¬ág. 173- 8/1/01 12:05 Página 173

Psicothema
2001. Vol. 13, nº 1, pp. 173-178

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2001 Psicothema

El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa

Africa Borges del Rosal, Concepción San Luis Costas, J. A. Sánchez Bruno e Isabel Cañadas Osinski
Universidad de La Laguna

El contraste de significación de la hipótesis nula lleva años levantando polémica en el seno de la comunidad científica de los investigadores de la conducta, ya que interpretaciones inadecuadas del procedimiento han producido un uso incorrecto del mismo. En el presente trabajo se presenta una revisión de las últimas aportaciones de los metodólogos, con las diversas posturas enfrentadas a favor y en contra, a la vez que se exponen las directrices que el comité técnico de investigación de la APA (American Psychological Association), la Task Force on Statistical Inference, acaba de emitir (Wilkinson, 1999) referente a las líneas a seguir en la investigación en la Ciencia de la conducta.

The judgment against null hypothesis. Many witnesses and a virtuous sentence. Null hypothesis significance testing has been a source of debate within the scientific community of behavioral researchers for years, since inadequate interpretations have resulted in incorrect use of this procedure. In this paper, we present a revision of the latest contributions of methodologists of different opinions, for and against, and we also set out the guidelines to research within behavioral science recently issued by the A.P.A. (American Psychological Association) Task Force in Statistical Inference (Wilkinson, 1999).

El contraste de significación de la hipótesis nula es un tema que ha hecho correr ríos de tinta. La controversia, como se verá, abarca una amplia gama de frentes de batalla y se alimenta de réplicas y contra réplicas. Es más, con la mayor utilización de la red, la polémica se ha trasladado allí, concretamente a la revista de difusión por Internet *Psycoloquy*. Las interpretaciones inadecuadas y, por tanto, las conclusiones improcedentes, han proliferado. Si se profundiza en su génesis resulta lógico que haya sido así, pues el contraste de hipótesis que se ha consolidado y que se maneja en la actualidad supone un maridaje extraño entre dos posturas que nunca pretendieron aunarse: la de Fisher, de una parte, y la de Neyman y Pearson de otra. De ambas surgió lo que Gigerenzer (1993) ha denominado, no sin razón, la «lógica híbrida». No vamos a detenernos en ello, ni tampoco a profundizar en los problemas que tal nacimiento ha producido en la metodología de las ciencias del comportamiento, puesto que estas reflexiones ya se expusieron en un trabajo anterior (Borges, 1997). Sin embargo, sí cabe mencionar el estatus que adquiere este procedimiento en la investigación en Psicología. Gigerenzer y Murray (1987) hablan de la «revolución inferencial»: desde la década de los 40 hay un incremento considerable en la utilización del contraste de hipótesis. Hubbard, Parsa y Luthy (1997), revisando las publicaciones del *Journal of Applied Psychology*, confirman la instauración del procedimiento, de tal manera que llega a ser sinónimo de análisis empírico. Los autores señalan que el contraste de significación recoge los cinco factores que hacen que se expanda una idea en un sistema social o académico (Rogers y Shoemaker, 1971):

a) Ventaja relativa con respecto a otros procedimientos: las decisiones basadas en los contrastes de significación hacen la vida más fácil al investigador.

b) Compatibilidad: resultan congruentes con los valores establecidos y las necesidades de los investigadores que los adoptan.

c) Baja complejidad: Antes de los años 60 y el advenimiento de la informática, el cálculo de los contrastes de significación era una tarea ardua. La simplificación de la misma produce un incremento en su uso (desde un 80% al comienzo de los 60 hasta el 94% de la década de los 90)

d) Posibilidad de ensayos (grado en que una metodología puede ser muestreada antes de adoptarla o rechazarla): El acceso vía informática lo hace accesible, tanto en los campus como en ordenadores personales.

e) Posibilidad de observación (grado en que los beneficios de una idea son visibles a otros): La adopción del contraste de hipótesis hace más fácil que se publique un trabajo.

Y la tendencia continúa. Vacha-Haase y Ness (1999), en una revisión de la revista *Professional Psychology: Research and Practice* entre los años 90-97, encontraron que el 77 % de los 265 trabajos de investigación cuantitativos realiza contrastes de significación. Además, aunque los autores parecen irse adecuando al estilo que la APA preconiza en su manual (APA, 1994), pues el 81,9 % de ellos utiliza en sus informes grados de libertad, niveles de alfa y los valores de sus tests estadísticos, no parece ocurrir lo mismo con la información referida a los tamaños del efecto, que sólo aparece reflejada en un 14,8 % de los trabajos revisados. En el trabajo de Clark-Carter (1997), únicamente un 18,52% informa del tamaño del efecto, en el de Keselman y cols. (1998), sólo en un 16,1% aparecen datos relativos a tamaño del efecto y/o potencia

Correspondencia: Africa Borges del Rosal
Facultad de Psicología
Universidad de La Laguna
38205 La Laguna (Spain)
E-mail: ABorges@ull.es

¬ág. 173- 8/1/01 12:05 Página 174

del contraste, mientras que Thompson y Snyder (1998), revisando la publicación de *Journal of Counseling and Development* del año 1996, encontraron que sólo uno de los 25 artículos con estudio experimental presentaba tamaños del efecto. Además, es posible que los investigadores se hayan hecho dependientes de los paquetes estadísticos informatizados, pues Kirk (1996) en su revisión de 4 revistas pertenecientes a la APA (*Journal of Applied Psychology, Journal of Educational Psychology, Journal of Experimental Psychology, Learning and Memory* y *Journal of Personality and Social Psychology*), encontró que se presentaban los estadísticos que miden el tamaño del efecto que los paquetes estadísticos analizan casi por defecto, como es el caso de R^2, mientras que resulta muy infrecuente el cálculo de otras medidas del tamaño del efecto que no se encuentran habitualmente en tales paquetes. Esto se repite en la literatura en castellano; así, revisando los dos primeros números de la revista *Psicothema* de 2000, todos los estudios que han realizado análisis de regresión múltiple informan de los valores de R^2 (Aznar, Amador, Freixa y Turbany, 2000; Castro y Sánchez, 2000; Comeche, Vallejo y Díaz, 2000; Martínez-Iñigo, 2000; Rial, Valera, Braña y Lévy, 2000), mientras que no se presentan datos de ningún otro índice de tamaño del efecto en otros trabajos experimentales que utilizan otros contrastes, cuestión que ya se ha puesto de manifiesto en algún estudio de meta-análisis (Sánchez Meca, Rosa y Olivares, 1999).

Sin embargo, las razones de la amplia difusión de los contrastes de significación estadística no son todas tan *razonables*. Influyen también las interpretaciones erróneas de lo que un contraste de hipótesis aporta. Fundamentalmente, la simplicidad en la toma de decisión que supone el rechazo de la hipótesis nula ofrece una garantía de objetividad, algo que los investigadores de las Ciencias Sociales están deseosos de obtener (Schmidt, 1997; Schmidt y Hunter, 1998)

Vacha-Haase y Ness (1999) ofrecen un irónico ejemplo con el fin de poner en guardia contra la tan querida pequeña probabilidad de aparición de un fenómeno:

> «… los hechos improbables no son siempre intrínsecamente interesantes. Por ejemplo, si uno lanza un dólar de plata, es improbable que la moneda caiga de canto. Y si este hecho extremadamente improbable ocurre, no obstante, pocas vidas pueden verse afectadas por ello de forma notable» (pág. 104)

Las críticas, que se inician temprano, en la década de los años 60 (Rozeboom, 1960; Cohen, 1962; Binder, 1963; Bakan, 1966), dan lugar a diversos textos recogiendo las controversias suscitadas (Morrison y Henkel, 1970; Lieberman, 1971; Kirk, 1972) o a amplios espacios en las revistas científicas (el monográfico editado por Thompson del *Journal of Experimental Education* del volumen de verano de 1993; el número 2 del volumen 21 de la revista *Behavioral and Brain Science*, 1998), y culminan en la década de los 90, con una virulencia tal que The American Psychological Association's Board of Scientific Affairs crea en 1996 la denominada «Task Force in Statistical Inference», cuya misión fundamental es analizar la problemática del contraste de significación de la hipótesis nula y del uso de *p*, probabilidad asociada. Por ello, no es de extrañar que la mayor virulencia de las críticas aparezca a partir de 1995, siendo frecuente hablar del fallecimiento del procedimiento (Falk y Greenbaum, 1995; Kirk, 1996; Abelson, 1997; Harlow, Mulaik, y Steiger, 1997). Pero no sólo hay comentarios referentes a una utilización inadecuada de los análisis estadísticos y de las conclusiones que estos derivan, sino que también hay comprobación empírica de ello. Así, Dar, Serlin y Omer (1994) revisan tres décadas de publicaciones, poniendo de manifiesto los errores cometidos por los investigadores:

- Confirmación de la H_0: por ejemplo, concluyendo equivalencia de los grupos a línea base, siempre que no hay diferencias significativas entre los grupos.
- Utilización de diversos niveles de significación a lo largo de un estudio.
- Pocos artículos informan de tamaño de efecto, sobre todo los más antiguos y los que usan Anova. No se informa de intervalos de confianza.
- El error tipo I no se controla, ya que se usa el mismo para familias de contrastes.
- Utilización de estadísticos univariados después de Manova[1]

Pero, como en todo juicio, no sólo hay fiscales. Algunas voces se levantan a favor del contraste de hipótesis. En ocasiones, como es el caso de Macdonald (1990), su defensa es parcial: el contraste estadístico es una forma de lenguaje, que permite a los investigadores formalizar sus hallazgos y posibilitar la replicación de los efectos encontrados, pues la significación de los mismos expresa que no se deben al azar. Desde su perspectiva, el contraste de hipótesis es sólo un instrumento que tiene su papel en colaboración con la teoría (Macdonald, 1997). Biskin (1998) defiende la utilización de los contrastes de significación bajo determinadas circunstancias, como en los primeros estadios de investigación de un campo de conocimiento; no obstante, cuando el área de conocimiento ha madurado suficientemente, de tal manera que hay consenso en tomo a que la hipótesis nula es falsa, los contrastes de significación se tornan innecesarios. Frick (1996) considera que el contraste de hipótesis es válido en lo que él denomina corroboración de afirmaciones ordinales, esto es, una afirmación que no tiene en cuenta el tamaño del efecto, sino sólo especifica el orden de las afirmaciones. Por ejemplo, sería especialmente adecuado en el contraste de teorías, que se realiza de forma ordinal. No obstante, en aplicaciones prácticas conviene establecer también el tamaño del efecto.

Hagen (1997) presenta una defensa más profunda. Afirma que no sólo es el único procedimiento disponible, sino que es el mejor. A este artículo se da respuesta en el volumen de julio de 1998 de la revista *American Psychologist* (Tryon; McGrath; Malgady; Falk; Thompson; Graanas). En su contra-réplica (Hagen, 1998) afirma que la interpretación inadecuada del procedimiento inferencial se debe a los investigadores, y no al procedimiento en sí. Aún aceptando que tenga una parte de razón, hay que reconocer que la «lógica híbrida» no es, desde luego, un modelo de claridad conceptual. Wainer (1999) afirma que en muchas ocasiones la ciencia ha avanzado mediante decisiones binarias como las que se toman en el contraste de significación. El autor matiza, no obstante, que la tarea del científico no debería detenerse en tal decisión, sino que este habría de constituir un inicio de la línea de investigación.

Hagen (1997), por otra parte, defiende que los intervalos de confianza aportan la misma información que el contraste de hipótesis. Dejando de lado que la información que producen ambos procedimientos puede ser la misma, rechazo o no de la hipótesis nula, tal afirmación es claramente equivocada, pues el intervalo de confianza permite una cuantificación de los valores entre los que se espera que esté el parámetro, lo cual supone superar la respuesta dicotómica que aporta el contraste de hipótesis (Kirk, 1996; Schmidt y Hunter, 1997; Thompson, 1998). McGrath (1998) afir-

¬ág. 173- 8/1/01 12:05 Página 175

ma que contraste de hipótesis e intervalos de confianza no son equivalentes, ya que, además de que el primero aporta menos información que los segundos, el tipo de interrogante al que responden es distinto. Así, el contraste de significación de la hipótesis nula da respuesta a la siguiente cuestión: «Basado en nuestra muestra, ¿cuál es nuestra mejor estimación acerca de si r es o no igual a 0?», mientras que lo que plantea el intervalo de confianza es: «Basado en nuestra muestra, ¿cuál es nuestra mejor estimación acerca del valor de r?»(pág. 797).[2]

Es más, incluso hay voces que admiten la posibilidad de que la hipótesis nula sea cierta. Así, Chow (1999) afirma que H_0 puede ser cierta bajo dos condiciones: a) la teoría que implica la hipótesis experimental es falsa; y b) los datos se han recogido adecuadamente de acuerdo con la regla inductiva que subyace al diseño experimental. En una defensa de la hipótesis nula *per se*, Greenwald (1993) considera que la hipótesis nula puede ser aceptada. Desde su punto de vista, la Ciencia avanza de forma más poderosa por la falsación. Así, si los datos de un estudio bien realizado niegan la existencia de una relación predicha, es decir, aceptan la hipótesis nula, el rechazo de la teoría subyacente supone una fuente no despreciable de información.

Hasta aquí se ha expuesto la visión de los abogados defensores que apoyan la hipótesis nula tal cual se ha venido utilizando. Alternativamente, se proponen diversas soluciones para superar el problema planteado, lo que se podría denominar defensores reformistas. Entre ellos también se puede establecer una doble categoría. En un primer bloque se podrían enmarcar las soluciones «conservadoras», en el sentido de que mantienen la utilización prioritaria del contraste de hipótesis, pero con mejoras o procedimientos alternativos. Dentro de este apartado, se pueden señalar cuatro. El ejemplo más elocuente de esta posición es la solución del «principio suficientemente bueno» de Serlin y Lapsley (1985, 1993): Una H_0 afirma que el valor de un determinado parámetro δ es δ_0. Ahora bien, aún en el caso de ser cierta H_0, dada la inexactitud de los procedimientos experimentales en general, se obtendrá un valor muestral δ^* que diferirá de δ_0. Dado un tamaño muestral lo suficientemente grande, dicha diferencia (espúrea) se hará significativa. Lo que este procedimiento sugiere es incluir en la hipótesis nula unos márgenes que determinen *a priori* la magnitud de error Δ aceptable; así, el intervalo $\delta_0 \pm \Delta$ sería el intervalo de valores «suficientemente buenos» y H_0 quedaría expresada como $|\delta - \delta_0| \leq \Delta$.

Además, dos autores aportan procedimientos de cálculo para asegurar que los resultados que se obtienen, además de significativos desde el punto de vista estadístico, sean relevantes. Así, Rosenthal y Rubin (1994) presentan el cálculo de un estadístico a tal fin, el «contra nulo»: El valor contra nulo de un tamaño de efecto obtenido es la magnitud no nula de tamaño de efecto que es sostenida por exactamente la misma cantidad de evidencia como lo es el valor nulo del tamaño del efecto. En otras palabras, si el valor contra nulo fuera tomado como la hipótesis nula, el valor de *p* resultante debería ser el mismo para la hipótesis nula real.

En una línea similar, Ares (1999), señala que el problema se agrava al usar muestras grandes, ya que cualquier resultado, por escaso que sea, resultará significativo. Propone el cálculo de un estadístico, n, que, en esencia, trata de ponderar cuánto de *significativo* es haber obtenido un resultado significativo. La ventaja, posible, que suponen los dos estadísticos comentados es que permiten dar un valor objetivo. Sin embargo, cabe preguntarse si la determinación sin más del tamaño del efecto, que se lleva recomendando como práctica habitual en los informes de investigación, no produce suficiente información, más intuitiva y, por ende, más próxima y adecuada para ser aportada por cualquier investigador, tenga o no grandes conocimientos de metodología.

En último extremo, las aportaciones de Rosenthal y Rubin (1994) y de Ares (1999) parecen ir en apoyo de una supuesta objetividad: además de tener un punto de corte si/no en el contraste de significación de la hipótesis nula, se poseerá un punto de corte si/no en la credibilidad de cuán significativo es el resultado hallado. Y, de esta forma, el investigador se limita a actuar de forma mecánica (asimilable a objetiva).

Por último, se puede enmarcar aquí a los que han definido el denominado error tipo III, que se refiere al signo del contraste, esto es, que las diferencias o relaciones encontradas vayan en la dirección contraria a la predicha. Diversos autores hacen hincapié de una u otra forma sobre el particular (Leventhal y Huynh, 1996; Harris, 1997a, 1997b).

El segundo bloque de los defensores reformistas presenta una visión más «aperturista», entendiendo por ello que no se limitan a dar soluciones puntuales, sino que abren más posibilidades. De una parte, estaría el planteamiento que sugiere Grayson y su equipo (Grayson, 1998; Grayson, Pattison y Robin, 1997), en el sentido que el contraste de hipótesis es sólo uno de los procedimientos inferenciales con los que se cuenta. Señalan que, más que abandonar esta forma frecuentista de enfrentar la inferencia, se debería optar por ampliar el marco de referencia, dando cabida, cuando la investigación lo haga idóneo, a la inferencia bayesiana.

Otros autores que abogan por la inferencia bayesiana son Krueger (1998b, 1998c, 1999) y Rindskopf (1997, 1998), si bien ambos consideran que esta forma de inferencia resulta superior al contraste de hipótesis. La inferencia bayesiana permite establecer cual de las dos hipótesis, nula o alternativa (consideradas exhaustivas y mutuamente excluyentes) resulta más acorde a los datos. La dificultad de la inferencia bayesiana estriba en el establecimiento de probabilidades a priori; no obstante, los estudios previos, incorporando tamaños del efecto, pueden solucionarlo. McCauley (1998), por su parte, considera que la investigación se ha centrado tanto en el contraste de significación de la hipótesis nula que se ha dado de lado a la inferencia bayesiana.

El otro tipo de soluciones se mantiene dentro de la inferencia clásica (Cohen, 1990, 1994; Kirk, 1996; Ciccehetti, 1997; Hubbard y Armstrong, 1997; Lunt y Livingstone, 1989; Snyder y Thompson 1998; Tryon, 1998), proponiendo alternativas al contraste de hipótesis: información referente a intervalos confidenciales, tamaños del efecto, potencia del contraste, uso de *alfa* marcado por los objetivos de la investigación, y replicación de resultados. La mayor relevancia que adquieren el tamaño del efecto explica que algunos investigadores hayan dedicado sus estudios al respecto en los últimos años (Fern y Monroe, 1996; Snyder y Lawson, 1993; Tatsouka, 1993).

Abundando en el interés de la replicación, resulta esclarecedora la afirmación de Krueger (1998a):

> «Un estudio individual no decide el problema, pero docenas de efectos en la misma dirección sí lo hace, incluso aunque no haya logrado resultados significativos»(pág. 4).

Entre los fiscales encarnizados, cabe citar a Schmidt (1996), quien se declara partidario de prescindir de la significación estadística, eliminando así tanto los problemas de robustez como de potencia de los contrastes, aportando dos alternativas: la estimación del tamaño del efecto mediante intervalos de confianza y el uso del meta-análisis. Al igual que Krueger, Schmidt considera

¬ág. 173- 8/1/01 12:05 Página 176

que en muy raras ocasiones un único estudio permite responder a preguntas científicas; cada estudio no puede responder por sí solo a un problema científico, sino que no es más que un punto para contribuir a un estudio de meta-análisis. En un estudio posterior (Schmidt y Hunter, 1997) se rebaten aquellas razones que parecen justificar la utilización de los contrastes de significación, cuestiones todas ellas referidas al engañoso estatus de objetividad y de única alternativa posible, que ha primado entre los investigadores.

Otras soluciones se señalan como alternativa al contraste de hipótesis. Graanas (1998) considera que el ajuste de modelo, dentro de la estimación de parámetros, permitirá tener siempre el mejor estimador del parámetro que sea posible en las condiciones presentes de la investigación. Esto supone comparar un modelo reducido o compacto (Modelo C) con un modelo completo o aumentado (Modelo A), para determinar cuál es el que mejor representa, o se ajusta, a los datos. En su formato más simple, el modelo reducido especifica un valor para el parámetro que está siendo estimado; el modelo aumentado mejora el compacto estimando el mismo parámetro de los datos. Se calcula la reducción proporcional del error del modelo A respecto al C para comprobar si el A es mejor que el C. En tal caso, A reemplaza a C y se convierte en el nuevo mejor estimador para el parámetro.

El problema, desde nuestro punto de vista, es más profundo que simplemente un cuestionamiento metodológico, y se enlaza con las formas de investigar. Ya se han comentado las dudas que plantea llegar a conclusiones sobre un trabajo individual (Schmidt, 1996; Krueger,1998a). Se aboga por una investigación más sustantiva y más acumulativa. Y esto no es nuevo (Rossi, 1990; Cohen, 1990; 1994; Macdonald, 1993; Gigerenzer, 1993; Greenwald, 1993; Harris, 1997; por citar sólo trabajos publicados en la década).

La Task Force on Statistical Inference (Wilkinson, 1999) ha completado ya su misión, y va en la línea que se acaba de comentar. Su veredicto ha sido virtuoso, salomónico y consecuente con la parsimonia que exige la Ciencia. No da en sus directrices nada distinto a lo que los profesionales de las Ciencias Sociales han aprendido en los manuales básicos de metodología, siguiendo punto por punto los apartados de un informe de investigación. No se decanta por santificar el contraste de significación de la hipótesis nula, ni en abogar por su desaparición. Sus recomendaciones no son novedosas, puesto que mucho de lo que se sugiere como deseable en el análisis de datos ya se ha puesto sobre el tapete (Cohen, 1990, 1994; Gigerenzer, 1993, 1998; Ciccehetti, 1997; Hayes, 1998; Rossi, 1998). No obstante, lo más destacable de las recomendaciones de la Task Force, desde nuestra óptica, es la de trabajar con el fin de apuntalar la investigación, sugiriendo una mayor coherencia en el análisis de los datos: análisis exploratorio de los mismos, añadir contenido teórico que ayude a interpretar los tamaños del efecto encontrados, la inclusión de intervalos de confianza para todos los tamaños de efecto de resultados centrales, así como comprobación de los supuestos. Ahondando en este último punto, hay información fehaciente de los resultados devastadores que para el contraste tiene el incumplimiento de sus supuestos, sobre todo si se viola más de uno (Borges, San Luis y Sánchez-Bruno, 1993; Borges, 1994; Borges, Sánchez-Bruno y Cañadas, 1996; Sánchez-Bruno y Borges, 1997; Sánchez-Bruno, Borges, San Luis, C. y Cañadas, 1999; Sánchez-Bruno, Borges y Cañadas, 1999; Borges, San Luis, Cañadas y Sánchez Bruno (en prensa); Cañadas, Borges, A y Sánchez-Bruno (en prensa)). Por ello, resulta especialmente estimulante encontrar trabajos que se preocupen por comprobar los supuestos previamente a realizar los contrastes de interés (por ejemplo, Del Barrio y Gutiérrez, 2000; Iglesias, De la Fuente y Martín, 2000).

Al fin y a la postre, su juicio va en el sentido del aforismo que se atribuye a Ramón y Cajal: «la investigación supone un 10 % de inspiración y un 90% de transpiración». Esto, en nuestro campo de conocimiento, se manifiesta en sentido contrario, primando una forma de hacer en investigación demasiado volcada hacia los resultados y a la rapidez. Tal vez ello pueda deberse a dos causas. De una parte, las facilidades que supuso el uso de la estadística inferencial (frente al más tedioso estudio del sujeto individual, como era la forma de proceder de autores tan fuera de duda en su cientificidad como Skinner o Pavlov) y, de otra, el «advenimiento» de la informática, con lo que ello supone de ejecución de programas de análisis estadístico sin tener que dedicarle un tiempo a su elaboración.

Queda, no obstante, un interrogante sobre el tapete, ¿hasta qué punto los investigadores obedecerán las recomendaciones hechas? Y la duda tiene total sentido. Hay claros ejemplos de hasta que punto la comunidad científica es reacia a incorporar nuevas formas de hacer. Como muestra, un botón: Cohen (1962) llama por primera vez la atención sobre la poca potencia estadística de los trabajos, pero las revisiones posteriores son completamente concordantes con esos primeros resultados (Sedlmeier y Gigerenzer, 1989; Rossi, 1990; Clark-Carter, 1997; Valera, Sánchez, Marín y Velandrino, 1998). La solución, tal vez, pueda venir desde las editoriales de las revistas científicas. Además de dar recomendaciones a los autores (Thompson, 1994), sería conveniente esperar que velasen para que la labor de la Task Force vea resultados. La comunidad científica al completo saldrá beneficiada con ello.

Por otra parte, queda aún un problema de fondo que trasciende a la investigación en sí misma: la construcción de teorías en psicología. La tradición de investigación en nuestro campo de conocimiento, deudora de una forma de hacer inductiva, no ha prestado toda la atención que merece la construcción sólida del conocimiento. No conviene olvidar que la evidencia empírica sólo contribuye a establecer teorías psicológicas, no siendo un fin en sí mismo (Macdonald, 1997). Y el contraste de significación no evalúa directamente hipótesis sustantivas, sino que ayuda a descartar hipótesis competidoras (Ruscio, 1999). Además, en ocasiones la mejora tecnológica no aporta únicamente ventajas. Loftus (1993), en su lúcido trabajo, compara el estado de desarrollo de las teorías psicológicas al que tenía la Astronomía en el siglo XVII. El autor alerta contra la utilización de la simulación como una forma de hacer viables teorías demasiado complejas.

Macdonald (1997) afirma: «Un segmento de investigación requiere inspiración en su concepción, meticulosidad en su ejecución y elocuencia en su exposición. El análisis de datos ideal requiere las tres cualidades simultáneamente» (pág. 345). Al fin y a la postre, la teoría es más que investigación, investigación supone asimismo más que análisis de datos y, en todo el proceso, sin descartar un procedimiento sistemático en el mismo, no se puede obviar la parte de arte que conlleva la ciencia. En palabras de Yela (1996) «El científico es, literalmente, un poeta humilde, un inventor que somete sus invenciones a comprobación rigurosa. Sin invención, sin poesía, no hay ciencia; sin comprobación, tampoco» (pág. 354).

¬ág. 173- 8/1/01 12:06 Página 177

Notas

[1] Keselman y cols. (1998) confirman este dato: En un 84% de los estudios con Manova revisados por ellos, *nunca* se interpretan los resultados del Manova para explicar los efectos de agrupamiento de la/s variable/s; se limitan a comentar los análisis univariados.

[2] Un revisor de este trabajo nos hizo notar que la finalidad del intervalo de confianza es la de buscar entre qué valores se encuentra el del parámetro (conocido el valor del estadístico). Aún cuando, efectivamente, tiene razón, creímos que debíamos mantener la cita tal como figura, por ser textual.

Referencias

Abelson, R.P. (1997) On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

American Psychological Association (1994) *Publication manual of the American Psychological Association.* (4ª ed.) Washington, DC: Author.

Ares, V.M. (1999). La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta. *Metodología de Encuestas, 1,* 47-68.

Aznar, J.A., Amador, J.A., Freixa, M. y Turbany (2000). Consumo atencional en la estimación de la profundidad retrovisual. *Psicothema, 12,* 71-78.

Bakan, D. (1966) The test of significance in psychology. *Psychological Bulletin, 66,* 423-437.

Barrio, JA y Gutiérrez, J.N. (2000). Diferencias en el estilo de aprendizaje. *Psicothema, 12,* 180-186.

Biskin, B.H. (1998) Comment on significance testing. *Measurement and Evaluation in Counseling and Development, 31,* 58-62

Binder, A. (1963) Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70,*

Borges, A. (1994) Un estudio mediante simulación del contraste de medias a través de técnicas de aleatorización. Comunicación presentada en la IV Conferencia española de Biometría. Sitges.

Borges, A. (1997) Algunos problemas frecuentes en la interpretación de los contrastes de hipótesis estadísticas en psicología. *Iberpsicología,* 2:3:7
http://fs-morente.filos.ucm.es/publicaciones/iberpsicologia/iberpsicologia.htm

Borges, A., San Luis, C., Cañadas, I. y Sánchez Bruno, J.A. (en prensa). El contraste de hipótesis en tres grupos: alternativas al Anova frente a la violación de sus supuestos. *Psicothema.*

Borges, A., San Luis, C. y Sánchez-Bruno, A. (1993) Contraste de la hipótesis nula para la diferencia de muestras: Alternativa frente al problema de Berhrens-Fisher. Poster presentado al III Simposio de Metodología de las Ciencias Sociales y del Comportamiento. Santiago de Compostela.

Borges, A., Sánchez-Bruno y Cañadas, I. (1996) El contraste de las diferencias de medias con grupos pequeños, con escalas ordinales y en ausencia de normalidad. *Psicológica, 17,* 455-466.

Cañadas, I., Borges, A. y Sánchez-Bruno, A. (en prensa) La t de Student y sus alternativas, ante la violación de los supuestos. *Psicothema.*

Castro, A. y Sánchez, M.P. (2000). Objetivos de vida y satisfacción autopercibida en estudiantes universitarios. *Psicothema, 12,* 87-92.

Charter, R.A. (1997). Effect of measurement error on tests of statistical significance. *Journal of Clinical and Experimental Neuropsychology, 19,* 458-462.

Chow, S.L. (1999). In defense of significance tests. Commentary on Krueger on social-bias. *Psycoloquy*: 10 (6)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?10.006
ftp://ftp.princeton.edu/pub/harnad/Psycholoquy/1999.volume.10/psyc.99.10.006.social-bias.15.chow

Ciccehetti, D.V. (1998). Role of the null hypothesis significance testing (NHST) in the design of neuropsychologic research. *Journal of Clinical and Experimental Neuropsychology, 20,* 293-295.

Clark-Carter, D. (1997) The account taken of statistical power in research published in the *British Journal of Psychology. British Journal of Psychology, 88,* 71-83.

Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1994). The Earth is round (p<. 05). *American Psychologist, 49,* 997-1003.

Comeche, M.I., Vallejo, M.A. y Díaz, M.I. (2000). Tratamiento psicológico de la cefalea. Predicción de la mejoría en un acercamiento activo-pasivo. *Psicothema, 12,* 55-63.

Dar, R., Serlin, R.C. y Omer, H. (1994) Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75-82.

Falk, R. y Greenbaum, C.W. (1995). Significance tests die hard. *Theory and Psychology, 5,* 75-98.

Falk, R. (1998) In criticism of the Null Hypothesis Statistical Test. *American Psychologist, 53,* 798-799.

Fern, E.F. y Monroe, KB. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23,* 89-105.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379-390.

Gigerenzer, G. (1993) The Superego, the Ego and the Id in statistical reasoning. En G. Keren y Ch. Lewis (Eds.) *A handbook for data analysis in the Behavioral Sciences. Methodological issues.* Hillsdale, NJ: LEA.

Gigerenzer, G. (1998) We need statistical thinking, no statistical rituals. *Behavioral and Brain Sciences, 21,* 199-200.

Gigerenzer, G. y Murray, D.J. (1987) *Cognition as intuitive statistics.* Hillsdale, NJ: LEA.

Graanas, M.M. (1998) Model fitting: A better approach. *American Psychologist, 53,* 800-801.

Grayson, D.A. (1998). The frequentist facade and the flight from evidential inference. *British Journal of Psychology, 89,* 325-345.

Grayson, D.A., Pattison, P. y Robins, G. (1997). Evidence, inference and the «rejection» of the significance test. Grayson, D.A. *Australian Journal of Psychology, 49,* 64-70.

Greenwald, A.G. (1993) Consequences of prejudice against the null hypothesis. En G. Keren y Ch. Lewis (Eds.) *A handbook for data analysis in the Behavioral Sciences. Methodological issues.* Hillsdale, NJ: LEA.

Hagen, R.L. (1997) In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15-24.

Hagen, R.L. (1998) A futher look at wrong reasons to abandon statistical testing. *American Psychologist, 53,* 801-803.

Harlow, L.L., Mulaik, S.A. y Steiger, J.H. (1997) *That if there were no significance tests?* Mahwah, NJ: LAU.

Harris, R.J. (1997a) Reforming significance testing via three- valued logic. En L. Harlow, S. Mulaik y J. Steiger (Eds.) *What if there were no significance tests?* Mahwah, N.J.: LEA.

Harris, R.J. (1997b) Significance tests have their place. *Psychological Science, 8,* 8-11.

Hayes, A.F. (1998) Reconnecting data analysis and research design: Who needs a confidence interval? *Behavioral and Brain Sciences, 21,* 203-204

Hubbard, R. y Armstrong, J.S. (1997) Publication bias against null results. *Psychological Reports, 80,* 337-338.

Hubbard, R. y Armstrong, J.S. (1997) Publication bias against null results. *Psychological Reports, 80,* 337-338.

Hubbard, R., Parsa, R.A. y Luthy, M.R. (1997) The spread of statistical significance testing in Psychology. *Theory and Psychology, 7,* 545-554.

Iglesias, S., De la Fuente, E.I. y Martín, I, (2000). Efecto de las estrategias de decisión sobre el esfuerzo cognitivo. *Psicothema, 12,* 267-272.

Keselman, H.J., Huberty, C.J., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C. y Levin, J.R. (1998) Statistical practices of educational research: An analy-

¬ág. 173- 8/1/01 12:06 Página 178

sis of their Anova, Manova, and Ancova analysis. *Review of Educatio - nal Research, 68*, 350-386

Kirk, R.E. (1972) *Statistical issues*. Monterey, CA.: Brooks/Cole.

Kirk, R.E. (1996) Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Krueger, J. (1998a). The bet on bias: A foregone conclusion? *Psycoloquy: 9* (46)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.46
ftp://ftp.princeton.edu/pb/harnad/Psycoloquy/1998.volume.9/psyc.98.9 .46.social-bias.1.krueger

Krueger, J. (1998b). Getting to the core of the data by testing against alternative hypotheses. *Psycoloquy*: 9 (70)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.70
ftp://ftp.princeton.edu/pub/harnad/Psycholoquy/1998.volume.9/psyc.9 8.9.70.social-bias.8.krueger

Krueger, J. (1998c). Theoretical progress requires refined methods and then some. *Psycoloquy*: 9 (73)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.73
ftp://ftp.princeton.edu/pub/harnad/Psycholoquy/1998.volume.9/psyc.9 8.9.73.social-bias.10.krueger

Leventhal, L. Y Huynh, C-L (1996). Directional decisions for two tailed tests: power, error rates and sample size. *Psychological Methods, 1*, 278-292.

Lieberman, B. (Ed) (1971) *Contemporary Problems in Statistics*. Nueva York: Oxford University Press.

Loftus, G.R. (1993) Computer simulation: Some remarks on theory in psychology. En K. Gideon y C. Lewis (Eds.) *A handbook for data analy - sis in the behavioral sciences. Statistical issues*. Hillsdale, NJ: LEA.

Lunt, P.K. y Livingstone, S.M. (1989) Psychology and statistics: Testing the opposite of the idea you first thought of. *The Psychologist, 12*, 528-531.

Macdonald, R.R. (1990) Language truth and Statistic. *The Psychologist, 3*, 125-126.

Macdonald, R.R. (1993) On statistical testing in Psychology. *British Jour - nal of Psychology, 88*, 337-347.

Macdonald, R.R. (1997). On statistical testing in Psychology. *British Jour - nal of Psychology, 88*, 337-347.

Malgady, R.G. (1998) In praise of value judgments in null hypothesis testing… and of «accepting» the null hypothesis. *American Psychologist, 53*, 797-798.

Martínez-Íñigo, D. (2000). Contrastación del modelo de Rusbult en una muestra de casados y divorciados. *Psicothema, 12*, 65-69.

McCauley, C. (1998) The bet on bias is cockeyed optimistic. *Psycoloquy*: 9 (71)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?9.71
ftp://ftp.princeton.edu/pub/harnad/Psycholoquy/1998.volume.9/psyc.9 8.9.71.social-bias.9.mccauley

McGrath, R.E. (1998) Significance testing: Is there something better? *American Psychologist, 53*, 796-797

Morrison, D.E. y Henkel, R.E. (Eds.) (1970) *The significance test contro - versy*. Chicago: Aldine.

Rial, A, Valera, J., Braña, T. y Lévy, J.P. (2000) El valor de la marca a partir de su relación con el consumidor. *Psicothema, 12*, 247-254.

Rindskopf, D. (1998) Null-hypothesis tests are not completely stupid, but Bayesian statistics are better. *Behavioral and Brain Sciences, 21*, 215-216.

Rindskopf, D. (1997) Testing «small», not null, hypothesis: Classical and Bayesian approaches. En L. Harlow, S. Mulaik y J. Steiger (Eds.) *What if there were no significance tests?* Mahwah, N.J.: LEA.

Rogers, E.M. y Shoemaker, F.F. (1971) *Communications of innovations: A crosscultural approach*. Nueva York: Free Press.

Rosenthal, R. y Rubin, D.B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science, 5*, 329-334.

Rossi, J.S. (1990) Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646-656.

Rossi, J.S. (1998) Meta-analysis, power analysis, and the null-hypothesis significance-test procedure. *Behavioral and Brain Sciences, 21*, 216-217.

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

Ruscio, J. (1999) Statistical models and strong inference in social judgment research. *Psycoloquy*: 10 (027)
http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?10.027
ftp://ftp.princeton.edu/pub/harnad/Psycholoquy/1998.volume.10/psyc. 99.10.027.social-bias.17.ruscio

Sánchez-Bruno, A. y Borges, A. (1997). Violación del supuesto de normalidad en contrastes estadísticos para grupos pequeños. Comunicación presentada al V Simposio de Metodología de las Ciencias del Comportamiento. Sevilla.

Sánchez-Bruno, A., Borges, A. y Cañadas, I. (1999) El contraste de las medias recortadas ante la violación de los supuestos paramétricos. Comunicación presentada en el VI Congreso de Metodología de las Ciencias Sociales y de la Salud. Oviedo.

Sánchez-Bruno, A., Borges, A., San Luis, C. y Cañadas, I. (1999) Algunos problemas del análisis de datos en la investigación en psicología del deporte: potencia de los contrastes en grupos pequeños. Poster presentado al I Congreso Internacional de Psicología Aplicada al Deporte. Madrid

Sánchez Meca, J., Rosa, A.I. y Olivares, J. (1999). Las técnicas cognitivo-conductuales en problemas clínicos y de salud: meta-análisis de la literatura española. *Psicothema, 11*, 641-654.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psycho - logical Methods, 1*, 115-129.

Schmidt, F.L. y Hunter, J. (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. En L. Harlow, S. Mulaik y J. Steiger (Eds.) *What if there were no sig - nificance tests?* Mahwah, N.J.: LEA.

Sedlmeier, P. y Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Serlin, R.C. y Lapsley, D.K. (1985) Rationality in psychological research. The good-enough principle. *American Psychologist, 40*, 73-83.

Serlin, R.C. y Lapsley, D.K. (1993) Rational appraisal of psychological research and the good-enough principle. En G. Keren y Ch. Lewis (eds.) *A handbook for data analysis in the Behavioral Sciences. Methodolo - gical issues*. Hillsdale, NJ: LEA.

Snyder, P. y Lawson, S. (1993) Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334-349.

Snyder, P.A. y Thompson, B. (1998) Use of tests of statistical significance and other analytic choices in a School Psychology Joumal: Review of practices and suggested alternatives. *School Psychology Quarterly, 13*, 335-348.

Tatsuoka, M. (1998) Effect size. En G. Keren y Ch. Lewis (Eds.) *A hand - book for data analysis in the Behavioral Sciences. Methodological is - sues*. Hillsdale, NJ: LEA.

Thompson, B. (1998) In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800

Thompson, B. (1994) Guidelines for authors. *Educational and Psycholo - gical Measurement, 54*, 837-847.

Thompson, B. y Snyder, P.A. (1998). Statistical significance and reliability analysis in recent JCD research articles. *Journal of Counseling and De - velopment, 76*, 436-441.

Tryon, W.W. (1998) The inscrutable null hypothesis. *American Psycholo - gist, 53*, 796.

Vacha-Haase, T. y Ness, C.M. (1999). Statistical significance testing as it relates to practice: Use within Professional Psychology: Research and Practice. *Professional Psychology: Research and Practice, 30*, 104-105.

Valera, A., Sánchez Meca, J., Marín, F. y Velandrino, A. (1998) Potencia estadística de la *Revista de Psicología General y Aplicada*. *Revista de Psicología General y Aplicada, 51*, 233-246.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psy - chological Methods, 4*, 212-213.

Wilkinson, L. and Task Force on Statistical Inference APA Board of Scientific Affairs (1999) Statistical Methods in Psychology journals: Guidelines and explanation. *American Psychologist, 54*, 594-604.

Yela, M. (1996). El problema del método científico en Psicología. *Psicot - hema, 8*, 353-361.

Aceptado el 20 de octubre de 2000