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(2):113-116
DOI: 10.20986/revesppod.2018.1522/2018

Intervalos de confianza *vs.* valor *p* (II). Pensamiento meta-analítico *vs.* pensamiento dicotómico

Confidence intervals vs. *p values (II). Meta-analytic thinking* vs. *dichotomous thinking*

Javier Pascual-Huerta

Clínica del Pie Elcano. Bilbao

"I suspect that the main reason they (confidence intervals) are not reported is that they are so embarrassingly large!"
Jacobs Cohen (1994)

El valor *p* aplicado al test de contraste de hipótesis (TCH) desarrollado por Neyman y Pearson es, sin duda alguna, la técnica estadística más utilizada en la investigación biomédica de los últimos 80 años. Ya hemos hablado de ello en anteriores cartas[1,2] y de los problemas asociados al valor *p* y a su interpretación en los TCH. Curiosamente, la altísima mayoría de los investigadores científicos nunca se han preocupado por la base filosófica que se esconde detrás de los TCH; de hecho, la mayoría lo usa como un proceso puramente técnico para obtener los resultados de una investigación sin reparar en su verdadero significado. Existe, de hecho, la idea generalizada dentro de la comunidad científica de que los resultados de los TCH son a menudo malinterpretados[3,4]. Una alternativa útil a los TCH es el uso de los intervalos de confianza (IC). En la anterior carta de esta sección ya revisamos el papel de los intervalos de confianza (IC) en la interpretación de los resultados de las investigaciones científicas en contraposición al uso de exclusivo del valor *p* aplicado al TCH, y en la presente carta continuaremos revisando esta idea.

Un intervalo de confianza para un determinado parámetro (por ejemplo, una media poblacional μ) se define como un rango de valores generado mediante un procedimiento estadístico que en un sistema de muestreo repetitivo aleatorizado tiene una probabilidad fija de contener el parámetro. Si la probabilidad de que el procedimiento estadístico genere un intervalo que incluya el parámetro es 0.5, entonces el IC es al 50 %. Igualmente, si la probabilidad de que el procedimiento estadístico genere un intervalo que incluya el parámetro es 0.95, el IC es del 95 %, y así sucesivamente. ¿Qué significa esto? Imaginemos que López y cols. quieren saber cuál es el valor máximo de flexión dorsal de la primera articulación metatarsofalángica (1.ª MTF) en la fase propulsiva de la marcha en pacientes operados de *hallux valgus* en la población española. Ese valor máximo de fle-

* Cohen J. The earth is round (p < 0,05). Am Psichol 1994;49(12):997-1003.

Recibido: 16/09/2018
Aceptado: 21/09/2018

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

xión dorsal (que es una media poblacional μ) es imposible de calcular de forma exacta, ya que López y cols. no tienen ni el tiempo, ni el dinero, ni los medios para poder estudiar ese valor de todos los pacientes operados de *hallux valgus* en la población española. Sin embargo, López y cols., sin poder saber de forma exacta cuál es ese valor, sí que pueden realizar una estimación de dónde es posible que esté ese valor estudiando únicamente una muestra reducida de la población española. López y cols. realizan entonces un estudio de la marcha a 123 sujetos españoles operados de *hallux valgus* mediante un sistema de análisis cinemático con marcadores reflectantes que mide la flexión dorsal de la 1.ª MTF durante la fase propulsiva de la marcha. El valor medio que obtienen de flexión dorsal máxima durante la propulsión, es de 19.3°, y calculan el IC al 95 % que es de 10.8° a 27.8°. López y cols. estiman que es muy posible que el valor máximo de flexión dorsal de la 1.ª MTF durante la propulsión en pacientes operados de *hallux valgus* en la población española esté entre 10.8° y 27.8°. ¿Cómo llegan a esa conclusión? Bien si López y cols. repitieran ese mismo estudio en otros 123 sujetos operados de *hallux valgus* y obtuvieran otra media (posiblemente algo diferente) con un IC al 95 % (también algo diferente), y volvieran a repetir el estudio en otros 123 sujetos, y volvieran a repetirlo, y así sucesivamente hasta hacerlo 100 veces, se estima que en 95 de los 100 estudios realizados por López y cols. los IC obtenidos contendrían la media poblacional real μ y en cinco de los 100 estudios realizados por López y cols. la media poblacional no estaría dentro del IC calculado en esos cinco estudios. López y cols. únicamente han hecho un estudio con 123 sujetos intervenidos de *hallux valgus*, pero entienden que lo más probable es que su estudio sea uno de los 95 en los que el IC calculado contenga la media poblacional. No es correcto decir que existe un 95 % de probabilidad de que la media de flexión dorsal máxima de la 1.ª MTF durante la marcha en pacientes intervenidos de *hallux valgus* esté entre 10.8° y 27.8° (esta es una interpretación errónea muy común de los IC). El 95 % se refiere a la probabilidad de que el estudio realizado por López y cols. (de los 100 que podían haber hecho) sea uno de los que contenga la media poblacional en su IC, y eso no significa que el IC calculado en su estudio concreto tenga un 95 % de probabilidades de tener el valor poblacional real.

Este concepto de IC puede utilizarse para comprobar una hipótesis. De hecho, existe una relación directa y estrecha entre los IC y los TCH. López y cols. creen que el valor que han obtenido en su estudio (que consideran excesivamente pequeño) es diferente entre sujetos que han sido operados por cirugía abierta y sujetos que han sido operados por cirugía de mínima incisión (MIS). Creen que podría haber diferencias en el rango máximo de flexión dorsal de la 1.ª MTF durante la propulsión en pacientes operados por medio de una u otra técnica. Además, ya que consideran el valor obtenido muy pequeño, les gustaría también compararlo con gente que no ha sido operada. Para ello vuelven a analizan los 123 casos según los pacientes hayan sido operados por una técnica de cirugía abierta (76 casos) o por una técnica MIS (47 casos), y también analizan casos que tienen estudiados en su laboratorio (236 casos) de pacientes no operados de *hallux valgus* para ayudarles a interpretar mejor los resultados (Tabla I) (Figura 1). Si el IC excluye un determinado valor (que simbolice la hipótesis nula), el resultado es equivalente a rechazar la hipótesis nula siempre que el nivel de significación establecido sea igual al valor de confianza calculado en el IC (un IC al 90 % equivale a un valor de significación a de 0,1; un IC al 95 % equivale a un valor de significación a de 0,05; y así sucesivamente...). En este caso, la hipótesis nula viene a decir que no existen diferencias y eso significaría que los IC no deberían solaparse o solaparse menos de un 50 % de una de sus ramas. Eso no ocurre comparando los pacientes operados por cirugía abierta con los pacientes operados por cirugía MIS (los IC al 95 % se solapan en su gran mayoría), por lo que no podemos rechazar la hipótesis nula ($p = 0.843$). Sin embargo, los IC al 95 % no se solapan si comparamos ambos grupos con pacientes no operados. En este caso, mirando los IC (que no se solapan) se puede decir que existen diferencias entre los casos operados por cirugía abierta y los casos no operados ($p < 0.001$) y entre los casos operados por cirugía MIS y los casos no operados ($p < 0.001$). Los investigadores no deberían de tener miedo de reportar únicamente el IC sin mencionar el valor *p* obtenido en la prueba, ya que esa información ya está incluida en el IC.

Una de las principales ventajas de los IC sobre los TCH es que los IC favorecen la acumulación de evidencia sobre los experimentos: favorecen el metanálisis a través de la estimación de los tamaños de efecto de los tratamientos. Esta idea ha sido denominada como "pensamiento metanalítico"[5,6]. La investigación centrada en los TCH promueve las decisiones dicotómicas de aceptación o no aceptación, limitando o restringiendo la forma en que los investigadores piensan. Es lo que se denomina como *pensamiento dicotómico* reforzado por los TCH que limita las cuestiones que los investigadores preguntan e incluso las teorías que se desarrollan. Autores contrarios a esta corriente[5-7] argumentan que, usando técnicas de inferencia basadas en los TCH cuyo objetivo principal es afirmar si el resultado es o no es estadísticamente significativo, lleva a los investigadores a formular cuestiones igualmente empobrecidas (por ejemplo, ¿funciona este tratamiento?). Sin embargo, los IC proveen intervalos de estimación que ayudan a promover a los investigadores a formular mejores cuestiones cuantitativas y a desarrollar mejores teorías cuantitativas (por ejemplo, ¿cuánta es la mejoría que aporta este tratamiento? –de la cual, evidentemente, la respuesta puede ser ninguna o negativa–).

Los IC pueden ayudar a mover a los investigadores del "pensamiento dicotómico" dominante en los TCH por el "pensamiento metanalítico" o de estimación, que pone más énfasis en el tamaño de efecto del cambio que si únicamente existiera un cambio o no existiera. Obtener un resultado "estadística-

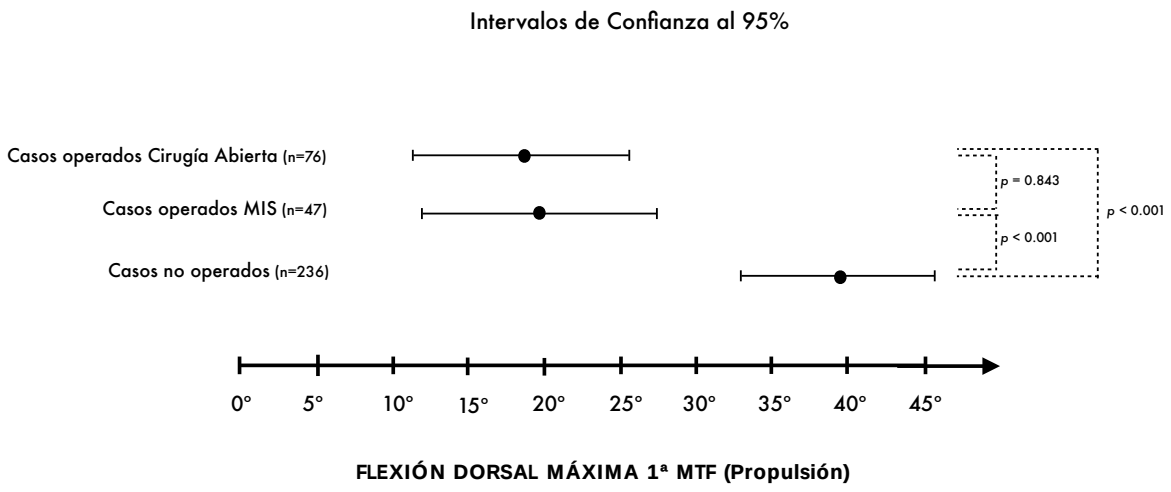

Figura 1. Gráfica de los IC al 95 % de los tres grupos estudiados por López y cols. Nótese que los IC entre los casos operados por cirugía abierta y los casos operados por MIS se solapan en un porcentaje muy amplio por lo que el valor $p > 0.05$. Sin embargo, los IC de confianza de los casos operados por cirugía abierta y los casos no operados no se solapan. Esto también ocurre al comparar los IC de los casos operados por cirugía MIS y los casos no operados. El valor *p* de ambos casos es $p < 0.001$. En términos generales, el valor *p* de la comparación de dos medias es menor del valor convencional de 0.05 cuando el solapamiento de los intervalos de confianza es la mitad o menor del 50 % del margen de error medio (es decir, la mitad de la distancia desde el punto medio hasta uno de los brazos del IC). Igualmente, el valor *p* de la comparación de dos medias es menor de 0.01 si los intervalos de confianza no se solapan. Esta regla es únicamente aplicable para comparar medias de muestras independientes. Una información más detallada de este proceso en Cumming y Finch[5].

Tabla I.

	Pacientes intervenidos de HV mediante cirugía abierta (n = 76)	Pacientes intervenidos de HV mediante cirugía MIS (n = 47)	Pacientes NO intervenidos de HV (n = 236)
Media (flexión dorsal máxima 1.ª MTF durante propulsión)	18.7°	19.6°	39.4°
IC al 95 %	11.4° - 26°	11.7° - 27.5°	33.2° - 45.6°

mente significativo" puede dar una sensación ficticia de seguridad (aunque muy seductiva). Afirmar que un resultado es "significativo" puede fácilmente sugerir que existe un efecto que es importante o grande. Es fácil pasar por encima de la variabilidad del muestreo y de la posibilidad de que la decisión del TCH sea errónea. Por el contrario, la "anchura" que aportan los intervalos de confianza puede entenderse como una medida de precisión del efecto que se mide y que aporta una cuantificación bastante exacta de la certeza o no de la comparación realizada. Un IC estrecho justifica que tenemos un conocimiento bastante preciso sobre el efecto del que estamos estudiando. Un intervalo de confianza grande indica que el nivel de incertidumbre sobre la variable de estudio es alto (independientemente de que sea estadísticamente significativo o no).

Este concepto de amplitud del intervalo de confianza es importante más que ver únicamente que si el intervalo cruza el valor o no de significancia estadística. El hecho de obtener IC muy anchos, como es el caso del estudio de López y cols., ayuda a los investigadores a conocer el nivel de incertidumbre de la variable (que es alto) y ayuda a diseñar mejores estudios futuros que estudien de forma más específica la incertidumbre que existe sobre la variable de estudio. El pensamiento dicotómico de los TCH únicamente nos ayuda

a ver si existe efecto o no existe. El pensamiento metanalítico que aportan los IC ayuda a entender que es muy raro que un único estudio sea adecuado para responder a una pregunta de investigación. Se necesitan varios estudios y cada uno de ellos aporta un punto de estimación sobre la variable que posteriormente puede ser interpretado mediante metanálisis[7]. Un único estudio contribuye a la evidencia, pero necesita ser considerado en el contexto de otros estudios, pasados y futuros, que pueden ser tratados mediante técnicas de metanálisis.

Nota final: El estudio de López y cols. es un ejemplo hipotético cuyos datos han sido completamente inventados.

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

FINANCIACIÓN
Ninguna.

BIBLIOGRAFÍA

1. Pascual Huerta J. Inferencia estadística y aproximación al valor p. Parte I. Rev Esp Podol. 2016;27(1):42-4. DOI: 10.1016/j.repod.2016.04.002.
2. Pascual Huerta J. Inferencia estadística y apoximación al valor p. Parte II. Contraste de hipótesis. Rev Esp Podol. 2016;27(2):86-7. DOI: 10.1016/j.repod.2016.08.001.
3. 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.
4. 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.
5. Cumming G, Finch S. Inference by eye: confidence intervals, and how to read pictures of data. Am Psychol. 2005;60(2):170-80. DOI: 10.1037/0003-066X.60.2.170.
6. Cumming G, Finch S. A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. Educ Psychol Meas. 2001;61(4):532-74. DOI: 10.1177/0013164401614002.
7. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. Psychol Methods. 1996;1:115-29. DOI: 10.1037/1082-989X.1.2.115.