p-value ‹ 0.05 will not generate knowledge by itself; it is also necessary to inform the effect size.

An example is often cited (Sullivan & Feinn, 2012) on the effect size issue. In a study involving more than 22,000 people, aspirin was associated with a reduction in myocardial infarction, with a p-value = 0.00001 (Bartolucci, Tendera, & Howard, 2011). After that, the drug was recommended for general prevention. However, the effect size (practical significance) was very small—a difference in the risk of myocardial infarction of 0.77%, with $R^2 = 0.001$, which caused the later revision of this medical recommendation due to the side effects of aspirin. Studies that present conclusions exclusively from the p-value may simply be showing negligible results in practice. It is essential, in parallel, to check the effect size.

The measures of effect size allow us to know if a given effect has importance in its area of study. For example, in experiments on

   ISSN 0034-7590; eISSN 2178-938X

retail stores, if the conversion to sales of a given product increased from 31% to 32% (p-value = 0.001) when playing a certain music genre in the store, the result, despite being statistically proven, is irrelevant to managers. On the other hand, if the conversion to sales increased from 31% to 52% (with the same p-value = 0.001), we have an effect size of music in stores that is now worth noting. The effect size can be explained in many different ways according to the statistical technique used (differences between means in two subpopulations, correlations, coefficient of determination R^2, regression coefficients, odds ratio, and several others). However, there are more common measures that facilitate further meta-analysis studies, for example, Cohen's d (which is nothing more than the standardized difference between two means) and Pearson's correlation (Borenstein, 2011).

HUNTING FOR THE P-VALUE: P-HACKING AND HARKING

Richard Bettis, a professor at the Kenan-Flager Business School at the University of North Carolina, once asked a doctoral student at a major American business school, "So what are you studying?" The answer was, "I look for asterisks" (Bettis, 2012, p. 108). The R statistical software marks p-values lower than 0.05, 0.01, and 0.001 with asterisks, namely * for 0.05, ** for 0.01, and *** for 0.001 (Navarro, 2017, p. 339).

Two main incentives are added that make researchers behave like "hunters" of p-values ‹ 5% sometimes, regardless of being aware if this is good science or not. These are a) publication in good journals is vital for researchers to have jobs, promotions, and grants; b) positive and original results are more likely to be published compared to negative results or experiment reproduction (Nosek, Spies, & Motyl, 2012; Witteloostuijn, 2015). To obtain a p-value ‹ 5%, a researcher will have much freedom to conduct his/her analysis, making it easy to publish statistically significant results. The freedom relates to time to stop collecting data, criteria for excluding observations, use of control variables, transformation and combination of measures, and more (Brodeur, Lé, Sangnier, & Zylberberg, 2016; Meyer, Witteloostuijn, & Beugelsdijk, 2017; Simmons, Nelson, & Simonsohn, 2011).

P-hacking (Starbuck, 2016), also known as data fishing, refers to the search for a p-value ‹ 5% through variations in the analysis method used. In Brazil, p-hacking is also known as "torturing data until they confess". HARKing means hypothesizing after the results are known (Kerr, 1998). These two practices do not necessarily indicate a malevolent intention of the researcher since they arise from the desire to obtain a statistically significant result and given the excessive freedom to conduct the analysis, in addition to some cognitive phenomena (Munafò et al., 2017). Such cognitive phenomena include: a) apophenia (perception of patterns or connections in purely random data); b) confirmation bias (focus on what is in line with previous expectations); c) retrospective bias (tendency to see an event, which has just occurred, as having been predictable).

However, P-hacking and HARKing have serious consequences for good science. When using the level of significance $\alpha = 5\%$ in each analysis of an experiment, the 5% probability of obtaining a false positive (or Type I Error) in that analysis is acceptable. However, the probability that at least one of the several analyses carried out will produce a false positive conclusion can be much greater than 5%. If 100 different analyses are performed, looking for some p-value ‹ 5%, the probability of obtaining at least one false positive will be $1 - (1\text{-}0{,}05)^{100} = 99{,}4\%$, that is, almost a certainty. Therefore, on the contrary, to make a set of 100 studies, with a probability of 5% of the occurrence of false positive, the significance level $\alpha = 0.05\%$ should be used in each of the individual analyses (Bettis, 2012; Benjamini & Braun, 2002).

When analyzing collections of published quantitative studies, one can analyze the distribution of all p-values. Due to p-hacking and HARKing, there is an unexpectedly high concentration of p-values just below 5% (and an unexpectedly low concentration of p-values just above 5%). This phenomenon is reported by Brodeur et al. (2016) in economics journals (American Economic Review, Quarterly Journal of Economics, and Journal of Political Economy), by Masicampo and Ladance (2012) in psychology journals (Journal of Experimental Psychology, Journal of Personality and Social Psychology, and Psychological Science), and by Meyer et al. (2017) in an analysis of the 2015-2016 publications in the Journal of International Business Studies, Organization Science, and Strategic Management Journal.

A first possible remedy for p-hacking and HARKing is indicated by the American Statistical Association as fundamental to the use of p-values: "Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed" (Wasserstein & Lazar, 2016, p. 132). Although accurate and correct, this heavy recommendation does not seem to be easy to implement and control.

A second remedy for p-hacking and HARKing is the replication of research to be able to generalize the results obtained and generate a genuine contribution to science. However, Evanschitzky, Baumgarth, Hubbard, and Armstrong (2007) report a very low percentage of published papers with replications in the

   ISSN 0034-7590; eISSN 2178-938X

marketing field in the Journal of Marketing, Journal of Marketing Research, and Journal of Consumer Research (1.2% in the period 1990-2004, against 2.4% in the period 1974-1989). Several major journals (Strategic Management Journal, Administrative Science Quarterly, and Organization Science), although formally encouraging replication work, have rare examples published in recent years (Witteloostuijn, 2015).

Researchers should aim to make real science. Therefore, there is a need for adequate incentives so that they must: a) record all previous analyses performed, failing to neglect negative results; b) publish experiments that generated negative results; and c) publish reproductions of known experiments. One way to do so is through pre-registration ("Promoting reproducibility", 2017), in which the journal approves and guarantees publication after having analyzed the complete protocol, regardless of the result to be obtained, as long as the researcher followed the protocol. The possibility of pre-registration has become common in clinical medicine, thus avoiding p-hacking and HARKing, but it is not yet widespread in the social sciences (Meyer et al., 2017; Munafò et al., 2017; Witteloostuijn, 2015).

Simmons, Nelson, and Simonsohn (2013) suggest that researchers declare their work as free from p-hacking, writing: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (p. 775).

THE FILE DRAWER EFFECT AND THE POTENTIAL SOLUTION THROUGH META-ANALYSIS

Meyer et al. (2017), using a good sense of humor, say that academics should be able to predict the future since they obtain empirical evidence for the vast majority of their hypotheses. According to the authors, of the 711 hypotheses tested in papers published by the Journal of International Business Studies, Strategic Management Journal, and Organization Science in 2016, approximately 89% obtained favorable evidence, with statistical significance. In fact, there was a publication bias, that is, most scientific studies that generated negative or inconclusive results were simply not published (because either the authors discarded them, or the editors did not accept them). This publication bias was called the "file drawer effect" by Rosenthal (1979), who wrote, "The extreme view of the drawer effect is that journals are filled with the 5% of studies that show Type I errors, while the file drawers back in the lab are filled with 95% of studies that show non-significant results [...]" (p. 638).

A technique called meta-analysis gathers, quantitatively, the results of several previous studies to estimate the effect size with better precision. These previous studies are considered a sample of all those that could be conducted and, therefore, the conclusions of the meta-analysis usually consider the part that is common to all the individual studies involved (Card, 2012).

From the perspective of meta-analysis, there are techniques capable of measuring the file drawer effect and even making corrections to minimize its impact on the effect size estimates. One of them, the funnel plot technique (Sterne, Becker, & Egger, 2005), attempts to detect the file drawer effect of studies from the distribution of effect size in smaller studies that should be expected from the effect size found in the studies with larger samples.

Unfortunately, meta-analyses, which are well established in the fields of medicine and psychology, are rarely published in business and management journals, namely, in the Academy of Management Journal, Administrative Science Quarterly, and Journal of Management, perhaps due to the lack of similarity between studies in the area (Witteloostuijn, 2015).

REFLECTIONS

Given the problems inherent in the inappropriate use of the p-value, by itself, to assess the scientific relevance of a study, some movements have occurred. More radically, a journal recently decided to ban the use of the p-value in its papers completely. At the time, the editors of Basic and Applied Social Psychology (Trafimow & Marks, 2015) anticipated some possible questions. Among them, the first was,

"Will manuscripts with p-values be desk rejected automatically?" (p. 1). The answer was, "No. If manuscripts pass the preliminary inspection, they will be sent out for review. But, prior to publication, authors will have to remove all vestiges of the NHSTP [null hypothesis significance testing procedure]" (p. 1).

For this movement, there was a counterattack by García-Pérez (2016), who argued that there are, indeed, important criticisms about the use of the NHST, but these appear due to false beliefs, incorrect interpretations, and unrealistic expectations of researchers. The problem is not the p-value concept itself, but its occasional incorrect use. The author also contends that the only way to ban the NHST would be to restrict studies to colossal samples sizes in which descriptive statistics could be used with ease.

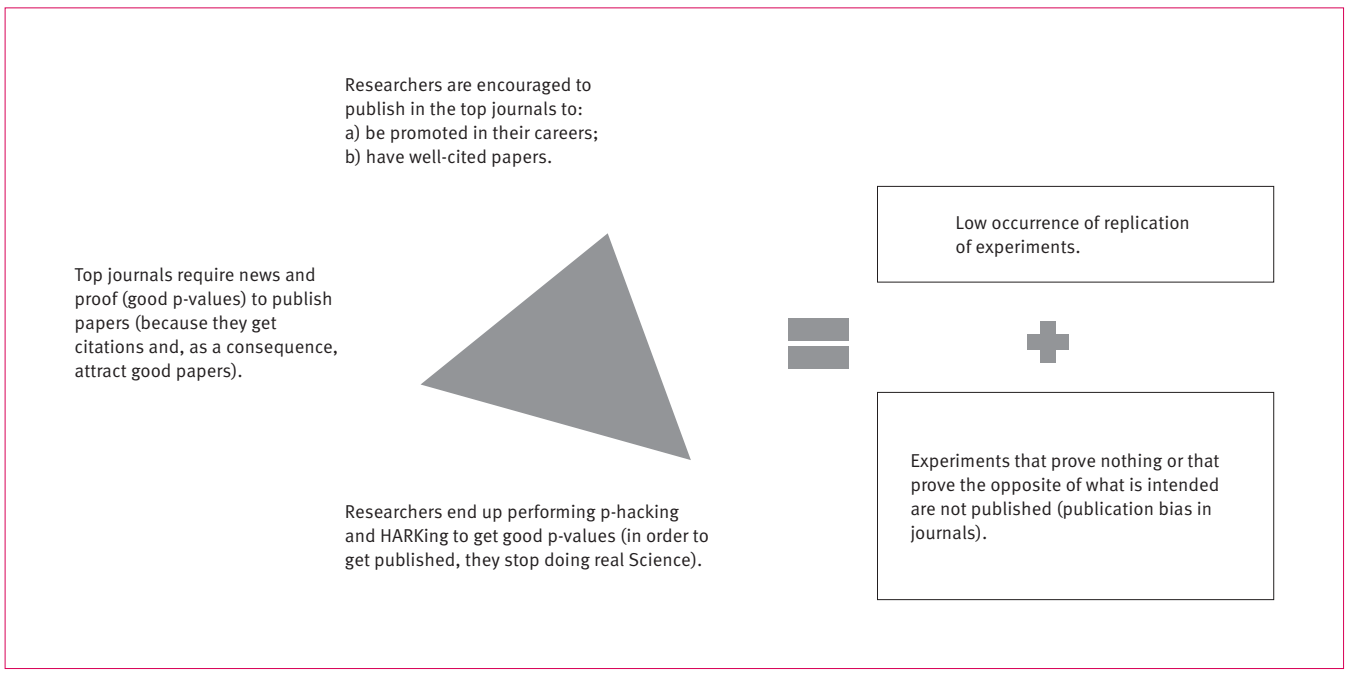
The tangibilization of this movement appeared in 2019, in a long editorial of The American Statistician (Wasserstein, Schirm, & Lazar, 2019), where there are some established consensuses

   ISSN 0034-7590; eISSN 2178-938X

on what not to do. For example, although maintaining the p-value as a continuous variable, the directive is to no longer use the expression "statistically significant" in any situation, thus avoiding inaccurate interpretation about its meaning. There is also a large collection of recommendations on what to do for authors, reviewers, and editors. However, as it is clear in the text, the scientific community is far from unanimous. A comment in Nature (Amrhein, Greenland, & McShane, 2019) also strongly recommends banning the expression "statistically significant," suggesting presenting p-values with adequate precision, without comparisons of type $p < 0.05$ and without tagging asterisks.

The issue of replicability in research urgently needs an incentive policy. Somehow, the triangle in Figure 1 needs to be broken. As a background to the current state of quantitative research, it is difficult to avoid the inappropriate use of the p-value as a final arbiter of truth. For Goodman (2019), the problem is neither scientific nor philosophical, but sociological, since the p-value is used due to its value in attesting knowledge, allowing publication, obtaining research funds, and obtaining academic promotions. In other words, there is a need for change in academic institutions, journals, agencies providing funding for research, and regulatory agencies. Colquhoun (2019, p. 192) advises, "In the end, the only way to solve the problem of reproducibility is to do more replication and to reduce the incentives that are imposed on scientists to produce unreliable work".

Figure 1. **The incentive cycle goes against Science**



Wasserstein et al. (2019) discuss the proper pace at which journals should implement their requirements for papers that use statistical inference. In fact, breaking the incentive cycle shown in Figure 1 is not simple. The authors of this essay suggest that Brazilian management journals begin to take small and important steps until the scientific community manages to reach clear and unanimous standards. The first step is to require that authors declare in their papers, clearly and formally, that they did not perform p-hacking or HARKing. The second step is to prohibit authors from classifying their findings as statistically significant based on the p-value. The third step is to require mandatory analysis of the effect size. The fourth step is to encourage studies that use pre-registration (as a way to avoid publication bias). The fifth step is to actively encourage the publication of replication studies (that is, do not favor only innovative research) and meta-analyses.

The guidelines for authors in all Brazilian journals published in the national territory, in the fields of public and business administration, accounting sciences, and tourism, with at least Qualis A2 classification (February 2019) were examined (Brito, Luca, & Teixeira, 2017). These are: Advances in Scientific and Applied Accounting, Brazilian Administration Reviews, Brazilian Business Review, *Cadernos EBAPE*, Estudios y Perspectivas en Turismo, Contabilidade Vista & Revista, Organizações & Sociedade, Review of Business Management (*Revista Brasileira*

de Gestão de Negócios [RBGN]), Brazilian Journal of Tourism Research (*Revista Brasileira de Pesquisa em Turismo*), Accounting & Finance Review (*Revista Contabilidade & Finanças*), *Revista Contemporânea de Contabilidade*, Journal of Contemporary Administration (*Revista de Administração Contemporânea*), RAUSP Management Journal (*Revista de Administração da USP*), Journal of Business Management (*Revista de Administração de Empresa*s [RAE]), Brazilian Journal of Public Administration (*Revista de Administração Pública* [RAP)], *Revista de Contabilidade e Organizações*, and Revista Universo Contábil. The guidelines for authors in these 17 journals were analyzed in August 2019 and compared to the five steps mentioned above. It was found that the steps recommended were not found explicitly in any of them. Nevertheless, in some of these journals, generic warnings about the fabrication and falsification of data and results were found (Byington & Felps, 2017). The authors of this essay strongly advocate that the authors of scientific papers should themselves declare that they specifically did not perform p-hacking or HARKing.

Here, we see an opportunity for innovation for these Brazilian journals in the field of public and business administration, accounting, and tourism. After all, "Novelty and positive results are vital for Publishability but not for Truth" (Nosek et al., 2012, p. 617).

REFERENCES

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

AUTHORS' CONTRIBUTIONS

The authors declare that they participated in all stages of development of the manuscript: conceptualization, theoretical-methodological approach, theoretical review, data collection, writing, and final revision of the Essay.