# Las pruebas de significación estadística: seis décadas de fuegos artificiales

## Tests of statistical significance: six decades of fireworks

## As provas de significação estatística: seis décadas de fogos artificiais

*Luis C. Silva-Aycaguer*[1]

[1] Doctorado en (), Escuela Nacional de Salud Pública, La Habana, Cuba. Correo electrónico: lcsilva@infomed.sld.cu

Recibido: 12 de febrero de 2016. Aprobado: 20 de mayo de 2016. Publicado:

Silva-Aycaguer LC. Las pruebas de significación estadística: seis décadas de fuegos artificiales. Rev. Fac. Nac. Salud Pública 2016; 34(3):

## Resumen

Tras varios decenios de críticas a las técnicas inferenciales basadas en las pruebas de significación estadística orientadas al rechazo de la llamada "hipótesis nula" y, a pesar del notable consenso alcanzado entre los estadísticos profesionales, este recurso se mantiene vigente tanto en las publicaciones biomédicas, entre ellas las de Salud Pública, como en cursos introductorios de estadística. Entre las muchas deficiencias señaladas por los más prominentes especialistas se destacan tres por ser las más obvias y fáciles de comprender: que no contribuyen a cumplimentar la encomienda de la ciencia, que se conocen de antemano las respuestas a las preguntas que se encaran por su conducto y que los resultados que producen dependen de un elemento ajeno a la realidad estudiada: el tamaño muestral. El artículo discute en detalle tales limitaciones, ilustra su perniciosa presencia en la investigación actual y valora las razones para la subsistencia de la sinrazón en esta materia.

----------*Palabras clave:* inferencia estadística, prueba de significación estadística, intervalo de confianza, tamaño muestral, valores *p*

## Abstract

After decades of criticism against inferential techniques based on statistical significance tests, which mainly reject the so-called "null hypothesis", and in spite of the remarkable consensus among professional statisticians, this resource remains prevalent in both biomedical publications (including public health journals) and introductory statistics courses. Among the many problems identified by the most prominent specialists, three of them are the most obvious and easy to understand: that these tests do not contribute to the actual enterprise of science, that the answers to the questions that are addressed are known in advance and that their results depend critically on an element that is external to the domain that is being studied: sample size. This paper discusses in detail these limitations, illustrates their pernicious presence in current research and evaluates the reasons for the survival of the senselessness in this matter.

----------*Keywords:* statistical inference, significant statistical tests, confidence intervals, sampling size, p-values.

## Resumo

Trás vários decênios de críticas as técnicas inferenciais baseadas nas provas de significação estatística orientadas ao rejeito da chamada "hipótese nula" e, embora do notável consenso alcançado entre os estatísticos profissionais, este recurso se mantem vigente tanto nas publicações biomédicas, entre elas as de Saúde Pública, como nos cursos introdutórios de estatística. Entre as muitas deficiências assinaladas pelos mais proeminentes especialistas se destacam três por ser as mais obvias e fácies de compreender: que não contribuem a complementar a encomenda da ciência, que se conhecem de antemão as respostas ás perguntas que se encaram pelo seu conduto y que os resultados que produzem depende dum elemento alheio á realidade estudada: o tamanho amostral. O artigo discute em detalhe tais limitações, ilustra a sua perniciosa presença na investigação atual e valora as razões para a subsistência da sem-razão em esta matéria.

---------*Palavras chave:* inferência estatística, prova de significação estatística, intervalo de confiança, tamanho amostral, valores p

## Introducción

"Nunca se construyeron templos con filigranas,
ni se ganaron batallas con fuegos artificiales"
*José Ingenieros*

En la década de los 20 del siglo pasado se introdujo un recurso estadístico que vino a revolucionar las técnicas inferenciales. Inicialmente, se trató de un procedimiento ideado por Ronald Fisher para medir el grado de incompatibilidad de los datos con una hipótesis; con él se instituyeron los famosos "valores *p*". Pocos años más tarde, Jerzy Neyman y Egon Pearson proponen un procedimiento emparentado con la propuesta fisheriana, aunque operativa y epistemológicamente diferente [1], concebido para elegir una de dos hipótesis complementarias. Desde los años 50, de una manera más bien inercial, se arraiga el método híbrido, basado en la fusión de ambos enfoques, que regularmente se emplea en la actualidad y al que aquí llamaremos "pruebas de significación estadística" (PSE).

Retroalimentadas por una sostenida presencia en las revistas científicas y abonadas por el acceso universal a poderosos recursos computacionales que facilitan su aplicación, las PSE inundan la investigación biomédica contemporánea.

Tal recurso, sin embargo, se ha visto crecientemente desacreditado entre los estadísticos profesionales. Las numerosas críticas a cargo de renombrados especialistas apuntan tanto a sus endebleces lógicas, como a sus limitaciones epistémicas y prácticas [2-11]. Algunos han sido especialmente cáusticos al valorarlas; tal es el caso de Marks Nester cuando hace unos años afirmaba en una revista altamente especializada que "la aceptación generalizada de las PSE es uno de los aspectos más desafortunados de la ciencia aplicada en el siglo XX" [12]. Pese a la extendida opinión de que las PSE no solo son inútiles sino que pueden y suelen ser dañinas para alcanzar los propósitos de la ciencia [13-17], ese rico acervo crítico es sistemáticamente omitido en los libros de texto y cursos introductorios de estadística.

Tal realidad ha traído consigo no pocas conductas rituales y estériles. La reproducción inercial de patrones produce una retroalimentación mutua entre profesores, autores, árbitros y editores que ha ralentizado notablemente el proceso natural que, paulatinamente, se ha ido imponiendo en la práctica: su erradicación y suplantación por recursos racionales y fecundos.

El enfoque de las PSE ignora que el conocimiento se construye de manera colectiva y crea la ilusión de que cada trabajo puede dar una respuesta definitiva. Sin embargo, nuestras convicciones científicas siempre son provisionales y han de estar abiertas a cambios y perfeccionamientos en la medida en que nuevos datos lo aconsejen. La consolidación del nuevo conocimiento es gradual y cualquier aporte metodológicamente riguroso es bienvenido. Unos serán más trascendentes y otros menos, pero todos pueden hacer alguna contribución en el cambiante proceso de construcción racional del conocimiento. Tal es la plataforma conceptual sobre la que se erigen, tanto el meta-análisis para la optimización de estimaciones [18], como el enfoque bayesiano [19], dos alternativas llamadas a complementarse [20]. Ambos incorporan el conocimiento acumulado en lugar de sobrevalorar los aportes aislados a la vez que se desentienden de la pueril dicotomía "significación-no significación".

Lo más inquietante no reside tanto en las manifiestas deficiencias de las PSE como en el hecho de que su predominio es enorme a pesar de ellas. La avalancha de sólidas objeciones que se produjo a lo largo de varias décadas no ha sido suficiente para superar la fuerte sedimentación del método. En uno de los artículos críticos más citados se daba cuenta de ello: "Con cientos de artículos ya publicados que critican acerbamente a las PSE, tuve dudas acerca de la pertinencia de escribir uno más" [21].

Tres lustros después, me embarga un sentimiento similar; pero la situación es tal que sigue siendo preciso el esclarecimiento en el marco de los investigadores no especializados en el tema [22]. El propósito que me anima no es, sin embargo, ofrecer un recuento de las diversas

insuficiencias que padecen las PSE. Partiendo de que el lector domina los códigos operativos de su aplicación, me concentraré en los que a mi juicio constituyen sus tres principales y más graves imperfecciones, a saber: que no responden a las preguntas que realmente se formula la ciencia empírica, que ya conocemos la respuesta a la pregunta que ellas encaran, y que sus resultados dependen vitalmente de un elemento ajeno a la realidad que se examina: el tamaño muestral. Mientras que la mayoría de las críticas tienen cierta complejidad técnica, las mencionadas son quizás las más sencillas, y su comprensión exige apelar a poco más que al sentido común y la intuición.

**La función de las pruebas de significación**

Independiente de la variedad de situaciones en que se aplican, así como de las cuantiosas expresiones concretas que pueden adoptar, la única función que cumplen las PSE es la de valorar si existe o no suficiente evidencia muestral como para rechazar la validez de cierta conjetura: la llamada "hipótesis nula", frecuentemente denotada por $H_0$ lo general, $H_0$ expresa que no hay diferencia alguna entre varios parámetros (por ejemplo, que son idénticas las probabilidades de recuperación asociadas a respectivos tratamientos médicos). El caso en que $H_0$ alude a una desigualdad que se maneja en algunos textos, pero es virtualmente inexistente en la práctica real, en virtud de lo cual no es objeto de análisis en el presente artículo.

Luego de observar el resultado que arroja una muestra concreta, se calcula la probabilidad que, bajo el supuesto de que la hipótesis nula es cierta, se haya obtenido dicho resultado o uno más alejado de él en dirección opuesta a lo que dicha hipótesis afirma. Tal probabilidad condicional es lo que se conoce como "valor *p*". Para decidir si "se rechaza" o no $H_0$, p se compara con determinado umbral α prefijado (usualmente, igual a 0,05): si es menor que α, se rechaza la hipótesis nula, en tanto que si no lo supera, el investigador se abstiene de rechazarla.

Para comprender cabalmente las ideas que subsiguen, resulta crucial mantener en mente que la valoración que se realiza a través de una PSE no es si determinado efecto es "significativo" en el sentido convencional del término (es decir, "importante", "trascendente" o que tiene algún significado cualitativo relevante) sino, estrictamente, si hay motivos para descartar que dicho efecto es nulo.

**Respondiendo la pregunta equivocada**

Una de las limitaciones más notables de las PSE es que están concebidas para resolver un problema cuya solución no interesa y, por lo tanto, nos distrae de los verdaderos objetivos de la ciencia: centrarnos en nuestras hipótesis, en la magnitud de los efectos y su significación práctica, y en la acumulación gradual de conocimientos [23]. Un importante estadístico francés recalcaba hace pocos años: "La PSE es un método inadecuado para el análisis de datos experimentales […] simplemente porque no atiende a las preguntas que la investigación científica reclama" [24].

Veámoslo con un ejemplo sencillo. Supongamos que en el manejo del cáncer de mama se consideran dos opciones: mastectomía radical y tratamiento con citostáticos. Llamemos $E_1$ y $E_2$ a las magnitudes de las supervivencias que corresponden respectivamente a uno y otro procedimiento y llamemos Δ a la magnitud Δ = $E_1 - E_2$ que expresa cuánto mayor es la esperanza de vida corrêspondiente a la primera terapia que la asociada a la segunda.

¿Cuál es la pregunta de interés para las pacientes o para el gestor de salud? Lo que interesa a un salubrista es si vale la pena sugerir el uso generalizado de una terapia por encima de la otra; lo que resultaría útil a una paciente es contar con datos que le ayuden a decidirse por una u otra alternativa terapéutica. Una PSE aplicada para comparar las estimaciones de $E_1$ y $E_2$ podría permitir el rechazo de la nulidad; pero el juicio que merezcan los tratamientos solo podrá realizarse si se contempla la magnitud de Δ. Por ejemplo, saber que dicha diferencia asciende a 7,2 años puede llevar a una decisión y si se supiera que es de 2,1 meses, el proceso racional de decantarse por una de las dos opciones podría producir la contraria. Conocer que Δ=0 también podría ser útil, pero las PSE no tienen la capacidad de corroborar la nulidad; solo permiten en el mejor de los casos, descartarla. Sin embargo, el dato de que Δ≠0 solo informa que hay infinitos valores posibles y, por ende, no resulta útil en sentido alguno.

Sintetizando, lo que importa es conocer la magnitud de la diferencia y el valor de *p* no aporta nada ni para estimarla, ni para enjuiciar la calidad de dicha estimación, ni para valorar si ella es o no relevante en determinado contexto.

Esta endeblez medular del método fue una de las razones principales que motivaron la advertencia reiterada a mediados del siglo pasado, entre otros, por cuatro de los más prominentes especialistas de la estadística [25-28], estas pruebas son totalmente irrelevantes para el progreso de la ciencia.

**La nulidad no pasa de ser una ilusión**

El rasgo más chocante de las PSE no estriba tanto en que responderían a la pregunta que no interesa como en que ya se conoce la respuesta a dicha pregunta sin necesidad de realizar prueba alguna. Más concretamente, la mayor evidencia de la esterilidad de una PSE la aporta el hecho de que solo sirve para enjuiciar la validez de algo cuya falsedad ya se sabe de antemano. En efecto, prácticamente siempre se tiene certeza anticipada de que la hipótesis nula es insostenible. Así lo señalaba persuasivamente

David Bakan hace más de 40 años: "Es un hecho objetivo que casi nunca hay buenas razones para esperar que la hipótesis nula sea verdadera. ¿Por qué razón la media de los resultados de cierta prueba habría de ser *exactamente* igual al este que al oeste del río Mississippi?, ¿por qué deberíamos esperar que un coeficiente de correlación poblacional sea igual a 0,00? ¿Por qué esperar que la razón mujeres/hombres sea *exactamente* 50:50 en una comunidad dada? o ¿por qué dos drogas habrán de producir *exactamente* el mismo efecto?" [29].

Si dos tratamientos rivales para determinada dolencia (por ejemplo, medicamentos con principios activos diferentes) produjesen *exactamente* el mismo efecto, estaríamos ante algo prodigioso. Como expresaba el célebre profesor Edward Deming de la *Universidad de Columbia*: "Un experimento no se realiza para delimitar si dos variedades de trigo o dos drogas son iguales. Sabemos de antemano, sin gastar un solo dólar, que no lo son" [30]. Efectivamente, ningún fármaco producirá el mismo efecto que un placebo, ni se hallarán dos arbustos idénticos, ni una moneda de curso legal perfectamente equilibrada. No se necesita prueba alguna para saberlo: la nulidad no se presenta en la naturaleza, ni en la sociedad, ni en la tecnología. Si dos entes cualesquiera parecen iguales, bastaría examinarlos con suficiente esmero —o más detenidamente, o con tamaños de muestras mayores— para disipar cualquier duda al respecto.

**La realidad científica no depende de los recursos de que dispone quien la examina**

Un tercer importante demérito de los valores *p* se deriva de que la magnitud del efecto observado se "mezcla" inseparablemente con el tamaño muestral: a un pequeño efecto en un estudio con un tamaño de muestra grande puede corresponder el mismo valor *p* que a un gran efecto encontrado en una muestra pequeña. Más específicamente: el valor *p* se puede reducir tanto como se desee virtualmente en cualquier situación práctica; basta con tomar una muestra suficientemente grande. En palabras de Thompson: "Las PSE se reducen a una búsqueda tautológica de suficientes sujetos para alcanzar significación estadística. Si no se consigue el rechazo, ello es debido exclusivamente a que hemos sido demasiado perezosos para conseguir suficientes participantes" [31].

Una de las primeras declaraciones que advirtieron con toda claridad esta realidad fue debida a Leonard Savage, uno de los estadísticos más brillantes del siglo pasado, de quien el premio Nobel Milton Friedman dijo en sus memorias de 1998 [32] que era "una de las pocas personas que conocía a quien calificaría sin dudarlo como un genio". En las propias palabras de Savage: "Con extrema frecuencia se sabe de antemano que las hipótesis de nulidad son falsas sin necesidad de recoger los datos; el rechazo o la aceptación de la hipótesis nula es entonces, un mero reflejo del tamaño de la muestra y no hace, por tanto, contribución alguna a la ciencia" [33].

En síntesis, basta que haya una diferencia, por minúscula o intrascendente que sea (incluso el más mínimo sesgo), para que la diferencia cuya nulidad es proclamada por $H_0$ pueda ser declarada "estadísticamente significativa". La circularidad de las PSE (consistente en usar muchísimos datos para luego corroborar que se usaron muchísimos datos) no sólo es bien conocida en el mundo académico, sino que desde años atrás ocupa un lugar en el marco mediático. Por ejemplo, un periodista de *Wall Street Journal* advertía a sus lectores: "Ud. puede probar cualquier hipótesis, por estúpida que sea, llevando adelante una prueba estadística con toneladas de datos" [34].

Esta es una muy grave imputación, pues nos dice que un juicio sobre la realidad examinada queda en manos de un elemento completamente ajeno a ella: los recursos disponibles. No es casual que una afirmación tal como "No encontramos significación, pero con un tamaño de muestra mayor probablemente la hubiéramos hallado" aparezca recurrentemente en la literatura. La intrascendencia de tal afirmación dimana de que *siempre* se puede echar mano de esa verdad. Como ocurre con toda tautología, enunciarla no agrega absolutamente nada de interés. Incidentalmente, lo que no se hallará jamás en la literatura es la afirmación complementaria, a pesar de ser tan válida como la anterior. Es preciso decir que no existe trabajo alguno donde se haya informado al lector que: "Si bien encontramos significación, téngase en cuenta que, con un tamaño de muestra menor, probablemente no la hubiéramos hallado". Si no hubiera otras razones, bastaría este doble rasero para reclamar que se recupere la sensatez.

Considere el lector el siguiente consejo: "Puesto que Ud. sabe que con una lupa adecuada se podrá constatar que estas dos hormigas, muy parecidas entre sí, no son exactamente iguales, emplee una lupa más pequeña que mantenga en pie la hipótesis de que son idénticas".

Una conducta así de irracional se sugiere en un artículo muy reciente [35], destinado a dar indicaciones prácticas acerca de cómo manejar las PSE. Allí se plantea que resulta peligroso o dañino trabajar con muestras "demasiado" grandes. Ello produce, según los autores, una "exagerada tendencia a rechazar la hipótesis nula cuando las diferencias son clínicamente despreciables" y para conjurarlo llegan, incluso, a dar una orientación insólita: cuando se tienen muestras grandes en un estudio retrospectivo, el investigador debe seleccionar una submuestra al azar y aplicar las PSE desdeñando el resto de la información.

Sugerir que no se tome una muestra demasiado grande porque, de hacerlo, el investigador se enteraría de la verdad, es simplemente pueril. Es como poner

balas de salva a un rifle con la finalidad de que sobreviva la presa a la que se le dispara. Estas sugerencias entrañan una confesión implícita de la insuficiencia capital que nos ocupa, a la vez que nos convocan a sacrificar el sentido común en el altar de un dogma metodológico.

Situaciones como esta traen a la mente un *dictum* que se atribuye a Winston Churchill: «En ocasiones, el hombre tropieza con la verdad, pero, casi siempre, evita caerse y sigue adelante». Las tecnologías no pueden ser vistas como una finalidad en sí mismas, que han de salvarse independientemente de lo que se pueda conseguir con ellas. Este sonambulismo tecnológico ignora una regla tan básica que cabe rememorar la ya cincuentenaria reflexión [36] que, en relación con este tema, advertía: "Cuando se llega al extremo en que los procedimientos estadísticos pasan a suplir nuestros pensamientos en lugar de estar en función de ellos, y somos conducidos así al reino del absurdo, entonces es el momento de regresar al sentido común".

## Las alternativas

A la luz de las diversas insuficiencias y, en particular, a raíz de las tres expuestas en las secciones precedentes, a lo largo de los últimos 20 años se han propuesto diversos procedimientos alternativos a las PSE. Uno de los más connotados es el empleo del enfoque bayesiano y el llamado "factor de Bayes" en particular. Sin embargo, solo uno de ellos, el más simple, ha contado con un respaldo generalizado e inequívoco: comunicar la magnitud del efecto y acompañarlo de una medición del error asociado a su estimación. De hecho, prácticamente todos los autores de juicios críticos que se han citado en el presente artículo sugieren prescindir de los valores *p* y, en su lugar, ceñirse al empleo de intervalos de confianza (IC). El argumento central estriba en que los IC proveen más información que las PSE, a la vez que no obligan a dicotomizar las conclusiones. Constituyen un recurso para resumir tanto la magnitud de un efecto como el grado en que el conocimiento de la verdadera diferencia es adecuado. De modo que, a diferencia de los valores *p*, abren el camino para responder a la pregunta que interesa, aunque, este terminará de transitarse cuando se incorporen otras consideraciones clínicas o salubristas ajenas a la estadística. Por otra parte, si bien los IC también dependen vitalmente del tamaño muestral, ello no entraña problema alguno. Cuando se emplean ramplonamente como meros sucedáneos de las PSE (se rechaza $H_0$ si el IC no contiene la nulidad y viceversa), padecen de los mismos defectos que las PSE porque, con un tamaño de muestra suficientemente grande, el IC se podrá estrechar tanto como se quiera hasta dejar la nulidad fuera de sus límites. Sin embargo, si se usan para complementar las estimaciones puntuales con una medida del error que pudieran afectarlas, entonces, cuanto mayor sea el tamaño de muestra, con más precisión conocerá el investigador la magnitud del efecto.

Aunque reivindicada desde mucho antes, esta propuesta alcanzó gran notoriedad en el campo de la ciencias de la salud a raíz de un artículo publicado en *British Medical Journal* [37] que supuso un impulso medular para que en 1988 el llamado *International Committee of Medical Journal Editors* (ICMJE, conocido como el *Vancouver Group*) la incluyera dentro de sus sugerencias en los siguientes términos: "Siempre que sea posible, cuantifique los resultados y preséntelos con indicadores apropiados de error o incertidumbre de la medición (por ejemplo, intervalos de confianza). Evite la dependencia exclusiva de las pruebas de hipótesis estadísticas, tal como el uso de los valores *p*, que no expresan ninguna información cuantitativa importante" [38].

Esta recomendación, que desvaloriza inequívocamente a las PSE, ha llegado a revistas tan encumbradas como *Nature* [39] y ha sido convalidada en una reciente guía para la comunicación de resultados estadísticos elaborada por experimentados metodólogos [40]. Por su parte, la poderosa Asociación Americana de Psicología (APA, por sus siglas en inglés), tras algunos titubeos al respecto en versiones anteriores de sus "*guidelines*", establece actualmente que, al valorar hipótesis, la comunicación apropiada de los tamaños de los efectos con sus intervalos de confianza es la "expectativa mínima" que deben satisfacer los artículos publicados en todas las revistas amparadas por la APA [41]. Una de ellas, *Basic and Applied Social Psychology*, ha ido incluso más lejos y acaba de comunicar que los valores *p* ya no podrán ser incluidos en trabajo alguno [42]. Y muy recientemente, la *American Statistical Association* (ASA) ha hecho pública una declaración [43] donde se resume el hondo malestar prevaleciente con el modo en que se aplican cotidianamente los valores *p*. Algunos la han interpretado como una advertencia sobre el uso incorrecto de este instrumento; otros, como una clara invitación a abandonar su empleo. Por ejemplo, el famoso profesor Norman Matlof de la Universidad de California, está en el segundo caso y en el debate producido en torno al tema [44] afirma que nadie ha podido poner un solo ejemplo convincente de la utilidad pasada o presente de los valores *p* en el proceso de construcción de nuevos conocimientos.

## Filigranas y fuegos artificiales en la práctica

En los estudios descriptivos, la esterilidad de las PSE es particularmente evidente, pero, en esos casos, son relativamente inofensivos. En el contexto de los estudios explicativos, en cambio, especialmente en los ensayos clínicos controlados, el impacto negativo del dogma de la significación puede ser particularmente dañino. Como la posibilidad de hallar significación ("detectar efectos" es el engañoso término que suele emplearse) se incrementa con el tamaño muestral, es obvio que solo

los muy poderosos —en especial, la farmaindustria— adquieren privilegios a la hora de hacer "aportes" al valorar los fármacos [45].

El hallazgo de significación estadística, casi siempre al amparo de muestras enormes, es usado por la industria para legitimar y promover comercialmente medicamentos que carecen de relevancia práctica.

Consideremos en detalle un ejemplo especialmente expresivo a los efectos de ilustrar el espurio *modus operandi* de las empresas aprovechando la presunta capacidad de los "*p*-values" para convalidar las generalizaciones. Se trata de un ensayo clínico publicado en *The Lancet* sobre el valor preventivo de dos inhibidores de la agregación plaquetaria: Plavix® (clopidogrel) y Aspirina® (ácido acetil salicílico) [46].

En el artículo se "prueba" que el Plavix® (comercializado por *Sanofi*, laboratorio que financió el estudio) es más eficaz que la Aspirina® para reducir la aparición de eventos vasculares graves en población con alto riesgo de padecerlos.

La población objeto de estudio fue la de aquellos sujetos con historia reciente de infarto de miocardio dentro de los 35 días anteriores, o con antecedentes de un accidente cerebrovascular isquémico durante los 6 meses anteriores, o signos neurológicos residuales de por lo menos una semana de evolución, o con enfermedad arterial periférica objetivamente establecida.

Tras una asignación aleatoria, los dos tratamientos se aplicaron a unas 20 mil personas con los rasgos mencionados. El *endpoint* fue la ocurrencia de un nuevo accidente cerebrovascular isquémico o un nuevo infarto al miocardio, fatales o no, o cualquier muerte de origen vascular.

El megaestudio permitió estimar que el riesgo ascendió a D1=583 eventos por cada 10 mil "años-personas" de tratamiento con Aspirina® y a D2=532 por cada 10 mil "años-personas" cuando se empleó Plavix®. Con los datos que se incluyen en el artículo se pueden calcular los intervalos de confianza respectivos; ellos son: 499-568 por 10 mil años-personas para el Plavix® y 548-619 por 10 mil años-personas para la Aspirina®. Tras aplicar una PSE, el estudio arriba triunfalmente a un valor $p = 0{,}043$.

Esto significa que ambas tasas de incidencia se han estimado con altísima precisión, como era de esperar, pues se obtuvieron con tamaños de muestras enormes (alrededor de 18 mil años-personas en cada uno de los brazos). La extraordinaria precisión salta a la vista: es fácil corroborar que el error relativo correspondiente a cada una de las estimaciones es aproximadamente igual al 6%, muy por debajo del 10-15% de error relativo exigido en los libros clásicos para considerar que estamos ante una "buena" estimación. Vale decir: conocemos entonces la magnitud del efecto virtualmente sin error y podemos confiar, por tanto, en que los efectos son prácticamente indistinguibles desde la perspectiva salubrista. Cabe señalar que el artículo (y también las reseñas de prensa) resaltaron que el riesgo fue un 9% mayor para la Aspirina® que para el Plavix® (632/583=1,09). De tal suerte, se oculta que el riesgo atribuible, el que realmente importa para la salud pública, resultó ser nimio. En efecto, la colectividad de pacientes de alto riesgo se ahorraría apenas D1-D2=51 eventos por cada 10 mil años de tratamiento preventivo si este se basa en Plavix® en lugar de en Aspirina®.

Consecuentemente, en principio un decisor racional de la salud pública, supuesto que se base en este estudio, se abstendría de recomendar el Plavix®. Pero incluso, si alguien considerara que tal diferencia es "importante", lo que está fuera de discusión es que el valor de *p* no participa ni tendrá jamás sentido que participe en la recomendación que se adopte. Cualquier decisión se puede (y se debe) tomar sobre la base de los valores de D1 y D2, acompañados de otros datos tales como el costo y los efectos adversos. El único papel que desempeñó la *p* en el artículo (y luego en sus numerosas citaciones) es dar cobertura a quienes tratan de vendernos el Plavix®, aprovechando la inercia acrítica en el uso de la estadística y la sumisión a las PSE, así como en el truco de escamotear el significado real en términos prácticos de aquello que recomiendan.

Obviamente, el tratamiento patrocinado por las empresas no necesariamente será siempre mejor que aquel con el cual se compara, aunque las transnacionales suelen arreglárselas para que así sea, al suprimir algunos pacientes en uno de los grupos, rebajar o incrementar una dosis según convenga, definir muestras sesgadas, etc., de todo lo cual hay sobradas pruebas [47]. Lo que sí sabemos es que los fármacos siempre serán diferentes. La Aspirina®, por ejemplo, no puede producir exactamente el mismo efecto que el Plavix® por la sencilla razón de que el ácido acetil salicílico y el clopidogrel son dos principios activos diferentes. Pero en general, la diferencia que, valorada con un tamaño de muestra suficientemente grande producirá una *p* tan pequeña como queramos, puede favorecer a uno o al otro medicamento. Cuando no puede o no quiere manipular los datos indeseables, la industria suele resolver este percance de una manera muy simple: no publicarlos. En este ejemplo, si el resultado hubiera favorecido a la Aspirina®, con altísima probabilidad, el artículo no hubiera visto la luz.

El reporte de este tipo de resultados no es una excepción. Por más señas, el filósofo y economista Germán Velásquez, señalaba en octubre de 2015 que: "A la industria farmacéutica lo que menos le preocupa es el paciente. Trata de asegurar sus beneficios. Un estudio de la revista *Lancet,* que analiza los 70 medicamentos contra el cáncer que fueron puestos en el mercado en los últimos 10 años en EEUU, demuestra que su único beneficio es prolongar la vida del

paciente una media de dos meses. Algunos de estos fármacos cuestan hasta 100.000 dólares" [48]. La dictadura de las PSE conserva plena vigencia en la literatura biomédica [49, 50]. Gracias a ella se produce lo que señala el profesor finlandés Esa Läärä: "La comunicación de 'resultados' triviales y sin sentido que no proveen información cuantitativa adecuada que tenga interés científico, es masiva" [51].

## A modo de resumen: la situación actual

Los editores de muchas revistas médicas y otras prominentes figuras académicas hacen una contribución especialmente perniciosa para que las PSE se perpetúen como el recurso por antonomasia sobre el que reposaría la buena ciencia [52,53]. Esta es una de las razones para que muchos sigan viéndolas como una salvaguarda efectiva contra hallazgos espurios. "La mayor ironía —escribía el especialista de *Intel Corporation*, Charles Lambdin— reside en que nuestras revistas arbitradas, nuestro juez supremo de lo que cuenta como escritura científica, es parcialmente culpable al mantener viva la tiranía de las PSE" [54].

Sociológicamente, se trata de un interesante ejemplo de endogamia metodológica que explica el altísimo coeficiente de rozamiento con que nos desplazamos hacia un paradigma racional y lógico. Los comités editoriales y grupos académicos suelen estar presididos o dirigidos por personajes de edad avanzada, acaso en la etapa final de su carrera y atenazados por una enorme pereza para reconsiderar sus rutinas. Muchos son profesores de estadística que aplican y explican por decenios el ritual consagrado, al usar las mismas diapositivas, con escasas variaciones en sus ejemplos y diciendo los mismos chistes en clase. Y, por añadidura, son los que más poder académico tienen para perpetuar sus tradiciones y el dogma que les inmoviliza. He tenido oportunidad de presenciar reacciones irritadas por parte de árbitros y de participantes en foros que se desarrollan en las redes sociales, donde la exasperación desplaza al pensamiento racional. A los propios autores les resulta mucho más cómodo el enfoque convencional que una aproximación que exija un pensamiento profundo y sopesado [55]. No en balde, Guttman señalaba hace 4 décadas que "la experiencia muestra que la intolerancia suele venir de los firmes creyentes en prácticas sin fundamento" [56]. Lo más curioso es que este tema ya ha dejado de ser controversial entre los estadísticos; el consenso crítico es, simplemente, clamoroso [57], aunque no pocos insisten en la idea de que, si bien los valores *p* "se usan mal" a diario, ellos son útiles cuando "se usan bien". El problema está en que mientras proliferan miles de ejemplos de un "mal uso" cada semana, nadie parece poder dar con una explicación de cómo pueden "usarse bien", ni mucho menos de ilustrarlo de forma transparente y persuasiva.

No obstante, se observa un sostenido crecimiento del porcentaje de artículos que prescinden de los valores *p* y solo usan intervalos de confianza, u otras técnicas que se desentienden de la inútil dicotomía "significación - no significación" [58-59].

Las endebleces intrínsecas del método, unidas al reclamo para que cese el uso adocenado que suele dársele a diario y al espaldarazo que han recibido las posturas "disidentes" por parte de prominentes autoridades y comités especializados, permiten vaticinar que más temprano que tarde se consolide una nueva era [60-61] en la que el uso de los intervalos de confianza y otros recursos desplace definitivamente a las PSE.

## Referencias


1 Gerrodette T. Inference without significance: measuring support for hypotheses rather than rejecting them. Mar Ecol 2011; 32: 404-418.

2 Berkson J. Some difficulties of interpretation encountered in the application of the chi-squaretest. J Am Stat Assoc 1938; 33:526-542.

3 Rozeboom WW. The fallacy of the null hypothesis significance test. Psychol Bull 1960; 56:26-47.

4 Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol 1986; 44: 1276–1284.

5 Chernoff H. A comment. Am Stat 1986; 40(1): 5–6.

6 Berger J. Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. J Am Stat Assoc 1987; 82: 112.

7 Thompson B. In praise of brilliance: Where that praise really belongs. Am Psycholt 1998; 53: 799–800.

8 Goodman SN. Toward evidence-based medical statistics (1): The p value fallacy. Ann Intern Med 1999; 130: 995-1004.

9 Nicholls N. The insignificance of significance testing. B Am Meteorol Soc 2001; 82(5): 981-986.

10 Armstrong JS. Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. Int J Forecasting 2007; 23: 335–336.

11 Hubbard R, Lindsay RM. Why p values are not a useful measure of evidence in statistical significance testing. Theor Psychol 2008; 18: 69–88.

12 Nester MR. An applied statistician's creed. Appl Stat 1996; 45: 401-410.

13 Rozeboom WW. Good science is abductive, not hypothetico-deductive. En Harlow LL, Mulaik SA, Steiger JH (Eds.), What if there were no significance tests? Hillsdale, NJ: Erlbaum; 1997 (pp. 366–391).

14 Ioannidis JPA. Why most published research findings are false. PLoS Med 2005; 2(8): e124.

15 Skipper JK, Guenther AL, Nass G. The sacredness of 0.05: a note concerning the uses of statistical level of significance in social science. Am Sociol 1967; 2: 16–18.

16 Nelder JA. Comment. J Roy Stat Soc A Sta 1985; 148(3): 238.

17 Kelley J. The perils of p-values: Why tests of statistical significance impede the progress of research. Handbook of Evidence-Based Psychodynamic Psychotherapy 2009; 367-377.

18 Cumming G. Understanding the new statistics: Effect sizes confidence intervals and meta-analysis. New York: Routledge; 2012.

19 Matthews WJ. What might judgment and decision making research be like if we took a Bayesian approach to hypothesis testing? Judg Dec Mak 2011;6(8): 843–856.

20 Kruschke JK, Liddell TM. The Bayesian new statistics: two historical trends converge. Judg Dec Mak 2014; 9 (6), 523-547.

21 Johnson DH. The insignificance of statistical significance testing. J Wildlife Manage 1999; 63(3):763-772.

22 Hauer E. The harm done by tests of significance. Accident Anal Prev 2004; 36: 495-500.

23 Kirk RE. The importance of effect magnitude. In S.F. Davis (Ed.), Handbook of research methods in experimental psychology (pp. 83–105). Oxford, UK: Blackwell, 2003.

24 Lecoutre B. Training students and researchers in Bayesian methods for experimental data analysis. J Data Scien 2006; 4: 207-232.

25 Berkson J. Tests of significance considered as evidence. J Am Stat Assoc 1942; 37: 325–335.

26 Yates F. The influence of Statistical Methods for Research Workers on the development of the science of statistics. J Am Stat Assoc 1951; 46: 19-34.

27 Anscombe FJ. Discussion on Dr. David's and Dr. Johnson's Paper. J Roy Stat Soc B Met 1956; 18: 24-27.

28 Savage IR. Nonparametric statistics. J Am Stat Assoc 1957; 52(279):331–344.

29 Bakan D. The test of significance in psychological research. Psychol Bull 1966; 66: 423-437.

30 Deming WE. On probability as a basis for action. Am Stat 1975; 29(4): 146-152.

31 Thompson B. In praise of brilliance: where that praise really belongs. Am Psycholt 1998; 53: 799–800.

32 Friedman M. Two lucky people: Memoirs. Chicago: University of Chicago Press; 1998.

33 Savage IR. Nonparametric statistics. J Am Stat Assoc 1957; 52: 332-333.

34 Albert J. The numbers guy. Periódico Wall Street Journal. 7 de diciembre 2007, Nueva York.

35 Faber J, Martins L. How sample size influences research outcomes. Dental Press J Orthod 2014; 19 (4): 27-29.

36 Bakan D. The test of significance in psychological research. Psychol Bull 1996; 66: 423-437.

37 Gardner MJ. Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Brit Med J 1986; 292, 746-750.

38 International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Ann Intern Med 1988; 108: 258-265.

39 Nuzzo R. Scientific method: statistical errors. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature 2014; 506:150-152.

40 Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL guidelines. En: Science Editors' Handbook. EASE, 2013.

41 American Psychological Association. Publication manual of the American Psychological Association (6th ed). Washington DC, 2010.

42 Trafimow D, Marks M. Editorial. Basic Appl Soc Psych 2015; 37(1): 1-2.

43 Wasserstein RL, Lazar NA. The ASA´s statement on p-values: context, process, and purpose. Am Statist 2016, DOI: 10.1080/00031305.2016.1154108.

44 Matloff N. After 150 years, the ASA says no to p-values. [Internet] Disponible en https://matloff.wordpress.com/2016/03/07/after-150-years-the-asa-says-no-to-p-values/ Consultada el 16 de mayo de 2016.

45 Silva LC. Una pincelada estadística con repercusiones extra-metodológicas. Salud Colectiva 2012; 7(3): 399-400.

46 CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events. Lancet 1996; 348: 1329-1339.

47 Smith R. The trouble with medical journals. J Roy Soc Med 2006; 99:115–119.

48 Flotats A. Entrevista a Germán Velásquez. Periódico El País, 25 de octubre de 2015, Madrid.

49 Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25:225–230.

50 Gigerenzer G. Mindless statistics. SocioEcon 2004; 33: 587–606.

51 Läärä E. Statistics: reasoning on uncertainty, and the insignificance of testing null. Ann Zool Fenn 2009; 46 (2): 138-157.

52 Savitz DA. Tolo KA. Poole C. Statistical significance testing in the American Journal of Epidemiology, 1970-1990. Am JEpidemiol 1994; 139 (10): 1047-1052.

53 Tressoldi PE, Giofré D, Sella F, Cumming G. High impact=high standards? Not necessarily so. PLoS One. 2013; 8(2): e56180

54 Lambdin C. Significance tests as sorcery: significance tests are not. Theor Psychol 2012; 22(1): 67 –90.

55 Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals but can't make them think. Psychol Sci 2004; 15: 119-126.

56 Guttman L. What is not what in statistics? Statistician 1977; 26: 81-107.

57 Gross JH. Testing what matters (If you must test at all): A Context-Driven Approach to Substantive and Statistical Significance. Am J Polit Sci 2015; 59 (3): 775–788.

58 Silva LC, Suárez P, Fernández A. The null hypothesis significance test in health sciences research (1995-2006): Statistical analysis and interpretation. BMC Med Res Methodol BMC Med Res Methodol 2010; 10: 44-53.

59 Fidler F, Burgman MA, Cumming G, Buttrose R. Thomason N. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conserv. Biol. 2006; 20(5):1539–1544.

60 Sedlmeier P. Beyond the significance test ritual. J Psychol 2009; 217(1): 1-5.

61 Odgaard EC, Fowler RL. Statistical reporting practices can be reformed confidence intervals for effect sizes: Compliance and clinical significance in the Journal of Consulting and Clinical Psychology. J Consult Clin Psych 2010; 78: 287–297.