Lezana. *Kinesiología (2022); 41(2): 295-299*
ISSN: 0716-4173 / e-ISSN: 2735-7481

SERIE DE GUIAS PRACTICAS DE METODOLOGÍA DE LA INVESTIGACIÓN

Parte 7: Significancia Estadística en Ciencias de la Salud: Limitaciones y Alternativas

Melissa Lezana Zúñiga[1]

Resumen

La prueba de significancia de la hipótesis nula (PSHN) constituye la herramienta más usada para evaluar hipótesis científicas y tomar decisiones al respecto, en especial en ciencias de la salud. Sin embargo, por décadas ha estado en el centro del debate, ya que se han identificado varios problemas conceptuales y de interpretación. Se realizó una revisión de artículos científicos que ilustran las críticas de esta controversia y su relevancia en el ámbito de la investigación en salud. Algunas alternativas para la PSHN son una adecuada interpretación del valor *p*, uso de intervalos de confianza, incluir el tamaño del efecto y adoptar un marco de inferencia bayesiana. En todos los casos en que se utilice PSHN, su uso debe ser claramente justificado.

Palabras claves: Prueba de Significancia; Prueba de Significancia de la Hipótesis Nula; Valor *p*; Inferencia Estadística; Estadística en Salud; Educación.

Introducción

Las pruebas de significación de hipótesis nula (PSHT) han estado bajo discusión durante décadas. La literatura muestra evidencia de distintos problemas que afectan el uso de PSHT. La mayoría de las investigaciones en el área de las ciencias de la salud incluyen pruebas de significancia estadística (rechazando o no una hipótesis de nulidad) en las secciones de resultados o discusión. Dentro de las pruebas que se suelen utilizar, las pruebas de t de student, U de Mann-Whitney, Análisis de Varianza (ANOVA), entre otras son las más conocidas.

La PSHT está muy arraigada en las mentes y práctica actual de investigadores de distintas áreas, como la biología, psicología y ciencias sociales, quienes otorgan un rol preponderante a la PSHN, con el fin de poder decidir cuán válidos son sus resultados. Actualmente, existen distintos enfoques para decidir su uso, siendo los principales los enfoques de Fisher y el de Neyman-Pearson.

Existen diferencias filosóficas y conceptuales entre los enfoques de Fisher y Neyman y Pearson. Al realizar las pruebas de significancia, Fisher estaba interesado en encontrar resultados importantes al evaluar la solidez de dicha evidencia. Usando su enfoque, el investigador está preparado para prestar atención a los resultados estadísticamente significativos e ignorar el resto[1].

Por el contrario, el interés de Neyman y Pearson era decidir qué hipótesis aceptar como más probables. Llamaron a su procedimiento "pruebas de hipótesis estadísticas", una denominación que difiere del de Fisher, pero que no crea una separación conceptual clara entre ambos enfoques. Después de todo, Fisher también usa hipótesis estadísticas, sin embargo, ninguno de los procedimientos prueba las hipótesis, sólo los datos de investigación contra las hipótesis estadísticas que se suponen verdaderas. La referencia de "probar" hipótesis es, por tanto, engañosa y el procedimiento se beneficia al ser rebautizado como pruebas de aceptación de Neyman-Pearson.[2] Estas distinciones son de relevancia para la disciplina, sobre todo en aumentar la formación de uso de lógica e inferencias en la práctica basada en la evidencia.

Desarrollo

La PSHT es la prueba más utilizada en la actualidad, bajo la falsa suposición de probar hipótesis sustantivas. La PSHT es una mezcla de las teorías de Fisher y Neyman-Pearson la cual no está claramente definida y depende del autor o investigador que lo usa, si es que tiende más hacia el enfoque de Fisher o hacia el enfoque de Neyman y Pearson. Desafortunadamente, los enfoques de Fisher y Neyman-Pearson son incompatibles en varios aspectos (ver Tabla 1). En general, la mayoría de las amalgamas siguen a Neyman-Pearson procedimentalmente, pero a Fisher filosóficamente.

Para explorar el estado del arte de esta discusión, se reportan los resultados de una revisión narrativa de artículos publicados desde el año 2015 a las fechas sobre PSHT.

En general, los investigadores interpretan el valor *p* exacto y lo utilizan como una medida de evidencia contra H_0, como hizo Fisher. Un resultado "altamente significativo" con un valor *p* pequeño se percibe como una evidencia mucho más sólida que uno débilmente significativo[3]. Generaciones de científicos alentados por interpretaciones incorrectas dependen exclusivamente del valor *p* en sus decisiones, incluso si esto significa descuidar su conocimiento.[4]

El valor *p* es prácticamente lo único que la PSHN calcula, pero lo que les interesa generalmente a los científicos es saber cuál es la probabilidad de que su teoría sea verdadera o falsa según los datos, es decir, están interesados en la probabilidad post-experimental de la hipótesis de nulidad (H_0) e hipótesis alternativa (H_1)[5]. Por ejemplo, la publicación "Retirar la significancia estadística" de la Revista *Nature*, sugiere que todo concepto de significación estadística debe abandonarse con el fin de detener el uso de valores *p* de la manera convencional y dicotómica, para decidir si un resultado refuta o apoya una hipótesis[6]. De forma simultánea, la Asociación Estadounidense de Estadística publicó una declaración en la que rechaza las prácticas que rodean la dependencia y la interpretación de las pruebas de significancia y solicita la aplicación de métodos de inferencia alternativos[3].

La PHSN se utiliza para determinar la "significación estadística" del efecto observado, generalmente definido por un valor de $p < 0{,}05$. Sin embargo, los valores de *p* a menudo se malinterpretan[7]. Por lo tanto, en lugar de centrarse en la significación estadística, los investigadores deben proporcionar estimaciones plausibles sobre la magnitud del efecto en la población de la que se tomaron los datos[8].

En el marco de Fisher, el valor *p* es una métrica de evidencia contra H_0; y cuando H_0 se considera falso, parece que la hipótesis alternativa debe ser cierta. Pero ¿y si encontramos un efecto no significativo (por ejemplo: $p = 0{,}3$)? ¿Es esa evidencia de que no tenemos ningún efecto? Aparte de la variabilidad de muestreo, hay una posible explicación para un valor no significativo: cuando el tamaño de la muestra es pequeño. A medida que crece el tamaño de la muestra, un valor *p* no significativo aumenta cada vez más, lo que sugiere que el tratamiento realmente no tuvo un efecto, o al menos solo uno demasiado pequeño como para ser relevante[9]. Se podría creer que, si "valores p" más bajos proporcionan más evidencia contra H_0, valores mayores de p deberían proporcionar más evidencia a favor de H_0. Es un error frecuente pensar que un valor p más bajo siempre significa evidencia más sólida independientemente del tamaño de la muestra y tamaño del efecto[10].

Esta limitación inherente del *p*-valor impide sacar la conclusión de que, por ejemplo, un tratamiento no tiene efecto - y por lo tanto que una vía o circuito cerebral particular no está involucrado, o que la dimensión de un estímulo particular no importa para la actividad cerebral. Mientras que la evidencia y experiencia de los investigadores indica que no es así. En cierto sentido, el conocimiento se vuelve como la luna: el lado que nos mira nos es familiar, pero el otro permanece en la oscuridad[9].

Algunas alternativas para complementar las PSHN, por ejemplo, informar intervalos de confianza y tamaño del efecto o aplicar el enfoque bayesiano. Recuerde que un intervalo de confianza se interpreta en términos de repeticiones hipotéticas del estudio, por ejemplo, si se repitiera muchas veces el muestreo, se aplicara el mismo procedimiento y se calcularan respectivos intervalos de confianza al 95% según las fórmulas conocidas, esto implica que 95 de cada 100 intervalos incluirían al verdadero parámetro que está siendo estimado[8].

La interpretación que usa IC como una estimación del rango es más consistente con las hipótesis estadísticas en comparación a la utilización del valor *p* de la PSHN. Los resultados que utilizan IC en lugar de los valores *p* son más confiables ya que los IC indican el tamaño esperado del efecto[8].

Tabla 1 Equivalencia de constructos en las teorías de Fisher, Neyman-Pearson y PSHN.

Concepto		Fisher		Neyman y Pearson
Objetivo de la prueba	PSHN	Datos—P(D|H_0)	= Datos como si se estuviera probando una hipótesis falsificable hipótesis —P (H_0|D)	Datos—P(D|H_M)
Enfoque	PSHN	A posteriori	≠ A posteriori, a veces ambos	A priori
Objetivo de la investigación	PSHN	Importancia estadística de los resultados	≠ Significación estadística, también utilizada para decidir entre hipótesis	Decidir entre hipótesis contrastadas
H0 bajo la prueba	PSHN	H_0, para ser anulado con evidencia	≈ Ambas, $H_0 = H_M$	H_M, a favor de la H_1
Hipótesis alternativa	PSHN	No es necesario (implícitamente, "No H_0 ")	≠ H_A usada como "No H_0 "	Necesario, aporta Error estándar y β
Estadístico de interés	PSHN	valor p, como evidencia contra H_0	≠ valor p, utilizado como evidencia contra H_0 y un proxy para aceptar H_A	CVtest (el valor p no tiene significado inherente, pero se puede utilizar como un proxy en su lugar)
Interpretación de resultados en la región crítica	PSHN	Ocurrió un evento raro o H_0 no explica los datos de la investigación	≠ H_A ha sido probado / dado que es verdadera; o H_0 ha sido rechazada/dado que es falsa; o ambos	H_A explica mejor los datos de investigación que H_M (dado α)

$P(D|H_0)$: Distribución de probabilidad de los datos dada una hipótesis de nulidad verdadera, $P(D|H_M)$: Distribución de probabilidad de los datos dada una hipótesis principal verdadera, H_0: Hipótesis de nulidad, H_A: Hipótesis alternativa (enfoque de Neyman-Pearson), H_M: Hipótesis principal (enfoque de Neyman-Pearson), PSHN: Prueba de significancia de hipótesis nula, CVtest: Valor crítico de la prueba. Fuente: Perezgonzalez, JD. A reconceptualization of significance testing. Theory & Psychology. 2014; 24(6):852-859.

El valor *p*, a diferencia del IC, no entrega información respecto al rango en el que se encuentra la magnitud del efecto de una determinada intervención kinésica (valor real), por lo que sólo indica diferencias estadísticas significativas, sin permitir evaluar si esta diferencia es importante para el paciente. Por ejemplo, un resultado significativo (p<0,05) podría incluir diferencias clínicamente irrelevantes, y resultados no significativos (p>0,05) podrían esconder una diferencia clínicamente importante entre 2 intervenciones si el estudio no incluye un tamaño muestral adecuado (un estudio con bajo poder puede no mostrar una diferencia que realmente existe) [8].

Otra medida que ayuda a cuantificar y comprender los resultados de una prueba de hipótesis es el tamaño del efecto. El tamaño del efecto describe la magnitud de la relación entre una variable y otra, por ejemplo, cuantificar el tamaño de la diferencia entre la media de un grupo control y la de un grupo experimental[10].

Los investigadores no deben centrarse simplemente en la significación estadística, sino que también deben informar el tamaño del efecto observado. Las estimaciones puntuales de los tamaños del efecto deben ir acompañados de toda la gama de valores plausibles para cuantificar esta incertidumbre. Esto facilita la evaluación de cuán grande o pequeño podría ser realmente el efecto observado en la población de interés y, por lo tanto, cuán clínicamente importante podría ser. Algunos resultados pueden identificarse como estadísticamente significativos porque sus valores *p* están por debajo del umbral de p=0,05, pero que han observado tamaños del efecto por debajo del obtenido en un cálculo de tamaño de muestra[7].

Estas PSHN, bajo la cual se calculan valores de *p* con el fin de rechazar una H_0, se aplican bajo el enfoque de la estadística frecuentista, sin embargo, durante las últimas décadas, se ha comenzado a utilizar el método bayesiano, el cual, a diferencia del método frecuentista clásico, permite combinar los conocimientos previos del investigador con los datos observados para realizar inferencia acerca de los parámetros de interés[11].

Véase de esta forma: El clínico tiene un criterio a priori sobre un paciente; realiza evaluaciones complementarias y actualiza su visión inicial sobre el diagnóstico que corresponde a ese paciente al conjugar las dos cosas (visión inicial e información complementaria) en lo que constituye un proceso de inducción integral. Este razonamiento es similar a lo que realiza un investigador o investigadora; esa es exactamente la forma en que opera el pensamiento bayesiano. La diferencia esencial entre el pensamiento clásico y el bayesiano radica en que aquel se pronuncia probabilísticamente sobre los datos a partir de supuestos (la "p" no es otra cosa que eso); en tanto que éste se pronuncia (también probabilísticamente) sobre los supuestos partiendo de los datos.

Conclusión

Actualmente se han presentado argumentos polarizados a favor y en contra del uso de PSHN. Es importante que las disciplinas entregan recomendaciones claras sobre la aceptabilidad de métodos de investigación.

Los científicos parecen preferir publicar resultados estadísticamente significativos. Esto distorsiona la evidencia publicada y puede tener consecuencias adversas para la atención del paciente porque las decisiones clínicas se basan, al menos en parte, en investigaciones publicadas.

La solución a este problema de la PSHN incluye la aplicación de métodos de inferencia distintos de las pruebas estadísticas. Si bien estos métodos han estado disponibles durante años, no han logrado disminuir la popularidad de la PSHN[5]. Cabe destacar que la literatura sugiere no prohibir el uso de la PSHN y debe usarse siempre y cuando esté bien justificado. El uso de medidas como el IC y los tamaños del efecto son un aporte sustantivo para este análisis.

Por otra parte, se considera relevante la traslación de los conocimientos (*Translational Medical Research*) a todo nivel (clínico, usuario, creadores de políticas públicas, entre otros). Independiente de los análisis estadísticos realizados, se debe incentivar a la obtención de una conclusión narrativa en términos sencillos, que no tenga una interpretación ambigua, ya que generalmente se usan términos estadísticos (no manejados por todos los lectores) para, por ejemplo, favorecer una conclusión que no es clara o contundente en los análisis.

Es probable que ningún método proporcione una solución que reemplace a la PSHN. En cambio, es necesario reeducar a la comunidad científica para revertir su mal uso.

Agradecimientos

El presente artículo fue realizado bajo supervisión de los académicos Felipe Medina Marín y Sandra Flores Alvarado del programa de Magíster en Bioestadística de la Escuela de Salud Pública de la Universidad de Chile.

Financiamiento

Ninguno.

Conflicto de interés

El autor declara no tener conflicto de interés

Detalles de los autores

[1] Escuela de Salud Pública, Núcleo Desarrollo Inclusivo, Universidad de Chile.

Correspondencia a:

Melissa Rosalba Lezana Zúñiga

Escuela de Salud Pública, Núcleo Desarrollo Inclusivo, Universidad de Chile. Avda. Independencia 1027, Independencia. Santiago, Chile

melissa.lezana@ug.uchile.cl

Recibido: Junio 2022

Publicado: Marzo 2022

Referencias

1. Perezgonzalez, JD. A reconceptualization of significance testing. Theory & Psychology. 2014; 24(6):852-859.
2. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol. 2015; 6:223.
3. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129-33.
4. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Front Hum Neurosci. 2017; 11:390.
5. Perneger TV, Combescure C. The distribution of P-values in medical research articles suggested selective reporting associated with statistical significance. Journal of Clinical Epidemiology. 2017; 87:70-7.
6. Lovell DP. Null hypothesis significance testing and effect sizes: ¿can we 'effect' everything … or … anything? Current Opinion in Pharmacology. 2020; 51:68-77.
7. Schober P, Bossers SM, Schwarte LA. Statistical Significance Versus Clinical Importance of Observed Effect Sizes: What Do P Values and Confidence Intervals Really Represent? Anesth Analg. 2018;126(3):1068-72.
8. Candia B R, Caiozzi A. G. Intervalos de Confianza. Revista médica de Chile. 2005; 133:1111-5.
9. Keysers C, Gazzola V, Wagenmakers E-J. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. Nature neuroscience.2020; 23(7): 788-799.
10. Harrison AJ, McErlain-Naylor SA, Bradshaw EJ, Dai B, Nunome H, Hughes GTG, et al. Recommendations for statistical analysis involving null hypothesis significance testing. Sports Biomech. 2020;19(5):561-8.