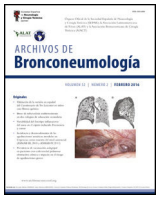
Arch Bronconeumol. 2020;56(10):613–614

Sociedad Española de Neumología y Cirugía Torácica SEPAR

ARCHIVOS DE Bronconeumología

www.archbronconeumol.org



Editorial

P of significance: Is it better to avoid it if it is poorly understood?☆

p de significación: ¿mejor no usarla si se interpreta mal?

This editorial follows on from a previously published editorial which explained the role of inferential statistics in the scientific method.[1] The aim of this second editorial is to highlight the most common errors in the interpretation of the p-value and statistical significance, in line with recent articles and comments in impact journals such as *Nature* that echo initiatives such as that of more than 800 prestigious scientists who call for an end to the use of significance thresholds and the dichotomous notion of statistical significance.[2–5]

To understand the above, we must remember that the aim of so-called inferential statistics is to evaluate the role of chance in our results. This can be quantified or estimated by obtaining the standard error, calculating the probability that the results can be explained by chance under the null hypothesis or H0, giving us a p-value in statistical significance tests. This approach, known as null hypothesis significance testing (NHST), was invented in the 1920s and 1930s by Ronald Aylmer Fisher (recognized as the father of inferential statistics), in order to determine which fertilizer increased maize production to the greatest extent. NHST involves a dichotomous approach, as follows: if the p-value is less than a statistical significance threshold (0.05 based on the consensus of an alpha risk of 5%), the null hypothesis is rejected and the alternative hypothesis is therefore accepted.

This has resulted in a reductionist interpretation, in which if $p<0.05$, a result is considered significant (e.g. a 120 ml difference in FEV1 between groups in favor of a new inhaled therapy molecule versus another standard treatment) and "there are differences between the two treatments", whereas if the same treatment with the same 120 ml difference has a p of, for instance, 0.06, it is considered non-significant.

The main objective of this editorial is to make clear that non-statistically significant differences are not synonymous with equivalence. The fact that a result is not statistically significant does not necessarily imply that the interventions are equivalent. However, the authors of a published study were alarmed to find that in more than 50% of articles, when p is non-significant, it is erroneously concluded that "there are no differences between the 2 treatments" or, worse still, both drugs or interventions are considered to be "equal or equivalent".[2,6–9]

This editorial does not aim to provide a comprehensive explanation of statistics, but we should remember that when we accept the null hypothesis (Ho), a beta error emerges, which is the probability of not having found differences when they actually exist, that is, the probability of not rejecting the null hypothesis when it is false. The complementary aspect is statistical power (1 — beta error), which is the probability of finding statistically significant differences if they really do exist.

There is an example in English where a researcher is compared to Michael Jordan (the basketball player)[10] and another, adapted to Spanish, where the ability of a researcher and Leo Messi (the soccer player) to shoot penalties is compared.[11]

In the latter example, both shoot 8 penalties from the same positions with a defensive wall of 5 players. Messi scores 8 goals, all in the back of the net, and the researcher scores 4 and misses another 4. Later, at home that night, the researcher enters the data in the computer to check whether statistically there is much difference between their scores and those of Messi, and calculates the p-value using Fisher's exact test (2-tailed). The p-value is 0.077. In other words, the difference is not statistically significant.

If the researcher goes to bed, happy in the knowledge that there are no differences between their penalty shootout results and Messi's, he is being easily fooled, because in reality it is clear that there are differences between the two. Therefore, if we accept the null hypothesis we will be falling into the beta-type error, which in this case is high because the power of the study to detect differences is low due to the low sample size (number of penalties shot).

Let's not forget that the standard error can be used in both the p-value approach to significance and also in the construction of 95% confidence intervals (95% CI). The latter also support the rejection of the null hypothesis, but the width of the intervals, whether narrow or wide, reports the so-called "effect size", and as such the precision of the study.

Logically, in the case of the example of Messi, the 95% CI of the difference in percentage of goals will be very wide, that is, very imprecise. If we increase the number of penalty shots to, for example, 80, we would see how the standard error decreases because the sample size increases and the same difference in the percentage of goals (100% for Messi and 50% for the researcher) becomes statistically significant ($p<0.001$), with a much more precise 95% CI.

Finally, the International Conference on Harmonization (ICH) defines an equivalence trial as a clinical trial in which the main objective is to show that the response to the 2 treatments differs

☆ Please cite this article as: Santibáñez M, García-Rivero JL, Barreiro E. p de significación: ¿mejor no usarla si se interpreta mal? Arch Bronconeumol. 2020;56:613–614.

by an amount that is not clinically important.[12] Thus, in order to truly compare a hypothesis of equivalence between Messi and the researcher, you would need to: (a) have set non-inferiority and non-superiority limits (which would establish the percentage differences in goals scored that would be considered as equivalent); b) have determined the 95% CI of the percentage difference instead of the p-value of significance, and (c) have verified that the 95% CI was within these limits.

References

1. Santibáñez M, García-Rivero JL, Barreiro E. Don't put the cart before the horse (if you want to publish in a journal with impact factor). Arch Bronconeumol. 2020;56(2):70–1, http://dx.doi.org/10.1016/j.arbr.2019.05.019.
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–7.
3. Hurlbert SH, Levine RA. Utts. Coup de Grâce for a Tough Old Bull: "Statistically Significant" expires. Am Stat. 2019;73(S1):352–7.
4. McShanea BB, Galb D, Gelmanc A, Robertd C, Tackette JL. Abandon statistical significance. Am Stat. 2019;73(S1):235–45.
5. Wasswerstein RL, Schirm AL, Lazar NA. Moving to a world beyond p < 0.05. Am Stat. 2019;73(S1):1–19.
6. Schatz P, Jay KA, McComb J, McLaughlin JR. Misuse of statistical tests in Archives of Clinical Neuropsychology publications. Arch Clin Neuropsychol. 2005;20:1053–9.
7. Fidler F, Burgman MA, Cumming G, Buttrose R, Thomason N. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conserv Biol. 2006;20:1539–44.
8. Hoekstra R, Finch S, Kiers HA, Johnson A. Probability as certainty: dichotomous thinking and the misuse of p values. Psychon Bull Rev. 2006;13:1033–7.
9. Bernardi F, Chakhaia L, Leopold L. Sing me a song with social significance: the (Mis)Use of Statistical Significance Testing in European Sociological Research. Eur Sociol Rev. 2017;33:1–15.
10. Vickers AJ. Michael Jordan won't accept the null hypothesis: notes on interpreting high P values. Mescape. 2006;7:3.
11. Pascual-Huerta J. Yo no tiro las faltas como Leo Messi, porque no rechazar la hipótesis nula no es aceptarla. Rev Esp Podol. 2017;28:119–20.
12. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Confer- ence on Harmonisation E9 Expert Working Group. Stat Med. 1999;18:1905–42.

Miguel Santibáñez,[a,*] Juan Luis García-Rivero,[b] Esther Barreiro[c,d,e]

[a] *Grupo de Investigación de Salud Global, Universidad de Cantabria, Instituto de Investigación Marqués de Valdecilla (IDIVAL), Santander, Cantabria, Spain*

[b] *Servicio de Neumología, Hospital de Laredo, Cantabria, Spain*

[c] *Servicio de Neumología-Debilidad muscular y caquexia en las enfermedades respiratorias crónicas y el cáncer de pulmón, IMIM-Hospital del Mar, Barcelona, Spain*

[d] *Departament de Ciències Experimentals i de la Salut (CEXS), Universitat Pompeu Fabra (UPF), Parc de Recerca Biomèdica de Barcelona (PRBB), Barcelona, Spain*

[e] *Centro de Investigación en Red de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III (ISCIII), Barcelona, Spain*

* Corresponding author.
E-mail address: santibanezm@unican.es (M. Santibáñez).