Rev Chil Salud Pública 2006; Vol 10 (1):47-51

TEMAS DE ACTUALIDAD

El Valor P en Epidemiología

CAROLINA MENDOZA(1)

El valor p es una herramienta ampliamente usada en Epidemiología. Forma parte del proceso de inferencia científica, entendiendo por éste el proceso reflexivo de evaluar teorías a partir de observaciones[1]. Sin embargo, a pesar de su uso masivo, no ha estado libre de cuestionamientos, dando pie a interesantes debates entre los expertos.

Este trabajo tiene como objetivo revisar la historia del valor p, desde su origen hasta hoy, destacando los antecedentes que permitan comprender sus usos y abusos en Epidemiología, y así promover el uso adecuado de este importante elemento de la inferencia científica.

ORIGEN Y EVOLUCIÓN DEL *VALOR P*

El valor p tiene su origen en la propuesta de Ronald A Fisher llamada Dócima de Significación. Esta fue planteada alrededor de 1920 y su objetivo era establecer si un resultado era significativo. Para ello, Fisher propuso el valor p o probabilidad de significación, que fue pensado como el indicador que permitiría evaluar la significación. Luego 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]. Así, 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 valor p correspondía a una medida de la discrepancia entre los datos y la hipótesis nula[2-4].

Sin embargo, Fisher fue claro al plantear que este indicador debía ser utilizado con flexibilidad dentro de los procesos complejos de descripción e inferencia de la investigación científica. 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 transformaba al valor p en un indicador informal que no formaba parte de un método formal de inferencia, dejando finalmente su interpretación en manos del investigador[2].

Fisher, quien compartía intereses entre la estadística y la genética, deseaba 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

(1) Programa de Doctorado en Salud Pública. Escuela de Salud Pública. Facultad de Medicina Universidad de Chile. Becaria proyecto MECESUP UCH 0219. caromendoza@med.uchile.cl

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 aquellos que puede cometer el investigador en el proceso de Dócima de Hipótesis, siendo el tipo I referido a la obtención de resultados falsos positivos (plantear que hay diferencia entre los grupos cuando no la hay), mientras que el 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 su definición 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 conceptuales, 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 Valor p 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 valor p en base a las observaciones y finalmente rechazar la hipótesis nula si el valor p 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 valor p. 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 (Figura 1), es clara la similitud entre la probabilidad de error tipo I (•'5f) y el valor p, al referirse ambos a áreas de la cola de la curva. Sin embargo, mientras el área bajo la curva para •'5f es definida antes del experimento, el área definida para el valor p es establecida sólo después de realizadas las observaciones. Ello permitió que éste fuera interpretado como un tipo especial de probabilidad de error tipo I

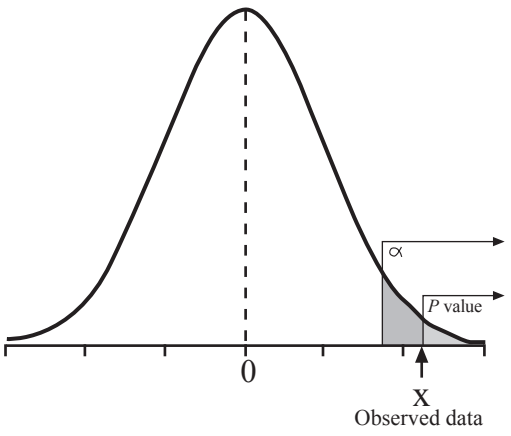

Figura 1. Curva que representa la probabilidad de todos los resultados posibles bajo la hipótesis nula. (X: estadística de trabajo)

(•'5f), el error tipo I asociado a los datos. El valor p 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, era un tipo especial de probabilidad de error tipo I, el error asociado a los datos. Luego el valor p 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].

EL VALOR P EN EPIDEMIOLOGÍA

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 valor p representa la probabilidad de obtener resultados iguales o más extremos que el observado asumiendo que no hay diferencia entre los grupos (hipótesis nula), el valor p se transforma en la probabilidad de obtener resultados igual o más extremos que el observado por efecto del azar, ya que éste es la principal fuente de variabilidad al asumir que no hay diferencia entre los grupos. Por lo tanto el valor p 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. Ante ello, 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 valor p 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 valor p 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 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 valor p de manera fácil y rápida, lo que explica su uso excesivo e inapropiado 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 valor p"[10].

El valor p 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 valor p obtenido, olvidando el proceso descriptivo, reflexivo e interpretativo requerido en la investigación científica.

El reconocimiento de este mal uso llevó a importantes revistas epidemiológicas a desaconsejar enérgicamente el uso del valor p[7]. Probablemente una de las primeras fuera British Medical Journal, quien en 1986 publicó un artículo titulado "Intervalos de confianza en lugar de valor p: estimación en lugar de dócimas de hipótesis"[11]. Este desaconsejaba su uso, argumentando que existen mejores métodos

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 valor p, luego de reconocer que su utilización 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 ese porcentaje 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 se calculara el Intervalo de Confianza, el 95% incluiría el valor real en la población[11]. Por ello, 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 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 éste alcanza a 95% de confianza incluye el valor nulo, entonces es posible establecer que el resultado no es estadísticamente significativo a un nivel de •'5f 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 valor p.

Estas características transforman a los Intervalos de Confianza en una alternativa más adecuada para presentar los resultados en los estudios epidemiológicos, ya que entregan más información que el valor p, 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 valor p como complemento[11].

CONCLUSIÓN

Desde su origen, el valor p ha sufrido un proceso de transformación conceptualmente controvertido, ya que implicó la combinación de propuestas incompatibles entre sí. Esto hace pensar que es un elemento problemático de la inferencia, 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 valor p 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 su uso como principal método 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 fueron menos estrictas[11]. Los Intervalos de Confianza fueron entonces propuestos como el método de la inferencia más adecuada a utilizar, ya fuera como complemento al valor p 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 valor p.

Hoy, tal vez reconociendo que los métodos no son tan culpables como quienes los utilizan[14], los llamados son a hacer un uso reflexivo de ellos en lugar de prohibirlos. 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 los métodos disponibles para poder determinar cuales de entre ellos son adecuados 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 los distintos métodos y en la interpretación de los resultados que tanto daño le hacen al desarrollo de la ciencia.

REFERENCIAS

1.- ROTHMAN KJ. Significance questing. Ann Intern Med 1986, 15(3): 445-47.

2.- GOODMAN SN. Toward evidence-based medical statistics. 1: the p value fallacy. Ann Intern Med 1999, 130(12): 995-1004.

3.- GOODMAN SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993, 137(5): 485-96.

4.- STERNE JAC, SMITH GD, y COX DR. Sifting the evidence {---} what's wrong with significance tests? BMJ 2001, 322(7280): 226-31.

5.- BODMER W. RA Fisher. statistician and geneticist extraordinary: a personal view. Int J Epidemiol 2003, 32(6): 938-42.

6.- CHIANG CL. Jerzy Neyman. Statisticians in history. Disponible en: (consultado en diciembre 2005).

7.- SARRIA M, y SILVA L. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Rev Panam Salud Pública 2004, 15(5): 300-06.

8.- BRENNAN P, y CROFT P. Interpreting the results of observational research: chance is not such a fine thing. BMJ 1994, 309(6956): 727-30.

9.- LANG JM, ROTHMAN KJ, y CANN CI. That confounded p-value. Epidemiology 1998; 9(1): 7-8.

10.- THE VALUE OF P. EPIDEMIOLOGY 2001, 12 (3): 286.

11.- GARDNER M, y ALTMAN D. Confidence intervals rather than p values: estimation rather than hypothesis testing. BMJ 1986, 292: 746-50.

12.- CLARK M. Los valores p y los intervalos de confianza: ¿en qué confiar? Rev Panam Salud Pública 2004, 15(5): 293-96.

13.- ROTHMAN K. Writing for Epidemiology. Epidemiology 1998, 9(3): 333-37.

14.- WEINBERG CR. It´s time to rehabilitate the p-value. Epidemiology 2001, 12(3): 288.