Un punto de inflexión histórico para las pruebas de significación en 2019

Silva Aycaguer, Luis Carlos, Escuela Nacional de Salud Pública de Cuba_

An error does not become truth by reason of multiplied propagation,
nor does truth become error because nobody will see it.
– Mahatma Gandhi

Resumen

Se hace un repaso histórico de los episodios más notables por los cuales han pasado las pruebas de significación estadística desde su surgimiento en los años 20 del siglo pasado hasta la actualidad. Se expone la sostenida aparición de críticas a cargo de famosos estadísticos a lo largo de los años, las cuales se han vertebrado tanto en torno al empleo inapropiado de tales pruebas como a las deficiencias epistémicas que se fueron descubriendo con el paso de los años. Este proceso experimentó un giro notable en marzo de 2019, cuando las voces en defensa de dicho procedimiento inferencial se tornaron definitivamente inaudibles en la literatura especializada. Se da cuenta de las enormes dificultades que entraña todo cambio de paradigma científico, del cual la situación con las pruebas de significación estadística es un ejemplo vívido y actual. Se demuestra que el proceso que se está viviendo en esta materia, al comprometer la línea de flotación de técnicas universalmente empleadas en la producción científica pasada e incluso actual, está llamado a tener un significado epistemológico de gran calibre. Su impacto acaso sea tan notable como el que supuso la propia irrupción de esta técnica hace poco menos de una centuria.

Palabras clave: *Karl Pearson, Ronald Fisher, pruebas de significación estadística, paradigma científico, epistemología*

1. Introducción

A inicios de la segunda década del siglo pasado se introdujeron las *pruebas de significación*, un recurso estadístico que vino a revolucionar las técnicas inferenciales. Pocos años más tarde, se propone una familia de procedimientos emparentados con las anteriores, conocida como *pruebas de hipótesis*.

Desde los años 50, estos enfoques se amalgamaron para desembocar en el procedimiento que regularmente se empleó a lo largo de la segunda mitad del siglo XX: las "pruebas de significación estadística" (PSE) (Silva, 2009). En efecto, alrededor de la Segunda Guerra Mundial, los dos sistemas se unen de manera anónima (Christensen, 2005) en torno a un elemento común, el "valor p", para dar lugar a un recurso inferencial que no satisfaría a ninguno de los estadísticos que crearon los respectivos "ingredientes". Sin embargo, las PSE tuvieron un éxito enorme en la práctica, y su empleo sobrevive hasta nuestros días.

Pero, a la vez, desde muy temprano los *valores p* y las PSE merecieron críticas, que han venido produciéndose sostenidamente a lo largo de más de siete décadas. Su aplicación, en cambio, declinaba a un ritmo mucho menor que el aumento de las objeciones. Sin embargo, no fue hasta 2019 que la polémica

experimentó un cambio verdaderamente medular: virtualmente saturado el platillo de las críticas, el de las recomendaciones de abandonarlos alcanza un peso sustancial y, con él, la balanza refleja una situación cualitativamente nueva. La presente contribución procura dar cuenta de los hitos fundamentales de ese proceso.

2. Desarrollo

2.2 El surgimiento de los valores p: décadas de 1920 y 1930.

La generación formal de los procedimientos que nos ocupan se remiten al primer tercio del siglo XX; sin embargo, como suele ocurrir, existían algunos antecedentes históricos de tales desarrollos. Aparentemente, la primera publicación que contiene una rudimentaria prueba estadística de hipótesis data del siglo XVIII y se atribuye al médico escocés John Arbuthnot (1667-1735), miembro desde 1704 de la elitista *Royal Society* y famoso por haber traducido en 1692 el tratado de Christiaan Huygens sobre probabilidades, primera obra sobre el tema publicada en inglés. Este galeno advirtió la pequeña pero consistente supremacía de nacimientos masculinos sobre los femeninos, registrada a lo largo de 82 años consecutivos considerados en su estudio (Arbuthnott, 1710). Su razonamiento era muy similar al de hoy en día: la probabilidad condicional de tal resultado bajo el supuesto de que en cada año fuese tan probable un desenlace global favorable a los hombres como a las mujeres, resultaba extremadamente pequeña. Consecuentemente, el alto índice empírico de masculinidad al nacimiento, se concluía, sería obra de la intervención divina y no del azar. Fue la primera aplicación de la teoría de probabilidades a la solución de interrogantes de naturaleza social.

Otro importante antecedente, según se ha afirmado (Hogben, 1957), es un trabajo del siglo XIX (Gavarret, 1840) donde se habría consagrado el uso del “error probable” como una prueba de significación en el campo biológico; este mismo autor atribuye al famoso matemático y lógico británico John Venn el haber empleado por vez primera (Venn, 1888) los términos “prueba” y “resultado significativo”. Casualmente, Venn falleció el mismo año en que tales términos aparecieran en lo que pudiéramos llamar la era moderna de las PSE.

Pero esa manera de razonar demoraría decenios en asentarse como un procedimiento formal. En la segunda mitad del siglo XIX y los primeros 20 años del XX se produjo un enorme crecimiento de la literatura científica, especialmente en el campo de la medicina. La estadística, disciplina que ya había irrumpido con claridad en el ámbito de las aplicaciones sociales y epidemiológicas, desempeñaba un papel en esta eclosión: la comparación entre repeticiones de ciertas experiencias, el análisis de diseños factoriales y el empleo de tablas y gráficos estadísticos figuraban con profusión en tales publicaciones (Smith, Best, Cylke y Stubbs, 2000). Las revistas médicas solían dar cuenta de numerosas narraciones acerca de dolencias y tratamientos aplicados, basadas por lo general en testimonios sobre éxitos y fracasos cosechados con unos pocos pacientes. Los testimonios sobre acaecimientos clínicos más o menos aislados abarrotaban las revistas médicas y sus editores reclamaban de los estadísticos que les proveyeran de recursos para distinguir entre las anécdotas y las confirmaciones de las leyes que éstas parecían sugerir.

Estadísticos y biometristas eran por entonces poseedores de notable reputación. Basta reparar en que la primera revista destinada exclusivamente al manejo cuantitativo de datos biológicos, *Biometrika*, fue fundada en 1901, mérito histórico que comparten Karl Pearson, Walter Weldon, y Francis Galton. La revista estaba dedicada al estudio de los problemas teóricos y prácticos planteados a la estadística por la biología, y a la postre sería la publicación más influyente en esta materia a lo largo del siglo XX. El más encumbrado y carismático de sus fundadores, sin duda alguna, fue Pearson, director de la revista desde su fundación hasta su muerte en 1936. El número inaugural rindió homenaje a Charles Darwin y resaltaba sus palabras "*Ignoramus,*

in hoc signo laboremus", una divisa acorde con la propia vida de Karl Pearson, batallador natural a favor de variadas causas tales como la emancipación de la mujer, la ética de la libre expresión, la eugenesia y el socialismo (Williams et al., 2003). Entre los aportes más trascendentes de Pearson se halla el coeficiente que aún se emplea a diario y que lleva su apellido: el "coeficiente de regresión lineal de Pearson", llamado a desempeñar un papel en los turbulentos acontecimientos que se exponen a continuación.

Los *valores p* fueron introducidos por un matemático genial, Sir Ronald Fisher (1890-1962), quien tenía solo 10 años de edad cuando se fundara *Biometrika*. El joven Ronald pronto se destacó, no solo por su brillantez como matemático sino también por su osadía intelectual. En ancas de la vocación irreverente que siempre le caracterizó, demostró que Pearson se había equivocado al deducir la distribución de su famoso coeficiente de correlación lineal. De hecho, muchas de los aportes de Fisher fueron enmiendas y perfeccionamientos de la obra de Karl Pearson. Al publicar esas ácidas evaluaciones, se ganó la animadversión de por vida del más encumbrado biometrista de la época.

En este caldeado ambiente irrumpe la contribución de Fisher. Fue una aportación epistemológicamente trascendente, pues con ella se estaría aportando una medida de la discrepancia de los datos con una hipótesis, lo que permitiría valorar su plausibilidad. El silogismo subyacente proclamaba que si la probabilidad condicional propuesta fuera muy pequeña, entonces tal hipótesis merecería ser puesta en cuestión y demandaría análisis más profundo.

Fisher también dio vida al concepto de "hipótesis nula" H_0 que, por lo general, establece que ciertos parámetros no difieren entre sí. El *valor p* serviría para enjuiciar si vale o no la pena profundizar en la posible validez de tal hipótesis. Su libro *Statistical Methods for Research Workers* (Fisher, 1925) fue sumamente exitoso y le valió a Fisher para ostentar el título informal de "padre de la estadística moderna". El creador de esta teoría nunca respaldó, sin embargo, la aplicación de umbrales rígidos con los cuales comparar el *valor p*, que luego fueran casi universalmente adoptados (aunque la introducción del 0,05 como elemento orientativo tiene también su paternidad).

Pocos años más tarde de la aparición de la teoría de Fisher, Jerzy Neyman y Egon Pearson, hijo de Karl, publicaban un artículo (Neyman y Pearson, 1928) donde se introduce un procedimiento emparentado con la propuesta fisheriana, aunque operativa y epistemológicamente diferente, a la que denominaron *prueba de hipótesis*. Mas tarde, los propios autores (Neyman y Pearson, 1933) refinaron sus argumentos en favor del método concebido para facilitar la *elección* entre dos conjeturas complementarias, bautizadas como hipótesis nula e hipótesis alternativa (Gerrodette, 2011).

Procede recordar que, en este contexto histórico, surgió y se consolidó el influyente cuerpo teórico debido al filósofo Karl Popper (1902-1994). Algunos han sugerido que su obra constituyó fuente de inspiración para la lógica de las PSE, ya que Popper abogaba por el "enfoque falsacionista" que considera el rechazo de las hipótesis como el único camino válido para el avance de la ciencia empírica. Sin embargo, se trata de una falacia; en realidad la influencia mutua entre Popper y los estadísticos contemporáneos fue casi nula, como demuestra inequívocamente García (2003).

2.3 Las críticas más precoces. Período 1940-1970

Desde el surgimiento mismo de las PSE surgieron reparos y recelos. Ya a comienzos de los años 40 aparecen las primeras dudas acerca de la validez del método cuando algunos estadísticos cuestionaron las bases lógicas y la utilidad práctica de los valores p de Fisher.

Tal fue el caso de afamados estadísticos como Joseph Berkson (1899 – 1982), quien no apreciaba que las PSE permitieran separar las evidencias de los deseos (Berkson, 1942), y como el británico Frank Yates (1902-1994), quien escribía un poco más tarde que: "el énfasis que se ha dado las PSE ha producido una concentración del esfuerzo de los estadísticos en métodos que entrañan muy poca o ninguna importancia práctica" (Yates, 1951). Tal cuestionamiento tenía una connotación especial, toda vez que Yates había trabajado como estadístico auxiliar en la *Estación Experimental de Rothamsted* bajo la égida de Fisher, con quien publicara 13 años antes las famosas *Statistical Tables for Biological, Agricultural and Medical Research* (Fisher y Yates, 1938).

La naturaleza problemática del método motivó que destacados investigadores de la época consideraran que el empleo de los *valores p* era "innecesario su empleo por ser ellos, simplemente, irrelevantes" (Lipset, Trow y Coleman, 1956).

Otro crítico especialmente prominente fue Leonard Savage, uno de los estadísticos más brillantes del siglo pasado; por más señas, el premio Nobel de Economía, Milton Friedman (1912-2006) le catalogó en sus memorias (Friedman, 1998) como "una de las pocas personas que conocía a quien calificaría sin dudarlo como un genio". Savage apuntó a lo que probablemente sea la objeción menos cuestionada que merecen las PSE; textualmente, señala: "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" (Savage, 1957).

Quien fuera el muestrista de referencia en aquellos años, el profesor Leslie Kish de la *Universidad de Michigan* recomendaba "desentenderse de la noción de "prueba de significación" y concentrarse en la significación sustantiva que puedan tener los resultados" (Kish, 1959), afirmación refrendada por otros destacados investigadores de diversos campos (Rozeboom , 1960).

En esa línea, para muchos colegas de entonces resultó especialmente persuasiva la siguiente reflexión del psicólogo David Bakan profusamente citada:

> "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 Mississipi? ¿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?" (Bakan, 1966)

Algunas importantes figuras de la época fueron bastante más estridentes en sus embates contra la ritualización en el empleo de los métodos que nos ocupan y en la denuncia de su empleo de una manera casi religiosa (Skipper, Guenther y Nass, 1967).

2.4 Un primer giro trascendente a cargo de las autoridades metodológicas en la década de los 80

En un artículo premonitorio, (Guttman, 1977) señalaba que "la experiencia muestra que la intolerancia suele venir de los firmes creyentes en prácticas sin fundamento".

En consonancia con ello, algunas voces muy respetadas advertían que, si bien las PSE tienen muchas ventajas, entre ellas la de ejercer una peligrosa seducción consistente en que nos relevan del esfuerzo de tener que pensar, toca ahora abandonarlas (Barnard, 1982) o que los valores p y la evidencia guardaban una relación irreconciliable (Berger y Sellke). Más subidas de tono eran las descalificaciones que merecían aquellos cursos y textos de estadística, cuyo "grotesco énfasis en las PSE dificultan la comprensión de que ellas son peores que irrelevantes" (Nelder, 1985).

En 1986 ya las cosas habían madurado como para desembocar en una decantación realmente significativa: un artículo donde se exhortaba a los autores a prescindir de las PSE y operar con intervalos de confianza (IC) (Gardner y Altman, 1986). La singularidad dimanó de que Martin Gardner y Douglas Altman eran los"gurúes" de la muy influyente *British Medical Journal*.

Como consecuencia, diversas revistas punteras de la producción científica internacional tienden crecientemente a no admitir trabajos en los cuales aparezcan pruebas de este tipo como único recurso inferencial. Por ejemplo, *la British Heart Journal* anunció en un editorial de 1988 que se unía a la política sugeridas por aquellos autores. También se sumaron otras revistas encumbradas como *American Journal of Public Health*, *The Lancet* y *Annals of Internal Medicine*. Y en ese propio año 1988, la recomendación fue adoptada por el llamado "Grupo de Vancouver" (*International Committee of Medical Journal Editors*, ICMJE), creado en 1979 para velar por la calidad y uniformidad de los artículos científicos. En el apartado de requisitos técnicos que dicho comité dedica al empleo de la estadística, dentro de sus recomendaciones, se consignaba textualmente:

> "Siempre que sea posible, cuantifique los resultados y preséntelos con indicadores apropiados de error o la incertidumbre de la medición (por ej., intervalos de confianza). No dependa exclusivamente de las pruebas estadísticas de comprobación de hipótesis, tales como el uso de los valores p, que no transmiten información importante." (ICMJE, 1988).

Cabe señalar que esta declaración ha sido ratificada, años tras año, con apenas alguna variación marginal, hasta la actualidad, y también subrayar que tales recomendaciones no son declaraciones cosméticas sino sustantivas, así como que las enfáticas exhortaciones de destacados especialistas no eran ejercicios intelectuales a cargo de metodólogos ingeniosos e incómodos, sino muy serias advertencias epistemológicas.

A pesar de estar inscritos en la órbita de las mismas matemáticas frecuentistas que las pruebas de significación, los IC no desembocan en una interpretación automática, propia de los *valores p*. Constituyen un recurso indirecto para resumir tanto la magnitud de los efectos, como el grado en que la estimación de la verdadera diferencia es adecuada. El argumento central estriba en que los IC proveen más información que las PSE, a la vez que no obligan a dicotomizar las conclusiones.

Importantes asociaciones científicas, se adhirieron a esa política. Se destaca por ejemplo la poderosa *Asociación Americana de Psicología* que, tras algunos titubeos al respecto y luego de la creación de una comisión *ad hoc* (*Task Force on Statistical Inference*) para encarar el problema, recomendó unos años más tarde en sus "guidelines" el empleo de IC en lugar o como complemento de las PSE (Wilkinson, 1999).

Sin embargo, estos autorizados pronunciamientos no se tradujeron en una modificación decisiva de los patrones enraizados en la literatura (Silva, Suárez y Fernández, 2010). A ese proceso aún le esperaba un largo recorrido.

2.5 El declive indetenible de la defensa de los valores p en la última década del siglo XX

Hasta 1990, los numerosos disensos contra "la dictadura de los valores p" se ubicaban de manera esencialmente dispersa en recomendaciones editoriales o en artículos de revistas. Pero bajo el principio de que "forzar la elección entre significación y no significación oscurece la incertidumbre presente siempre que se quiera realizar inferencias a partir de una muestra", la rebelión llega a los libros de texto (Altman, 1991),

Entre tanto, las numerosas descalificaciones a cargo de renombrados especialistas se siguen sucediendo: apuntan tanto a sus endebleces lógicas, como a sus limitaciones epistémicas y prácticas. Algunos autores eran especialmente cáusticos al valorarlas. Tal es el caso de Marks Nester cuando reafirmaba 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" (Nester, 1996). Otros llegaron a conceptuar las PSE como una expresión de pseudociencia (Johnson, 1998). Y algunos consideraban que la mayoría de los estadísticos daría la bienvenida a un cambio ordenado que se oriente al abandono de las PSE (Nix y Barnette, 1998). Son ejemplos típicos de la realidad prevaleciente a finales de siglo.

Uno de los trabajos críticos más citados refleja ese clima cuando el autor expresaba: "Con cientos de artículos ya publicados que critican acerbamente a las PSE, tuve dudas acerca de la pertinencia de escribir uno más". (Johnson, 1999).

Algunos de tales artículos revelaban incluso un notable grado de exasperación, mientras que otros recurrían al sarcasmo. He aquí algunos ejemplos:

> "Es difícil imaginar una manera menos apropiada para traducir los datos en conclusiones" (Loftus, 1991)
>
> "Luego de haber coleccionado datos de cientos de sujetos, agotados por el esfuerzo, los investigadores realizan una PSE para evaluar si los sujetos eran una gran cantidad, aunque esto es algo que los investigadores ya sabían, ya que están agotados debido precisamente a la gran cantidad de datos recogidos." Thompson (1993)
>
> "Después de cuatro décadas de duras críticas, el ritual de probar la significación de hipótesis nulas – mecánicas decisiones dicotómicas alrededor del sacralizado criterio de 0,05- aún persiste. ... ¿Cuál es el problema con las pruebas de significación? Bueno, entre otras cosas, que no nos dicen lo que queremos saber; sin embargo, es tan intenso el deseo de saber lo que queremos saber que, por desesperación, ¡creemos que lo hace!" Cohen (1994)
>
> "Las PSE constituyen con toda seguridad el más idiota proceder jamás institucionalizado en el entrenamiento maquinal de los estudiantes de ciencia" (Rozeboom, 1997)
>
> "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" (Thompson, 1998)

En síntesis, con el fin del siglo, las PSE como estándar metodológico mostraban claros síntomas de agotamiento y estaban inmersas en una inocultable crisis. Continuamente, más estadísticos e investigadores captan la racionalidad de las críticas de que han sido objeto durante años y simpatizan con la idea de prescindir del método convencional (Nickerson, 2000). La polémica desarrollada en torno a las PSE como recurso inferencial a todo lo largo de su historia, que ya por entonces iba concluyendo como tal, labró el camino para llegar a un punto singular.

2.6 Siglo XXI: acercándonos al punto de inflexión

Ya en nuestro siglo, se produjo una verdadera catarata de objeciones, cada vez más abrasivas. Algunas eran muy punzantes tales como "la utilidad de los valores p es completamente limitada y nosotros nos mantenemos reclamando eutanasia para tales procedimientos" (Anderson y Burnham, 2002), como la que elige Jeff Gill para iniciar su trabajo sobre el tema: "Las PSE no deberían siquiera existir … ya que entrañan una bancarrota intelectual y son profundamente inconsistentes tanto desde una perspectiva lógica como práctica." (Gill, 2004), o como la afirmación de que las PSE son inútiles, incluso cuando se interpretan correctamente (Armstrong , 2007).

La circularidad de las PSE -consistente en usar muchísimos datos para luego corroborar que se usaron muchísimos datos- no sólo era bien conocida en el mundo académico especializado, sino que por entonces ocupaba un lugar en el marco mediático. Por ejemplo, un periodista de *Wall Street Journal* prevenía a sus

lectores: “Ud puede probar cualquier hipótesis, por estúpida que sea, llevando adelante una prueba estadística con toneladas de datos” (Albert, 2007).

Paralelamente, aunque sobrevivía como un legado la duradera reticencia de muchos investigadores a abandonar o reformar tales métodos, lo cierto es que las voces en defensa de las PSE desaparecían o eran crecientemente inaudibles. Así lo consignaba (Senn ,2001) : “…es muy difícil hallar a un estadístico que defienda a los valores p. Los bayesianos en particular las hallan ridículas, pero incluso los frecuentistas modernos no las tienen en muy alta estima”

2.7 El punto de inflexión.

En 2016, la *American Statistical Association* (ASA) se pronuncia oficialmente por primera vez sobre algún tema en su centenaria existencia. Desmarcándose de su tradicional cautela, hizo pública un “statment” donde se resume el hondo malestar prevaleciente con el modo en que se aplican cotidianamente los *valores p* (Wasserstein y Lazar, 2016). Sin embargo, la declaración fue algo ambigua y no disipó con nitidez las dudas sobre el asunto central: ¿se deben erradicar las PSE o sólo se debe enmendar el desastroso empleo que se ha hecho de ellas? Algunos lo interpretaron 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 Matloff de la Universidad de California, está en el segundo caso (Matloff, 2016), y en el debate producido en torno al tema 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.

En un detallado examen histórico sobre la significación estadística (Hurlbert y Lombardi, 2009) habían anticipado prácticamente todas las recomendaciones de ASA, aunque comunicaron, además, la convicción de que esta noción debía ser radicalmente extirpada y que, más temprano que tarde, se llegaría a ese punto.

No obstante, esta declaración, si bien no llega a declarar una sentencia de muerte para las PSE, las desvalorizó inequívocamente y, sobre todo, abrió definitivamente la caja de Pandora, dejando allí, como en el mito griego, solo el espíritu de la esperanza para sus defensores. En vista de esa situación, la propia ASA convoca tres años después a una centena de estadísticos del más alto nivel para que plasmen los puntos de vista promovidos por su histórico “statement”

Así fue cómo, en el primer trimestre de 2019, este proceso franqueó un punto de inflexión trascendental. Un número especial del órgano de la *American Statistical Association*, publicado en marzo de ese año, recoge la posición de los estadísticos convocados, plasmados a través de 43 artículos científicos que contienen diversas contribuciones sobre el tema.

Si bien esta producción ofrece un panorama donde coexisten elementos controversiales, unánimemente convocan a evitar los dos errores más extendidos. Por una parte, el de considerar que una etiqueta de “significación estadística” quiera decir o implique que un efecto sea altamente probable, plausible, real, verdadero, influyente o importante y, por otra, creer que cuando no se ha hallado dicha significación, es apropiado afirmar que el efecto es improbable, intrascendente o inexistente. Por otra parte, y este es el dato más relevante: todos los participantes en este esfuerzo colectivo coinciden en que convertir el *valor p* a una escala binaria para calificar los resultados como "significativos" o "no significativos" resulta ilógico, profundamente engañoso e inapropiado. Sugieren que es hora de abandonar enteramente el uso del término “estadísticamente significativo” y que han de erradicarse variantes tales como “significativamente diferente”, “$p<0,05$” y “no significativo”, se expresen en palabras, mediante asteriscos en las tablas o de cualquier otra forma. En síntesis, en ese momento las PSE recibieron lo que se ha calificado como “un golpe de gracia” (Hurlbert, Richard, Levine y Utts, 2019).

Si se quisiera resumir el resultado de todo este esfuerzo en una sola oración, esta sería la que se eligió para titular el editorial de ese singular número de *The American Statistician* que cobijó esas decenas de artículos: “*Moviéndonos hacia un mundo que esté más allá que el $p<0,05$*”.

Un trabajo publicado ese mismo mes de marzo en la revista Nature, debido a tres sobresalientes personalidades de la estadística (Amrhein, Greenland y McShane, 2019), refrendado a su vez por un editorial de la propia revista (Nature, 2019) y avalado por 800 destacados metodólogos e investigadores de todo el mundo, fue una lanza más clavada en el cuerpo de las PSE.

2.8 El coeficiente de rozamiento inherente al cambio de un paradigma

Indiscutiblemente, se ha asistido a un punto de no retorno, pero no a un funeral. La muerte, por definición, es un fenómeno binario. Y es un hecho que, 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 consiguieron inundar la investigación biomédica contemporánea hasta el punto de conseguir por esa vía un seguro de vida.

Los editores de muchas revistas médicas y otras prominentes figuras académicas hacen una contribución especialmente perniciosa para que las PSE se mantengan para muchos como el recurso por antonomasia sobre el que reposaría la buena ciencia. 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” (Lambdin, 2012).

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 (Nelder, 1985; Ioannidis, 2005; Kelly, 2009), el rico acervo crítico con que contamos desde hace tantos años, ha sido sistemáticamente omitido en los libros de texto y cursos introductorios de estadística. Es decir, 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, especialmente en el siglo actual, no fue suficiente para superar la fuerte sedimentación del método (Stang, Poole y Kuss, 2010). Y podría decirse que la guerra no ha terminado del todo (Dirnagl, 2019).

En su *Estructura de las revoluciones científicas*, el gran filósofo vienes Thomas Kuhn dibujó con claridad una realidad a la que se ajusta el momento que actualmente viven las PSE, abocadas a un cambio de paradigma de profundas implicaciones. Kuhn fundamentó que los paradigmas científicos pueden ir agonizando, pero sobreviven mientras sus reveses no sean suficientemente reiterados e incontrovertibles. (Khun, 1962). La resistencia que se opone a los cambios arraigados durante decenios suele ser enorme.

No hay duda de que estamos en un punto de no retorno, pero la transición suele ser conflictiva, porque ha de superar extraordinarias fuerzas inerciales y necesita, además, de enfoques alternativos. Estos suelen forjarse a partir de la irreverencia ante la obra, por lo general meritoria, de los antepasados, pero exigen la aparición de aportes que consigan un apoyo relativamente generalizado.

Alvan Feinstein lo vaticinaba desde 1985:

> “Puesto que la historia de la investigación médica también muestra una larga tradición de mantenerse por mucho tiempo fiel a doctrinas establecidas después de que tales doctrinas hubieran sido desacreditadas, o de haberse demostrado su escaso valor, no podemos esperar un súbito cambio en esta materia por el mero hecho de que haya sido denunciada por connotados conocedores de la estadística.” (Feinstein, 1985)

A la conducta acomodaticia de profesores, revisores, tribunales y editores de la vieja escuela, enquistados en una pereza que acaso esté justificada por la edad y por muchos años de apacible convivencia con métodos indiscutidos, se adicionan elementos objetivos que obstaculizan la transición. Cabe consignar al menos dos.

En primer lugar, está el arraigo de decenas de programas informáticos vertebrados en torno a dichos métodos; si estos se descalifican, se derrumbarían al menos parcialmente empresas dedicadas a comercializarlos a precios muy considerables.

En segundo lugar, tal descalificación tendría un efecto dominó sobre protocolos de actuación, guías de práctica editorial y normas docentes. Un efecto, incluso, que afectaría a la propia disciplina estadística, ya que diversos procedimientos (e.g. la determinación de tamaños de muestra en ensayos clínicos, el uso del ANOVA, y los métodos de selección de variables en el marco de la regresión múltiple) se basan en los valores p.

3. Conclusiones

Resulta obvio que no será posible completar un giro de 180 grados a corto plazo. Las pruebas de significación y los *valores p* seguirán siendo considerados por muchos como un único y coherente par de conceptos necesarios para la inferencia. Hasta hace muy poco, numerosos investigadores ignoraban (o conocían de manera vaga) que las pruebas de significación estadística han sido objeto de crecientes y fundamentados cuestionamientos desde su creación. Y tampoco alcanzan a captar la magnitud del cisma que se ha producido.

Pero la atmósfera actual es novedosa y difícil de desconocer. Una convincente y abarcadora revisión a cargo de seis afamados estadísticos (Greenland et. al., 2016) ha colocado recientemente al alcance de cualquier lector los argumentos más relevantes que la caracterizan. Dos artículos del investigador greco norteamericano de la Universidad de Stanford, el más citado dentro de la literatura médica del mundo, John Ioannidis, han conmovido al mundo académico y marcan la inevitabilidad de un cambio de paradigma de gran alcance. El primero demuestra que la mayoría de los hallazgos científicos que se publican son falsos (Ioannidis, 2005) mientras que el segundo fundamenta que "en general, no sólo la mayoría de los resultados de investigaciones son falsos sino que es peor: la mayor parte de los resultados verdaderos no son útiles" (Ioannidis, 2016).

Por iniciativa del propio Ioannidis, en colaboración con el estadístico Steven Goodman, un viejo crítico de las PSE, de la *Universidad John Hopkins*, se creó el *Meta-Research Innovation Center at Stanford* (METRICS), un órgano dedicado a la investigación cuyo objeto de estudio es la propia investigación, y cuyo cometido es la evaluación y el mejoramiento de la reproducibilidad y la transparencia de la producción científica (Silva, 2015). Acaso METRICS sea la iniciativa que mejor encarna tanto los desafíos que hoy enfrentan los estadísticos como las avenidas para superar la crisis que se vive.

Referencias

ALBERT, J. (2007). "The numbers guy". *Periódico Wall Street Journal*. 7 de diciembre, Nueva York.

ALTMAN, D. G. (1991). *Practical Statistics for Medical Research*, London: Chapman and Hall.

AMRHEIN V, GREENLAND S, MCSHANE B. (2019) "Scientists rise up against statistical significance" en *Nature*, 567, 305–307.

ANDERSON DR, BURNHAM KP (2002) "Avoiding pitfalls when using information–theoretic methods" en . *Journal of Wildlife Management*, 66, 912–918.

ARBUTHNOTT, J. (1710). "An argument for divine providence taken from the constant regularity in the births of both sexes" en *Philosophical Transactions of the Royal Society*, 27, 186–190.

ARMSTRONG, J.S. (2007).“Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries” en *International Journal of Forecasting*, 23, 335–336.

BAKAN D. (1966). “The test of significance in psychological research” en *Psychological Bulletin*, 66, 423-437.
BARNARD, G. A. (1982), “Conditionality Versus Similarity in the Analysis of 2 × 2 Tables,” en *Statistics and Probability: Essays in Honor of C.R. Rao, eds. G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh*, New York: North Holland Publishing.

BERGER, J. y SELLKE, T. (1987). “Testing a point null hypothesis: the irreconcilability of P-values and evidence” en *Journal of the American Statistical Association*, 82, 112.

BERKSON J. (1942). “Test of significance considered as evidence” en *Journal of the American Statistical Association*, 37, 325-335.

CHRISTENSEN, R. (2005). “Testing Fisher, Neyman, Pearson, and Bayes” en *The American Statistician*, 59, 121-126.

COHEN, J. (1994). “The earth is round (p < .05)” en *American Psychologist*, 49, 997-1003.

DIRNAGL, U. (2019). “The p value wars (again)” en *European Journal of Nuclear Medicine and Molecular Imaging*, 46, 2421–2423.

FEINSTEIN, A.R. (1985).“Clinical epidemiology: The architecture of clinical research”. Philadelphia: W.B. Saunders Company.

FISHER, R. y YATES, F. (1938). “Statistical tables for biological, agricultural and medical research”. London: Oliver & Boyd.

FISHER, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

FRIEDMAN M. (1998). *Two lucky people: Memoirs*. Chicago: University of Chicago Press.

GARCÍA, F.M. (2003). “Popper, el contraste de hipótesis y el método crítico” en *Revista Cubana de Salud Pública*, 29, 52-60.

GARDNER M.J. y ALTMAN D.G. (1986). “Confidence intervals rather than P values: estimation rather than hypothesis testing” en *British Medical Journal*, 292, 746-750.

GAVARRET, J. (1840). *Principes generaux de statistique medicale*. Paris.

GERRODETTE, T. (2011). “Inference without significance: Measuring support for hypotheses rather than rejecting them” en *Marine Ecology*, 32, 404-418.

GILL, J. (2004). “Grappling with Fisher’s legacy in social science hypothesis testing” en *Journal de la Société Française de Statistique*. 145 , 4, 39-46.

GOODMAN, S.N. (1999). “Toward evidence-based medical statistics (1): The p value fallacy” en *Annals of Internal Medicine*, 130, 995-1004.

GREENLAND, S., SENN, S.J., CARLIN, J.B., POOLE, C, GOODMAN, SN, ALTMAN, D.G. (2016). “Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations” en *European Journal of Epidemiology*, 31, 337–350.
GUTTMAN, L. (1977). “What is not what in statistics?” en *Statistician*, 26, 81-107.

HOGBEN, L. (1957). *Statistical theory*. London: Allen and Unwin.

HURLBERT, S. H. y LOMBARDI, C.M. (2009). “Final collapse of the neyman-pearson decision-theoretic framework and rise of the neoFisherian” en *Annales Zoologici Fennici*, 46, 311–349.

HURLBERT, S., RICHARD, A. LEVINE, R.A. y UTTS, J. (2019). “Coup de grâce for a tough old bull: “statistically significant” expires” en *The American Statistician* 73, sup1, 352-357.

ICMJE (1988). Uniform requirements for manuscripts submitted to biomedical journals” en *Annals of Internal Medicine*, 108: 25865.

IOANNIDIS, J.P.A. (2005). "Why most published research findings are false" en PLoS Medicine, 2, 8, e124.

IOANNIDIS, J.P.A. (2016). "Why most clinical research is not useful" en *Plos Medicine*,13, 6, e1002049.

JOHNSON, D.H. (1998). "Hypothesis testing: statistics as pseudoscience**"**. *Fifth Annual Conference of the Wildlife Society*, Buffalo, New York, September 22–26.

JOHNSON, D.H. (1999). "The insignificance of statistical significance testing" en *Journal of Wildlife Management* , 63, 3, 763-772.

KELLEY, J. (2009). "The perils of p-values: Why tests of statistical significance impede the progress of research" en *Handbook of Evidence-Based Psychodynamic Psychotherapy,* 367-377.

KISH, L. (1959). "Some statistical problems in research design" en *American. Sociology Review* 24, 328-338.

KUHN, T.S. (1962). "The structure of scientific revolutions". Chicago: University of Chicago Press.

LAMBDIN, C. (2012). "Significance tests as sorcery: significance tests are not" en Theoretical Psychology, 22, 1, 67 –90.

LIPSET, S.M., TROW, M.A. y Coleman, J.S. (1956). "Statistical problems" Apéndice 1-B en *Union Democracy, Glencoe*, Illinois: Free Press.

LOFTUS, G. R. (1991). "On the tyranny of hypothesis testing in the social sciences" en *Contemporary Psychology, 36,* 102-105.

MATLOFF, N. (2016). "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/**.

NATURE. (2019). "It's time to talk about ditching statistical significance". Editorial. en *Nature*, 567 (7748), 283.

NELDER, J.A. (1985). "Comment" en *Journal of the Royal Statistical Society*, 148, 3, 238.

NELDER, J. A. (1971). "Discussion on the papers by Wynn and Bloomfield, and O'Neill and Wetherill" en *Journal of the Royal Statistical Society B*, 33, 244-246.

NESTER, M.R. (1996). "An applied statistician's creed" en *Applied Statistics*, 45, 401-410.

NEYMAN, J. y PEARSON, E. (1928) "On the use and interpretation of certain test criteria for purposes of statistical inference" en *Biometrika*, 20, 175-240.

NEYMAN, J. y PEARSON, E. (1933) "On the problem of the most efficient tests of statistical hypotheses" en *Philosophical Transactions of the Royal Society.* A, 231, 289-337.

NICKERSON, R.S. (2000). "Null hypothesis significance testing: A review of an old and continuing controversy" en *Psychological Methods*, 5, 241-301.

NIX, T.W., BARNETTE, J.J. (1998). "The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing" en *Research in the Schools* 5: 3-14.

RONALD, L. WASSERSTEIN, R.L., SCHIRM, A.L. y LAZAR, N.A. (2019). "Moving to a world beyond "$p < 0.05$". *The American Statistician* 73(sup1), 1-19.

ROTHMAN, K. (1998). "Writing for Epidemiology" en *Epidemiology*, 9, 98-104.

ROZEBOOM, W.W. (1960). "The fallacy of the null hypothesis significance test" en *Psychological Bulletin,* 57, 416-428.

ROZEBOOM, W.W. (1997). "Good science is abductive, not hypothetico-deductive" en *LL Harlow, SA Mulaik, & JH Steiger (Eds.), What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

SAVAGE, I.R. (1957). "Nonparametric statistics" en *Journal of the American Statistical Association*, 52: 332-333.

SENN, S. (2001) . "Two cheers for p-values and two cheers for Bayes" en *Journal of Epidemiology and Biostatistics,* 6,193-210.

SILVA, L.C. (2009). *Los laberintos de la investigación biomédica. En defensa de la racionalidad para la ciencia del siglo XXI*. Madrid: Díaz de Santos.
SILVA, L.C., SUÁREZ, P. y FERNÁNDEZ A. (2010) "The null hypothesis significance test in health sciences research (1995-2006): Statistical analysis and interpretation" en *BMC Medical Research Methodology*, 10, 44-53.

SILVA, L.C. (2015) "La meta-investigación: en defensa del rigor y la transparencia informativa" en *Revista Cubana de Información en Ciencias de la Salud*,26, 2.

SKIPPER JK, GUENTHER AL, NASS G. (1967). "The sacredness of 0.05: a note concerning the uses of statistical level of significance in social science" en *American Sociology*, 2, 16–18.

SMITH, L.D., BEST, L.A., CYLKE V.A. y STUBBS, L. (2000). "Psychology without p values. Data analysis at the turn of the 19th century" en *American Psychologist,* 55, 260-263.

STANG, A., POOLE, C. y KUSS, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, 25, 225–230.

THOMPSON, B. (1993). "The use of statistical significance tests in research: Bootstrap and other alternatives" en *Journal of Experimental Education*, 6, 361-377

THOMPSON, B. (1998). "In praise of brilliance: where that praise really belongs" en *American Psychologist*, 53, 799–800.
[1] VENN, J. (1888). "Cambridge anthropometry" en *The Journal of the Anthropological Institute* 18, 140-154.

WASSERSTEIN, R.L. y LAZAR, N.A. (2016). "The ASA's statement on p-values: context, process, and purpose" en *The American Statistician* 70, 129–133.

WILKINSON, L. (1999). "Task Force on Statistical Inference, APA Board of Scientific Affairs Statistical Methods in Psychology Journals: Guidelines and Explanations" en *American Psychologist*, 54, 594–604.

WILLIAMS, R.H. et al. (2003) "On the intellectual versatility of Karl Pearson" en *Human Nature Review* 3: 296-301.

YATES F. (1951). "The influence of statistical methods for research workers on the development of the science of statistics" en Journal of the American Statistical Association, 46, 19-34.