Avances en Medición, 4, 9-22
2006

LA PRUEBA DE SIGNIFICACIÓN DE LA HIPÓTESIS NULA Y SUS ALTERNATIVAS EN EL MARCO DE LA EVALUACIÓN DE LOS RESULTADOS DE INVESTIGACIÓN EN PSICOLOGÍA[1]

Nekane Balluerka[2]
Universidad del País Vasco, España

Juana Gómez[3]
Universidad de Barcelona, España

Mª Dolores Hidalgo[4]
Universidad de Murcia, España

Resumen

La prueba de significación de la hipótesis nula constituye en la actualidad el método de inferencia inductiva más utilizado en la investigación psicológica. Ya desde su implementación ha sido sometida, por una parte, a numerosas críticas y, por otra, a fervorosas defensas respecto a su validez y utilidad. Dado que un porcentaje superior al 90% de la investigación empírica en Psicología utiliza valores de significación para interpretar los resultados obtenidos, en este trabajo se pretende hacer balance de los argumentos a favor y en contra de la prueba de significación de la hipótesis nula, así como de las principales alternativas que se han planteado y de sus respectivas posibilidades de implantación o de generalización en el marco de la evaluación de los resultados. Se ofrecen, además, algunas recomendaciones sobre la adecuada contextualización e interpretación de la prueba de significación de la hipótesis nula y sobre las ventajas de integrar su información con la de procedimientos alternativos que aportan una evaluación complementaria sobre los datos de la investigación.

Palabras clave: Hipótesis nula, Prueba de hipótesis, inferencia estadística

Abstract

The null-hypothesis significance testing is currently the most widely used method of inductive inference in psychological research. Since it was first introduced it has been the subject of both numerous critiques and fervent support as regards its validity and usefulness. Given that over 90% of empirical research in psychology uses significance values to interpret the results obtained, the present study aims to weigh up the arguments in favour of and against the null-hypothesis significance testing, as well as those regarding the main proposed alternatives; the respective likelihood of the latter being introduced and generalised within the context of results evaluation is also considered. In addition, a number of recommendations are made regarding when to use and how to interpret the null-hypothesis significance test, as well as about the advantages of combining the information it provides with that derived from other procedures that offer a complementary evaluation of research data.

Key words: null-hypothesis, Hypothesis test, statistical inference

---

[1] Este trabajo ha sido financiado, en parte, por el Ministerio de Ciencia y Tecnología de España (SEJ2005-09144- C02-02) y, en parte, por la Generalitat de Catalunya (2005SGR00365).

[2] Dpto. de Psicología Social y de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Av. De Tolosa, 70, 20018-San Sebastián, España. E-mail: nekane.balluerka@ehu.es

[3] Dpto. de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Paseo Valle Hebrón, 171, 08035-Barcelona, España. E- mail: juanagomez@ub.edu

[4] Dpto. Psicología Básica y Metodología, Facultad de Psicología, Universidad de Murcia, Campus de Espinardo, 30080-Murcia, España. E-mail: mdhidalg@um.es

La prueba de significación de la hipótesis nula se convirtió en el método de inferencia inductiva más utilizado en Psicología a partir de la década de los 40, período en el que se produjo la revolución inferencial en el ámbito de dicha disciplina. Se trata de un método fácil de aplicar y que conduce a un resultado claramente interpretable que, aparentemente, favorece la objetividad. Probablemente estos factores son los que han provocado el fuerte incremento en la utilización de valores *p* en la investigación empírica; del 70% en la década de los 50 a un uso superior al 90% a partir de los 90 (Hubbard, Parsa y Luthy, 1997; Hubbard y Ryan, 2000). Aún así, la prueba de significación de la hipótesis nula no ha estado, ni está, exenta de controversia. Diversos autores han vertido críticas sobre la prueba a lo largo de diferentes décadas (p.ej. Bakan, 1966; Berkson, 1938; Carver, 1978; Cohen 1990, 1994; Rosnow y Rosenthal, 1989; Thompson, 1996), y algunos otros han elaborado magníficos trabajos defendiendo la utilización del procedimiento (Abelson, 1997; Chow, 1988, 1996; Cox, 1977; Dixon, 1998).

En el presente artículo pretendemos analizar dichas críticas y evaluar las ventajas y los inconvenientes de los procedimientos alternativos a la prueba de significación que se han propuesto a lo largo de este tiempo. A tal fin, empezaremos con una breve introducción acerca de la naturaleza de la prueba. Continuaremos exponiendo las críticas más importantes que ha recibido, así como los argumentos que defienden la validez de la estrategia. Proseguiremos con el análisis de las ventajas y los inconvenientes de las alternativas que se han propuesto como complementarias o sustitutivas. Y, para terminar, expondremos las recomendaciones que hace la *Task Force on Statistical Inference* (TFSI) de la APA (Wilkinson y TFSI, 1999) sobre la utilización de tales alternativas y realizaremos una reflexión acerca de las posibilidades que éstas presentan para ser adoptadas como prácticas habituales en la comunidad científica.

El concepto de prueba de significación estadística de la hipótesis nula

La forma en la que actualmente se utiliza la prueba de significación constituye un híbrido entre la propuesta de Fisher (1925) y la de Neyman y Pearson (1928a, 1928b, 1933). En este apartado abordaremos los aspectos esenciales de ambas propuestas con la intención de clarificar conceptos y expondremos brevemente la forma en la que se utiliza la prueba de significación en la actualidad.

La prueba de significación de la hipótesis nula (H_0) propuesta por Fisher toma en consideración dos probabilidades, la crítica y la calculada, y un tipo de error (el Error Tipo I). La *p* crítica, denominada también alfa (α), hace referencia a la probabilidad asociada a la decisión de rechazar la H_0 cuando es verdadera, es decir, al Error Tipo I. La *p* crítica se establece en base a un juicio subjetivo respecto a las consecuencias de cometer un Error Tipo I y el valor máximo aceptado en la comunidad científica es de *0.05*. Por su parte, la *p* calculada o nivel de significación *p* del resultado empírico expresa la probabilidad de obtener un estadístico muestral igual o mayor al obtenido, dado un determinado tamaño muestral y asumiendo que la muestra ha sido extraída de una población en la que la H_0 es exactamente verdadera. Esta probabilidad depende de los valores de los parámetros reales de la población de la que se ha extraído la muestra y del tamaño muestral. Así, considerando que la H_0 es verdadera en la población, cabe deducir que los estadísticos muestrales serán menos probables a medida que aumenta el tamaño de la muestra y que, por lo tanto, la *p* calculada será menor. En este contexto, cuando la *p* calculada es igual o menor que la *p* crítica se rechaza la H_0, lo que significa que dado nuestro tamaño muestral y asumiendo que la muestra ha sido extraída de una población en la que la H_0 es verdadera, los resultados que hemos obtenido no son consecuencia del error muestral. Debido a que el cálculo de la *p* calculada es bastante tedioso, se suele sustituir por

algún índice estadístico equivalente. Tomando en consideración los conceptos arriba definidos, para someter a prueba una H_0 se debe plantear dicha hipótesis sobre el valor de un parámetro en la población, determinar una distribución muestral o un estadístico como estimador de la función poblacional, obtener una muestra aleatoria de tamaño *n* y calcular en ella el valor del estimador y, por último, calcular la *p* calculada ($p(D/H_0)$). Si la $p(D/H_0)$ es igual o menor que la *p* crítica, se rechaza la H_0; si es mayor, se mantiene la H_0 de forma provisional.

Por otra parte, la prueba de la hipótesis de Neyman y Pearson (1928a, 1928b, 1933) partió de un enfoque distinto. Así, tomando como base la teoría de la probabilidad, estos autores formalizaron una regla de decisión entre dos hipótesis complementarias, la hipótesis nula (H_0) y la hipótesis alternativa (H_1), cada una de ellas con su propia distribución. Bajo este enfoque, además de tenerse en cuenta tanto la $p(D/H_0)$ como la $p(D/H_1)$, se distinguen dos tipos de errores: el Error Tipo I (aceptar la H_1 cuando la H_0 es verdadera), con una probabilidad α que debe fijarse antes de realizar el análisis de datos, y el Error Tipo II (aceptar la H_0 cuando la H_1 es verdadera) con una probabilidad β.

Así, Fisher estableció la prueba de significación de la hipótesis nula como un método de inferencia inductiva, mientras que Neyman y Pearson la concibieron como una regla de decisión. A pesar de esta gran diferencia, Chow (1996) destaca algunas premisas metateóricas que subyacen a ambos enfoques, a saber, la consideración de que la prueba de significación es un procedimiento científico y la de que ambos planteamientos se refieren a distintas fases de un mismo proceso que, en última instancia, nos permite decidir acerca de la influencia o no del azar como posible explicación de los datos obtenidos.

Detractores y defensores de la prueba de significación: puntos en discordia

Tal y como hemos señalado previamente, la prueba de significación ha sido sometida a numerosas críticas y éstas, a su vez, han recibido contra críticas. A continuación procedemos a exponerlas.

1.- Tipo de información proporcionada. Diversos autores (Carver, 1978; Cohen, 1990, 1994; Rozeboom, 1960) defienden la idea de que la inferencia estadística y la prueba de significación de la hipótesis nula responden a diferentes objetivos. El objetivo de la primera es conocer la probabilidad de que la H_0 sea verdadera dados los resultados o los datos obtenidos en la muestra: $p(H_0/D)$; mientras que la segunda nos informa acerca de la probabilidad de obtener unos datos igualmente o más discrepantes que los obtenidos en caso de que la H_0 sea cierta: $p(D/H_0)$. Adheriéndose a esta crítica, Lindley (1957) pudo demostrar que, en algunos casos, la $p(D/H_0)$ y la $p(H_0/D)$ pueden ser muy dispares, por lo que el valor de *p* no refleja la probabilidad de que la H_0 sea cierta. Este fenómeno se conoce como la "paradoja de Lindley". Así, el hecho de obtener un resultado estadísticamente significativo no tiene por qué implicar que la H_0 sea improbable. De hecho, algunos autores defienden que la única forma de obtener esta probabilidad posterior es mediante la estadística bayesiana, determinando la probabilidad de la H_0 antes de realizar la investigación y entendiendo la probabilidad como el grado de creencia en una determinada hipótesis.

Ante esta crítica, autores como Hagen (1997) realizan la siguiente defensa: en los artículos que plantean la crítica, la probabilidad de la H_0 se asocia a frecuencias relativas de base empírica y susceptibles de ser cuantificadas. De esta forma, la H_0 y la H_1 se relacionan directamente con la muestra, cuando las hipótesis estadísticas siempre deben hacer referencia a la población. Así, para Hagen, si partimos de la concepción de la probabilidad basada en frecuencias relativas (como hacen los autores que plantean la crítica), realmente la prueba de significación no aporta la información que se pretende obtener, pero si se equipara la probabilidad de la H_0 con el grado de creencia subjetiva, dicha prueba sí proporciona tal información. Por otra parte, Cortina y Dunlap (1997) afirman que esta crítica no constituye una deficiencia de la prueba de significación en sí misma, sino un problema de interpretación respecto a la información que aporta dicha prueba.

En este punto también cabe citar las contra-críticas de Frick (1996) y Chow (1996) acerca de la adecuación de la estadística bayesiana como solución al problema. En opinión de estos autores, la estadística bayesiana no constituye un procedimiento mejor que la prueba de significación, dado que presenta los siguientes problemas: (a) el cálculo de las probabilidades a priori es subjetivo y arbitrario y la respuesta proporcionada por el análisis bayesiano es transitoria, (b) el análisis bayesiano ignora si una hipótesis explicativa es consistente con el fenómeno que pretende explicar, y (c) la concepción del teorema bayesiano como mera regla inductiva lleva a establecer una equivalencia entre la investigación científica y el puro formalismo. Por último, citaremos la reciente contra-crítica de *Nickerson (2000)*, quien afirma que cuando la $p(H_0)$ es tan grande como la $p(H1)$ y la $p(D/H1)$ es mucho mayor que la $p(D/H0)$, la obtención de un pequeño valor en la $p(D/H_0)$ permite predecir con mucha precisión que la $p(H0/D)$ también tendrá escaso valor, lo que nos lleva a afirmar que, en algunas situaciones, la $p(D/H0)$ y la p(H0/D) no difieren sustancialmente.

2.- Validez lógica de premisas de carácter probabilístico. La base de esta crítica está en el supuesto de que la regla de razonamiento silogístico deductivo conocida como "Modus Tollens" (negar el antecedente por la negación del consecuente) no puede aplicarse a premisas de carácter probabilístico. En este sentido, algunos autores opinan que la prueba de significación está basada en una aplicación incorrecta de dicha regla. Esta regla de razonamiento incorrecto se conoce también como la "ilusión de obtener improbabilidad" y se asocia a varias creencias erróneas, entre las que destacan la creencia de que el valor p es la probabilidad de que la H_0 sea verdadera y la de que $1-p$ (valor complementario de p) expresa la probabilidad de que la H_1 sea verdadera así como la de obtener resultados estadísticamente significativos en caso de replicar el experimento.

Repasemos la secuencia de premisas propia de la prueba de significación de la hipótesis nula: Si la H_0 fuera verdadera, una muestra extraída de la población asociada al valor nulo proporcionaría un estadístico ubicado dentro de un determinado rango de valores. El estadístico no se halla dentro de dicho rango de valores; en consecuencia, probablemente la muestra no procede de una población asociada al valor nulo.

Frente a esta crítica, Cortina y Dunlap (1997) demuestran que la secuencia de premisas no viola las reglas del razonamiento silogístico en aquellos casos en los que las veracidades del

antecedente y del consecuente de la primera premisa están relacionadas de forma positiva. Desde otra perspectiva, Hagen (1997) añade que un argumento puede ser razonable aún careciendo de validez lógica en sentido formal. Respecto a las falsas creencias derivadas del razonamiento incorrecto en el que se basa la presente crítica, Chow (1996) afirma que la interpretación errónea de *p* no se daría si los investigadores comprendieran que *p* es una probabilidad acumulada condicional que depende de que la H_0 sea verdadera.

3.- Contrastación de teorías psicológicas y avance del conocimiento. Cohen (1994) y Rozeboom (1960) defienden que la decisión de aceptar o rechazar la H_0 no permite poner a prueba una teoría psicológica, ya que consideran que es erróneo creer que la prueba de significación nos informa acerca de la probabilidad de que la hipótesis de investigación sea verdadera. Otros autores como Carver (1978), Erwin (1998), Nickerson (2000) y Snow (1998) añaden que, aunque se rechace objetivamente una H_0, es necesario excluir otras hipótesis alternativas rivales antes de aceptar la veracidad de la hipótesis de investigación.

Ante dicha crítica, Chow (1996) argumenta que se halla sustentada en una confusión entre cuestiones estadísticas y no-estadísticas. El autor está de acuerdo en que el resultado de la prueba de significación no aporta suficiente evidencia como para confirmar la hipótesis de investigación; lo que la prueba realmente nos dice es si existe o no base racional para descartar factores aleatorios desconocidos como causantes de los resultados. A partir de ahí, dejando de lado la estadística, hay que realizar una inferencia inductiva para determinar qué factores no aleatorios específicos han sido los responsables de los datos obtenidos. El autor añade que, en el sentido propuesto por los autores de la crítica, ninguna prueba estadística permite comprobar una teoría, ya que tal comprobación constituye un proceso que va mucho más allá de la refutación de una hipótesis estadística.

En la misma dirección que la primera crítica planteada en este punto, muchos autores señalan que la prueba de significación de la hipótesis nula no aporta información acerca de la importancia práctica, ni de la magnitud de los efectos observados. En consecuencia, no permite conocer la verdadera relación existente entre los parámetros poblacionales a partir de los estadísticos muestrales e infravalora la importancia de la magnitud de los fenómenos estudiados, información que, según Tukey (1969, 1991), es esencial para el avance del conocimiento. En estrecha relación con esta última idea y con la concepción de que la H_0 siempre es falsa (ver crítica 5), varios autores (Grant, 1962; Hunter, 1997; Schmidt, 1996) sostienen que el hecho de no rechazarla sólo indica que el investigador es incapaz de especificar la dirección de la diferencia entre determinadas condiciones, mientras que rechazarla indica que tal dirección puede establecerse con cierto grado de confianza. A juicio de dichos autores, el simple conocimiento acerca de la dirección de una diferencia no permite desarrollar una teoría psicológica. Frick (1996) y Chow (1996) se muestran de acuerdo con que la información proporcionada por la prueba de significación no es suficiente para extraer conclusiones en aquellos experimentos cuyo fin radica en aplicar inmediatamente los resultados, ya que en ellos el tamaño del efecto es importante. Así mismo, la prueba tampoco resulta válida para someter a prueba modelos teóricos en los que se realizan predicciones de carácter cuantitativo. Sin embargo, ambos autores sostienen que la prueba resulta muy útil cuando se pretende esclarecer si una determi-

nada diferencia entre condiciones es de carácter positivo o negativo, lo que constituye la base de la mayoría de las teorías y de las leyes que se someten a comprobación en Psicología.

4.- Replicabilidad de los resultados. Otra crítica planteada por varios autores (Cohen, 1994; Falk y Greenbaum, 1995; Shaver, 1993) es la que hace referencia a la falsa creencia de que el valor complementario de *p*, (*1 - p*), expresa la probabilidad de que los resultados sean replicables ("la falacia de la replicación"). Esto sería cierto si la *p* calculada nos permitiera saber cuál es la probabilidad de que la H_0 sea verdadera en la población pero, como ya se ha comentado anteriormente, la *p* calculada no aporta tal información. De hecho, ya hemos visto que para poder calcular la probabilidad de los estadísticos muestrales es necesario asumir que la H_0 es verdadera en lo que respecta a los correspondientes parámetros poblacionales.

En realidad, esta crítica sólo es válida cuando la H_0 es verdadera, ya que si la H_0 es falsa y el tamaño del efecto y el tamaño muestral de la réplica coinciden exactamente con los del estudio original, sí existe una relación monotónica creciente entre replicabilidad y valores *p* (Greenwald, Gonzalez, Harris y Guthrie, 1996). Chow (1996), por otra parte, opina que la naturaleza de la distribución muestral del estadístico, base matemática de la prueba de significación, pone de manifiesto que no existe nada inherente a la prueba de significación que incite a la interpretación errónea que subyace a la falacia de la replicación.

5.- Utilidad informativa. Otra de las críticas importantes a la prueba de significación se hace en base a la consideración de que se trata de una prueba inútil, ya que las distintas muestras observadas, aún procediendo de la misma población, siempre diferirán entre sí en cualquier variable que se mida y, por lo tanto, en sentido literal, la H_0 siempre es falsa (Binder, 1963; Cohen, 1994; Meehl, 1967; Murphy, 1990; Pollard, 1993). Así, el hecho de rechazarla sólo refleja que el diseño de la investigación es lo suficientemente bueno como para detectar un efecto que simplemente existe, sean cuales sean su magnitud y utilidad. Dado que, partiendo de este hecho, los resultados dependen en mayor medida del tamaño muestral que de si la hipótesis de investigación y/o la teoría en la que se apoya son verdaderas o falsas, varios autores (Cohen, 1990, 1994; Kirk, 1996; Nickerson, 2000) plantean lo irónico que resulta preocuparse tanto del Error Tipo I en las investigaciones (dado que no puede producirse, ya que todas las H_0 son falsas) y tan poco del Error Tipo II (aceptando niveles tan altos como *0.5* o *0.8*), cuyas consecuencias, en muchos contextos aplicados, son mucho más graves que las derivadas de cometer un Error Tipo I. A partir de la idea de que la H_0 siempre es falsa, Cohen (1994) y Meehl (1967) concluyen que la utilización de la prueba de significación sólo sería válida en experimentos verdaderos con aleatorización o cuando cualquier desviación del azar puede ser importante.

Para Cortina y Dunlap (1997), sin embargo, la H_0 (o la utilización del valor cero asociado a ella) no carece de utilidad informativa a pesar de que, literalmente, sea falsa. De hecho, el valor cero puede considerarse como el punto medio de un intervalo que: (a) incluye todos los valores que se pueden considerar triviales y (b) es lo suficientemente pequeño como para que los cálculos basados en el valor cero proporcionen una buena estimación de los cálculos basados en otros valores del intervalo. Desde esta perspectiva, el rechazo de la H_0 puede aportar información relevante en una investigación.

Partiendo de otro planteamiento, Hagen (1997) rechaza la crítica argumentando que la H_0 postula que las muestras han sido extraídas de la misma población, no que tengan que ser iguales. Además, añade que, cuando las muestras pertenecen a la misma población, el aumento del tamaño muestral no va directamente ligado a que la probabilidad de rechazar la H_0 se aproxime a 1. Yendo todavía más allá, Baril y Canon (1995) y Frick (1995) sostienen que la H_0 puede ser verdadera, sobre todo en los experimentos en los que la teoría tiene un gran peso y se manipula una sola variable, aunque admiten que en los experimentos de carácter puramente aplicado en los que se manipulan variables complejas es muy difícil que la H_0 sea verdadera. De cualquier forma, dado que la estadística no permite comprobar que la H_0 sea verdadera, Frick propone utilizar todas aquellas estrategias metodológicas que incrementen la probabilidad de detectar un efecto que realmente existe (criterio del esfuerzo adecuado).

6.- Decisión dicotómica de rechazo/no rechazo de la H_0. Por último, otra de las críticas que ha recibido la prueba de significación hace referencia a que el investigador transforma un continuo que va de la probabilidad *0* a la probabilidad *1* en una decisión dicotómica de rechazo/no-rechazo de la H_0. Además, el criterio para la elección del nivel de significación que establece si los resultados son o no estadísticamente significativos es totalmente arbitrario (Glass, McGraw y Smith, 1981; Johnson, 1999; Rosnow y Rosenthal, 1989; Rozeboom, 1960).

Cox (1977) y Frick (1996), por el contrario, consideran que la elección del valor α no es arbitraria, sino que responde a una decisión establecida por la comunidad científica para garantizar la objetividad y eliminar la influencia de las opiniones del investigador en el momento de interpretar los datos. Asimismo, Chow (1996) afirma que se trata de un valor matemático que nada tiene que ver con el bagaje teórico del investigador. A esto, Cowles y Davis (1982) añaden que, a pesar de la opinión generalizada acerca de su arbitrariedad, la elección del valor α derivó de convenciones científicas centradas en la noción del azar.

Por otra parte, pero en el mismo sentido, Gigerenzer (1993) critica la institucionalización de la prueba de significación, que ha hecho que, a menudo, un resultado estadísticamente significativo se asocie a haber realizado correctamente la investigación. No obstante, en opinión de Chow (1996), este argumento no debería llevarnos a abandonar el hábito de rechazar o no rechazar la H_0, aunque sí debería hacernos tomar conciencia de que la prueba de significación ha de utilizarse en el seno de una investigación que tenga una adecuada validez.

Análisis de las principales alternativas a la prueba de significación de la hipótesis nula

Como hemos ido viendo a lo largo del artículo, la prueba de significación de la hipótesis nula ha sido objeto de muchas críticas y, también, de múltiples defensas. Sus adeptos postulan que ha sido mal interpretada y mal usada, defendiendo, al mismo tiempo, la elegancia y creatividad de la lógica en la que se basa, así como su buena integración en el proceso de inferencia estadística (Hagen, 1997) y el hecho de que aporta información que permite responder a preguntas relevantes en el ámbito de la investigación (Abelson, 1997; Dixon, 1998; Krueger, 2001). Sus críticos más radicales, sin embargo, consideran que las alternativas que expondremos en el presente apartado deberían sustituir a la prueba de significación. En este artículo nos centramos básicamente en aquellas alternativas cuyo uso recomienda la *Task Force on Statis-*

tical Inference (TSFI) de la *American Psychological Asociation* (APA) para mejorar la calidad de las investigaciones en lo que respecta al análisis estadístico e interpretación de los datos.

1.- Intervalos de confianza en torno a las estimaciones puntuales. Muchos autores (Brandstätter, 1999; Kirk, 2001; Loftus, 1996; Schmidt, 1996) consideran que el cálculo de los intervalos de confianza en torno a las estimaciones es un buen complemento (o un sustituto) de la prueba de significación. Los intervalos de confianza nos permiten comprobar, de acuerdo con una distribución de probabilidad, si un determinado valor poblacional se encuentra comprendido dentro de un rango de estimaciones. Así, además de la información incluida en la prueba de significación, un intervalo confidencial proporciona un rango de valores dentro de los cuales se halla comprendido, con un determinado nivel de probabilidad, el verdadero parámetro de la población. De esta forma, nos aporta información tanto acerca de la hipótesis cero como de aquellas H_0 que no adoptan el valor 0. Además, el intervalo confidencial refleja la precisión en la estimación del parámetro poblacional (los intervalos estrechos nos dan estimaciones más precisas que los anchos).

Otra ventaja, en el caso de los intervalos para diferencias entre parámetros, es que nos proporcionan información acerca de la dirección y de la magnitud de la diferencia. A esto hay que añadir que se expresan (igual que las estimaciones puntuales) con la misma unidad de medida que los datos, lo que facilita la interpretación. Otras dos ventajas destacables de los intervalos confidenciales respecto de las pruebas de significación son que permiten mantener el nivel de error real en 0,05 y que proporcionan información útil para realizar estudios meta-analíticos en el futuro. Así, además de proporcionar más información y ser más fáciles de interpretar, los intervalos confidenciales evitan problemas inherentes a la prueba de significación de la hipótesis nula, dado que no requieren formular hipótesis a priori ni someten a prueba hipótesis triviales.

Sin embargo, cabe señalar que no todos los autores están de acuerdo con esta serie de ventajas. En este sentido, Abelson (1997), Hagen (1997) y Hayes (1998) rechazan que los intervalos confidenciales superen los problemas de la prueba de significación, ya que parten de la misma lógica. De hecho, en lugar de partir de un parámetro hipotético y establecer una distribución muestral con la cual comparar el estadístico muestral, se establece un intervalo de confianza y se contrastan un número infinito de parámetros tomando dicho intervalo como referencia. Para Cortina y Dunlap (1997), los intervalos confidenciales son tan imperfectos como la prueba de significación, ya que cuando se establece un intervalo confidencial también existe una probabilidad α de cometer un error. Chow (1996) señala que los críticos no explican qué aportaciones diferentes a la prueba de significación nos proporcionan los intervalos de confianza para la comprobación de teorías y Frick (1996) añade que la información que nos da el valor *p* es lo suficientemente importante como para no remplazarla por un intervalo confidencial.

2.- Tamaños del efecto. El tamaño del efecto, según Cohen (1988), es el grado en el que el fenómeno se halla presente en la población o el grado en el que la H_0 es falsa (si estamos en el contexto de la prueba de significación). Para Snyder y Lawson (1993), expresa la medida en la que la variable dependiente puede ser controlada, predicha o explicada en función de la(s) variable(s) independiente(s). Los tamaños del efecto nos informan acerca de la magnitud del

efecto observado, además de permitirnos comparar directamente los resultados de diferentes investigaciones, ya que estos índices son transformaciones a una escala común (hecho que resulta muy útil para obtener información sobre la significación práctica de los resultados de una investigación cuando las escalas de medida de las variables no son familiares). También sirven para realizar análisis de potencia en el marco de las pruebas de significación y para llevar a cabo estudios meta-analíticos.

Pero los tamaños del efecto tampoco están exentos de críticas. De hecho, varios autores (Brandstätter, 1999; Dooling y Danks, 1975) afirman que deben interpretarse con cautela, puesto que dependen de la variabilidad de las medidas y de las manipulaciones experimentales utilizadas en una determinada muestra. Además, en algunos casos, la estandarización puede llevar a distorsiones en el orden o en la intensidad de los efectos observados (Greenland, 1998). Por otra parte, resulta muy complejo decidir qué índice es el más adecuado para calcular el tamaño del efecto en cada contexto, y los investigadores suelen discrepar al respecto (Gorsuch, 1991; Parker, 1995; Strahan, 1991). Para Chow (1996) no supone ningún problema evaluar los resultados de la investigación con criterios no estadísticos (como son los tamaños del efecto) una vez descartado el azar como explicación plausible de los datos (mediante la prueba de significación), pero no le encuentra sentido a evaluar el impacto de los resultados en la vida real sin haber determinado si el suceso es azaroso. Por otra parte, añade que las decisiones tomadas en base a los tamaños del efecto son igual de arbitrarias que las basadas en el valor *p*.

3.- Intervalos de confianza para el tamaño del efecto. Los autores que recomiendan el cálculo de los intervalos de confianza para el tamaño del efecto, lo hacen basándose sobre todo en que proporcionan información fácil de comprender que ayuda a la buena interpretación de los resultados, en que resultan muy útiles para acumular información aportada por diferentes investigaciones (facilitando la realización de estudios meta-analíticos) y en que la anchura de los intervalos proporciona información acerca de la precisión en la estimación; además, suelen argüir que existe una asociación entre los intervalos de confianza y la prueba de significación de la hipótesis nula que puede favorecer una mejor comprensión de la lógica que atañe a las dos estrategias (cuando un intervalo no incluye un determinado valor cabe rechazar la H_0 que afirma que ese valor es verdadero).

Sin embargo, a pesar de las ventajas que comporta el cálculo de los intervalos confidenciales para el tamaño del efecto y de que muchos autores (Cumming y Finch, 2001; Fidler y Thompson, 2001; Smithson, 2001) se hayan adherido a la recomendación de la TSFI de la APA respecto a la necesidad de realizar dicha estimación para todos aquellos tamaños del efecto asociados a resultados principales, esta práctica todavía es poco habitual. Probablemente ello se deba a que el cálculo de los intervalos de confianza para el tamaño del efecto requiere el uso de distribuciones "no-centrales" y la utilización de software especializado.

En el mismo sentido, Rosenthal y Rubin (1994) proponen adquirir el hábito de calcular los "intervalos contranulos" para los tamaños del efecto como alternativa a los intervalos de confianza. Estos autores definen el valor contranulo del tamaño del efecto como la magnitud no nula del tamaño del efecto que está apoyada por la misma cantidad de evidencia que el valor nulo de dicho tamaño. El uso del valor contranulo junto con el valor *p* eliminaría el error

de considerar que no rechazar la H_0 equivale a obtener un tamaño del efecto igual a *0* y ayudaría a erradicar la creencia de que la obtención de un resultado estadísticamente significativo se asocia a un logro científico importante.

4.- Análisis de potencia. Neyman y Pearson (1928a, 1928b), en el contexto de la perspectiva de la decisión estadística, plantearon que se podría determinar el tamaño muestral necesario para detectar un efecto realmente existente en la población, usando para ello el tamaño del efecto en la población hipotética y fijando valores para las probabilidades asociadas al Error Tipo I (α) y al Error Tipo II (β). También consideraron la posibilidad de fijar α y el tamaño muestral a fin de calcular β o su complemento, la potencia del contraste (1-β), que es la probabilidad de rechazar la H_0 cuando es falsa. Este tipo de análisis es especialmente importante cuando se pretende concluir que no existe efecto, o que su magnitud no es relevante, a pesar de no resultar posible rechazar la H_0 (Meehl, 1991; Schafer, 1993). Respecto a esta alternativa, Hagen (1997) considera que sigue exactamente la misma lógica que la prueba de significación y Nickerson (2000) advierte de que obtener un resultado estadísticamente significativo también depende de la variabilidad de los datos y que ésta no se tiene en cuenta en el análisis de potencia.

La crítica más dura a esta alternativa la plantea Chow (1996), quien considera que el valor de potencia que se establece como adecuado también ha sido escogido de forma arbitraria por la comunidad científica. Además, cree que los siguientes argumentos cuestionan su validez: (a) cuando se considera el análisis de potencia se modifica el significado del Error Tipo II y (b) no se puede representar gráficamente la potencia estadística sin representar erróneamente la prueba de significación.

5.- Replicación. Varios autores defienden la necesidad de replicar las investigaciones para alcanzar conocimiento científico (Allen y Preiss, 1993) y para que los resultados superen la especulación (Hubbard y Armstrong, 1994). A pesar de la cantidad de autores que defienden el valor y la necesidad de la replicación, ésta se usa muy poco en psicología debido, en gran medida, al alto coste que supone. Sin embargo, se ha de señalar que la replicación (externa o interna) constituye el método más objetivo para comprobar si el resultado de una investigación es fiable.

Recomendaciones y posibilidades de desarrollo de las alternativas a la prueba de significación en la investigación psicológica

A lo largo del artículo se han presentado las críticas y las contra-críticas que han rodeado a la prueba de significación de la hipótesis nula y se han examinado las ventajas e inconvenientes de las principales alternativas que han sido propuestas para complementar o sustituir la información proporcionada por dicha prueba. A pesar de las discrepancias que puedan existir entre los autores, consideramos necesario recordar que la TFSI de la APA recomienda el uso de algunas pruebas alternativas para complementar, aunque no para sustituir, la información que aporta la prueba de significación. En este último apartado expondremos tales recomendaciones y realizaremos una breve reflexión acerca de las posibilidades que éstas presentan para ser adoptadas como prácticas habituales en la comunidad científica.

La TFSI recomienda el uso de los intervalos de confianza en torno a las estimaciones puntuales porque considera que proporcionar un valor *p* real y, en mayor medida, un intervalo de confianza siempre es preferible a tomar una decisión dicotómica de rechazo/no-rechazo. En nuestra opinión, el valor añadido de los intervalos confidenciales respecto a la prueba de significación es que aportan información referida a la precisión en la estimación de los parámetros poblacionales. Cabe señalar que el uso de esta alternativa cuenta con un alto grado de seguimiento en la comunidad científica debido a que el cálculo de los intervalos de confianza es muy sencillo y, además, está incorporado en la mayoría de los programas estadísticos de uso habitual.

En lo que concierne a la alternativa referida a los tamaños del efecto, la TFSI insta a presentarlos para todos los resultados principales. Además, destaca la necesidad de interpretarlos dentro de un contexto práctico y teórico y señala la importancia que tienen tales índices para realizar análisis de potencia y meta-análisis en el futuro. A nuestro juicio, siempre que se parta de un profundo conocimiento teórico sobre el objeto de estudio, el cálculo de los tamaños del efecto puede resultar muy útil para obtener información acerca de la importancia práctica de los datos derivados de la investigación. Además, si se contextualizan dentro de un conjunto más amplio de estudios, nos permiten incrementar la precisión en la estimación de los parámetros poblacionales. A pesar de sus múltiples ventajas y de que el cálculo de la mayoría de los índices del tamaño del efecto es relativamente sencillo, la revisión de la literatura muestra claramente que su uso está poco extendido en la comunidad científica. Sin embargo, hemos de señalar que existe software específico y, en muchos casos gratuito, para calcular distintos índices del tamaño del efecto y que algunos programas estadísticos de uso habitual (por ejemplo, el SPSS) también proporcionan la opción de calcular directamente, o mediante macros, algunos de estos índices. Además, cada vez son más las revistas que exigen el cálculo de los tamaños del efecto como requisito para la aceptación de cualquier trabajo. Por estas razones, consideramos que la estimación de los tamaños del efecto llegará a generalizarse, aunque, debido a la confusión que existe al respecto, mantenemos ciertas reservas en cuanto a que los investigadores lleguen a calcular aquellos índices que resulten más adecuados en cada situación de investigación.

En estrecha relación con la alternativa que acabamos de examinar, la TFSI también recomienda calcular los intervalos de confianza asociados a los tamaños del efecto en el caso de los resultados principales. Aún cuando, como se ha puesto de manifiesto en el apartado anterior, las ventajas asociadas a esta práctica son múltiples, consideramos que está lejos de convertirse en una rutina habitual en la comunidad científica. A nuestro juicio, ello se debe básicamente a que los programas estadísticos de uso común no proporcionan la posibilidad de calcular directamente los intervalos de confianza para el tamaño del efecto y a que los investigadores aplicados no están habituados a trabajar con distribuciones no-centrales, las cuales son necesarias para realizar dicho cálculo.

En cuanto al análisis de potencia, para la TFSI adquiere sentido cuando se realiza antes de la recogida y del análisis de los datos. Por ello, recomienda calcular un rango de análisis de potencia para poder observar cómo cambian las estimaciones de la potencia en función de diferentes tamaños del efecto y niveles alfa. A su vez, sugiere que, en la descripción de los resul-

tados, la potencia calculada sea sustituida por los intervalos de confianza. En nuestra opinión, el análisis de potencia es un procedimiento muy útil para garantizar la confianza en los resultados obtenidos, por lo que debería constituir una práctica habitual. Además, existe software gratuito y fácil de utilizar que permite llevar a cabo distintos tipos de análisis de potencia, tanto antes como después de realizar la investigación. Por ello, consideramos que el hecho de que no sea una práctica muy extendida en la comunidad científica puede deberse, básicamente, a que los investigadores aplicados desconocen la utilidad de llevar a cabo este tipo de análisis. De hecho, aún cuando el software que permite realizar análisis de potencia se encuentra disponible desde hace mucho tiempo, la verdad es que, como demuestran varios autores (p.ej. Kazdin y Bass, 1989; Sedlmeier y Gigerenzer, 1989), los hábitos de los investigadores parecen inamovibles, lo que nos lleva a augurar que el uso de esta alternativa no llegará a generalizarse en un futuro inmediato.

Por último, respecto a la replicación, la TFSI sugiere llevar a cabo replicaciones del estudio original para no incurrir en el error de publicar teorías falsas derivadas del uso de una metodología inadecuada, aún cuando el análisis estadístico sea correcto. Al igual que el análisis de la potencia, consideramos que la replicación es un procedimiento esencial para garantizar que los resultados obtenidos son fiables. Sin embargo, dado el alto coste que supone llevarla a cabo (máxime en el caso de la replicación externa) y lo difícil que resulta publicar un estudio de replicación en las revistas científicas, dudamos seriamente de que llegue a convertirse en una práctica habitual en la comunidad científica.

En definitiva, al evaluar los resultados de la investigación psicológica deberíamos optar por un uso reflexivo e integrador de los procedimientos de los que disponemos; ello supone abandonar posturas dogmáticas y/o el uso sistemático de recetas estadísticas y procedimentales. Desde aquí, abogamos por seguir utilizando la prueba de significación de la hipótesis nula, dándole una contextualización y una interpretación adecuadas, y por complementar la información que ésta aporta con la de procedimientos alternativos que informen sobre la precisión de los parámetros y la importancia práctica de los resultados obtenidos y aumenten el grado de certeza de los hallazgos.

Referencias

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-144). Hillsdale, NJ: Erlbaum.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Allen, M., y Preiss, R. (1993). Replication and meta-analysis: A necessary connection. *Journal of Social Behavior and Personality, 8*(6), 9-20.

Bakan, D. (1966). The tests of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Baril, G. L., y Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist*, 50, 1098-1099.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70*, 107-115.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online, 4*(2), 33-46.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48* (3), 378-399.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103*, 105-110.

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility*. Beverly Hills, CA.: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45(12)*, 1304-1312.

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

Cortina , J. M., y Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2* (2), 161-172.

Cowles, M., y Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37*(5), 553-558.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4*, 49-70.

Cumming, G., y Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin and Review, 5*, 390-396.

Dooling, D., y Danks, J. H. (1975). Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society, 5*, 15-17.

Erwin, E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences, 21*, 197-198.

Falk, R., y Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*, 75-98.

Fidler, F., y Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement, 61*(4), 575-604.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver y Boyd.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition*, *23* (1), 132-138.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1* (4), 379-390.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren, y C. Lewis (Eds.), *A handbook for data analysis in the behavioural science: Volume 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Glass, G. V., McGaw, B, y Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gorsuch, R. L. (1991). Things learned from another perspective (so far). *American Psychologist, 46*(10), 1089-1090.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological review, 69*, 54-61.

Greenland, S. (1998). Meta-analysis. In K. Rothman, y S. Greenland (Eds.). *Modern epidemiology*. Philadelphia: Lippincot-Raven.

Greenwald, A. G., Gonzalez, R., Harris, R. J., y Guthrie, D. (1996). Effect sizes and *p*- values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175-183.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52 (1)*, 15-24.

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

Hubbard, R., y Armstrong, J.S. (1994). Replications and extensions in Marketing: Rarely published but quite contrary. *International Journal of Research in Marketing, 11*, 233-248.

Hubbard, R., Parsa, A. R., y Luthy, M.R. (1997). The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917-1994. *Theory and Psychology, 7*(4), 545-554.

Hubbard, R., y Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement, 60*(5), 661-681.

Hunter, J. E. (1997). Need: A ban on the significance test. *Psychological Science, 8*, 3-7.

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

Kazdin, A. E., y Bass, D.(1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138-147.

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement, 56 (5)*, 746-759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61 (2)*, 213-218.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist, 56 (1)*, 16-26.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187-192.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way to analyse data. *Current Directions in Psychological Science, 5*, 161-171.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. E. Snow, y D. E. Wilet (Eds.), *Improving inquiry in social science: A volume in honour of Lee J. Cronbach.* (pp. 13-59). Hillsdale, NJ: Erlbaum.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45*(3), 403-404.

Neyman, J., y Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A*, 175-263.

Neyman, J. y Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A*, 264-294.

Neyman, J., y Pearson, E.S. (1933). On the testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 28*, 492.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of and old and continuing controversy. *Psychological Methods, 5 (2),* 241-301.

Parker, S. (1995). The "difference of means" may not be the "effect size". *American Psychologist, 50*, 1101-1102.

Pollard, P. (1993). How significant is "significance"? In G. Keren, y C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences: Volume 1. Methodological issues*. Hillsdale, NJ: Erlbaum.

Rosenthal, R., y Rubin, D. B. (1994). The counternull value of an effect size: A new Statistic. *Psychological Science, 5(6)*, 329-334.

Rosnow, R. L., y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

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

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

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education, 61*, 383-387.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1 (2)*, 115-129.

Shaver, J. (1993). What statistical significance testing is, and what is not. *Journal of Experimental Education, 61*(4), 293-316.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: the importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61 (4)*, 517-531.

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

Snow, R. E. (1998). Inductive strategy and statistical tactics. *Behavioral and Brain Sciences, 21*, 219.

Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46*, 1083-1084.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25(2)*, 26-30.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83-91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Wilkinson, L., y the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Manuscrito recibido en agosto de 2005
Aprobado para publicación en octubre de 2005