Rev Esp Podol. 2016;27(2):86–87

Revista Española de Podología

www.elsevier.es/rep

RINCÓN DEL INVESTIGADOR

Inferencia estadística y aproximación al valor *p*. Parte II. Contraste de hipótesis

Statistiscal inference and aproximation to *p*-value. Part II. Hypothesis testing

Javier Pascual Huerta

Clínica del Pie Elcano, Bilbao, Vizcaya, España

Recibido el 10 de agosto de 2016; aceptado el 18 de agosto de 2016
Disponible en Internet el 5 de octubre de 2016

En el pasado número de esta sección del *Rincón del investigador* discutimos lo que significa el valor *p* en el test de significación de la hipótesis nula descrito por Fisher. El valor *p* de un determinado estudio representa la probabilidad de obtener resultados iguales o más extremos de los obtenidos, considerando que la hipótesis nula fuera cierta. Cuanto más bajo es el valor *p*, menor es la probabilidad de que la hipótesis nula sea cierta, y en un punto dado, la hipótesis nula es rechazada con bajo riesgo de equivocarnos[1]. Es por tanto una medida cuantitativa en forma de probabilidad de la evidencia que aportan los datos obtenidos en un estudio en contra de la hipótesis nula. Sin embargo, la historia del valor *p* no queda aquí. El concepto de esta idea se complicó algo más cuando aparecieron en escena Neyman y Pearson, y establecieron lo que se conoce actualmente como el *Test de contraste de hipótesis* dando una vuelta de tuerca más a la idea del test de significación de la hipótesis nula de Fisher[2]. Pensaron que no tiene sentido formular una hipótesis nula (H_0) si no cabe la posibilidad de imaginar una «hipótesis alternativa» (designada como «H_1») opuesta y contraria a la hipótesis nula formulada. Al fin y al cabo, si un investigador toma la decisión de rechazar H_0, está aceptando una hipótesis alternativa.

Eagon Pearson (1895-1980) —el contrincante intelectual de Fisher— y Jerzy Neyman (1894-1980) argumentaron que el test de significación de la hipótesis nula de Fisher nunca puede dar una seguridad absoluta al investigador para rechazar o no H_0. No es más que una probabilidad, y siempre cabe la posibilidad de que el investigador esté cometiendo un error al decantarse por una u otra decisión. Nunca se puede estar totalmente seguro de si un tratamiento es superior a otro, o de si 2 factores están relacionados basándose en el valor *p* de un único estudio. Sin embargo, el investigador sí puede determinar el riesgo que desea asumir al rechazar o no H_0 (y consecuentemente aceptar H_1) conforme al valor *p* obtenido en su estudio, y en esto se basa la idea del test de contraste de hipótesis.

Veamos un ejemplo, Frykberg et al.[3] estudiaron la presencia de limitación a la flexión dorsal de tobillo en pacientes diabéticos y pacientes no diabéticos para ver si existían diferencias entre ellos. Para ello estudiaron 43 pacientes diabéticos y 59 pacientes no diabéticos (con igual distribución de edad y sexos) a los que se le midió el rango de flexión dorsal de tobillo con un goniómetro. La hipótesis nula planteada es que no existen diferencias en el porcentaje de pacientes con limitación de la flexión dorsal de tobillo en ambos grupos (diabéticos y no diabéticos), y la hipótesis alternativa es que sí que existen diferencias. Los resultados mostraron que 16 de los pacientes diabéticos tenían limitación a la flexión dorsal de tobillo (37,2%), mientras que únicamente 9 de los pacientes no diabéticos (15,3%) tenían limitación a la flexión dorsal de tobillo. El valor *p* obtenido de la diferencia (21,9%) fue de $p = 0{,}011$. ¿Cómo interpretamos estos resultados en el contraste de hipótesis? El resultado nos muestra que si H_0 fuera cierta (H_0 = no existen diferencias en el porcentaje de pacientes con limitación a la flexión dorsal de tobillo en ambos grupos), la

Correo electrónico: javier.pascual@hotmail.com

http://dx.doi.org/10.1016/j.repod.2016.08.001

Tabla 1 Errores posibles de tipo I y II, de acuerdo con el test de contraste de hipótesis

	Realidad	
	H_0 es cierta	H_0 es falsa
Hallazgos del estudio		
Se rechaza H_0	Error tipo I (α generalmente 5%)	Rechazo verdadero
No se rechaza H_0	No rechazo verdadero	Error tipo II (β generalmente 0,1-0,2)

probabilidad de haber obtenido una diferencia como la que hemos obtenido de 21,9% o mayor, usando una muestra como la utilizada, ocurriría únicamente en 11 de 1.000 estudios. En este caso, los autores decidieron rechazar H_0 y aceptar H_1 concluyendo que (según sus datos) sí existen diferencias en el porcentaje de casos con limitación a la flexión dorsal de tobillo entre pacientes diabéticos y pacientes no diabéticos.

Sin embargo, al extraer esta conclusión del estudio, es posible que los autores estén cometiendo un error. Cabe la posibilidad de que la H_0 sea cierta (en realidad no existan diferencias en la flexión dorsal de tobillo entre ambos grupos) y el estudio sea uno de los 11 estudios sobre 1.000 en los que anómalamente se han encontrado diferencias, cuando en realidad estas no existen. Es raro, pero eso podría haber ocurrido, y a eso se le llama error tipo I (rechazar H_0, cuando en realidad es cierta) (tabla 1). Imaginemos ahora que este mismo grupo de investigadores, preocupados por no cometer ese error tipo I (afirmar que existen diferencias en el porcentaje de casos con limitación a la flexión dorsal de tobillo entre pacientes diabéticos y no diabéticos, cuando en realidad no existen), hubieran sido mucho más conservadores en sus conclusiones, y hubieran estimado el valor $p = 0{,}011$ como insuficiente para poder rechazar H_0. En definitiva, 11 de cada 1.000 no es un valor tan fuerte como poder descartar por completo y de forma tajante la posibilidad de que no existan diferencias en la flexión dorsal de tobillo en pacientes diabéticos y no diabéticos, y que los resultados hayan sido debidos únicamente al azar o a «la mala suerte» de los investigadores al tomar la muestra. En este hipotético caso los autores hubieran concluido que los datos obtenidos no son suficientes como para rechazar H_0 y aceptar H_1. Es muy improbable que al adoptar esta postura más conservadora los autores estuvieran cometiendo un error tipo I (rechazar la hipótesis nula, cuando en realidad es cierta), pero es probable que estén cometiendo otro tipo de error. Puede ser que realmente sí que existan diferencias entre ambos grupos, y por ser tan conservadores en su interpretación de los resultados no rechacen H_0, cuando en realidad deberían de hacerlo. Esto se denomina error tipo II (no rechazar H_0 cuando es falsa) (tabla 1).

Ya que es imposible librarse de la posibilidad de cometer uno de estos errores, al tomar una decisión con los resultados de una investigación (error tipo I o tipo II), el contraste de hipótesis de Pearson y Neyman establece unos márgenes de error determinados, que los investigadores aceptan para extraer conclusiones con los resultados de su estudio. Estos márgenes de error se establecen *a priori* antes de comenzar el estudio, y definen cuales van a ser las «normas de comportamiento» de los investigadores con respecto a rechazar H_0 y aceptar H_1 cuando tengan que interpretar los resultados de su investigación. Actualmente se ha universalizado en la literatura científica el valor de 5% ($p < 0{,}05$) como el valor de referencia para rechazar o no H_0. Este valor límite de 0,05 es el error tipo I que el investigador acepta como posible cuando rechaza H_0 (curiosamente Neyman y Pearson no señalaron el 5% como el umbral de error permitido para el error tipo I, aunque este valor ha calado profundamente en el mundo científico).

Es importante que el investigador entienda que el test de contraste de hipótesis no es realmente un método para conocer la realidad o realizar deducciones, sino para tomar una decisión con datos obtenidos prefijando unos valores de «seguridad» a partir de los cuales aceptaremos o rechazaremos H_0. «*Sin interés por conocer si cada hipótesis por separado es verdadera o falsa, podemos buscar reglas que gobiernen nuestro comportamiento, de tal forma que nos aseguremos que usando estas reglas, a la larga, no estaremos equivocados muy a menudo*»[4].

Nota importante para investigador, clínico y lector en general. Es importante no confundir ni mezclar las ideas del test de significación de Fisher y del contraste de hipótesis de Pearson y Neyman. En ocasiones, la investigación da un valor *p* con el que tenemos dudas razonables sobre si rechazar H_0 o no. Esas dudas razonables no desaparecen porque prefijemos un valor *p* a partir del cual vamos a rechazar H_0 antes de realizar la investigación, o porque nos empeñemos en que por debajo de una valor *p* determinado rechazaremos H_0 «caiga quien caiga». Esas dudas siempre van a estar ahí, por lo que es posible que en ocasiones después de hallar un valor *p* no obtengamos una respuesta clara a nuestra pregunta, y aplicar el contraste de hipótesis nos aportará una regla de comportamiento a seguir, pero el investigador sensato entenderá que puede que no nos esté ayudando a resolver nuestras dudas o a avanzar en nuestro conocimiento[5].

Bibliografía

1. Pascual Huerta J. Inferencia estadística y aproximación al valor *p*. Parte I. Rev Esp Podol. 2016;27:42–4.
2. Biau DJ, Jolles BM, Porcher R. Value and the theory of hypothesis testing. An explanation for new researchers. Clin Orthop Relat Res. 2010;468:885–92.
3. Frykberg RG, Bowen J, Hall J, Tallis A, Tierney E, Freeman D. Prevalence of equinus in diabetic versus nondiabetic patients. J Am Podiatr Med Assoc. 2012;102:84–8.
4. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A. 1933;231:289–337.
5. 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.