Rev. chil. endocrinol. diabetes 2012; 5 (3): 136-139

Evaluando la experiencia: el p-value

Gabriel Cavada Ch.[1,2]
[1]*Facultad de Medicina, Universidad de Los Andes.*
[2]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*

Evaluating the experience: the p-value

Tal vez el concepto más usado para evaluar la evidencia empírica y al mismo tiempo menos entendido por los generadores y lectores de la investigación cuantitativa en biomedicina es el p-value. En este artículo deseo presentar una revisión del concepto extraída de artículos epidemiológicos y hacer algunas reflexiones propias alusivas al tema.

El p-value es un concepto ampliamente usado en todas las áreas de investigación que requieren de la estadística para validar sus resultados. Es una componente escencial del proceso de inferencia científica, entendido como el proceso reflexivo que permite evaluar teorías a partir de observaciones[1]. Sin embargo, el p-value no ha sido bien entendido y muchas veces su uso ha sido discutido.

La historia del p-value, se remonta a la propuesta de Ronald A. Fisher llamada "Dócima de Significación". Esta propuesta data de 1920 y su objetivo era establecer si un resultado era significativo o no significativo. Para ello, Fisher propuso la probabilidad de significación, el que fue pensado como el indicador que permitiría evaluar la significación de los resultados en orden a abandonar el Status Cuo. Luego, el p-value fue definido como la probabilidad bajo la hipótesis nula de obtener valores de la estadística de trabajo iguales o más extremos que los observados en el experimento[2]. Por lo tanto, el p-value fue concebido como la medida de la evidencia en un único experimento, lo que reflejaba la credibilidad de la hipótesis nula a la luz de los datos. Dicho de otro modo, el p-value correspondía a una medida de la discrepancia entre los datos y la hipótesis nula[2-4].

Fisher clarificaba que este indicador debía ser utilizado con flexibilidad dentro de los procesos complejos de descripción e inferencia de la investigación científica. El p-value debía ser combinado con otras fuentes de información sobre el fenómeno en estudio y en caso de utilizar un umbral para evaluar significación, éste debía ser flexible y depender del conocimiento acumulado sobre el fenómeno en estudio. Esto transformó al p-value en un indicador informal que no constituía parte de un método formal de inferencia, dejando finalmente la interpretación del p-value en manos del investigador[2].

Fisher, quien compartía intereses entre la estadística y la genética, estaba interesado en resolver problemas reales y sus propuestas teóricas siempre estaban relacionadas con aplicaciones prácticas[5]. Estas características de su trabajo, permiten comprender mejor su propuesta de racionamiento inductivo para evaluar la evidencia de un experimento, propuesta que generó distintas reacciones entre sus contemporáneos.

Tal vez los más críticos de su propuesta fueron Jerzy Neyman y Egon Pearson, quienes plantearon en 1928 una nueva propuesta llamada "Dócima de Hipótesis" tendiente a reemplazar la Dócima de Significación ideada por RA Fisher.

Neyman se caracterizó por un mayor énfasis en el razonamiento lógico y matemático, aunque sin dejar de lado la importancia de la aplicación práctica, ya que planteaba que los problemas prácticos eran la fuente de inspiración para la teoría estadística[6]. Junto a Pearson criticaron duramente la propuesta de RA Fisher declarando que "ninguna dócima basada en la teoría de probabilidades puede proveer por sí sola alguna evidencia valiosa sobre la veracidad o falsedad de una hipótesis"[3].

La propuesta de Dócima de Hipótesis buscaba reglas que gobernaran el comportamiento relacionado a las hipótesis planteadas, de manera de reducir los errores a largo plazo[2]. Esto introdujo los conceptos de hipótesis alternativa junto al de hipótesis nula y al error tipo II junto al tipo I. Los errores tipo I y II fueron definidos como los errores que puede cometer el investigador en el proceso de Dócima de Hipótesis, siendo el error tipo I referido a la obtención de resultados falsos positivos (plantear que hay diferencia entre los grupos cuando no la hay), mientras que el error tipo II estaba referido a los resultados falsos negativos (plantear que no hay diferencia cuando los grupos son diferentes). La magnitud de estos errores se debía ajustar a cada experimento en particular y debía estar en función de las consecuencias de cometer cada uno de ellos. Con la definición de estos errores era posible identificar regiones críticas que permitían rechazar o no rechazar la hipótesis correspondiente. Si el resultado caía dentro de la región crítica, la hipótesis alternativa debía ser aceptada y rechazada la hipótesis nula. Por el contrario, si el resultado caía fuera de la región crítica, la hipótesis nula debía ser aceptada y rechazada la alternativa[3].

Por lo tanto, esta propuesta implicaba un razonamiento deductivo que buscaba disminuir los errores a lo largo de distintos experimentos, en oposición al razonamiento inductivo basado en un único experimento planteado por Fisher. Esto significaba un avance en términos matemáticos y conceptua-

les, pero implicaba dificultades para la práctica científica, ya que no incluía ninguna medida de evidencia[2].

Tiempo después de ser planteadas estas propuestas, comenzó a gestarse anónimamente el recurso híbrido surgido de la fusión de ambas, dando origen a lo que hoy conocemos como "Dócimas de Hipótesis Basadas en el Cálculo del p-value" o "Dócimas de Significación Estadística"[7]. Este método combinado consiste básicamente en establecer la magnitud del error tipo I y II previo al experimento, luego calcular el p-value en base a las observaciones y finalmente rechazar la hipótesis nula si el p-value es menor a la magnitud del error tipo I establecida previamente[2]. En este método, la magnitud de los errores se establece arbitrariamente, siendo utilizado en casi todos los casos 0,05 como magnitud del error tipo I, transformando al proceso en algo mecánico.

Es decir, este método combina elementos de ambas propuestas originales, aunque sin considerar las restricciones de Neyman y Pearson quienes planteaban la imposibilidad de evaluar la evidencia en un único experimento, ni la flexibilidad de Fisher quien requería la incorporación del conocimiento acumulado sobre el fenómeno en estudio en el proceso de inferencia.

Quien hizo posible la combinación de estas propuestas rivales fue el p-value. Al observar la curva que representa la probabilidad bajo la hipótesis nula de todos los valores posibles de la estadística de trabajo asociada al experimento, es clara la similitud entre la probabilidad de error tipo I (α) y el p-value, al referirse ambos a áreas de la cola de la curva. Sin embargo, mientras el área bajo la curva para α es definida antes del experimento, el área definida para el p-value es establecida sólo después de realizadas las observaciones. Esto permitió que el p-value fuera interpretado como un tipo especial de probabilidad de error tipo I (α), el error tipo I asociado a los datos. El p-value adquirió entonces una aparente doble función, ya que por un lado era una medida de la evidencia contra la hipótesis nula (como lo planteó Fisher) y por otro lado era un tipo especial de probabilidad de error tipo I, el error asociado a los datos. Luego el p-value fue aceptado como una medida de la evidencia en un único experimento que no se oponía la lógica de largo plazo de la Dócima de Hipótesis de Neyman y Pearson, permitiendo la fusión de ambas propuestas[2].

Las propuestas tanto de Fisher como de Neyman y Pearson se refieren principalmente a los estudios experimentales, ya que fueron motivadas por los problemas prácticos a los que se veían enfrentados en esa época los investigadores en sus experimentos 5,6. En los estudios experimentales, el investigador interviene directamente en el estudio, logrando controlar en gran medida la confusión y el sesgo a través de herramientas como la aleatorización y el enmascaramiento. Luego, dado que el p-value representa la probabilidad de obtener resultados iguales o más extremos que el observado asumiendo que no hay diferencia entre los grupos (hipótesis nula), el p-value se transforma en la probabilidad de obtener resultados igual o más extremos que el observado por efecto del azar, ya que el azar es la principal fuente de variabilidad al asumir que no hay diferencia entre los grupos. Por lo tanto, el p-value en los estudios experimentales evalúa el rol del azar en la obtención de los resultados, al estar controlados por el diseño la confusión y los sesgos.

En los estudios epidemiológicos observacionales con muestras probabilísticas, el investigador no está interesado en intervenir directamente, sino que pretende comprender a través de la observación los fenómenos de salud-enfermedad tal como ocurren en la realidad. Por lo tanto, los sesgos y la confusión son siempre explicaciones a evaluar, ya que difícilmente pueden ser controlados completamente en el diseño. En estas circunstancias, el uso e interpretación del p-value se hacen complejos, ya que el azar no es la principal explicación alternativa a evaluar.

K. Rothman define al azar como el "conjunto de etiologías demasiado complejas para nuestro poder de explicación" y justifica el uso de las Dócimas de Significación por el hecho de que "siempre parece haber mayor variabilidad de la que podemos predecir". Sin embargo, también plantea que el usar estas dócimas implica poner irracionalmente en el primer lugar al azar como principal explicación alternativa a evaluar, sin discutir la existencia de otras explicaciones alternativas más relevantes al problema[1].

Esto ha llevado a algunos autores a plantear que el p-value no debe utilizarse en los estudios observacionales, ya que no tendría una interpretación directa y por lo tanto no aportaría información válida para el proceso de inferencia[8]. Otros desaconsejan su uso, planteando que el p-value entrega información confusa y ambigua, ya que mezcla la magnitud del efecto observado con el tamaño del estudio[9].

Probablemente sea esta complejidad en la interpretación, ayudada por la utilización masiva de programas computacionales que permiten obtener el p-value de manera fácil y rápida, lo que explica el uso excesivo e inapropiado del p-value en la literatura epidemiológica. Tal vez la evidencia más clara sobre este fenómeno sea un editorial de la revista Epidemiology que señala que "de todas las herramientas de nuestra disciplina, probablemente no hay ninguna que haya sido más abusada que el p-value"[10].

El p-value se convirtió en una herramienta que llevaba al investigador a evaluar los resultados de manera mecánica, informando de forma dicotómica si los resultados eran significativos o no significativos en base al p-value obtenido, olvidando el proceso descriptivo, reflexivo e interpretativo requerido en la investigación científica.

El reconocimiento de este mal uso del p-value llevó a importantes revistas epidemiológicas a desaconsejar enérgicamente el uso del p-value[7]. Probablemente una de las primeras fuera British Medical Journal, quien en 1986 publicó un artículo titulado "Intervalos de confianza en lugar de p-value: estimación en lugar de dócimas de hipótesis"[11]. Este artículo desaconsejaba el uso del p-value, argumentando que existen mejores herramientas para interpretar los resultados de un estudio, como es el caso de los Intervalos de Confianza.

Los Intervalos de Confianza aparecen entonces como una alternativa al uso del p-value, luego de reconocer que

la utilización del p-value no estaba aportando al proceso de generar información que permitiera acumular conocimientos para mejorar la comprensión de los fenómenos en estudio[12].

Un intervalo con un nivel de confianza de 95%, indica que existe 95% de probabilidad de que el rango de valores del intervalo incluya al parámetro poblacional. Dicho de otro modo, si se realizara una serie de estudios idénticos en diferentes muestras de una misma población y para cada uno de ellos se calculara el correspondiente intervalo de confianza, el 95% de ellos incluiría el valor real en la población[11].

Por lo tanto, los Intervalos de Confianza entregan un rango de valores que parecen ser plausibles para la población de la que proviene la muestra, indicando a la vez la precisión de la estimación. Esta precisión corresponde a la amplitud del intervalo y es función del tamaño del estudio y del nivel de confianza establecido. Luego, los Intervalos de Confianza permiten realizar una estimación de la magnitud del efecto en la misma escala de medición de los datos, informando a la vez sobre la precisión de esta estimación, lo que facilita la interpretación de los resultados.

Además, es posible inferir el resultado de una Dócima de Significación a partir de un Intervalo de Confianza, ya que si el intervalo a 95% de confianza incluye el valor nulo, entonces es posible establecer que el resultado no es estadísticamente significativo a un nivel de α de 5%. Sin embargo, al interpretar los Intervalos de Confianza solamente como Dócimas de Significación para determinar si un resultado es significativo o no, se desprecia parte de la información contenida en él y no se diferenciaría demasiado de la interpretación mecánica y dicotómica del p-value.

Estas características transforman a los Intervalos de Confianza en una herramienta más adecuada para presentar los resultados en los estudios epidemiológicos, ya que entregan más información que el p-value, permitiendo una mejor interpretación de los hallazgos del estudio. Es por esto que se ha planteado que los intervalos de confianza deberían ser el método estándar para presentar los resultados de un estudio, aceptando el uso del p-value como complemento[11].

Conclusión

Desde su origen, el p-value ha sufrido un proceso de transformación conceptualmente controvertido, ya que implicó la combinación de propuestas incompatibles entre si. Esto hace pensar que el p-value es una herramienta problemática, ya que en su desarrollo existen aspectos conceptualmente cuestionables.

En el caso de los estudios epidemiológicos observacionales, a la complejidad conceptual se suma una interpretación especialmente delicada por el rol de la confusión y sesgos como explicaciones alternativas de los resultados a evaluar. Sin embargo su uso es frecuente pero no siempre adecuado, llevando al p-value a ser considerado como la herramienta más abusada en Epidemiología[10].

Este abuso generó un movimiento liderado por los cuerpos editoriales de las principales revistas epidemiológicas tendiente a disminuir el uso del p-value como principal herramienta del proceso de inferencia. Algunas revistas como Epidemiology adoptaron estrictas políticas editoriales que desaconsejaban fuertemente la publicación de artículos que incluyeran el uso de las Dócimas de Significación[13], mientras otras revistas fueron menos estrictas[11]. Los Intervalos de Confianza fueron entonces propuestos como la herramienta de la inferencia más adecuada a utilizar, ya fuera como complemento al p-value o en su reemplazo. Sin embargo, los Intervalos de Confianza también han sido objeto de mal uso al ser interpretados simplemente como Dócimas de Significación, lo que impide superar la interpretación mecánica y dicotómica que inducía el p-value.

Hoy en día, tal vez reconociendo que las metodologías no son tan culpables como quienes las utilizan[14], los llamados son a hacer un uso reflexivo de ellas en lugar de prohibirlas. Cada método tiene características propias que determinan su utilidad en el proceso de generar conocimiento científico. Esto implica que el investigador no sólo debe tener claro los objetivos del estudio que realiza, sino que además debe tener un conocimiento suficiente de las metodologías disponibles para poder determinar cuáles de entre ellas son adecuadas para cumplir los objetivos planteados. Luego, el uso de los diferentes métodos debe responder a las necesidades particulares de cada investigador y no sólo a una recomendación editorial determinada.

Tal vez sea el fomento de la reflexión y del razonar lo que logre disminuir los errores en el uso de las metodologías y en la interpretación de los resultados que tanto daño le hacen al desarrollo de la ciencia.

Referencias bibliográficas

1. Rothman KJ. 1986. Significance questing. Ann Intern Med 15 (3): 445-447.
2. Goodman SN. 1999. Toward evidence-based medical statistics. 1: the p value fallacy. Ann Intern Med 130 (12): 995-1004.
3. Goodman SN. 1993. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 137 (5): 485-496.
4. Sterne JAC, Smith GD, y Cox Dr. 2001. Sifting the evidence {---} what's wrong with significance tests? BMJ 322 (7280): 226-231.
5. Bodmer W, Fisher RA. 2003. Statistician and geneticist extraordinary: a personal view. Int J Epidemiol 32 (6): 938-942.
6. Chiang CL. Jerzy Neyman. Statisticians in history. Disponible en: http://www.amstat.org/about/statisticians/index.cfm?fuseaction=biosinfo&BioID=11 (consultado en diciembre de 2005).
7. Sarria M, Silva L. 2004. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Rev Panam Salud Publica 15 (5): 300-306.
8. Brennan P, Croft P. 1994. Interpreting the results of observational research: chance is not such a fine thing. BMJ 309 (6956): 727-730.

9. Lang JM, Rothman KM, Cann CI. That confounded p-value. Epidemiology 1998; 9 (1): 7-8.
10. The value of p. 2001. Epidemiology; 12 (3): 286.
11. Gardner M, y Altman D. 1986. Confidence intervals rather than p values: estimation rather than hypothesis testing. BMJ 292: 746-750.
12. Clark M. 2004. Los valores p y los intervalos de confianza: ¿en qué confiar? Rev Panam Salud Publica 15 (5): 293-296.
13. Rothman K. 1998. Writing for Epidemiology. Epidemiology 9 (3): 333-337.
14. Weinberg CR. 2001. It´s time to rehabilitate the p-value. Epidemiology 12 (3): 288.