Revista [IN]Genios, Vol. 7, Núm. 2, pp.1-10 (mayo, 2021)
ISSN#: 2374-2747
Universidad de Puerto Rico, Río Piedras

OVERVIEW ON THE OPINION OF CHANGING THE STANDARD P-VALUE FROM 0.05 TO 0.005 THROUGHOUT THE ACADEMIC DISCIPLINES

Erick Miguel Ivanovich Méndez
Departamento de Matemáticas
Facultad de Ciencias Naturales, UPR RP

Gabriela M. Lozano Pérez
Departamento de Matemáticas
Facultad de Ciencias Naturales, UPR RP

Recibido: 10/3/2020; Revisado: 16/3/2021; 18/4/2021; Aceptado: 29/4/2021

Abstract
The article "Redefine Statistical Significance" by Benjamin et al. (2018) proposed to change the p-value threshold from 0.05 to 0.005 to reduce the increasing rate of false positives. Therefore, we studied the question of how people from different disciplines feel about the current p-value threshold. Furthermore, we gathered data with a programing language from various scientific papers that cited Benjamin et al. (2018) via article keywords, then categorized and organized these citations. Finally, we concluded that a need for a new method surpassed that of those who wanted to only change the p-value to 0.005.

Keywords: p-value thresholds, changing p-values, statistical significance, Bayesian statistics

Resumen
El artículo "Redefine Statistical Significance" de Benjamin et al. (2018) propuso cambiar el límite del p-valor de 0.05 a 0.005 para reducir la tasa de falsos positivos. Por lo tanto, estudiamos cómo se sienten las personas de diferentes disciplinas sobre el límite del p-valor. Además, recopilamos datos con un lenguaje de programación de varios artículos científicos que citaban a Benjamin et al. (2018) vía palabras clave, que luego se categorizaron y organizaron. En conclusión, la opinión sobre la necesidad de un nuevo método superó la de aquellos que solo querían cambiar el p-valor a 0.005.

Palabras clave: límites del p-valor, cambiando los p-valores, significancia estadística, Estadística Bayesiana

Introduction

In all scientific disciplines, statistics has had an important role in assuring that experiments and scientific studies can be regarded as trustworthy. Moreover, Benjamin et al. (2018) has warned against this practice of the adoption of p-values as a unique or single standard for publication. For many years, the p-value statistics was being misused as a definitive statistical significance amidst growing concerns regarding reproducibility issues, which Ioannidis (2019) pointed out as the case in most biomedical articles. Moreover, Greenland et al. (2016) agreed with the notion of a misinterpretation. In his article, he prepared a list of 25 points on how different disciplines have erred with the p-value and other statistics. However, in the article "Redefining Statistical Significance" by Benjamin et al. (2018) argued that the lack of reproducibility in scientific studies in general was due to the current p-value threshold. Previously, the American Statistician Association published a statement regarding the definition of the p-value: "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value" (Wasserstein & Lazar, 2016, p.131). However, Wasserstein & Lazar (2016) pointed out that this statement was made with the purpose of satisfying a definition in which both Bayesian statisticians and frequentist statisticians could agree upon, not to tackle the underlying problem of misinterpretation brought by Ioannidis (2018) or underreporting and misinterpreting conclusion derived from a test hypothesis pointed out by Lakens et al. (2018).

In the light of the troubles surrounding p-values, many proposals have been made for the sole purpose of mitigating reproducibility. In the case of Benjamin et al. (2018) they proposed a solution deemed controversial that "for fields where the threshold for defining statistical significance for new discoveries is p-value < 0.05, we propose a change to p-value < 0.005" (p. 6). Additionally, Benjamin et al. (2018) argued that if Bayes factors were paired with p-values on individual test hypotheses, we would see an improvement by increasing categorical Bayes factor numbers from weak to substantial or strong. However, within different scientific disciplines there are conflicted views on the matter. For example, Amrhein and Greenland (2017) argued that the redefinition of a new arbitrary p-value was not viable and instead we

should be focusing on previous evidence from multiples studies to take decisions instead of relying on a new threshold. On the other hand, although some scientists, like Colquhoun (2017), agreed with the idea of mixing p-values with Bayesian methods for substantial evidence, some others like McShane et al. (2019) flat-out suggested abolishing statistical significance to an arbitrarily set threshold of p-values.

Although many disciplines have argued for many ways to solve the problem regarding reproducibility, there is no consensus on how to move forward. However, if you look at the number of times Benjamin et al. (2018) has been cited since his publication (more than 1400 by the time of writing), we know that the community has something to say about the issue, but no one has taken the time to recollect or study what their opinion is. Therefore, we want to question the overall opinion of the different scientific disciplines regarding what we should do with the p-value. Furthermore, we expect that the overall consensus disagrees with changing the threshold of the p-value from 0.05 to 0.005, and instead find broader support for new statistical methods in combination with p-values. As for new statistical methods, in general, we refer to different techniques already suggested by Ioannidis (2018), such as: abandoning p-values entirely for determining statistical significance; using alternative inference methods such as, Bayesian statistics; focusing on effect sizes and their uncertainty; training the scientific workforce, and addressing biases that led to inflated results. On the other hand, Lakens et al. (2018) in his article "Improving Inferences About Null Effects with Bayes Factors and Equivalence Tests" further supported the use of both frequentists and Bayesian statistics because it improved on statistical significance and decision making. In general, we are arguing that if we show that the overall consensus of the disciplines is willing to change the p-value, we will be adding concrete evidence that can help end the stalemate on how to approach this issue.

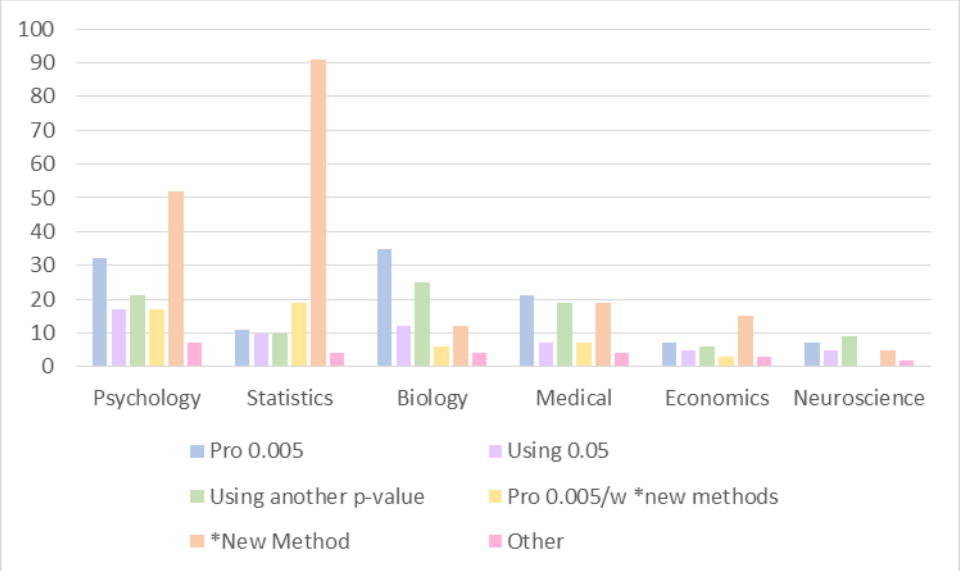
Methodology

We used the program Publish or Perish to obtain the articles that cited Benjamin et al. (2018). In doing so, we gathered around 904 results. Afterwards, we cleaned up the data, including the ones that were repeated or collected incorrectly. In the end, we ended up with 878 articles. Then, we used Python 3 to organize the articles based on keywords and the journals they were published. Some articles were added to more than one category. Furthermore, we noted that not every article had an identifying name nor journal; therefore, we organized those by hand. Then again, we looked at how in the articles was augmented the p-value < 0.005 threshold. Because of the

lack of information, we categorized the opinions by reading some of the articles. Moreover, another problem we encountered was the inaccessibility of some scientific articles due to paywalls which could affect our results.

Discussion

The main takeaway from the collected data was the number of papers that have cited "Redefining Statistical Significance;" around 900 articles have cited the work made by Benjamin et al. (2018). Importantly, publications made after the methodology was completed were not reported in this paper. Then, our final yield of articles collected resulted in 664 citations in total. First, we found that the scientific disciplines of Psychology, Statistics, Biology, and Medicine made up 65% of all the articles that referenced Benjamin et al. (2018). Additionally, we noted that the discipline of Psychology contributed to most of the articles invested in this issue while Statistics, the epicenter of this topic in question, was the second scientific discipline that issued the highest numbers regarding citations of the paper.

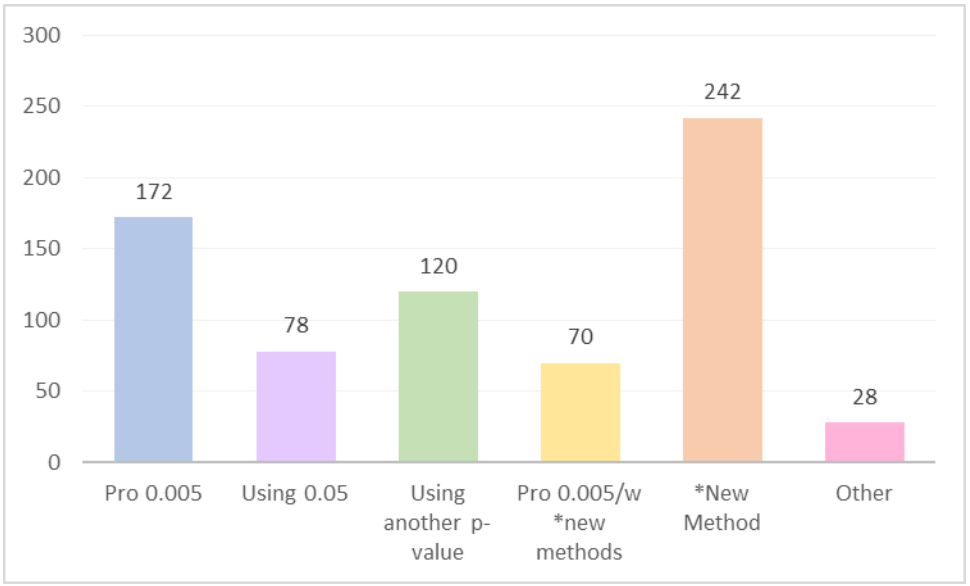


GRAPH 1: DISTRIBUTION OF OPINION WITHIN THE TOP 6 SCIENTIFIC DISCIPLINES

In Graph 1, we depicted the distribution of the top 6 fields that cited Benjamin et al. (2018). First, the discipline of Psychology was the top contributor with 146 citations. Second, the field of Statistics with 145. Next, the other 4 included Biology with 94 citations, Medicine related articles with 77, Economics with 39, and Neuroscience

with 28. Additionally, we divided the data into 5 different categories: the agreement of changing the p-value to 0.005, agreement of changing the p-value to 0.005 but only if it has more information, advocating for the use of another p-value, such as 0.005 or 0.001, and advocating for the need of a new method and the final category for articles with no set opinion on the matter.

For the top two fields, Psychology and Statistics, the opinion of using a new method was higher than the rest. Furthermore, we realized that between these two categories, even though they both suggested new methods with a higher frequency, there was a bigger acceptance of changing the p-value to 0.005 in Psychology than there was in Statistics. Moreover, we noticed that they agreed in that there is a need for a new and more effective statistical method. On the other hand, with Biology, Medicine and Neuroscience, we saw that the proposed change in p-value had a decrease when compared to Psychology and Statistics. Interestingly, this sudden decrease was traced back to reproducibility. As argued by Ioannidis (2018), when a p-value gets lowered, the samples' size must increase to compensate and fall within the margin of confidence, which means the cost for experiments increase. Thus, we reasoned that the disinterest in disciplines like Biology, Medicine and Neuroscience come as a direct result of an economic issue.



GRAPH 2: OVERVIEW OF THE OPINION OF THE SCIENTIFIC COMMUNITY REGARDING P-VALUES

In Graph 2, we depicted the general overview of the whole scientific community that made up the data set. In this, the number of articles arguing in favor of a new method

are 242 or 34.4%; modifying the p-value to 0.005 and having a new method is 70 or 10.5%; those advocating for a change of p-value to 0.005 has 172 or 25.90%, and those who want the p-value at 0.05 or maintaining the status-quo has 77 or 11.6%. Furthermore, when we add the total of articles that supports a new method and/or changing the p-value to 0.005 (242, 70, and 172) has 484 or 72.89% approval. On the other hand, if we add those who support a new method, we see that it is 312 or 46.99%. Therefore, we argue that the overall approval of needing a change in statistical significance for a new method outpaces that of simply changing the p-value to 0.005, even though there is no consensus on what that change should be. Notably, the articles denoted as Other are those that do not explain their stance clearly or do not have a stance in any p-value. Nevertheless, that number, even if it coalesced behind maintaining a p-value of 0.05 (106 or 15.96%), would not change the fact that most of the support is in favor of a change.

Conclusion

The biggest takeaway and importance of this investigation is that there is an overall acceptance of the need for a new method and/or changing the p-value. Initially, we did expect that the overall consensus disagrees with changing the threshold of the p-value from 0.05 to 0.005, and instead advocate for a new method. Indeed, the data points to the vast majority of articles agreeing with the idea of applying new methods. However, we did not expect that change to the p-value to a new threshold of 0.005 would be as overwhelming. In our findings, the need for new methods overpowers the rest of the opinions by a large margin, since most of them either wants to change the p-value to 0.005 with a new method or simply have a new method.

The acceptance of 0.005 and other p-values can be predominantly seen in Science related fields. Compared to the fields of Statistics, Psychology, and Economics, the sum of the articles that suggests keeping use of p-values is greater than those who would rather abolish it. Consequently, we suggested that this could be because these disciplines have no interest in developing new techniques but rather applying them. Moreover, using new methods may imply learning complex ideas that those disciplines would rather avoid for convenience as Ioannidis (2018) details. Next, we need to address that possible issues and errors done throughout this investigation could be in the organization and identification of articles step and the lack of accessibility of articles. Nonetheless, we can improve this work by applying more human and computer resources to achieve a more accurate result. In conclusion, we

can continue developing the ideas of this paper by next asking who amongst the different disciplines are using different p-values or new methods and who remain in a 0.05 threshold.

References

Amrhein, V., & Greenland, S. (2017). Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, *2*(1), 4. https://doi.org/10.1038/s41562-017-0224-0

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, *4*(12), 171085. https://doi.org/10.1098/rsos.171085

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

Ioannidis, J. P. A. (2018). The proposal to lower P value thresholds to .005. *JAMA*, *319*(14), 1429–1430. https://doi.org/10.1001/jama.2018.1536

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with P values? *The American Statistician*, *73*(sup1), 20–25. https://doi.org/10.1080/00031305.2018.1447512

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence Tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E.

M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., & Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

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

Acknowledgement

We want to thank our mentor and guide doctor Luis Perrichi Guerra for tasking us on this relevant yet controversial topic; his support was of inspiration to better our work after presenting it on the SIDIM at UPR-Cayey in 2019 and finally publishing it.

Appendix A

Python Assorted Results

	Pro 0.005	Using 0.05	Using another p-value	Pro 0.005/w new methods	*New Method	Other	Total
Psychology	32	17	21	17	52	7	146
Statistics	11	10	10	19	91	4	145
Biology	35	12	25	6	12	4	94
Medical	21	7	19	7	19	4	77
Economics	7	5	6	3	15	3	39
Neuroscience	7	5	9	0	5	2	28
Computer Science	14	2	4	0	6	1	27
Science	5	2	3	3	9	2	24
Epidemiology	6	2	6	2	5	0	21
Others	6	4	5	3	2	0	20
Education	3	2	2	4	5	0	16
Physics	3	3	6	0	3	0	15
Social Science	5	2	0	0	6	1	14
Linguistics	6	1	2	1	2	0	12

Philosophy	2	3	1	2	3	0	11
Politics	2	0	0	3	3	0	8
Environmental Science	3	1	1	0	2	0	7
Ecology	4	0	0	0	2	0	6
Total	172	78	120	70	242	28	710

NOTE 1. *NEW METHOD STANDS FOR SUGGESTING A NEW METHOD OF STATISTICAL SIGNIFICANCE AND OTHER INCLUDES ARTICLES THAT DO NOT SHOW ANY PREFERENCE. IN HIS PAPER, "THE PROPOSAL TO LOWER P VALUE THRESHOLDS TO .005" IOANNIDIS (2018) SUGGESTS NEW METHODS FOUND IN THE TABLE TITLED "VARIOUS PROPOSED SOLUTIONS FOR IMPROVING STATISTICAL INFERENCE ON A LARGE SCALE" (IOANNIDIS, 2018, E2) SUCH AS ABANDONING P-VALUES ENTIRELY, USING ALTERNATIVE INFERENCE METHODS SUCH AS BAYESIAN STATISTICS, FOCUSING ON EFFECT SIZES AND THEIR UNCERTAINTY, TRAINING THE SCIENTIFIC WORKFORCE, AND ADDRESSING BIASES THAT LED TO INFLATED RESULTS.

SOURCE: TABLE WITH THE DATA GATHERED FROM ARTICLES THAT CITED REDEFINE STATISTICAL SIGNIFICANCE, THE TOTAL IS 710 AND NOT 664 BECAUSE SOME ARTICLES WERE CONSIDERED FOR MORE THAN ONE TYPE.