XXV Curso de actualización
Medicina Interna
Innovación, humanidad y excelencia
2025

UNIVERSIDAD DE ANTIOQUIA
Facultad de Medicina

El valor *p* en la literatura médica: más allá del "estadísticamente significativo"

Carlos José Atencia Flórez
Especialista en Medicina Interna
Magíster en Epidemiología Clínica
Facultad de Medicina
Universidad de Antioquia

Juan Pablo Bedoya Gallego
Residente de Medicina Interna
Facultad de Medicina
Universidad de Antioquia

Guía para el aprendizaje

a. ¿Qué debes repasar antes de leer este capítulo?

- Principios básicos de la inferencia estadística.
- Pruebas de hipótesis en estudios clínicos: definiciones de hipótesis nula (H_0) e hipótesis alternativa (H_a).
- Definición del valor de *p*.
- Significancia estadística y umbrales de significancia, diferenciándola de la significancia clínica.

b. ¿Cuáles son los objetivos del capítulo?

- Describir los orígenes, la evolución histórica y la definición actual del valor *p* dentro del marco de la prueba de significancia de la hipótesis nula (PSHN).
- Analizar las limitaciones del valor *p*, incluyendo su sensibilidad al tamaño de la muestra, su incapacidad para evaluar la relevancia clínica de un efecto y la arbitrariedad del umbral de significancia estadística.
- Identificar las interpretaciones erróneas del valor *p* y su papel en la interpretación equivocada de la evidencia científica.
- Explorar las recomendaciones de la comunidad científica para un mejor uso e interpretación del valor *p*.
- Discutir diversas alternativas o complementos al uso exclusivo del valor *p*.

Viñeta clínica

Un equipo de investigadores lleva a cabo un estudio de cohorte prospectivo para evaluar la eficacia del fármaco semaglutide en la reducción del peso corporal (PAS) en pacientes con obesidad. Se incluyeron 1000 pacientes, divididos en dos grupos: uno con semaglutide y el otro con cirugía bariátrica. El seguimiento se realiza durante 8 semanas.

Al final del estudio, se miden los cambios promedio del peso en kg en ambos grupos. Los resultados muestran un peso promedio de 70 kg con semaglutide y 105 kg en el grupo de cirugía bariátrica. Se realiza una prueba estadística para comparar las medias de ambos grupos, obteniendo un valor *p* de 0.0005. ¿Cómo se interpreta la significancia estadística en este caso?

Introducción

En la práctica de la medicina basada en la evidencia, la interpretación de los resultados de la investigación clínica es esencial para obtener conclusiones sobre la eficacia de las intervenciones, a través de la evaluación de la significancia estadística. El valor *p* ha sido históricamente una de las herramientas más utilizadas para esta evaluación, y gracias a sus resultados se han logrado cambios significativos en la práctica clínica. Sin embargo, en los últimos años, su papel como parámetro estándar de la significancia estadística ha sido objeto de escrutinio, así como su relación con la falta de replicabilidad de los resultados obtenidos en estudios iniciales (1, 2).

En este artículo, nos proponemos analizar de manera crítica el valor *p* en la literatura, abordando su definición, origen, limitaciones prácticas, interpretaciones erróneas comunes y las alternativas propuestas para una interpretación más adecuada de los resultados de la investigación.

¿Qué es el valor *p*?

Desde un punto de vista matemático, el valor *p* no es un axioma, de hecho, tampoco un teorema. Es decir, no es una verdad evidente por sí misma y tampoco es un resultado al cual llegamos después de aplicar una demostración matemática rigurosa. Tampoco es un número importante ni trascendental como lo es *pi* para la geometría, *e* para el cálculo o la cantidad imaginaria *i* para el análisis complejo. El valor *p* fue arbitrariamente establecido hace más de 100 años por un tabacalero que argumentaba que en ausencia de un diseño experimental era imposible demostrar una asociación entre tabaco y cáncer de pulmón, y que trabajaba con cultivos (agricultura, no pacientes), que era extremadamente beligerante (como todos los genios de Cambridge, *v.g.* Isaac Newton) y que además era un prodigioso matemático y padre de las pruebas de significación estadística: Ronald Fisher.

Para entrar en detalle del análisis del valor *p* necesitamos algunos conceptos sobre estadística frecuentista y sobre modelos. La inferencia estadística es el proceso analítico mediante el cual, a partir de una selección de una muestra aleatoria (cosa que casi nunca se cumple en investigación biomédica) y según un tamaño determinado de esa muestra, y partiendo de una hipótesis, utilizamos métodos y modelos para estimar estadísticos y hacer inferencias sobre el parámetro en la población general, con base en esta muestra representativa (Walpole). Adicionalmente, hemos de suponer que las muestras son independientes (la realización efectiva de la variable peso no depende de otra de las mediciones).

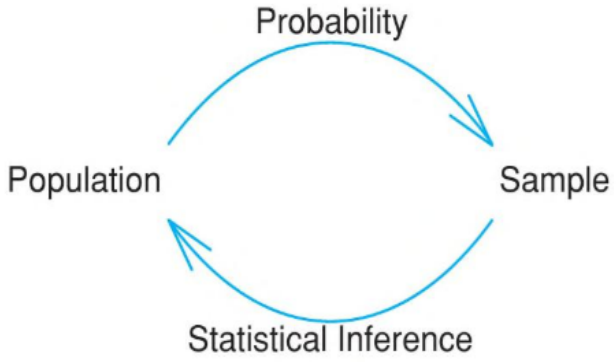

Figura 1.

Para hacer esta inferencia dependemos del tamaño de la muestra, pero esto lo facilitan las leyes de la naturaleza. Como Richard Feynman y Einstein lo han hecho notar, el universo sigue reglas simples que generan comportamientos complejos y que son explicados por principios matemáticos que se repiten análogamente a escala. Solo nótese cómo la tercera Ley de Newton de gravitación universal se parece a la Ley de Coulomb con asombrosa simplicidad. Aquí, las variables aleatorias (que toman el valor de cualquier número real) son las masas y las cargas, y G y K son los coeficientes (constantes universales), valores fijos que se calcularon experimentalmente.

Ambos son modelos matemáticos cuyo error es despreciable (la variabilidad del modelo está únicamente especificada por las variables medidas en él). Finalmente, fíjese en un supuesto adicional: las distancias entre las masas y cargas deben ser mayores a 0.

$$F_g = G\frac{m_1\, m_2}{r^2}\,, F_e = K_\varepsilon \frac{q_1\, q_2}{d^2}\,, r > 0\; y\; d > 0$$

Figura 2.

¿Pero qué tiene que ver la física con la investigación en medicina y la medicina basada en la evidencia y con el valor *p*? Mucho. Algunos de estos modelos los hemos utilizado para la bioestadística y la investigación en nuestro campo. Por ejemplo: el peso de los pacientes adultos (>18 años) también sigue reglas más o menos simples. De hecho, si nuestro tamaño muestral es lo suficientemente grande, el peso (una variable cuantitativa continua en el eje X) de un adulto sigue la siguiente distribución de probabilidad (ver Figura 3).

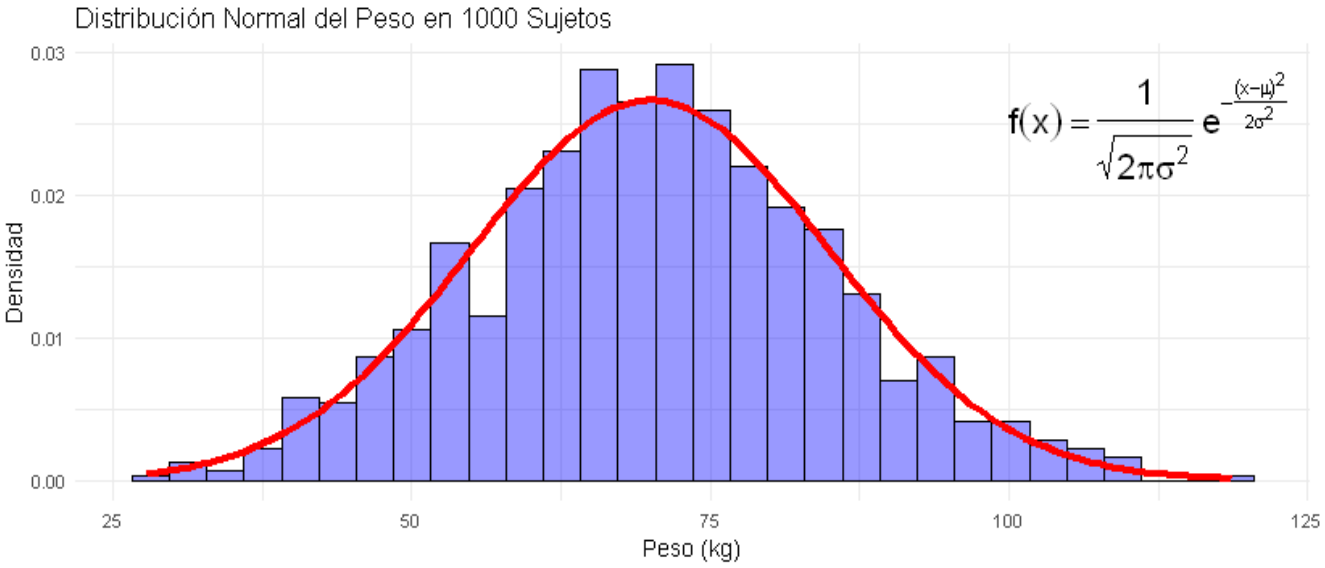

Figura 3.
Elaboración propia, generado en R

Ignore toda la expresión matemática y concentrémonos en el término del exponente (ya que 1, 2, *pi* y *e* son constantes, no le vamos a prestar atención). Si esta función la integramos en un intervalo definido, calculando el área debajo de la curva obtenemos la probabilidad total (que debe ser de 1, ya que las probabilidades de la suma de todos los sucesos —pesos, X para este caso— van de 0 a 1). Entonces, con solo calcular este estadístico (una función de esa variable aleatoria peso observado en la muestra y de la cual conocemos sus cantidades) podemos calcular el valor *p*. Pero antes de esto mencionemos el papel que tienen los parámetros μ y σ, que son la media y la varianza, y de tendencia central y dispersión de los datos. Es decir, nos da el promedio de todos los pesos y cuán dispersos (o variación) tienen esos pesos. En este caso: μ= 70 kg y σ^2= 15 kg. Gráficamente, están representados por la línea vertical azul y línea punteada amarilla de la Figura 4.

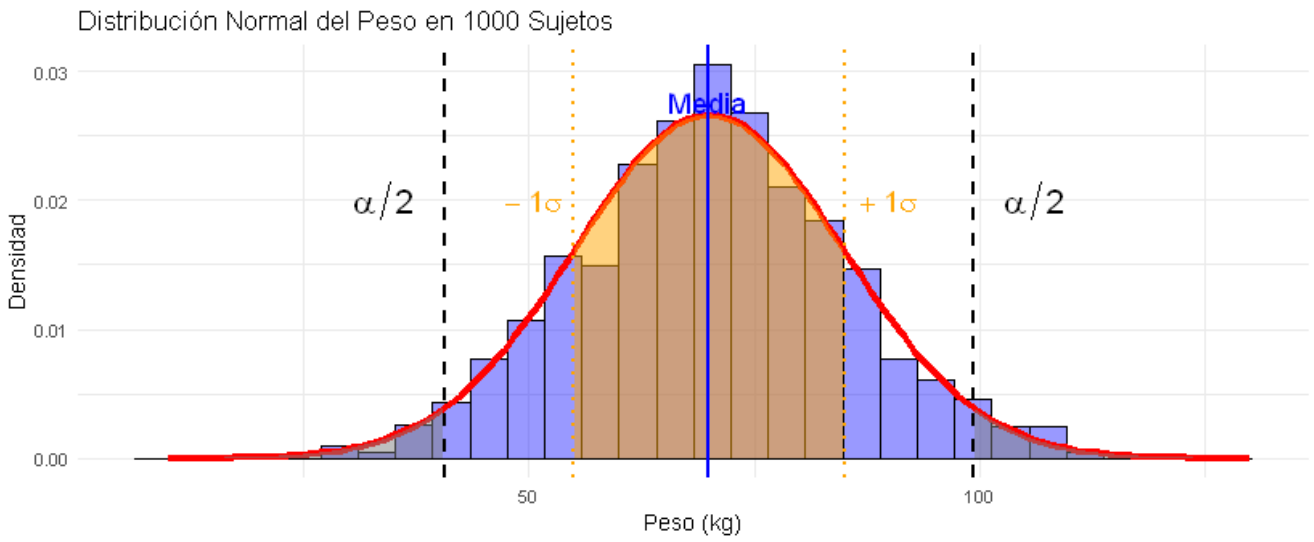

Figura 4.
Elaboración propia, generado en R.

Procedimiento de cálculo de la *p*

Los datos

Usemos los datos de la viñeta clínica: el estudio observacional de cohorte de 1000 pacientes obesos determinó que la reducción de peso usando el medicamento semaglutide en pacientes obesos es de 70 kg, con desviación estándar de 15 kg, y que el uso adicional de cirugía bariátrica generó una disminución en promedio de 105 kg de peso. ¿Son estos resultados estadísticamente significativos?

Los supuestos

Todos los modelos tienen supuestos, que son condiciones previas bajo las cuales "funcionan", y el modelo paramétrico de comparación de medias de variables para variables cuantitativas es el estadístico Z. Este modelo funciona si se cumplen los supuestos de distribución normal de μ y X, de que sus varianzas son iguales (σ) y de que existan al menos más de 30 observaciones *n*. Hemos dado por sentado que se cumplen, pero esto no siempre se cumple ni se explica en los estudios.

La hipótesis

La hipótesis H_0 es una proposición cuantitativa sobre el parámetro, y ya lo tenemos: H_0: X = la disminución del peso es de 105 kg.

Nivel de significancia

El nivel de significancia alfa (α) es escogido por el investigador como parte del diseño antes de la recolección de los datos. Si la hipótesis nula H_0 es verdad, α es la probabilidad con la cual vamos a rechazar la hipótesis nula. Note que, si hay incertidumbre con relación con el parámetro, una buena idea es usar una prueba de 2 colas (y dividir $\alpha/2$).

Test estadístico

Es el estimador, regla o algoritmo matemático para calcular el estadístico. Con este, al usar la integración de la función normal, computamos una cantidad de los datos de una probabilidad que es el valor *p*.

$$estadistico\ Z^* = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} = 84{,}33\ \ valor\ de\ p < 0{,}0005$$

Figura 5.

¿Y tenemos que integrar? No, eso lo hacen las computadoras. ¡Ah!, y nuestro amigo Fisher ya lo hizo por nosotros. Fíjese al final de los libros de estadística (al menos los decentes) que hay unas tablas que resultan de integrar numerosas distribuciones de probabilidad (gaussiana o normal, chi cuadrado, *student*, entre otras). Estas tablas fueron calculadas por Fisher.

La regla de decisión

La regla de decisión se basa en que si el valor de *p* es inferior a α, entonces podemos rechazar la hipótesis nula. Observe que, en este caso, 105 está contenido en el área sombreada de $\alpha/2$ a la derecha.

$$p = Pr\ Pr\ [Z^* < Z|\ H_0] < \alpha$$

Figura 6.

¿Pero qué ignoramos al solo basarnos en el valor *p*?

Este es precisamente el problema del valor *p*. El modelo que escogimos se aplicó en un diseño no experimental y por ello es factible que exista un conjunto de variables no controladas (el problema de la confusión). Es decir, el resultado pudo ser efecto de variables de confusión no medidas o mal medidas. ¿Qué tal si a los pacientes de cirugía bariátrica también se les ofreció el acompañamiento psicológico, nutricional y de actividad física, y a los pacientes del medicamento no? Los resultados hubiesen podido cambiar si ajustamos por estas variables de confusión. Entonces el modelo utilizado no toma en cuenta las múltiples variables que son causas comunes de la intervención (exposición) y del desenlace, y confunden el fenómeno. Inclusive, modelos como los de la física necesitan de varios tipos de variables (recuerde el ejemplo de la fuerza eléctrica y gravitacional, ambas toman en cuenta el cuadrado de la distancia, pero mire que las cargas y las masas también se deben tener en cuenta. Este modelo es multivariable).

Otro problema que parece desapercibido por la *p* son los errores sistemáticos o sesgos. Estos son cometidos al seleccionar la población y al medir las variables involucradas en el estudio. Suponga que los pacientes fueron seleccionados de una consulta de cirugía bariátrica o de un consultorio de atención primaria, o de un programa de entrenamiento y acondicionamiento. ¿Cuál cree, de estos escenarios, podría representar un sesgo potencial al momento de su inferencia estadística la población de estudio de su interés? Los criterios de inclusión establecidos al principio del estudio tienen mucho que ver con a qué tipo de sujetos podré aplicar los resultados del estudio. Por otro lado, suponga que el peso se midió con una báscula mal calibrada o con básculas diferentes o en momentos distintos, antes o después de la intervención. En este caso, nuestras observaciones y mediciones estarán sesgadas y esto

representa un error en la medición (por muy bajo que sea nuestro valor de la *p*).

Ya mencionamos que es inusual en investigación clínica tomar muestras probabilísticas (aleatorias, por estratos o *clusters*), y lo es más asegurar que el muestreo fue independiente. ¿Qué hubiera ocurrido si cada paciente invita a su familia o amigos obesos a unirse al estudio? ¿No hay razones para pensar que las mediciones de los amigos y familiares puede estar correlacionada? Estas correlaciones, al introducirlas a los modelos estadísticos, crean dependencias que deben ser tratadas apropiadamente para evitar subestimar los errores del modelo.

Existen muchas pruebas de significación estadística (prueba *t*, prueba F), y aunque no es la idea dar una taxonomía completa de estos, existen test exactos (test exacto de Fisher), asintóticos (test de Wald), bidireccionales (prueba *t*) o unidireccionales (test de Chi^2), que rechazan la hipótesis si el valor $p > 0{,}05$ (Kolmogorov-Smirnoff), que rechazan la hipótesis si el valor $p < 0{,}05$ (prueba Z), y todos ellos tienen supuestos diferentes, hipótesis nulas diferentes y se computan de una manera distinta. Esto, además de aumentar la dificultad en su correcta interpretación, no se explica en muchos modelos estadísticos que se publican.

Los valores *p* pueden cambiar si en estudios con muestras pequeñas suponemos pequeños cambios en los desenlaces. El texto clásico de ensayos clínicos de Piantadosi ofrece un ejemplo hipotético con las variables dicotómicas y el uso de la prueba exacta de Fisher en las tablas de contingencia. Como se observa, suponiendo un control perfecto de las variables de confusión, el efecto de la exposición a ondas electromagnéticas (EM) y riesgo de leucemia, el OR de ambos estudios es de 0,083 (IC95 % 0,001-0,86) y 0,083 (IC95 % 0,001-0,93), pero con pequeños cambios en los desenlaces, el valor de la *p* varió casi un 66 %.

Tabla 1.

	Expuesto a ondas EM	No expuesto
Leucemia	1	7
No leucemia	13	7

Valor *p* 0,0328

Tabla 2.

	Expuesto a ondas EM	No expuesto
Leucemia	1	6
No leucemia	13	6

Valor *p* 0,0261

Y esto se ha verificado en diversos metaanálisis de ensayos clínicos sobre la fragilidad de los resultados basados en significancia estadística (3).

Finalmente, y para nuestros pacientes, ¿son los resultados clínicamente relevantes? Qué tal si además de la reducción del peso los pacientes, estos se siguieron y la mortalidad fue mayor en el grupo de la cirugía bariátrica. Es decir, la intervención redujo mayor peso en comparación con el medicamento, pero por complicaciones del perioperatorio (infecciones, sangrado y tromboembolia pulmonar) 5 % de los pacientes de cirugía murieron en comparación con 1 % de los pacientes del brazo de medicamento. Esta es información relevante que nos permite reflexionar entre lo estadísticamente significativo y lo clínicamente significativo.

Correcta interpretación y orígenes del valor *p*

El valor *p* se define como la probabilidad de obtener un resultado tan extremo como el observado (o más extremo) si la hipótesis nula fuera verdadera (2, 4). Según la American Statistical Asociation, también puede entenderse como la probabilidad bajo un modelo estadístico especificado de que un compendio estadístico de los datos (por ejemplo, la diferencia de medias muestrales entre dos grupos comparados) sea igual o más extremo que su valor observado (5). En otros términos, puede considerarse como aquella evaluación de la compatibilidad entre los datos observados y lo que se esperaría si el modelo estadístico (todos los supuestos utilizados para el cálculo del valor *p*, no solo la hipótesis nula) fuera correcto (6).

Entendiendo que el valor *p* refleja la incompatibilidad entre los datos observados y los esperados por el modelo estadístico, históricamente se ha interpretado que valores cercanos a 1 indican una mayor consistencia entre los datos observados y los esperados, mientras que valores cercanos a 0 sugieren una mayor inconsistencia. Tradicionalmente, se ha establecido un umbral de 0.05 para considerar un resultado como estadísticamente significativo (6).

El concepto de valor *p* surgió de la combinación de dos enfoques estadísticos diferentes. Ya habíamos mencionado como Ronald Fisher propuso en 1925 las pruebas de significancia como una herramienta para evaluar la consistencia de los datos con la hipótesis nula. Fisher consideraba el valor *p* como una medida continua de evidencia en contra de la hipótesis nula. Según su interpretación, la significancia estadística se evaluaba en relación con un nivel de significancia preestablecido, interpretándose como estadísticamente significativo si el valor *p* era igual o menor a dicho nivel, y estadísticamente no significativo si el valor *p* era mayor. En su concepción original, este nivel de significancia no era fijo ni determinado por conveniencia, sino que debía ser flexible y ajustarse según el conocimiento acumulado en el campo (4).

Por otro lado, Jerzy Neyman y Egon Pearson desarrollaron la teoría de la prueba de hipótesis, introduciendo el concepto de hipótesis alternativa, los errores tipo I y tipo II, y un umbral predefinido (alfa, α) para tomar decisiones sobre el rechazo o no rechazo de la hipótesis nula (4). En la década de 1940, tras múltiples debates entre ambas escuelas de pensamiento, surgió un modelo que combinó varios elementos, dando origen al modelo actual del valor *p*, conocido como la prueba de significancia de la hipótesis nula (PSHN). Esta combinación de enfoques opuestos probablemente ha sido la fuente de las diversas controversias actuales en torno al valor *p* (2).

¿Cuáles son las limitaciones del uso del valor *p*?

Mezones e*t al.* (2) nos orientan sobre las principales limitaciones en el uso del valor *p*, categorizándolas de la siguiente manera:

- **Dependencia del tamaño de la muestra**: el valor *p* es altamente sensible al tamaño de la muestra. Esto no es de extrañarse, ya que muchos estadísticos, incluyendo el Z, tienen en su denominador el número de tamaño de muestra (√n). En estudios con muestras grandes, incluso efectos mínimos y clínicamente irrelevantes pueden alcanzar significancia estadística (valor $p < 0.05$), lo que facilita el rechazo de la hipótesis nula, independientemente de la magnitud real del efecto. Por el contrario, en estudios con muestras pequeñas, incluso efectos grandes y potencialmente importantes pueden no alcanzar significancia estadística

  (valor $p > 0.05$), lo que puede llevar a conclusiones erróneas de no efecto.

- **Discordancia con el tamaño del efecto**: relacionado con el punto anterior y con el tamaño de la muestra, la significancia estadística tiende a sobrestimar el efecto, mientras que lo que no es estadísticamente significativo puede corresponder a efectos importantes. Recuerde el ejemplo de las ondas electromagnéticas y las leucemias.

- **No correspondencia con la importancia clínica**: este parámetro debe ser evaluado dentro del contexto clínico de la pregunta de investigación.

- **Dicotomización arbitraria**: la categorización binaria entre resultados "estadísticamente significativos" ($p \leq 0.05$) y "no estadísticamente significativos" ($p > 0.05$), basada en un umbral arbitrario, conduce a una pérdida de información importante. Por lo tanto, idealmente debería evaluarse como un parámetro continuo, considerando el contexto específico de la investigación.

- **Pobre replicabilidad de los hallazgos científicos**: frecuentemente, los hallazgos iniciales estadísticamente significativos han sido cuestionados en estudios posteriores. Estos falsos positivos pueden representar hasta un 13 % de los estudios publicados en revistas indexadas. Este fenómeno se explica por lo que se conoce como la "maldición del ganador", en la cual los estudios iniciales que encuentran significancia estadística tienden a sobrestimar la magnitud real del efecto, lo que dificulta la replicación en estudios posteriores.

- **Susceptibilidad a malas prácticas**: el valor *p* es susceptible a fenómenos relacionados con la realización de múltiples pruebas estadísticas para evaluar cuáles de ellas reportan significancia estadística, lo que puede llevar a falsos positivos. Miremos cómo, si los autores del estudio de la viñeta hubiesen incluido un sinnúmero de desenlaces, diez por ejemplo (satisfacción con el peso, diámetro de cintura y cadera, escalas de actividad física, etc.), si nuestro $\alpha = 0.05$, y si lleváramos a cabo 10 test de significación estadística, suponiendo que cada uno fuese independiente, la probabilidad de no poder rechazar una hipótesis de no diferencia sería:

  $$(0{,}95)^{10} = 0{,}598$$

  Y la posibilidad de rechazar una hipótesis nula (así esta fuera cierta) sería:

  $$1 - (0{,}95)^{10} = 0{,}402$$

  Esta peligrosa práctica de usar desenlaces múltiples y altamente correlacionados aumenta las probabilidades de encontrar una diferencia estadísticamente significativa solo por azar (fenómeno de P-Hacking).

- Otro problema que puede ocurrir con las pruebas de hipótesis estadísticas son fenómenos de reporte selectivo de los resultados significativos. Se tiende a trabajar arduamente para lograr que el valor de *p* sea

lo suficientemente bajo como para ser significativo. Gotzsche (2006) utilizó un enfoque ingenioso para cuantificar si los resultados se presentaban de manera honesta. Si los valores de *p* se informaran de manera honesta, el número de valores de *p* entre 0.04 y 0.05 deberían ser similares al número de valores de *p* entre 0.05 y 0.06. Sin embargo, al analizar 130 resúmenes de artículos publicados en 2003, se encontró que había cinco veces más valores de *p* entre 0.04 y 0.05 que entre 0.05 y 0.06, lo que sugiere un posible sesgo en la presentación de los resultados (fenómeno de P-Harking) (7).

¿Cuáles son las interpretaciones erróneas más comunes?

Greenland e*t al.* (6) nos proporcionan una serie de interpretaciones erróneas comunes del valor *p*, algunas de las cuales son:

- **El valor *p* es la probabilidad de que la hipótesis nula sea verdadera. Falso.** El valor *p* considera la hipótesis nula como verdadera, pero no mide la probabilidad de que esta sea verdadera. Más bien, indica la probabilidad de observar los datos obtenidos (o algo más extremo) bajo la suposición de que la hipótesis nula es cierta, reflejando la compatibilidad entre los datos y el modelo estadístico.
- **El valor *p* es la probabilidad de que el azar por sí solo produzca la asociación observada. Falso.** Similar al punto anterior, el valor *p* asume que el azar es parte del modelo estadístico, pero no mide específicamente la probabilidad de que el azar sea el único factor que explique la asociación observada.
- **Un resultado estadísticamente significativo ($p \leq 0,05$) significa que la hipótesis nula es falsa o debería ser rechazada. Falso.** Un valor *p* bajo indica que los datos observados son poco compatibles con la hipótesis nula, pero no proporciona evidencia concluyente de que la hipótesis nula sea falsa. Simplemente sugiere que, si la hipótesis nula fuera cierta, los datos obtenidos serían poco probables.
- **Un resultado no estadísticamente significativo ($p > 0,05$) significa que la hipótesis nula es verdadera y no debería ser rechazada. Falso.** Un valor *p* alto no prueba que la hipótesis nula sea verdadera. Solo indica que no hay suficiente evidencia en los datos para rechazarla. También puede deberse a un tamaño de muestra insuficiente o a un error aleatorio.
- **Un valor *p* grande es evidencia a favor de la hipótesis nula. Falso.** Un valor *p* grande indica que los datos son compatibles con la hipótesis nula, pero no significa que la hipótesis nula sea más probable que las alternativas. El valor *p* solo mide la compatibilidad entre los datos y el modelo bajo no prueba la probabilidad de la hipótesis en sí.
- **Un valor *p* mayor que 0,05 significa que se observó un no efecto o que se demostró la ausencia de un efecto. Falso.** Un valor *p* alto no demuestra que no haya efecto. Solo indica que no se ha encontrado suficiente evidencia para rechazar la hipótesis nula. Es importante recordar que un valor *p* mayor a 0,05 no excluye la posibilidad de una asociación.
- **La significancia estadística indica científicamente que una relación importante ha sido detectada. Falso.** Los hallazgos estadísticamente significativos pueden estar presentes en estudios con

tamaños grandes, incluso cuando los efectos son pequeños y no clínicamente relevantes. Evaluar los intervalos de confianza es crucial para interpretar correctamente la relevancia clínica de los resultados.

- **La ausencia de significancia estadística indica que el tamaño del efecto es pequeño. Falso.** En estudios con muestras pequeñas, incluso efectos grandes pueden no alcanzar significancia estadística debido a la baja potencia del estudio. Es fundamental considerar los intervalos de confianza para una evaluación más precisa del tamaño del efecto.

- **Un valor *p* = 0,05 significa lo mismo que un valor p ≤ 0,05. Falso.** Un valor *p* exactamente igual a 0,05 es considerado un resultado limítrofe, mientras que un valor *p* menor o igual a 0,05 puede incluir tanto una evidencia limítrofe como una evidencia más fuerte en contra de la hipótesis nula. No se debe considerar lo mismo un valor *p* de 0,05 que uno inferior, ya que la interpretación depende del contexto y de la magnitud de la diferencia.

- **Cuando una misma hipótesis es testeada en diferentes estudios, y ninguno o la mayoría de las pruebas son estadísticamente significativas (*p* > 0,05), entonces en promedio la evidencia apoya a la hipótesis nula. Falso.** La falta de significancia en estudios individuales no significa que la evidencia global apoye la hipótesis nula. Los estudios combinados pueden mostrar una asociación significativa, y no se debe concluir automáticamente que la hipótesis nula es verdadera por la falta de significancia en estudios individuales.

- **Cuando la misma hipótesis es testeada en dos poblaciones diferentes y los resultados del valor *p* son opuestos en función al umbral de 0,05, estos resultados son inconsistentes. Falso.** Las pruebas estadísticas son sensibles a las diferencias en las poblaciones de estudio. Es posible obtener resultados con diferentes valores de *p*, pero con asociaciones similares, ya que los contextos y características de cada población pueden influir en los resultados. Recuerde el ejemplo de las ondas EM y la leucemia.

- **Cuando la misma hipótesis es testeada en dos poblaciones diferentes y obtenemos los mismos valores *p*, entonces los resultados son concordantes. Falso.** A pesar de obtener los mismos valores p, las asociaciones observadas pueden ser diferentes en función de otros factores contextuales, como las características de cada población o el diseño del estudio.

Recomendaciones de expertos

En este sentido, la Asociación Americana de Estadística (ASA), reconociendo la creciente preocupación sobre el uso y la interpretación inadecuada del valor *p*, así como su papel en la crisis de reproducibilidad de la ciencia, publicó en 2016 una declaración sobre el valor *p* y la significancia estadística. En dicha declaración, la ASA presentó seis principios clave para un uso e interpretación adecuados de estas herramientas estadísticas (5). Estos principios fueron los siguientes:

1. **Los valores *p* pueden indicar cuán incompatibles son los datos con una hipótesis nula específica**. Cuanto más pequeño sea el valor de *p*, mayor es la incompatibilidad de los datos con la hipótesis nula, lo que puede interpretarse como evidencia en su contra.

2. **Los valores *p* no miden la probabilidad de que la hipótesis nula sea verdadera, ni la probabilidad de que los datos hayan sido producidos únicamente por azar.** En realidad, el valor *p* compara los datos

obtenidos con los predichos por una hipótesis preespecificada.

3. **Las conclusiones científicas, así como las decisiones clínicas, comerciales o políticas, no deben basarse únicamente en si un valor *p* cruza un umbral específico.** Utilizar umbrales estandarizados, como $p < 0.05$, puede llevar a conclusiones erróneas y distorsionar el proceso científico, ya que es inapropiado considerar la información como verdadera solo si está por debajo de este umbral o falsa si está por encima. Para llegar a conclusiones adecuadas, deben considerarse otros factores, como el diseño del estudio y la medición correcta de las variables, entre otros.

4. **Un proceso de inferencia apropiado requiere un reporte completo y transparencia.** No se deben realizar múltiples análisis estadísticos con el único fin de reportar aquellos que arrojan valores de *p* a favor de la significancia estadística. Los investigadores deben explicar detalladamente la elección de las pruebas y los motivos de su selección, así como proporcionar un informe completo de todos los análisis realizados, para asegurar una correcta interpretación de los resultados.

5. **El valor *p* o la significancia estadística no es una medida del tamaño del efecto o de la relevancia del resultado.** Es importante comprender que la significancia estadística no es lo mismo que la significancia clínica. Así como un valor *p* muy pequeño no implica un efecto grande, un valor *p* mayor no significa necesariamente una ausencia de efecto o relevancia. También debe tenerse en cuenta el papel del tamaño de la muestra, ya que muestras grandes pueden generar valores de *p* muy pequeños, mientras que muestras pequeñas pueden hacer lo contrario.

6. **El valor *p*, por sí mismo, no proporciona una buena medida de evidencia sobre un modelo o hipótesis.** Debe interpretarse en su contexto y asociarse con otras evidencias, ya que los valores *p* pequeños son solo evidencia débil en contra de la hipótesis nula, mientras que valores *p* grandes no necesariamente implican evidencia a favor de la hipótesis nula.

¿Qué otras opciones tenemos?

Se han propuesto diversas alternativas y complementos al valor *p* para la interpretación de la literatura médica, tales como:

- **Intervalo de predicción del valor de *p*:** esta alternativa busca proporcionar al lector información adicional sobre la variabilidad del valor de *p* y la incertidumbre asociada a esta estadística, lo que permite una mejor comprensión de los resultados (8).
- **Riesgo estimado de falsos positivos:** este enfoque tiene en cuenta el nivel de creencia previa sobre la hipótesis nula antes de realizar el estudio, considerando la probabilidad de que un valor *p* estadísticamente significativo sea el resultado de un falso rechazo de la hipótesis nula (8).
- **Tamaño del efecto e intervalos de confianza:** su principal ventaja es la capacidad de estimar la magnitud de la diferencia o asociación observada, así como los intervalos de confianza (IC) alrededor de estas estimaciones, los cuales reflejan la precisión de los resultados. Además, permiten una evaluación más completa a través de metaanálisis, brindando un enfoque más integral (8).
- **Factor de Bayes:** utilizando la estadística bayesiana, este enfoque permite evaluar simultáneamente la hipótesis nula y la

alternativa. Representa matemáticamente qué tan bien son explicados los datos recolectados por cada una de las hipótesis. Por ejemplo, un factor de Bayes de 4 indicaría que la evidencia a favor de la hipótesis alternativa es cuatro veces mayor que la evidencia a favor de la hipótesis nula (8).

- **Criterio de información de Akaike (AIC):** este método proporciona una estimación de la idoneidad de un modelo particular en relación con un conjunto de modelos explicativos. Se calcula a partir de *software* estadísticos y permite contrastar diferentes modelos para seleccionar el que mejor explique el fenómeno medido (8).

Mensajes indispensables

El valor *p* ha sido una herramienta central en la investigación médica durante décadas, pero sus limitaciones y el potencial de mal uso son cada vez más evidentes. Es fundamental que los internistas y médicos generales tengan una comprensión crítica del valor p y lo interpreten en el contexto del tamaño del efecto, los intervalos de confianza, el diseño del estudio y el potencial sesgo, con el fin de evitar interpretaciones erróneas sobre la investigación clínica. La interpretación basada en una dicotomía de umbrales de significancia arbitrarios es problemática, ya que puede llevar a una visión simplista de la evidencia. El uso de otras propuestas en conjunto con el valor *p* permitirá una comprensión más completa de los resultados en la investigación clínica.

Referencias bibliográficas

1. Darling HS. To “P” or not to “P”, that is the question: A narrative review on P value. *Cancer Res Stat Treat.* 2021 Oct;4(4):756–62.

2. Mezones-Holguin E, Al-kassab-Córdova A, Soto-Becerra P, Hernández-Díaz S, Kaufman JS. La prueba de significancia de la hipótesis nula y la dicotomización del valor *p.* Errare Humanum Est. *Rev Peru Med Exp Salud Pública*. 2024 Nov 26;422–30.

3. Demarquette A, Perrault T, Alapetite T, Bouizegarene M, Bronnert R, Fouré G, et al. Spin and fragility in randomised controlled trials in the anaesthesia literature: a systematic review. *Br J Anaesth*. 2023 May;130(5):528–35.

4. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front Psychol* [Internet]. 2015 Mar 3 [cited 2025 Mar 10];6. Available from: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00223/abstract

5. Wasserstein RL, Lazar NA. The ASA Statement on *p* -Values: Context, Process, and Purpose. *Am Stat.* 2016, Apr 2;70(2):129–33.

6. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016 Apr;31(4):337–50.

7. Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. *BMJ.* 2006, Jul 27;333(7561):231–4.

8. Halsey LG. The reign of the *p* -value is over: what alternative analyses could we employ to fill the power vacuum? *Biol Lett.* 2019 May 31;15(5):20190174.