Rev Esp Podol. 2016;27(1):42–44

Revista Española de Podología

www.elsevier.es/rep

RINCÓN DEL INVESTIGADOR

Inferencia estadística y aproximación al valor *p*. Parte I

Statistiscal inference and aproximation to *p*-value. Part I

Javier Pascual Huerta

Clínica del Pie Elcano, Bilbao, España

Disponible en Internet el 9 de mayo de 2016

Comenzamos una nueva sección de la revista que se denomina *Rincón del investigador* y en la que iremos explicando de forma regular temas relativos a la investigación biomédica de forma clara y sencilla. Los métodos estadísticos y su aplicación a la investigación biomédica han sido durante años el «talón de Aquiles» de los profesionales sanitarios que realizan investigación. Las pruebas estadísticas son el final del viaje investigador y, seguramente, lo más difícil de entender. Esto afecta no solo a los investigadores, sino también a los clínicos que tienen que escuchar, leer e interpretar críticamente los resultados de las investigaciones más relevantes en su campo para poder llevarlas a cabo en su práctica diaria. Prieto Valiente y Herranz Tejedor[1] resumen este endémico problema, con gran acierto, argumentando que la estadística es una ciencia matemática que suele ser explicada en un lenguaje matemático y que se convierte en imposible para la mayoría de los profesionales de la salud que no llegan a entender sus métodos, sus formulaciones, el porqué de sus pruebas y, lo que es peor, no llegan a comprender completamente que significan los resultados de las investigaciones que han hecho o que están leyendo. Ante esta situación los investigadores (y los clínicos en general) acaban «tirando la toalla» y abandonándose en simplificaciones mentales de lo que son procesos conceptuales mucho más complejos. Esto les hace sentirse relativamente cómodos, pero les lleva a cometer errores graves en la interpretación de las investigaciones que hacen o leen[1].

Con estas ideas en mente y tomando como ejemplo las cartas que en su día publicó *The Bristish Medical Journal* con los ya míticos artículos de Bland y Altman[2–4] y lo que actualmente realiza *The Journal of Foot & Ankle Surgery* en su *Investigators' Corner*[5,6] comenzamos esta nueva sección en la que iremos tratando de forma regular aspectos relacionados con la investigación con el objetivo de hacer más fácil a la comunidad podológica la lectura e interpretación crítica de los resultados de la investigación que se publiquen en la revista. Empezamos en este número con el famoso y archiconocido valor *p*.

La mayoría de los investigadores tienen graves dificultades para entender lo que verdaderamente indica el valor *p* de los test estadísticos. El valor *p* es un concepto intelectualmente complejo cuya lógica produce confusión entre los investigadores y que tiene su origen en lo que se conoce como «pruebas de significación de la hipótesis nula» y «pruebas de contraste de hipótesis». Ambos son conceptos que se han usado como sinónimos de forma incorrecta en la literatura científica durante décadas. Existen diferencias fundamentales en el significado de las conclusiones que se pueden extraer de una u otra y este malentendido ha sido uno de los motivos que ha enturbiado el proceso de comprensión del valor *p* en la literatura médica[1,7] del que hablaremos en el siguiente número.

Comencemos desde el principio. Fisher (1890-1962) fue un científico y matemático inglés que desarrolló su trabajo durante principios del siglo xx y se le considera el autor con más influencia en los métodos estadísticos modernos. Su tratado *The Design of Experiments* tiene planteamientos de análisis científico vigentes hoy en día. Fisher estaba especialmente interesado en saber si se podían sacar conclusiones que fueran válidas y aplicables a la población general a partir de los resultados obtenidos de un estudio concreto con una muestra determinada. Trataba de saber si el resultado obtenido en un estudio concreto era el fruto de una

Correo electrónico: javier.pascual@hotmail.com

http://dx.doi.org/10.1016/j.repod.2016.04.002

Tabla 1 Valores de función y dolor en el pie (medidos a través del *Foot Health Status Questionnarie*) al comienzo y a los 3 meses de tratamiento en ambos grupos

	Al comienzo		A los 3 meses			
	Plantillas a medida (n = 75)	Plantillas placebo (n = 79)	Plantillas a medida (n = 75)	Plantillas placebo (n = 78)	Diferencia (IC 95%)	Valor *p*
Dolor en pie	42 (19,5)	46,7 (18,1)	73,2 (24,6)	67 (22,2)	8,3 (1,2 a 15,3)	0,022
Función	57,3 (23,5)	60,2 (23,5)	82,9 (22,8)	74,8 (23,8)	9,5 (2,9 a 16,1)	0,005

Adaptada de Burns et al.[8].

realidad general aplicable a toda la población o si por el contrario era el resultado anecdótico de una muestra estudiada concreta. Esta es la idea que subyace en el concepto de inferencia estadística: ¿Puedo extrapolar, en función de los resultados obtenidos en mi estudio, una conclusión aplicable a todos los pacientes? o ¿Son estos resultados debidos al azar de una muestra concreta de pacientes que he estudiado y no representan una realidad biológica extrapolable a todos los pacientes? Para ayudarnos con esto Fisher inventó el test de significación de la hipótesis nula.

Al comenzar cualquier estudio Fisher partía de lo que actualmente se denomina «hipótesis nula» (designada como «H_0»). Esta es el punto inicial del análisis estadístico de un estudio y es la negación de la teoría o idea que se plantea al iniciar el experimento. Lo contrario. La hipótesis nula asegura que un tratamiento no es eficaz para una determinada enfermedad o que 2 factores no están relacionados, etc. La hipótesis nula inicial se debe creer «por defecto» mientras no haya evidencias de lo contrario. Digamos que es una postura inicial escéptica que adopta el método científico. Un análogo al «todo el mundo es inocente mientras no se demuestre lo contrario». A partir de aquí se realiza un experimento tomando datos para valorar si esa hipótesis nula puede ser cierta o no y es aquí donde interviene el famoso valor *p*.

Fisher utilizó el valor *p* como un valor de probabilidad que valora la credibilidad de la hipótesis nula a partir de los datos obtenidos en el experimento en cuestión. El valor *p* nos muestra la probabilidad de haber conseguido el resultado que hemos obtenido *si suponemos que la hipótesis nula es cierta*. Veamos un fantástico ejemplo de esto, ejemplo ilustrado por Jupiter[5] para entender esta idea. Pensamos que una moneda está trucada y queremos probarlo. Lanzamos la moneda 4 veces y obtenemos 3 caras. Quizá esté trucada pero nadie se sorprendería de que una moneda no trucada obtuviera 3 caras en 4 lanzamientos solo por azar. Ahora lanzamos la moneda 10 veces y obtenemos 8 caras y empezamos a pensar que podría estar trucada. Continuamos y lanzamos la moneda 100 veces y tenemos 81 caras y empezamos a tener una evidencia más fuerte de que la moneda podría estar trucada. Lanzamos la moneda 1.000 veces y obtenemos 797 caras. Aquí ya estamos convencidos de que la moneda está trucada. ¿Como podemos probarlo? Bien, la probabilidad de obtener los resultados que hemos tenido, *si la moneda no estuviera trucada*, son de aproximadamente 1 entre 4, 1 entre 10, algo menos de 1 entre un millón y algo menos de 1 entre un trillón de trillones respectivamente. El valor *p* es esta probabilidad. Concluimos que la moneda está trucada porque es muy raro que hubiéramos obtenido estos resultados (y aquí está la clave) *si la moneda fuera normal*.

Esta es la lógica del razonamiento del valor *p* y existen en la literatura miles de ejemplos de su aplicación en estudios. Veamos otro ejemplo más aplicado a la Podología. Burns et al. realizaron un estudio aleatorizado en 2006 sobre el efecto de las plantillas a medida y las plantillas placebo en el dolor y función en pacientes con pie cavo[8]. Para ello a 75 pacientes con pie cavo les hicieron plantillas a medida y a 79 pacientes con pie cavo les hicieron plantillas placebo (una lámina plana de EVA de 3 mm). Midieron los resultados del dolor y la función por medio de escalas en ambos grupos al comienzo y a los 3 meses de llevar las plantillas. Los resultados obtenidos del estudio aparecen en la tabla 1.

¿Cómo interpretamos los resultados? Partimos de la hipótesis nula en este estudio de que las plantillas a medida NO ofrecen una mejoría del dolor ni de la función superior a las plantillas placebo en pacientes con pie cavo y entendemos que esta hipótesis es la verdadera inicialmente a no ser que se demuestre lo contrario. Al realizar este estudio los autores encontraron una diferencia a los 3 meses en la reducción del dolor y de la función entre los grupos de plantillas a medida y plantillas placebo. No es una diferencia muy grande (de 8,3 puntos sobre 100 en el dolor y de 9,5 puntos sobre 100 en la función) pero es una diferencia al fin y al cabo. La pregunta clave aquí ahora es: ¿Estas diferencias de 8,3 y 9,5 puntos en el dolor y en la función en pacientes con pie cavo que usan plantillas a medida comparado con plantillas tipo placebo se ha obtenido porque realmente existe una diferencia real entre usar plantillas a medida en comparación con usar plantillas placebo y es extrapolable a toda la población de pacientes con pie cavo?, o ¿Esta mejoría del dolor y de la función en pacientes con pie cavo que usan plantillas a medida se ha obtenido por pura casualidad de la muestra que los autores tomaron porque en realidad no existen diferencias entre usar una plantilla u otra en estos pacientes? Para responder a estas preguntas usamos el valor *p* que en este caso es $p = 0{,}022$ para el dolor y de $p = 0{,}005$ para la función.

El valor *p* nos dice que si partimos de la base de que NO existen diferencias entre usar plantillas a medida o plantillas placebo para reducir el dolor en pacientes con pie cavo (hipótesis nula), la probabilidad de encontrar una diferencia de 8,3 puntos en el dolor como se ha obtenido o superior es de 22 por 1.000. Es decir, si se repitiera este mismo estudio 1.000 veces con diferentes muestras de pacientes con pies cavos, en 22 estudios se obtendría una diferencia

en la reducción del dolor como la que se ha conseguido (8,3 puntos o más) o superior *si la hipótesis nula fuera cierta*. A su vez, en 978 estudios se obtendría una reducción del dolor menor de 8,3 puntos *si la hipótesis nula fuera cierta*. Dicho de otra forma, si realmente las plantillas a medidas no son más eficaces que las plantillas placebo en pacientes con pie cavo, encontrar la diferencia en el dolor de 8,3 puntos que se ha encontrado en este estudio solo ocurre en 22 veces de 1.000 estudios que se hagan como este. Este mismo proceso se aplica para los resultados de función del pie con un valor $p = 0{,}005$ (únicamente en 5 de cada 1.000 estudios se hubiera encontrado una diferencia entre ambos grupos igual o mayor de 9,5 puntos en la función del pie *si la hipótesis nula fuera cierta*). Es en este punto donde nos planteamos la autenticidad de la hipótesis nula. Es muy difícil que los autores hayan tenido la «mala suerte» de que su estudio sea uno de esos 22 estudios de 1.000 en los que aparecen diferencias en el dolor de 8,3 puntos o mayores entre ambos grupos solo por azar si la hipótesis nula es cierta. Tendemos a concluir que posiblemente la hipótesis nula no sea cierta y nos basamos en el valor *p* para decir esto.

Así es como funcionan las pruebas de significación. Son un método deductivo a partir de una muestra de datos y de una teoría previa (hipótesis nula). Tratan de deducir si lo observado con los datos confirma o no la teoría previa de partida. Si en el estudio obtenemos datos compatibles con la hipótesis nula, mantenemos la hipótesis nula como posible. Si, por el contrario, en el estudio obtenemos datos que no son compatibles o muy difícilmente compatibles con la hipótesis nula, nos planteamos que la hipótesis nula no sea cierta. Es por esto que el conocimiento científico se sustenta en datos y no en teorías y en esto se basa el método científico.

Bibliografía

1. Prieto Valiente L, Herranz Tejedor I, editores. ¿Qué significa «estadísticamente significativo»? La falacia del criterio del 5% en la investigación científica. Madrid: Ed. Díaz de Santos; 2004.
2. Bland JM, Altman DG. Measurement error and correlation coefficients. BMJ. 1996;313:41–2.
3. Altman DG, Bland JM. Statistics notes: The normal distribution. BMJ. 1995;310:298.
4. Bland JM1, Altman DG. One and two sided tests of significance. BMJ. 1994;309:248.
5. Jupiter DC. Mind your p values. J Foot Ankle Surg. 2013;52:138–9.
6. Jupiter DC. Counting your chickens before they're hatched: Power analysis. J Foot Ankle Surg. 2014;53:519–20.
7. Rebasa P. Entendiendo la «p < 0,001». Cir Esp. 2003;73:361–5.
8. Burns J, Crosbie J, Ouvrier R, Hunt A. Effective orthotic therapy for the painful cavus foot: A randomized controlled trial. J Am Podiatr Med Assoc. 2006;96:205–11.