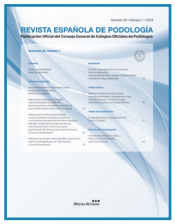

REVISTA
ESPAÑOLA DE PODOLOGÍA

Publicación Oficial del Consejo General de Colegios Oficiales de Podólogos

RINCÓN DEL INVESTIGADOR
Artículo en español

Rev Esp Podol. 2018;29(1):60-62
DOI: 10.20986/revesppod.2018.1517/2018

Intervalos de confianza *vs.* Valor *p*

Porque el gris es importante...

Confidence intervals vs. P value. When grey matters...

Javier Pascual-Huerta

Clínica del Pie Elcano. Bilbao

El progreso científico en las ciencias biomédicas está anclado de forma sólida en el concepto del valor *p* y del contraste de hipótesis. A pesar de todas las críticas que este modelo ha recibido[1-3], la "cultura del valor *p*" ha sido el vehículo por el que el conocimiento científico ha progresado durante décadas y todos los intentos que se han realizado para detenerlo han sido completamente inútiles. De hecho, según el número de publicaciones ha ido creciendo en los últimos años así también ha ido creciendo esta cultura del valor *p* hasta el punto de que la Sociedad Americana de Estadística se ha visto obligada a publicar recientemente una declaración sobre lo que es y lo que no es el valor *p*. Esta declaración señala las interpretaciones erróneas y los abusos de uso que se siguen produciendo con esta cultura en la literatura científica actual[4].

Unido a esto, cada vez son más las voces que critican el formato actual en la investigación biomédica de categorizar las variables en estadísticamente significativas o no estadísticamente significativas. Este es, sin duda, un abordaje demasiado simplista y una estrategia de análisis potencialmente dañina y contraproductiva para la interpretación válida de los datos de las investigaciones que debería abandonarse[5,6]. La pregunta es si existe una alternativa al valor *p* y, aunque no parece existir una solución definitiva, los intervalos de confianza parecen ser una buena solución a este abuso al que está sometida la investigación científica actual[3,5].

Un ejemplo. Supongamos que Fernández y cols. realizan un estudio sobre la eficacia del tratamiento mediante infiltraciones de 2 compuestos diferentes para la fasciopatía plantar. Para ello, seleccionan una muestra de casos con fasciopatía plantar (suficientemente grande como para mostrar una potencia adecuada del estudio) que dividen de forma aleatoria en 2 grupos: Grupo A (n = 34), que recibe una infiltración ecoguiada de la sustancia A, y Grupo B (n = 36), que recibe una infiltración también ecoguiada de la sustancia B. El grupo investigador de Fernández y cols. mide la eficacia de cada sustancia mediante una Escala Analógica Visual (EAV) de 0 a 10 puntos medida antes de la infiltración y a los 3 meses posteriores a la infiltración en todos los casos. Analizan los resultados valorando la reducción del dolor de los dos compuestos mediante un test de contraste de hipótesis tomando como valor límite 0.05 para rechazar la hipótesis nula. Los resultados son los siguientes: el Grupo A presentó una reducción del dolor a los 3 meses en la EAV de 2.19 (p = 0.013) y el Grupo B presentó una reducción del dolor a los 3 meses en la EAV de 1.54 (p = 0.11).

Según estos resultados, la infiltración de la sustancia A ha mostrado diferencias estadísticamente significativas en la EAV a los 3 meses de la infiltración mientras que la infiltración de la sustancia B no ha mostrado diferencias estadísticamente significativas en la reducción del dolor. Parece claro, por tanto, que "existen evidencias" a favor de la utilización de la sustancia A para la reducción del dolor a los 3 meses en pacientes con fasciopatía plantar, y que "no existen evidencias" que apoyen el uso de infiltraciones de la sustancia B para la reducción del dolor a los 3 meses en casos de fasciopatía plantar.

Simple, claro y fácil de entender... ¿No? Sin embargo, merece la pena explorar un poco más los datos y analizar los resultados

Recibido: 02/03/2018
Aceptado: 01/04/2018

Correspondencia:
Javier Pascual Huerta
javier.pascual@hotmail.com

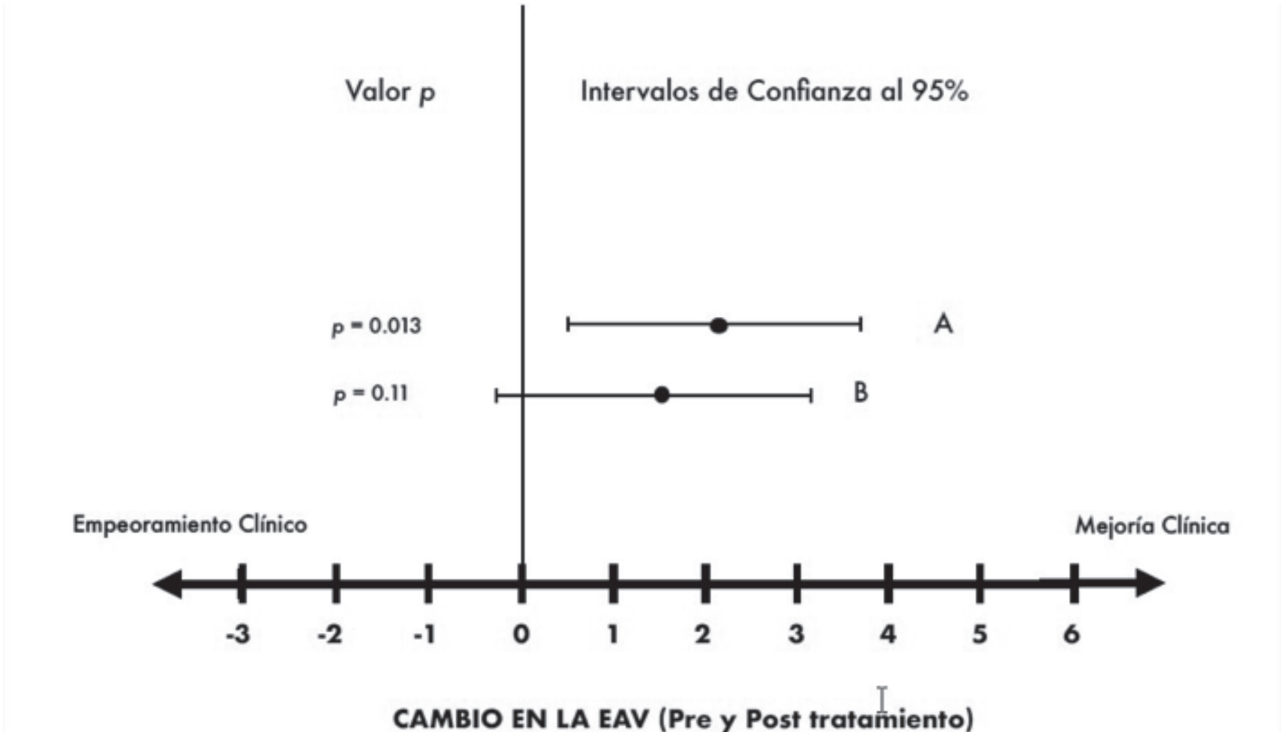

Figura 1.

mediante la aplicación de los intervalos de confianza. La Figura 1 muestra un gráfico de los resultados del estudio mediante la aplicación de los intervalos de confianza al 95 % (IC 95 %) para ambos compuestos A y B.

Lo primero, no olvidemos que Fernández y cols. han hecho un estudio tomando una muestra concreta y limitada para inyectarles las sustancias A y B y lo que estamos intentando hacer ahora es generalizar los resultados de Fernández y cols. sobre la eficacia de A y B a toda la población con fasciopatía plantar. Estamos haciendo una inferencia. Esto se puede hacer aplicando un contraste de hipótesis y calculando el valor *p* o estableciendo intervalos de confianza que definen un rango de valores entre los cuales se va a encontrar el valor real de mejoría de ambas sustancias aplicable a toda la población en 95 de cada 100 estudios realizados de la misma forma (generalmente el intervalo de confianza elegido es el 95 % pero también podemos elegir cualquier otro intervalo de confianza, p. e. 90 %, 95 %, 99 %...). Los intervalos de confianza muestran un intervalo o un rango de valores calculado por métodos estadísticos alrededor de la media de reducción del dolor que se obtendrá al aplicar estas sustancias en 95 de cada 100 estudios iguales que el realizado por Fernández y cols. La media de reducción del dolor en la EAV y su IC 95 % para la sustancia A es de 2.19 (0.48 a 3.82) y para la sustancia B es de 1.54 (-0.23 a 3.21). Esto significa que el valor real de mejoría del dolor con la infiltración de la sustancia A a los 3 meses va a estar entre 0.48 y 3.82, y el valor real de mejoría del dolor con la infiltración de la sustancia B a los 3 meses va a estar entre -0.23 y 3.21 en 95 de cada 100 estudios como el de Fernández y cols.

Mirando la gráfica podemos observar de forma más real cuál es la diferencia en la reducción del dolor entre ambos tratamientos, que, la realidad, no es mucha. Siendo sinceros, la infiltración de la sustancia A parece mejor que la infiltración de la sustancia B, pero ¿realmente hay tanta diferencia como para catalogar a A de "estadísticamente significativa" y a B de "no estadísticamente significativa"? Los IC 95 % son bastante similares y, mirando detenidamente la Figura 1, yo no me atrevería a afirmar rotundamente que A es claramente mejor que B ni que hay evidencias de que B no sea eficaz.

Este ejemplo nos ayuda a entender mejor las diferencias entre el valor *p* y los intervalos de confianza. Son 2 formas diferentes de presentar los mismos resultados, pero la realidad es que nos están dando dos mensajes muy diferentes. Uno es correcto y el otro no tanto. El test de significación de la hipótesis nula, especialmente aplicado al contraste de hipótesis, promueve el pensamiento en "blanco y negro", un pensamiento dicotómico (sí/no; existe significación estadística/no existe significación estadística, funciona/no funciona). Pero la realidad es que no podemos catalogar el conocimiento científico en un simple blanco o negro cuando en la práctica existen millones de grises. Los intervalos de confianza nos ayudan a ver las diferentes tonalidades grises de la escala. El valor *p* no.

Se ha debatido mucho sobre esto y parece que el pensamiento dicotómico simplifica y aporta seguridad al clínico, especialmente si no está muy versado en los métodos estadísticos y su interpretación. Esta es posiblemente una de las razones por las que ha sido imposible desterrar la cultura del valor *p* en la investigación biomédica. Francamente, es

una idea seductora y de apariencia de una fórmula mágica que nos hace separar lo que vale de lo que no vale. Con un simple número podemos saber si un factor está asociado a la enfermedad o no, si los tratamientos van a funcionar o no, etc. Sin embargo, esta seguridad que proporciona el valor *p* es ilusoria, ficticia y simplista.

Los intervalos de confianza aportan una mejor apreciación e interpretación de los resultados de una investigación. Pasamos de una decisión dicotómica a una estimación del rango de efecto que probablemente exista en la población, lo que supone un abordaje más real y más importante para la interpretación correcta de los resultados. Los valores por debajo del límite inferior y por encima del límite superior no son excluidos, pero son considerados como poco probables que ocurran en la población. En un IC 95 %, cada una de estas probabilidades (por arriba y por abajo) es menor de un 2.5 %. Además, los intervalos de confianza también permiten evaluar la significación estadística si el intervalo de confianza incluye o no un valor predeterminado y sabremos así también si ese resultado es estadísticamente significativo o no. En este caso, el IC 95 % de la infiltración de la sustancia A excluye el valor 0 y es por esto por lo que presenta un valor *p* menor de 0,05 (p = 0,013). Por su parte, el IC 95 % de la infiltración de la sustancia B abarca valores negativos y es por ello por lo que el valor *p* es mayor de 0.05 (p = 0.11).

Revistas médicas internacionales de altísimo prestigio como *Lancet* o *British Medical Journal*, así como el comité internacional de editores de revistas médicas (ICMJE), recomiendan el uso de los intervalos de confianza por encima del valor *p* en sus artículos[7]. El objetivo de esta carta no es hacer que los autores dejen de usar el valor *p* dentro de la evaluación de los resultados de su investigación. Sin embargo, sí queremos señalar que un uso más frecuente de intervalos de confianza ayudará a lectores e investigadores a una apreciación más real de los resultados obtenidos en una investigación y a un mejor entendimiento de la inferencia estadística.

CONFLICTO DE INTERESES
El autor no presenta ningún conflicto de intereses relevante con la presente carta.

FINANCIACIÓN
Ninguna.

BIBLIOGRAFÍA

1. Prieto Valiente L, Herranz Tejedor I, eds. ¿Qué significa "estadísticamente significativo"? La falacia del criterio del 5 % en la investigación científica. Madrid: Ed. Díaz de Santos; 2004.
2. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337-50. DOI: 10.1007/s10654-016-0149-3.
3. Ranstam J. Why the p-value culture is bad and the confidence intervals a better alternative. Osteoarthritis Cartilage 2012;20(8):805-8. DOI: 10.1016/j.joca.2012.04.001.
4. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose, Am Statistic 2016;70(2):129-33. DOI: 10.1080/00031305.2016.1154108.
5. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 2008;3(4):286-300. DOI: 10.1111/j.1745-6924.2008.00079.x.
6. Cumming G. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge; 2011.
7. Altman DG. Confidence intervals in practice. En: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. Statistics with Confidence. London: BMJ Books; 2002. p. 6-13.