

# Should we retire statistical significance?

## *Devemos aposentar a significância estatística?*

DOI 10.5935/2595-0118.20200199

Dear Editor,

The article entitled “*Retire Statistical Significance*” from the *Nature*<sup>1</sup> journal had great repercussion, criticizing statistical dogmatism and promoting an analysis on both sides of the same coin. On one hand, the value of the reflection brought by the authors. On the other hand, the unintended consequences of the retirement of the concept of statistical significance. The first point of view relates to the bias of value overestimation; the second point relates to the bias of positivism.

The concept of statistical significance is dichotomous, that is, it categorizes the analysis as “positive” or “negative”. Categorizing adds pragmatic utility, but all categorization is an arbitrary reductionism. When categorizing for pragmatic reasons, we should understand categories as something less important than the view of the whole. The **categorization paradox** occurs when we start to give more value to the information because it’s categorical rather than continuous<sup>2-4</sup>. Continuous information accepts the shades of gray, the intermediate, the doubt, while the categorical gives a definite (pseudo) tone to the statement.

Statistics is the exercise of recognizing uncertainty, doubt, chance. The definition of statistical significance was originally created to hinder claims arising from chance. The confidence interval was created to describe the inaccuracy of our statements. Statistics is the scientist’s exercise of integrity and humility.

However, the categorization paradox fosters a certain dogmatism. First, the authors of the above-cited *Nature* article highlight the overvaluation of negative results. A negative study is not the one that proves the nonexistence, which would be impossible; it is merely a study that has not proven the existence. Therefore, strictly speaking, “absence of evidence is not evidence of absence”, as Carl Sagan stated. That is, “the study proved that there is no difference” is not the best way to describe it, and it’s preferable to say, “the study did not prove a difference”.

Such a point should not be mistaken for the idea that a negative study means nothing. It has value and impact. The impact of a negative study ( $p > 0.05$ ) lies in reducing the likelihood of the phenomenon to exist. As good studies fail to prove it, the probability of the phenomenon drops progressively to the point where it becomes so low that nullifies the attempt to continue to try, making the null hypothesis the most likely path of thought.

A negative study is not necessarily contradictory in relation to a positive study. Perhaps, the result of the two may be the same when one failed to reject the null hypothesis, and the other was able to reject it. One could not see, and the other could see. In fact, most of the time, only one of the two studies are correct.

Finally, the categorization paradox makes us believe in any statistical significance, although most are false positives (Ioannidis). A  $p$ -value  $< 0.05$  is not irrefutable proof. Undersized studies, the multiplicity of secondary analyses and biases can fabricate false statistical significance.

In fact, the predictive value (negative or positive) of studies does not lie solely in statistical significance, as it depends on the quality of the study and analysis, the scientific ecosystem, and the pretest probability of the idea.

Therefore, the authors of the *Nature* article are right in criticizing the deterministic view of statistical significance.

### **But the question looms: should we really retire statistical significance?**

Such an achievement would mean retiring an advent that has historically been responsible for a major evolution in scientific integrity. However, all that is good tends to be “borrowed”. Artists of false positivation of studies “borrowed” the advent of the  $p$ -value (created to difficult the type I error) to prove false things.

While on one hand, the retirement of statistical significance would avoid the paradox of categorization, on the other hand, it would make room for the positivity bias, our tropism for creating or absorbing positive information. x

The criticism of statistical significance in this and other prominent articles<sup>5-7</sup> does not provide a better alternative. For example, the author of the *Nature* article acknowledges that other, more recent statistical approaches (Bayesian, for example) do not abandon the categorization paradigm. In certain passages the authors even mention that they do not propose a total abandonment of the notion of statistical significance. Perhaps the title that reflects the real content of the article should be rephrased with a question mark: “*Should We Retire Statistical Significance?*”? **Diferente do texto em português.**

Currently, much more is discussed about scientific integrity than two decades ago. However, in approaching this subject with more emphasis than in the past, it may appear that this problem is worse nowadays. That’s not the case. We experience a clear evolu-

tion in scientific integrity: concepts of multiplicity are much more discussed today than in the past, clinical trials have their designs published *a priori*, journals require CONSORT publication standards and there is more discussion about scientific transparency, open science, slow science. We are evolving. And the first step towards integrity was the creation of the notion of statistical significance in the first half of the last century by Ronald Fisher<sup>8</sup>.

A study published in PLoS One (Bob Kaplan)<sup>9</sup> analyzed, over a long period of years, the results of clinical trials funded by the National Institutes of Health (NIH). Before the year 2000, when there was no need to previously publish the protocol, the frequency of positive studies was 57%, dropping to only 7% of positive studies after the *a priori* publication rule. Previously, the authors used to positive their studies by multiple *a posteriori* analyzes. Today, this has become less frequent due to the obligation to publish *a priori*.

The impression is that it has become elegant to criticize the p-value, which seems to be a betrayal of an advent of great historical importance and which, until now, has not found a better substitute. We can't blame P for being "borrowed" by malicious researchers. Researchers are to blame.

Therefore, we propose to maintain the p-value and adopt the following measures:

- Describe the p-value only when the study has a suitable size for the hypothesis test. Otherwise, it would gain a more descriptive character, without using associations for testing of concepts. This would avoid false positives stemming from "small studies", which is the case of most published articles. For example, the median statistical power of biomedicine studies is 20%;
- Don't describe the p-value in secondary outcome analyzes;
- In (exploratory) subgroup analyzes, use only the interaction P (more conservative and difficult to give meaning), avoiding the p-value obtained by the comparison within a subgroup (small studies);
- Include in CONSORT the obligation for the authors to make explicit in the sub-studies title that it's an exploratory and secondary analysis of a previously published study;
- Abandon the term "statistical significance", replacing it with "**statistical accuracy**". Statistics is used to differentiate true causal associations from chance-mediated pseudo-causalities. Therefore, a p-value of <0.05 connotes accuracy. Whether the association is significant (relevant), it depends on the description of the numerical difference or the association measures of categorical outcomes. Thus, using "statistical accuracy" avoids the confusion between statistical significance and clinical significance. Finally, we propose the advent of the "**researcher integrity index**".


This index will be calculated by the ratio between the number of negative studies/number of positive studies. An integrity index <1 indicates a researcher of questionable integrity. This index is based on the assumption that the probability of a good hypothesis being true is less than 50%. Therefore, there should be more negative than positive studies. This does not occur due to the positivation techniques of studies (small papers, multiplicities, biases, spin of conclusions) and the publication bias that hides negative studies. A righteous author would be the one who does not use these practices and therefore would have several negative and few positive studies, resulting in an integrity index well above 1.

The *Nature* article is useful for promoting a reflection on the pros and cons of statistical significance. Nevertheless, it does not propose its retirement. Such a thing would be analogous to retiring a still very productive person. Conversely, statistical significance should remain active and progressively evolving in its use.

Let us also learn to value a  $p > 0.05$ . After all, the unpredictability of life is represented by this symbology, much of people's fate is mediated by chance.


Or nothing happens by chance?

*Luis Claudio Lemos Correia*


 <https://orcid.org/0000-0002-6910-1366>

Email: [luisclcorreia@gmail.com](mailto:luisclcorreia@gmail.com)

*Gabriela Oliveira Bagano*

 <https://orcid.org/0000-0002-6541-0372>

*Milton Henrique Vitória de Melo*

 <https://orcid.org/0000-0002-5130-1634>

Escola Bahiana de Medicina e Saúde Pública,  
Hospital Aliança, Salvador, BA, Brasil.

## REFERENCES

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-7.
2. Gigerenzer G. Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*. *Adv Meth Pract Psychol Sci*. 2018;1(2):198-218.
3. Greenland S. Invited Commentary: the need for cognitive science in methodology. *Am J Epidemiol*. 2017;186(6):639-45.
4. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. *Am Stat*. 2019;73(Suppl 1):235-45.
5. Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. *Int J Cardiol*. 2014;177(3):1089-90 (2014).
6. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p<0.05". *Am Stat*. 2019;73(Suppl)1-19.
7. Hurlbert SH, Levine RA, Utts J. Coup de Grâce for a tough old bull. "statistically significant" expires. *Am Stat*. 2019;73(Suppl):352-7. <https://doi.org/10.1080/00031305.2018.1543616>.
8. Fisher RA. The fiducial argument in statistical inference. *Ann Eug*. 1935;6:391-8.
9. Kaplan RM, Irvin VL. Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*. 2015;10(8):e0132382. <https://doi.org/10.1371/journal.pone.0132382>.

