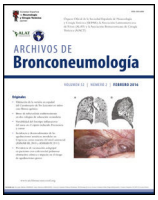
Arch Bronconeumol. 2020;56(10):613–614

Sociedad Española de Neumología y Cirugía Torácica SEPAR

ARCHIVOS DE Bronconeumología

www.archbronconeumol.org



Editorial

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

P of Significance: Is It Better to Avoid It if It Is Poorly Understood?

Este editorial es la continuación de un editorial publicado anteriormente, en el que se explicó el papel de la estadística inferencial en el ciclo del método científico[1]. En este segundo editorial se pretenden mostrar los errores más frecuentes en la interpretación de la p de significación, al hilo de los últimos artículos y comentarios en revistas de impacto como *Nature*, que se hacen eco de iniciativas como las de las más de 800 firmas de prestigiosos científicos recogidas para que se abandone el uso de umbrales de significación y el concepto dicotómico de significación estadística[2-5].

Para entender lo anterior, se debe considerar que el objetivo de la llamada estadística inferencial es evaluar del papel del azar en nuestros resultados. El papel del azar en nuestros resultados se puede cuantificar o estimar mediante la obtención del error estándar, calculando la probabilidad de que el azar explique los resultados bajo la hipótesis nula o H0, dándonos un valor p en las pruebas de significación estadística. Este abordaje en inglés se conoce como *null hypothesis significance testing* (NHST) y fue inventado por Ronald Aylmer Fisher en las décadas de 1920 y 1930 (y reconocido por ello como el padre de la estadística inferencial), para poder determinar qué fertilizante era el que en mayor medida incrementaba la producción de maíz. Es un abordaje dicotómico en el que si el valor p es menor de un umbral de significación estadística (0,05 con base en el consenso de un riesgo alfa del 5%), se rechaza la hipótesis nula y se acepta por consiguiente la hipótesis alternativa.

Esto ha derivado en una interpretación reduccionista, en la que si la p es < 0,05 se considera un resultado como significativo (por ejemplo: una diferencia entre grupos en el FEV1 de 120 ml a favor de una nueva molécula en terapia inhalada frente a otro tratamiento convencional) y se concluye que «existen diferencias entre ambos tratamientos», mientras que si el mismo tratamiento con la misma diferencia de 120 ml tiene una p de por ejemplo 0,06 se considera como no significativa.

Debe quedar claro, pues es el objetivo principal de este editorial, que si las diferencias no son estadísticamente significativas esto no es sinónimo de equivalencia. Que un resultado sea estadísticamente no significativo no implica necesariamente que las intervenciones sean equivalentes. Sin embargo, de forma alarmante según un estudio publicado, en más del 50% de los artículos, cuando la p es no significativa se concluye erróneamente que «no existen diferencias entre ambos tratamientos» o, lo que es aún peor, se considera que ambos fármacos o intervenciones «son iguales o equivalentes»[2,6-9].

Aunque este editorial no pretende abordar la estadística de una forma exhaustiva, conviene recordar que si aceptamos la hipótesis nula (Ho) tenemos un error beta, que es la probabilidad de no haber encontrado diferencias cuando en realidad las hay, es decir, es la probabilidad de no rechazar la hipótesis nula cuando esta es falsa. Su complementario es la potencia estadística (1 — error beta), que es la probabilidad de encontrar diferencias como estadísticamente significativas si de verdad estas existen.

Existe un ejemplo en inglés donde se compara a un investigador con Michael Jordan (el jugador de baloncesto)[10] y otro, adaptado al español, donde se compara la capacidad de tirar faltas entre un investigador y Leo Messi (el jugador de fútbol)[11].

En este último ejemplo, ambos tiran 8 faltas desde las mismas posiciones con una barrera estática de 5 jugadores. Messi anota 8 goles (todos dentro) y el investigador anota 4 y falla otros 4. Cuando llega a casa por la noche, el investigador introduce los datos en su ordenador para comprobar si estadísticamente hay mucha diferencia entre él y Messi tirando faltas, y calcula el valor p mediante la prueba del test exacto de Fisher (bilateral), obteniendo una p = 0,077. Es decir, no existen diferencias estadísticamente significativas.

Si el investigador se fuera a dormir contento sabiendo que no hay diferencias tirando faltas entre Messi y él, sería un incauto o inconsciente porque está claro que la realidad es que sí que hay diferencias entre ambos. Por consiguiente, si aceptamos la hipótesis nula estaremos cayendo en el error tipo beta, que en este caso será alto porque la potencia del estudio para detectar diferencias será baja porque, a su vez, el tamaño muestral (número de faltas tiradas) es bajo.

Recordemos que, además de utilizar el error estándar en el abordaje de la p de significación, con el error estándar se pueden construir también los intervalos de confianza al 95% (IC95%), que permiten asimismo rechazar la hipótesis nula con la ventaja de que su amplitud o estrechez informa del llamado «tamaño del efecto» o *effect size* en inglés y, por tanto, de la precisión del estudio.

Lógicamente, en el caso del ejemplo de Messi, el IC95% de la diferencia de porcentaje de goles será muy ancho, esto es, muy poco preciso. Si aumentáramos el número de faltas tiradas por ejemplo a 80 faltas, comprobaríamos cómo el error estándar disminuye porque aumenta el tamaño muestral y la misma diferencia en el porcentaje de goles (100% en Messi y 50% en el investigador) se vuelve estadísticamente significativa (p < 0,001), con un IC95% mucho más preciso.

Por último, la International Conference on Harmonisation (ICH) define el ensayo de equivalencia, *equivalence trial*, como un ensayo

https://doi.org/10.1016/j.arbres.2019.11.003

clínico que tiene como objetivo principal demostrar que la respuesta a los 2 tratamientos difiere en una cantidad que no es clínicamente importante[12]. Así pues, para contrastar realmente una hipótesis de equivalencia entre Messi y el investigador, habría que: *a)* haber puesto un límite de no inferioridad y de no superioridad (que demarcaría las diferencias en porcentaje de goles metidos que consideramos como equivalentes); *b)* haber hallado el IC95% de la diferencia de porcentajes en lugar del valor p de significación, y *c)* haber comprobado si el IC95% se encuentra dentro de estos límites.

Bibliografía

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. 2019;piiS0300-2896, http://dx.doi.org/10.1016/j.arbres.2019.05.019, 30279-0.
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. The American Statistician. 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 Conference on Harmonisation E9 Expert Working Group. Stat Med. 1999;18:1905-1942.

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

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

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

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

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

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

[*] Autor para correspondencia.
Correo electrónico: santibanezm@unican.es (M. Santibáñez).