REVISTA ARGENTINA
DE TERAPIA INTENSIVA
2018 - 35 Nº 3

El valor p. Interpretación, orígenes y su utilización actual

Ladislao Diaz Ballve,* Fernando Ríos**

* Gabinete de Apoyo para la Producción de Información Hospitalaria (GAPIH), Hospital Nacional "Prof. Alejandro Posadas", El Palomar, Buenos Aires.
Cátedra de Metodología de la Investigación Científica, Universidad Nacional de la Matanza, San Justo, Buenos Aires

** Servicio de Terapia Intensiva, Hospital Nacional "Prof. Alejandro Posadas", El Palomar, Buenos Aires.
Servicio de Terapia Intensiva, Sanatorio Las Lomas, San Isidro, Buenos Aires

Correspondencia:
Lic. Ladislao Diaz Ballve
pablodiazballve@yahoo.com.ar

Los autores no declaran conflictos de intereses.

El valor p (*p value*) se usa en todas las pruebas estadísticas, desde las pruebas t hasta el análisis de regresión. Todo profesional de la salud sabe que se usa el valor p para determinar la significación estadística en una prueba de hipótesis. De hecho, los valores p, con frecuencia, determinan qué estudios se publican y qué proyectos obtienen financiación. A pesar de ser tan importante, el valor p es un concepto difícil y muchos, a menudo, lo interpretan incorrectamente. ¿Cómo debemos interpretar los valores p?

En esta nota metodológica, intentaremos ayudar a comprender los valores p de una manera más intuitiva y evitar que el lector caiga en errores comunes de interpretación.

El valor p debe entenderse como la proporción de veces que el estadístico de contraste (media, desviación estándar, varianza, proporción, etc.) toma un valor más extremo (diferente) que el resultado del experimento realizado, puede entenderse como la probabilidad de encontrar un valor del estadístico de contraste más alejado o más extremo que lo observado en la muestra actual, si repitiéramos el experimento en iguales condiciones de forma infinita.[1-3]

Esta definición no suele aclarar mucho el panorama y esto se debe a que el valor p no es fácil de interpretar, sin entender qué busca probar específicamente. Para ello, es necesario introducirnos en el concepto de prueba de significación de la hipótesis nula.[4]

Creemos que, para comprender qué es el valor p, primero, se debe conocer la función de la hipótesis nula.

Prueba de significación de la hipótesis nula

Para empezar, pensemos en una comparación entre dos muestras (como podrían ser los grupos de un ensayo clínico) donde se desea medir el efecto de una variable independiente y se estipula *a priori* qué diferencia entre los grupos (diferencia clínica mínimamente importante) indica que un grupo es diferente del otro debido al factor intervención. Dicho factor podría ser la efectividad de un nuevo medicamento u otra intervención que supone beneficios. Desafortunadamente para los investigadores, siempre existe la posibilidad de que no haya ningún efecto, es decir, de que no haya diferencia entre los grupos (intervención y control). Esta falta de diferencia o no diferencia se debe hacer explícita al diseñar el estudio, como hipótesis nula o hipótesis de no diferencia.

Imaginemos un experimento para un medicamento que sabemos que es totalmente ineficaz. La hipótesis nula nunca podrá ser rechazada: no hay diferencia entre los grupos del estudio y lo mismo ocurre en la población. A pesar de que la hipótesis nula no puede ser rechazada, es muy posible que los datos de la muestra sean diferentes; esta incongruencia se explica debido a la presencia de error aleatorio (error debido al muestreo). De hecho, es extremadamente improbable que los grupos de muestra sean exactamente iguales.

Ahora sí, podríamos decir que, en la actualidad, el valor p suele interpretarse erróneamente como la probabilidad de que el valor observado en la muestra refleje el valor poblacional.

Orígenes del valor p

Mirando un poco hacia atrás en la historia, nos remontamos a mediados de la década de 1920 y hasta la década de 1930, cuando Ronald Fisher, y luego Egon Pearson y Jerzy Neyman establecieron las bases de lo que hoy conocemos como prueba de significación de la hipótesis nula.[4,5] Si bien existen diferencias sustanciales entre el planteo de Fisher y lo expuesto por Neyman-Pearson, actualmente, el proceso de decisión a través de la prueba de significación de la hipótesis nula es una variante que representa "una mezcla o híbrido" de los dos enfoques.[6]

Fisher fue el primer matemático que propició la utilización de la prueba de significación de la hipótesis nula para la toma de decisiones. Si asumimos lo expuesto por Fisher, entonces, la hipótesis nula puede o no ser rechazada (es decir, hay diferencias entre los estadísticos de contraste o no hay diferencias, respectivamente). El valor p, siguiendo a Fisher, es la probabilidad de encontrar un resultado igual o más alejado que el hallado en el estudio actual. Nada podemos decir de la veracidad de la hipótesis nula, plantea Fisher, solo es demostrable matemáticamente su falsedad con un grado de probabilidad que es arbitrario y que deberá ser definido por el investigador durante el diseño del experimento o prueba, aunque sabemos que, por convención, se utiliza frecuentemente el valor p <0,05 como significativo.

Queda claro que, para Fisher, el valor p es un criterio para definir la falsedad de la hipótesis nula; en la estadística de Fisher, el valor p obtenido aporta un grado de significación, cuanto más pequeño es el valor p, menor será la probabilidad de que la hipótesis nula sea verdadera. De esta manera, Fisher propone que valores p por debajo de 0,05 deberían ser interpretados como criterios de evidencia en contra de la hipótesis nula, pero no de forma absoluta. Por ejemplo, un valor p de alrededor de 0,05 no podría llevar ni al rechazo ni a la aceptación de la hipótesis nula, sino a la decisión de realizar otro experimento, que rechace o acepte la hipótesis nula, pero a medida que la zona de aceptación de la hipótesis nula se hace más pequeña (valor p más bajo), la evidencia a favor del rechazo o en contra de la hipótesis nula es cada vez más contundente (Figura 1).

Las ideas de Fisher, si bien propiciaron el comienzo de toma de decisiones mediante los tests de hipótesis, no son las que actualmente utilizamos en la estadística inferencial. Para llegar a esto, primero, debemos conocer otra propuesta que es la de Neyman-Pearson.

Las pruebas de hipótesis de Neyman-Pearson difieren sustancialmente del enfoque de Fisher. En primer lugar, se debe fijar un nivel de significación que, en general y por convención, se asigna entre 0,05 y 0,01 (el valor p fijado *a priori* y la magnitud arbitraria no difieren de lo planteado por Fisher), este valor servirá para definir el rechazo de la hipótesis nula comparando el valor p obtenido de los datos con aquel fijado previamente (si p <0,05 se rechaza la hipótesis nula, si p ≥0,05 no se rechaza la hipótesis nula), aquí ya se observa una diferencia entre los enfoques de Fisher y Neyman-Pearson. En el primero, el valor p debe ser interpretado según su valor para definir la posibilidad de rechazo; en cambio, para Neyman-Pearson, solo se tiene que comparar el valor p con respecto al valor de significación (valor de significación

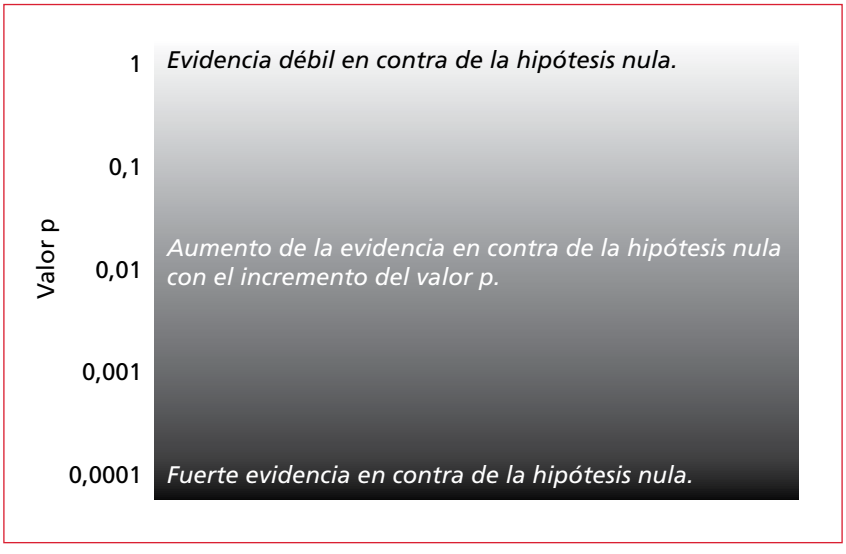

Figura 1. Interpretación del valor p sugerida por las publicaciones médicas.

estadística, por convención p <0,05) o como los autores lo denominaron valor de alfa.

Por otro lado, el enfoque de Neyman-Pearson se distancia aún más de lo expuesto por Fisher cuando los autores proponen que, además de la hipótesis nula, se debe formular, de la forma más precisa posible, una hipótesis alternativa, entonces, a la luz de estos autores, si la hipótesis nula cae en zona de rechazo (p ≤0,05), deberíamos aceptar la hipótesis alternativa o de diferencias.

Neyman y Pearson introducen, además de las pruebas de hipótesis, los conceptos de error tipo I, error tipo II y potencia del estudio, que actualmente son ítems obligatorios en cualquier programa de estadística inferencial (Tabla 1).

Las diferencias entre ambas teorías estadísticas no están aún saldadas, razón por la cual, además de otras circunstancias de costumbre o quizás de conveniencia editorial, las publicaciones científicas sobre salud decidieron unilateralmente utilizar este híbrido antes comentado, donde se toman las pruebas de significación estadística con los conceptos propuestos por Neyman-Pearson (nivel de confianza, error tipo I, error tipo II, potencia), pero a su vez, se publica el valor p exacto y no como un mecanismo fijo de aceptación y rechazo. Es decir que se le atribuye cierto grado de jerarquía al valor p encontrado, como defendía Fisher.[7]

A fin de colaborar con la interpretación correcta del valor p para aquellos que se inicien en la lectura crítica, podemos ahora definir una serie de ítems para tener en cuenta a la hora de su interpretación, pero sabiendo que existe cierta controversia, no solo estadística, sino también metodológica.

A modo de ejemplo, observemos los resultados del estudio de Kollef[8] (Tabla 2), el autor utiliza un modelo de análisis de regresión logística para definir qué factores se relacionaban con el desarrollo de neumonía asociada a la vía aérea. Como corresponde, durante el diseño, el autor debió elegir el valor p para aceptar o rechazar la hipótesis nula o de no diferencias que, por consenso, suele establecerse en p <0,05. Supongamos que el autor decide elegir el valor p para rechazo más alto, por ejemplo, p <0,1. A partir de esta elección, los editores, los pares y los lectores mirarán con escepticismo esta decisión y deberá estar firmemente sustentada o, de lo contrario, no hubiésemos conocido el estudio por haber sido rechazado para su publicación o mejor aún si se publica sabríamos la respuesta para esta elección y podríamos criticarla o apoyarla.

TABLA 1
Errores de interpretación de la prueba de significación de la hipótesis nula, según la propuesta de Neyman-Pearson[5]

Resultado del experimento-estudio	***"La Verdad"***	
	La hipótesis nula es verdadera *El tratamiento no funciona*	***La hipótesis nula es falsa*** *El tratamiento sí funciona*
Rechazo la hipótesis nula	Error tipo I (α o Confianza)	Decisión correcta
Acepto la hipótesis nula	Decisión correcta	Error tipo II (β o Potencia)

TABLA 2
Variables asociadas independientemente con neumonía asociada a la vía aérea mediante análisis de regresión logística

Variable	OR ajustado	IC95%	p
OSFI ≥3	10,2	4,5-23,0	<0,001
Edad ≥60 años	5,1	1,9-14,1	0,002
Antibióticos previos	3,1	1,4-6,9	0,004
Posición de la cabeza*	2,9	1,3-6,8	0,013

OR = *odds ratio* (cociente de posibilidades), IC = intervalo de confianza; OSFI = *organ system failure index*.
* Durante las primeras 24 horas de ventilación mecánica.
Tomada de Kollef MH. Ventilator-associated pneumonia. A multivariate analysis. JAMA 1003; 270(16): 165-170.

En la Tabla 2, se presentan los resultados primarios del estudio de Kollef. Allí se observa que las variables analizadas a través de la regresión logística nos muestran como el estadístico de contraste, el cociente de posibilidades (*odds ratio*), que puede leerse como las chances de desarrollar neumonía asociada a la vía aérea, si la variable se encuentra presente en la forma que está expresada. Se informa, además, el intervalo de confianza (IC) y el valor p.

El valor p, entonces, debe entenderse en este ejemplo, de la siguiente manera:

- Para el caso de la variable edad ≥60 años, se observa un valor p = 0,002, lo cual indica que la variable edad ≥60 años en la muestra aumenta en 5,1 veces la posibilidad de tener neumonía asociada a la vía aérea y la posibilidad de encontrar, mediante este modelo estadístico, un resultado más extremo (más chances de neumonía asociada a la vía aérea) tiene una probabilidad del 0,02%.
- En la Tabla 2, se informan también los IC95% que, si bien escapa a este escrito, su explicación nos sirve para entender que si repetimos infinitas veces este estudio, solo el 95% de ellos tendrán entre sus IC el valor del parámetro poblacional. El uso de IC nos facilita entender cuál es el verdadero valor del parámetro poblacional y su relación con los datos del estudio. Cuando un valor p es significativo, si además se nos informa el IC, esto nos permite conocer entre qué valores el estadístico estudiado podría representar el parámetro de la población, pero, al ser un contexto frecuentista, siempre debemos pensar que esta inferencia solo es correcta en el 95% de las repeticiones del estudio en idénticas condiciones.[9]

Errores frecuentes en la interpretación del valor p

- El valor p no nos da una probabilidad de que la hipótesis nula sea cierta, solamente nos permite definir su aceptación o rechazo, confrontándolo con el valor de alfa establecido (posibilidad de cometer error tipo I) .
- El valor p no nos informa la probabilidad de que la hipótesis alternativa sea cierta; de la misma manera que lo anterior, si el valor p es menor que el valor definido como valor de rechazo (generalmente <0,05) significa que, a partir de los datos obtenidos, no podemos aceptar la hipótesis de no diferencias. Pero esto no significa que el valor p asigne una probabilidad de que la hipótesis alternativa sea cierta.
- El valor p con magnitudes muy bajas no es una medida de efecto de la variable estudiada, es decir, el valor p muy bajo no demuestra "mayor efecto" que un valor p cercano a 0,05.[10]
- Tampoco debe entenderse que la significación estadística de una determinada prueba es evidencia de efecto, ni lo contrario, no es evidencia de no efecto la falta de significación estadística (p ≥0,05*).*[11]

Para concluir, el valor p surgió como un recurso matemático, una probabilidad de que los datos de una muestra sean coherentes con aquello que el investigador cree y diseña para fundamentar su creencia (experimento) y pasó a ser un valor que alcanzar y que suele relacionarse erróneamente con fuerza de verdad llegando incluso a superar su importancia a la magnitud de los resultados encontrados. Es importante que todo aquel que se inicia o quienes utilizan literatura científica desde hace rato conozcan y sepan interpretar este recurso como lo que realmente representa.

Bibliografía

1. Dawson B, Trapp RG. *Basic & clinical biostatistics*, 4th ed. New York: McGraw-Hill, Medical Pub. Division.; 2004: 438.
2. Dawson GF. *Interpretación fácil de la bioestadística*, London: Elsevier Health Sciences; 2009: 208.
3. Wasserstein RL, Lazar NA, Wasserstein RL, Lazar NA, Asa T. The ASA 's Statement on p -Values : Context, Process, and Purpose. Am Stat 2016; 70(2): 129-133.
4. Leenen I. La prueba de la hipótesis nula y sus alternativas: revisión de algunas críticas y su relevancia para las ciencias médicas. Inv Ed Med 2012; 1(4): 225-234.
5. Yáñez S. La estadística una ciencia del siglo XX. R.A. Fisher, El Genio. Rev Colomb Estadística 2000; 23(2): 1-14.
6. Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ 2001; 322: 226-231.
7. Milton SJ. *Estadística para biología y ciencias de la salud*, 3ª ed. Madrid: McGraw-Hill Interamericana de España S.L.; 2007: 722.
8. Kollef MH. Ventilator-associated pneumonia. A multivariate analysis. JAMA 1993; 270(16): 1965-1970.
9. Castañeda JA, Fabián J. Una mirada a los intervalos de confianza en investigación. Rev Colomb Psiquiatr 2004; XXXIII: 193-201.
10. Badenes-Ribera L, Frías-Navarro D, Monterde-I-Bort H, Pascual-Soler M. Interpretation of the p value: A national survey study in academic psychologists from spain. Psicothema 2015; 27(3): 290-295.
11. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008; 45(3): 135-140.