MEDwave

**Temas y controversias en bioestadística**
*Medwave* 2016 Sep;16(8):e6534 doi: 10.5867/medwave.2016.08.6534

# El valor de p en entredicho: significación estadística, clínica y práctica

## The questioned p value: clinical, practical and statistical significance

**Autora:** Rosa Jiménez Paneque**[1]**

**Filiación:**
**[1]** Subeditora y revisora estadística, *Medwave*

**E-mail:** rosa.jimenez@medave.cl

**Citación:** Jiménez Paneque R. The questioned p value: clinical, practical and statistical significance. *Medwave* 2016 Sep;16(8):e6534 doi: 10.5867/medwave.2016.08.6534
**Fecha de publicación:** 9/9/2016
**Origen:** no solicitado
**Tipo de revisión:** sin revisión por pares

## Resumen

El uso del valor de p y la significación estadística han estado en entredicho desde principios de la década de los 80 en el siglo pasado hasta nuestros días. Mucho se ha discutido al respecto en el ámbito de la estadística y sus aplicaciones, en particular a la Epidemiología y la Salud Pública. El valor de p y su equivalente, la significación estadística, son por demás conceptos difíciles de asimilar para los muchos profesionales de la salud involucrados de alguna manera en la investigación aplicada a sus áreas de trabajo. Sin embargo, su significado debería ser claro en términos intuitivos a pesar de que se basa en conceptos teóricos del terreno de la Estadística-Matemática. Este artículo intenta presentar al valor de p como un concepto que se aplica a la vida diaria y por tanto intuitivamente sencillo pero cuyo uso adecuado no se puede separar de elementos teóricos y metodológicos con complejidad intrínseca. Se explican también de manera intuitiva las razones detrás de las críticas que ha recibido el valor de p y su uso aislado, principalmente la necesidad de deslindar significación estadística de significación clínica y se mencionan algunos de los remedios propuestos para estos problemas. Se termina aludiendo a la actual tendencia a reivindicar su uso apelando a la conveniencia de utilizarlo en ciertas situaciones y la reciente declaración de la Asociación Americana de Estadística al respecto.

## Abstract

The use of p-value and statistical significance have been questioned since the early 80s in the last century until today. Much has been discussed about it in the field of statistics and its applications, especially in Epidemiology and Public Health. As a matter of fact, the p-value and its equivalent, statistical significance, are difficult concepts to grasp for the many health professionals some way involved in research applied to their work areas. However, its meaning should be clear in intuitive terms although it is based on theoretical concepts of the field of Statistics. This paper attempts to present the p-value as a concept that applies to everyday life and therefore intuitively simple but whose proper use cannot be separated from theoretical and methodological elements of inherent complexity. The reasons behind the criticism received by the p-value and its isolated use are intuitively explained, mainly the need to demarcate statistical significance from clinical significance and some of the recommended remedies for these problems are approached as well. It finally refers to the current trend to vindicate the p-value appealing to the convenience of its use in certain situations and the recent statement of the American Statistical Association in this regard.

## Introducción

En todo el mundo, el valor de p (*p value*, en inglés) ha sido y continúa siendo una preocupación para los investigadores que escriben y editores que publican artículos científicos.

Todavía hoy, tras 20 años de discusión, la literatura aborda este tema de la estadística y sus aplicaciones a la investigación y al reporte de sus resultados [1],[2],[3].

Por otro lado, la experiencia dice que, a pesar de lo extendido de su uso (y su desuso), muchos investigadores aún no saben bien qué significa. "Los estadísticos lo encuentran" o "los programas estadísticos me lo dan" podrían ser respuestas comunes a la pregunta "¿qué es el valor de p?". No se puede saber de todo, y ya es muy difícil saber de un tema en particular como para además conocer de estadística. Incluso después de asistir voluntaria o inducidamente a cursos de estadística, el significado del valor de p, si alguna vez se entendió, ya puede haber quedado olvidado.

### ¿Qué es el valor de p?

El sentido del valor de p es básico e intuitivo. Es similar al concepto de probabilidad que todos manejamos en la vida diaria. Por ejemplo, nadie puede asegurar que saldrá de su casa -un día cualquiera a sus labores habituales- y no tendrá un accidente aunque no haya indicios de que pueda ocurrir. En una situación de clima y ambiente normal de cualquier sociedad (que no esté en guerra y no esté ocurriendo algún desastre natural), esa probabilidad es muy baja y, por tanto, habitualmente nadie se queda en casa por esa razón. Aunque la probabilidad de que suceda un accidente no es 0, se toma la decisión de salir a la calle, bajo el supuesto de que tal probabilidad "es muy baja". Podríamos entonces pensar así: voy a salir a la calle pero primero me planteo la hipótesis "tendré un accidente" (solo una hipótesis), la valoro y pienso "no hay razones especiales para tener un accidente hoy (estoy bien de salud y todo funciona normalmente)" por tanto, la probabilidad de tener un accidente es sumamente baja, de modo que tomo la decisión de salir y la llevo adelante. He rechazado la hipótesis de tener un accidente.

Para el valor de p se puede aplicar el mismo razonamiento. Este valor de p identifica a la probabilidad de que ciertos resultados hayan ocurrido si se cumple la "hipótesis que se trata rechazar". Si esta p es muy bajita (cercana a cero, digamos), se rechaza la hipótesis. ¿Por qué decimos "hipótesis que se trata de rechazar"? Porque es de suponer que si alguien va a salir a la calle es porque eso quiere. Lo mismo pasa en la investigación.

Es muy fácil poner ejemplos de situaciones similares en la vida diaria ya que en realidad constantemente estamos rechazando o no rechazando hipótesis pues, aunque vivimos en el presente, siempre avanzamos hacia el futuro que es desconocido y solo interactuamos con él a base de probabilidades a las que algunos llaman probabilidades *subjetivas o intuitivas* [4] (porque no se someten estrictamente a las leyes de la teoría de probabilidades).

Particularizando en la evaluación de nuevas intervenciones sanitarias (que brindan los ejemplos más claros para el tema de la pruebas de hipótesis); cuando surge un nuevo tratamiento y este llega al momento de demostrar su eficacia, el investigador en general lo que intenta es "rechazar la hipótesis de que no es eficaz" para poder introducirlo en la práctica.

Ese es el significado simple del valor de p, la probabilidad de obtener ciertos resultados dado que se cumple la hipótesis que "se quiere" rechazar; si ese valor de p es muy pequeño, se rechaza la hipótesis y se logra lo buscado. Lo que no significa, por supuesto, que la hipótesis no sea cierta porque, al igual que en la vida diaria, solo hay adivinos en los cuentos de hadas.

### ¿Cómo se calcula (estima) el valor de p? Primero la intuición y luego la decisión

Otro aspecto que merece atención y puede interesar a los investigadores no estadísticos sobre el valor de p es cómo se calcula. Esto puede ser realmente complicado y lleno de cálculos matemáticos que no constituyen la intención de este artículo. No obstante, también podemos recurrir a la intuición para entender cómo se obtiene.

En términos más generales todas las probabilidades se estiman a partir de repetidas observaciones. Hago un paréntesis para cambiar el verbo calcular por estimar. Calcular indica algo que se puede saber "a ciencia cierta" - como calcular la velocidad necesaria que tendría que tener un vehículo para recorrer una distancia dada en cierto tiempo. Pero, en la práctica, las probabilidades se estiman (nos aproximamos a ellas) como la probabilidad de que "llueva hoy". No podemos calcularla pero sí podemos estimarla. Un observatorio meteorológico puede tener hoy día muchas (cientos, miles quizás) observaciones que le permiten calcular la proporción de veces que llueve en "días como hoy" y esa proporción (o porcentaje) será la estimación de la probabilidad de que llueva hoy. Seguro es más complicado pero, con perdón de los meteorólogos, se puede simplificar así.

Entonces, precisamente lo que necesitamos para esa estimación son observaciones. Esa es la base de la necesidad de realizar un estudio para estimar ese valor de p que implica, en efecto, recoger datos. Si vamos al contexto de la evaluación de una nueva intervención sanitaria, necesitaremos en primer lugar observaciones sobre resultados de esa intervención y también observaciones sobre la eficacia de la intervención que había hasta ese momento. En términos prácticos y generales, lo que se necesita es saber si la nueva intervención es más eficaz que la que había hasta el momento, en otras palabras, comparar la *nueva intervención* con la que existía, en cuanto a eficacia.

Entonces, la hipótesis que se quiere rechazar es que la nueva intervención es igual de eficaz que la anterior. Se necesitan observaciones que hablen de la eficacia de la nueva intervención y también observaciones sobre la

eficacia (resultado) de la intervención que existía hasta el momento (que podemos llamar intervención "convencional" o "control"). Se compararán entonces ambas "eficacias" y se obtiene la diferencia entre estas. Ahora se procede a estimar la probabilidad de que esa diferencia ya observada (o una mayor) pueda ocurrir si la nueva intervención fuera igual a la anterior y ese precisamente es el valor de p. En otras palabras, la probabilidad de que se observe esa diferencia (o una mayor) bajo la hipótesis de que realmente la diferencia es cero. Una vez estimada esa probabilidad, si es muy pequeña, decido rechazar la hipótesis de igualdad de eficacias, igual que se hace con la hipótesis "tendré un accidente".

**No es tan sencillo**

Todo lo anterior se ha simplificado en aras de la comprensión del razonamiento subyacente en el valor de p y su significado. Pero hay varios aspectos que lo complican y que constituyen la razón de la existencia de una rama de la matemática (la estadística inferencial) para formalizar e instrumentar todo este razonamiento.

- Para las estimaciones se necesitan al menos modelos teóricos que se ajusten a la realidad y también la teoría de probabilidades o sea, distribuciones probabilísticas de variables aleatorias que se ajusten a cada situación o realidad. Se trata de modelos probabilísticos, se trabaja con ellos y se cuenta además con herramientas para evaluar cuál es el modelo óptimo en cada una de las situaciones de la práctica.
- Se necesita decidir también a qué se le llama probabilidad "pequeña". Los investigadores, junto a los estadísticos, parecen haberse conformado con aceptar un valor de p menor de 0,05 como pequeño aunque ese valor también es amplio objeto de discusión y controversia [2],[5]. De aceptarlo sin discusión podríamos estar tomando la decisión de rechazar la igualdad de eficacias cuando una diferencia observada (o mayor) puede ocurrir con una probabilidad de 0,05 en el caso de eficacias iguales. Eso significa un riesgo, si aplicar la nueva intervención resultará muy costoso, un riesgo de 5% podría no ser compatible con una buena decisión. Por tanto, el valor de p debería ajustarse a las consecuencias de cometer un error en caso de que se rechazara la igualdad de eficacias. Por ejemplo, si implementar la nueva intervención costara un millón de dólares, se trataría de correr un riesgo menor que si solo costara diez mil. De modo que, aceptar siempre un riesgo de 5% no es compatible con el sentido práctico común. Respecto a esto, Ronald Fisher, el renombrado estadístico que introdujo el valor de p en los años 20 del siglo pasado, dijo lo siguiente [6]:

*It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result.*

- Se necesita formalizar y clasificar la forma en que se estima lo que hemos llamado "diferencia de eficacias". No todas las eficacias pueden medirse de igual manera, depende de cómo consideremos la medición del efecto de la intervención. En principio es sencillo al menos entender que existe una amplia gama de formas para medir eficacia de una intervención. No pueden medirse igual la eficacia de un tipo de ejercicio para aliviar la lumbalgia y la de un medicamento que debe disminuir el nivel de colesterol en la sangre.
- Se necesita lograr que las diferencias que se encuentren entre intervenciones sean debidas a las intervenciones mismas y no a la multiplicidad de factores que pueden incidir en la eficacia de una intervención. Esto concierne al diseño de la investigación y al reconocimiento de los sesgos que deben evitarse en las comparaciones. Por ejemplo, si estamos comparando la eficacia de un nuevo medicamento con uno ya conocido con el fin de aumentar la sobrevida de pacientes con algún tipo de cáncer, necesitaríamos que los pacientes del grupo en el que se aplica el nuevo medicamento y aquellos del grupo en el que se aplica el medicamento ya conocido tengan edades similares. Asimismo, ambos grupos de pacientes deberían ser similares respecto a estadio del cáncer en cuestión y probablemente a un número importante de características más. Si no es así, el resultado de la comparación podría verse confundido por las características de los grupos que se comparan dando lugar a un típico problema del diseño de estudios que comparan eficacias: el fenómeno de confusión.
- Se necesitan observaciones que representen a toda la realidad puesto que estamos en el campo de la investigación científica, no se necesita conocer si una intervención nueva es más eficaz que la anterior en un centro durante un mes (por ejemplo) sino en general. La pregunta práctica que habrá en el mejor de los casos es: ¿se introduce la nueva intervención en el arsenal de intervenciones sanitarias dirigidas a cierto fin o no? Por tanto, aunque parezca rimbombante, hay que aceptar que en general se quiere inferir a poblaciones o universos infinitos en el espacio y en el tiempo.
- Se necesita decidir cuántas observaciones de la realidad deberán tenerse para la decisión del rechazo (tamaño de la muestra). Este es un aspecto muy relacionado con el valor de p y también con la necesidad de evaluar la probabilidad de cometer el error "contrario" que sería no rechazar la hipótesis dado que esta fuera falsa. Altman y Bland realizan una ilustrativa discusión de este aspecto y enfatizan en la necesidad de no confundir la ausencia de rechazo (valor de p mayor de 0,05) con la inexistencia de diferencia entre eficacias sin tener en cuenta el tamaño de la muestra y la probabilidad de "no rechazar" cuando realmente la hipótesis no es cierta [7] .

De modo que, si bien el razonamiento que subyace detrás del valor de p y el interés que despierta en la investigación puede considerarse relativamente sencillo, su estimación y su implementación en la práctica contienen vericuetos realmente complejos y de difícil solución. Se han señalado solo algunos de esos aspectos, quizás los más connotados.

Vale añadir que ese valor de 0,05, mencionado antes, es el que respalda la conocida "significación estadística". Si el valor de p encontrado en una comparación de eficacias como la que se ha tomado de ejemplo, es menor de 0,05

se dice que la diferencia entre ambas eficacias es significativa o más precisamente, *estadísticamente significativa*. Quiere decir que tenemos una probabilidad alta (95%) de que esa diferencia que hemos obtenido (o una mayor) no provenga de una realidad donde la eficacia es igual (o peor); hemos rechazado esa hipótesis. En otras palabras como el valor de p es muy pequeño, se decide creer que la diferencia encontrada "no ha ocurrido por casualidad" en un contexto donde realmente la nueva intervención no tiene más eficacia que la convencional.

Pero, el resumen de aspectos que deben resolverse antes de que se puedan tomar decisiones a partir de un valor de p nos habla de que no se trata de un problema sencillo ni para estadísticos ni para investigadores. La buena noticia es que muchos de los aspectos señalados han sido abordados teóricamente y su solución está fundamentada e implementada a través de programas de computación ya comercializados y utilizados. Los aspectos de diseño para evitar sesgos, también se reportan y discuten con profusión en la literatura.

**Los problemas del valor de p, significación estadística vs significación clínica**

Entonces, cuál es el problema del valor de p que ha sido tan criticado en las últimas décadas [8],[9],[10]. La respuesta tiene una base teórica y también filosófica que parte de la convicción de que el valor de p o, mejor dicho, las pruebas de significación no fueron ideadas por su fundador (Ronald Fisher, 1890-1962) para ser utilizadas en la práctica tal como han sido utilizadas por décadas [11].

Pero, quizás la respuesta más sencilla y clara a esta pregunta se deriva del hecho de que valores de p pequeños pueden obtenerse en casos donde las diferencias encontradas son realmente pequeñas, o mejor dicho pequeñas para tener importancia práctica. Es decir que se dan situaciones en las que una diferencia sin significado práctico alguno puede ser "estadísticamente significativa".

La toma de decisiones en cualquier circunstancia está condicionada a una relación riesgo/beneficio, costo/beneficio o riesgo+costo/beneficio. Esto sucede en la práctica y también en la vida diaria. Continuando el parangón que se hizo al principio de este artículo, la decisión sobre salir o no se basa por supuesto en el riesgo (probabilidad) de tener un accidente pero ¿cuál es la importancia de la salida? No es lo mismo salir al cine por diversión o salir al trabajo so pena de que nos descuenten el día. Tenemos tendencia a aceptar mayores riesgos si el beneficio (o necesidad) de la decisión es mayor.

Algo similar sucede en la práctica sanitaria, no es lo mismo introducir una nueva intervención si esta generará una mejoría de 10% respecto a la que se tenía antes o si esta generará una mejoría de un 1%. Introducir una nueva intervención tiene un costo que solo vale la pena si la mejoría que se logrará será sustancial o importante para el contexto donde se evalúa. Entonces, el hecho de que el valor de p, que supuestamente debe ayudar a tomar decisiones, no sea sensible a la relevancia o las consecuencias de la decisión, representa un problema realmente mayúsculo.

El hecho es que, en buena ley, el valor de p depende del tamaño de la muestra de modo tal que solo con aumentar la muestra se puede obtener significación estadística para cualquier diferencia por muy pequeña que esta sea. No es un fenómeno raro, es natural que así sea, en la práctica no es posible que dos intervenciones por similar eficacia que tengan, actúen de manera idéntica, logrando eficacias (o efectos) exactamente iguales. Siempre existirá una diferencia y a medida que aumente la muestra, mayor será la probabilidad de encontrarla [12]. Entonces, en realidad el objetivo no es –ni puede ser- encontrar la diferencia a toda costa sino detectar la diferencia cuando esta puede tener algún beneficio para la práctica o al menos para la teoría. Por lo tanto, si tomamos la decisión de adoptar la nueva intervención solo porque su eficacia es significativamente mayor que la intervención anterior, corremos el riesgo de que todo el costo que significa la introducción práctica de la nueva intervención, efectivamente no valga la pena.

De modo que a partir de los años 80 del siglo pasado se tuvo conciencia de este problema y comenzó la discusión entre, por un lado, la *significación estadística* y, por otro, lo que muchos han dado en llamar*significación clínica* [9],[13],[14]. Este último apelativo, se debe a que el tema de la comparación de eficacias adquiere particular trascendencia en la evaluación de nuevos medicamentos mediante los conocidos ensayos clínicos pero, se podría sustituir por *significación práctica* ya que no es una cuestión solo propia de los ensayos clínicos ni de la clínica.

Había que alertar a investigadores y también editores de revistas científicas sobre la inoperancia de la significación estadística, el valor de p, sin una valoración de la significación práctica (o clínica) del resultado de las comparaciones.

**El remedio**

Encontrar un remedio (alternativo) para estas contradicciones detectadas cuarenta años después de utilizar el valor de p en las diversas pruebas de hipótesis que necesitaba la investigación y también la toma de decisiones, todavía es motivo de artículos y discusiones metodológicas. Varias alternativas de solución han aparecido para este problema en el mundo de la investigación y la estadística que se reflejan en la literatura científica.

Un remedio de enfoque radical consiste en cambiar todo el sistema de pruebas de hipótesis estadísticas para evaluar la eficacia de nuevas tecnologías o para probar hipótesis en cualquier esfera del conocimiento científico. Se trata de un nuevo paradigma que promueve sustituir las *técnicas frecuentistas* (como suelen llamar en general a todas estos métodos de pruebas de hipótesis que hemos mencionado) por las*técnicas bayesianas* [15]. Los métodos bayesianos para probar hipótesis, no se basan en el valor de p, obtenido a partir de un estudio en el que se controlan todos los factores que pueden afectar el cumplimiento de la

hipótesis, excepto el que se estudia. La idea que los alienta es que la experiencia acumulada sobre una hipótesis puede y debe contribuir a su verificación. Con la experiencia anterior (más la teoría de probabilidades, por supuesto) se construye lo que llaman "probabilidad a priori" y con los datos que aporta un nuevo estudio se llega a la "probabilidad a posteriori" que sería la que daría paso a la toma de decisiones.

Esta es a grandes rasgos la idea subyacente en las técnicas bayesianas que constituyen hoy un amplio arsenal que pugna por prevalecer como método principal de inferencia estadística pero que, por lo visto, no ha tenido el éxito que quizás se esperaba. Premonitoriamente, en 1987, en un artículo de la revista JAMA, Browner y Newman profundizaron en la similitud que había entre las pruebas de hipótesis en la investigación y la evaluación de pruebas diagnósticas en la clínica donde las probabilidades a priori y a posteriori cobran singular relevancia [16].

Otro remedio, quizás el más aceptado hoy, promueve el uso de los llamados intervalos de confianza. Estos intervalos supuestamente son capaces de cuantificar nuestra confianza en los resultados de un estudio y, sobre todo, nos permiten acercarnos a la magnitud del llamado "tamaño del efecto" [9]. El término "confianza" merece una explicación, podemos volver a nuestra salida a la calle. Como hemos salido cientos de veces en ciertas condiciones (climáticas, personales, sociales y otras) y no hemos tenido accidentes, tenemos confianza en que esta vez (dado que están presentes las condiciones anteriores) tampoco lo tendremos. Muy a menudo se confunden confianza y probabilidad pero, es claro que no conocemos la probabilidad de tener un accidente, podríamos estimarla pero no la conocemos. En realidad tenemos confianza en que no tendremos un accidente porque la probabilidad de tenerlo (dadas las condiciones) es muy baja. No obstante en estadística el término "confianza" sustituye al de probabilidad una vez que se ha obtenido la muestra y se tienen resultados, básicamente porque la probabilidad es un término ligado a lo desconocido. Una vez que tengo calculado el llamado intervalo de confianza, no puedo hablar de probabilidad porque ya el hecho pasó.

Más concretamente en el ejemplo de la comparación de eficacias entre una nueva intervención y la ya existente, la idea es obtener intervalos de confianza para la diferencia entre ambas intervenciones. Significa buscar un método con alta probabilidad de éxito (o sea, un método que proporcione una alta probabilidad para que la diferencia efectivamente se encuentre entre los dos extremos del intervalo). Una vez obtenido el intervalo se habla de confianza, es decir, en lugar de decir "este es el intervalo obtenido con un método que proporciona alta probabilidad de que la diferencia verdaderamente esté en este rango de valores" se dice "este es el intervalo de confianza". Hay que añadir que un intervalo de confianza para que tenga utilidad, no puede tener una amplitud muy grande. ¿Qué podría ganarse con saber, con alta confianza, que la diferencia en la eficacia de dos intervenciones está entre 1 y 50% de mejoría? Eso seguramente puede conocerse sin estudio alguno. Para intervalos de confianza estrechos se necesitarán tamaños de muestra adecuados. Por tanto, el obtener intervalos de confianza no elimina la necesidad de buscar el tamaño de muestra necesario, solo nos da mayor información que solo el valor de p.

Sin embargo, sustituir valores de p por intervalos de confianza no es realmente cambiar" lo malo por lo bueno" sino añadir información al resultado. Los intervalos de confianza utilizan las mismas técnicas estadísticas que el valor de p al punto de que de hecho son equivalentes [17]. Por tanto resulta engañoso pensar en los intervalos de confianza como algo diferente al valor de p, solo constituye una información adjunta. Ronald Fisher nunca propuso que este valor (de p) fuera tomado como una regla rígida para tomar una decisión sino como una ayuda valiosa para implementar en la inferencia estadística [18],[19].

Otras alternativas han sido propuestas en la literatura pero, en general, son más difíciles de entender y asimilar. Por ejemplo, el uso de cocientes de verosimilitud que comparan el grado de ajuste de dos modelos a la realidad [5],[20]. En el ejemplo general de comparación de eficacias de una nueva intervención con la convencional, se tendría un modelo que se basa en que la nueva intervención es más eficaz que la convencional y otro modelo se basa en que ambas eficacias son iguales. Se obtienen los datos (se realiza el estudio correspondiente) y se acepta el modelo que más se ajusta a los datos obtenidos. Parece atractivo, pero no ha tenido éxito en la sustitución de las pruebas de hipótesis aunque sí constituye un método básico para evaluar diferentes modelos de regresión y se utiliza con profusión en otras áreas del conocimiento.

**¿Se reivindica el valor de p?**

A modo de colofón, y como ejemplo de que la famosa expresión de Sócrates "solo sé que no sé nada" es -y probablemente lo será siempre- vigente, he de apuntar que varias voces en el ámbito de la ciencia, reivindican el uso del valor de p. Las más directas parecen haber sido las de Joseph Fleiss en 1986 [21] y la de Clarice R. Weinberg en 2001 [22]. Esta última en su comentario publicado en la revista *Epidemiology*, responde a la tendencia a calificar el uso del valor de p como "políticamente incorrecto" que prosperó a finales de la década de los 90 del siglo pasado. Ambos son pródigos en ejemplos de que, en muchos momentos de la aplicación de las técnicas estadísticas al análisis de datos, puede ser conveniente el uso de valores de p. Se pueden mencionar entre estas situaciones las que implican la evaluación conjunta de hipótesis múltiples o se pone en duda la pertinencia de cierto modelo.

La idea prevalente es que el valor de p ha de tomarse como lo que realmente es, una ayuda a la toma de decisiones que no puede abstraerse del contexto donde se realiza y mucho menos tomarse como sustituto de todo el razonamiento práctico/científico que rodea cualquier valoración del conocimiento y sus aplicaciones. La muy reciente declaración de la *American Statistical Association* (ASA) es elocuente y clara a este respecto [23].

### Conclusión

El valor de p unido a la significación estadística y la significación práctica de resultados encontrados en un estudio han sido motivo de discusión en el escenario de la investigación científica, ya por décadas. La balanza sin dudas se inclina hoy por confirmar que, en la mayoría de los casos, no debe haber interés en una mera significación estadística y que toda prueba de significación (si se realiza) deberá acompañarse de una demostración de su posible significación práctica. No obstante el tema es amplio y actual porque todavía no existe consenso general sobre cuál podría ser la mejor, menos riesgosa y más rápida, manera de inferir desde la muestra –que representa un estudio en particular- a la población donde se hacen las preguntas.

## Notas

### Agradecimientos
A Vivienne C. Bachelet, por su revisión crítica del manuscrito y aportes a la claridad del texto.

### Conflictos de intereses
La autora declara que no existe ningún conflicto de intereses relacionado con el tema que se aborda.

### Financiamiento
La autora declara que no ha recibido financiamiento alguno para la confección de este artículo.

## Referencias

1. Bertolaccini L, Viti A, Terzi A. Are the fallacies of the P value finally ended? J Thorac Dis. 2016 Jun;8(6):1067-8. | CrossRef | PubMed |
2. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods Nature Research. 2015 Feb 26;12(3):179-85. | Link |
3. Huber W. A clash of cultures in discussions of the P value. Nat Methods. Nature Research. 2016 Jul 28;13(8):607-607. | Link |
4. Woolfson M. M. 1.1. Probability and Everyday Speech. Everyday probability and statistics health, elections, gambling and war. 1st ed. London: Imperial College Press; 2008: 5-6.
5. Dixon P. The p-value fallacy and how to avoid it. Can J Exp Psychol. 2003 Sep;57(3):189–202.| PubMed |
6. Fisher RA. The Design of Experiments. 2nd ed. Edinburgh: Oliver and Boyd; 1937.
7. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995 Aug 19;311(7003):485. | PubMed |
8. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008 Jul;45(3):135-40. | CrossRef | PubMed |
9. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986 Mar 15;292(6522):746-50. | Link |
10. Goodman SN, Hopkins J. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999;130:995-1004. | Link |
11. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51(4):355-60. | PubMed |
12. Silva Ayçaguer LC. [Confidence intervals and p values]. Medwave. 2014;14(1):e5894. |CrossRef | PubMed |
13. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. J Am Stat Assoc. 1987 Mar;82(397):112-22. | Link |
14. Blackwelder WC. "Proving the null hypothesis" in clinical trials. Control Clin Trials. 1982;3(4):345-53. | PubMed |
15. Silva LC. 6.5.3. El advenimiento de la era bayesiana. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997:156–7.
16. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA. 1987;257(18):2459-63. | PubMed |
17. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ. 2011;343:d2304. | PubMed |
18. Buyse M, Hurvitz SA, Andre F, Jiang Z, Burris HA, Toi M, et al. Statistical controversies in clinical research: statistical significance-too much of a good thing ….Ann Oncol. 2016 May;27(5):760-2. | CrossRef | PubMed |
19. van Helden J. Confidence intervals are no salvation from the alleged fickleness of the P value. Nat Methods. Nature Research; 2016 Jul 28;13(8):605-6. | Link |
20. Glover S, Dixon PLikelihood ratios: a simple and flexible statistic for empirical psychologists. Psychon Bull Rev. 2004 Oct;11(5):791-806. | PubMed |
21. Fleiss JL. Significance tests have a role in epidemiologic research: reactions to A. M. Walker. Am J Public Health. 1986 May;76(5):559-60. | PubMed |
22. Weinberg CR. It's time to rehabilitate the P-value. Epidemiology. 2001 May;12(3):288-90. |PubMed |
23. Wasserstein RL, Lazar NA. The ASA's Statement on p -Values: Context, Process, and Purpose. Am Stat. Taylor & Francis; 2016 Apr 2;70(2):129-33. | Link |

**Correspondencia a:**
**[1]** Villaseca 21 Of. 702
Ñuñoa
Santiago
Chile