Revista de Historia de la Psicología
2000, Vol. 21, nº 1, pp. 9-26

EL PROCEDIMIENTO DE SIGNIFICACIÓN ESTADÍSTICA (NHST): SU TRAYECTORIA Y ACTUALIDAD

JUAN PASCUAL LLOBELL[1]
MARÍA DOLORES FRÍAS NAVARRO[1]
JOSÉ FERNANDO GARCÍA PÉREZ[1]
Universitat de València

RESUMEN

Toda investigación empírica requiere del tratamiento estadístico de los datos. Sin embargo el procedimiento usual de comprobación estadistica de hipotesis ha sido causa de confusión, de falacias lógicas y de interpretaciones incorrectas. El problema ha perdurado hasta la actualidad, siendo muchas las voces que reclaman procedimientos alternativos y/o complementarios de análisis, además de una pedagogia más exigente y una practica cientifica más rigurosa en la comprensión e interpretación de los datos.

Palabras clave: tamaño del efecto, significación estadística, prueba de hipótesis, procedimiento de hipótesis nula, tests estadísticos, metodología.

ABSTRACT

The standard hypothesis testing method has a number of well-known logical fallacies and the results of the procedures are often misinterpreted. Many scholars have suggested that, perhaps, NHST should be abandoned altogether in favor of other bases for conclusions such as confidence intervals and effect size estimates. Other researchers are often interested in testing the hypothesis that the effects of treatments, interventions, etc.

[1] Área de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universitat de València. Avenida Blasco Ibañez, 21, 46010 Valencia. E-mail: juanpa@uv.es, friasnav@uv.es, garpe@uv.es

are negligibly small rather than testing the hypothesis that treatments have no effects whatsoever.

We argue that we must question the "old" procedures to estimulate the application of new statistical procedures in the progress of scientific inference.

Key Words: effect size, statistical significance, hypothesis testing, null hypothesis testing, statistical tests, methodology

Toda investigación empírica requiere siempre del tratamiento estadístico de sus datos, al menos cuando se desea de verdad describir y comprender en su justa medida la relación entre las variables que definen la situación estudiada. Sea para comprobar teorías, sea para estimar los efectos de algún tratamiento, experimental o clínicamente aplicado, los investigadores están obligados a realizar un proceso de comprobación de teorías, hipótesis o supuestos, traduciendo la hipótesis científica a hipótesis estadística.

Históricamente el contraste y comprobación de hipótesis estadísticas ha sido causa de confusión, de crítica y controversia entre los investigadores, provocando entre ellos mismos interpretaciones a veces erróneas de los resultados que no han favorecido una acumulación científica y exenta de sesgo del conocimiento. El problema ha perdurado durante décadas, reavivándose en ocasiones, quedando latente en otras, hasta llegar al presente actual en el que la polémica está servida con opiniones claramente enfrentadas, en algunos casos de forma extrema, entre los defensores y detractores de las pruebas de significación estadística, bien sea acerca de la utilidad de su uso o, inclusive, acerca de la pertinencia de las mismas como instrumento válido para el progreso científico.

El principal objetivo de nuestro trabajo se centra en categorizar ambas posturas y tratar de descubrir otras líneas complementarias de actuación estadística que ayuden en el futuro al investigador a un mejor "modus operandi".

La aparición de nuevas técnicas estadísticas y su incorporación a la práxis científica durante los primeros años del siglo XX permitió un gran avance de las Ciencias del Comportamiento. Hasta ese momento la técnica convencional utilizada para analizar las diferencias entre grupos era la 'razón crítica' que se empleaba para determinar el grado de relación entre las variables independientes y dependientes de los experimentos. Entre los representantes de esta nueva era destacan los trabajos

de William Sealey Gosset (1876-1937) (quien utilizaba el seudónimo de Student), Karl Pearson (1857-1936), Egon S. Pearson (1895-1980) o Jerzy Neyman (1894-1981) y muy especialmente el modelo estadístico y de probabilidad que Sir Ronald Aylmer Fisher (1890-1962) desarrolló tomando como primera referencia la investigación agrícola y que la Psicología supo aprovechar para el análisis del comportamiento humano y animal. Su contribución fue decisiva para el desarrollo del diseño experimental y el análisis estadístico de la investigación psicológica (véase por ejemplo Fisher, 1925), destacando como sus aportaciones más relevantes:

1. La formulación de la distribución precisa de muchos estadísticos.
2. El empleo explícito del procedimiento de la aleatorización como garantía de la equivalencia previa de los grupos a comparar.
3. El desarrollo y aplicación moderna del Análisis de la Varianza cuya primera referencia la encontramos en la investigación agrícola que junto a MacKenzie realizó en 1923.

Muy brevemente, las pruebas de significación estadística facilitan al investigador un test o prueba que informa de la probabilidad de conseguir la diferencia obtenida (cuando se trata de comparación de grupos de observaciones), o mayor que la observada, si la hipótesis nula es cierta. La prueba estadística asume que la hipótesis nula es cierta en la población. Si el valor (p) de probabilidad es igual o menor que 0.05 se concluye que la probabilidad de obtener tales resultados por azar es insignificante, o sea, la probabilidad de que la variabilidad muestral aleatoria explique el resultado obtenido es muy baja y por lo tanto se rechaza la hipótesis de nulidad y se afirma que el resultado es estadísticamente significativo. En esencia, se trata de un procedimiento mediante el que se descarta el azar como explicación y se toma para ello un criterio de decisión que no es otro que el valor p asociado a los datos bajo el supuesto de azar.

En la conocida base de datos PsicLIT el término *Null Hypothesis Testing* fue introducido en 1973 como "*aplicación de pruebas estadísticas para determinar si una hipótesis nula puede ser aceptada o rechazada. Limitado a discusiones estadísticas*". Un análisis de la evolución de dicho término como *descriptor principal* permite observar el ascenso de su presencia en la base de datos con tres puntos claramente destacados al final del siglo XX: 1990 con 27 trabajos, 1995 con 12 trabajos y 1997 con 17 trabajos.

Aunque de introducción tardía, el crecimiento con los años del numero de publicaciones bajo dicho término habla bien a las claras del interés por el tema. Interés condimentado de debate, critica y búsqueda de alternativas, sean disyuntivas o complementarias.

LA PRUEBA DE SIGNIFICACIÓN DE LA HIPÓTESIS NULA: CRÍTICAS

El interés por el *test de significación de la hipótesis nula* (NHST) como modo de proceder en la investigación científica pronto fue criticado (Berkson, 1938, 1942), pero especialmente desde los años sesenta la lógica y el uso del contraste de hipótesis clásico ha merecido muchas críticas (véase una ilustración en *Cuadro 1*). Véanse entre otros los trabajos de Bakan, 1966, 1967; Binder, 1963; Carver, 1978, 1993; Cohen, 1990, 1994; Cronbach, 1975; Falk y Greenbaum, 1995; Grant, 1962; Folger, 1989; Greenwald, 1993; Guttman, 1985; Kirk, 1996; Lykken, 1968; Loftus, 1991, 1993a; Meehl, 1967, 1978, 1990; Nunnally, 1960; Oakes, 1986; Pollard, 1993; Rosenthal, 1990a, 1990b; Rosenthal y Rubin, 1985; Rosnow y Rosenthal, 1989; Rozeboom, 1960; Schmidt, 1992, 1996; Serlin y Lapsley, 1985, 1993; Shulman, 1970; Signorelli, 1974; Tukey, 1991.

Cuadro 1. Procedimiento de comprobación de la hipótesis nula (Carver, 1993)

La aplicación sistemática, rutinaria las mas de las veces, del procedimiento de comprobación de la hipotesis nula *no garantiza el descubrimiento de la verdad y en algunos casos puede llegar a corromper el método científico.* El siguiente ejemplo histórico es didáctico al respecto.

En 1887, Michelson y Morley recogieron datos acerca de la velocidad de transmisión de la luz para comprobar la hipótesis de si la luz se transmite a través de un medio que los científicos denominaban "ether". Caso de depender del eter, la luz se transmitirá a su través más rápidamente cuando la dirección de la luz coincide con la dirección del movimiento de la tierra (y del eter). Ejemplos análogos nos ayudarán a entender el contenido de la hipotesis: la barca se desplazará más rápidamente a favor de la corriente que contra corriente o el viaje de Madrid a New York de ida tardará más o menos que el de vuelta según se coincida o no con el movimiento de rotación de la tierra.

Pues bien los autores interpretaron sus resultados de observación, por cierto sin necesidad de aplicar ningún test de significación, en el sentido de que la luz se transmitia siempre a la misma velocidad, independientemente de la dirección que le hicieran tomar.

Sin embargo, sometido el protocolo básico de datos a un análisis ANOVA por el propio Carver, se encontraron diferencias significativas a favor del efecto de dirección de transmisión de la luz (p=0.001). Concluye Carver que si los autores hubieran utilizado el método de comprobación de la hipótesis nula tal como se aplica habitualmente en la investigación psicológica, el curso de la historia científica probablemente hubiera cambiado a peor. El trabajo posterior de Einstein sobre la relatividad se apoya sobre la idea de

que la velocidad de la luz se transmite a la misma velocidad siempre, sea cual sea la dirección tomada.

Por suerte los autores interpretaron sus datos a la luz de su hipótesis de investigación, sin testar para nada la probabilidad de esos mismos datos bajo el supuesto de la hipotesis nula.

Comentarios como "The Amazing persistence of a Probabilistic Misconception (Falk y Greenbaum 1995), o "one is tempted to conclude that NHST in an addictive affiliation of behavioural scientists" (Greenwald, Gonzales, Harris y Guthrie, 1996), "On the tyranny of hypothesis testing in the social sciences" (Loftus, 1991) o "NHST has not only failed to support the advance of psychology as a science but has also seriously impeded it" (Cohen, 1994) llenan las páginas de escritos que ante todo destacan el fracaso de tal procedimiento como el método ideal de análisis de los datos.

Durante la década de los noventa también ha habido momentos de destacadas reflexiones teóricas que han originado debates en foros como la revista *Journal of Experimental Education* (volumen 61 de 1993) o la *American Psychologist* (volumen 49 de 1994) o en *libros* monograficos (Chow, 1996; Gigerenzer, Swijtink, Porter, Daston, Beatty y Krüger, 1989; Harlow, Mulaik y Steiger, 1997; Kirk, 1972; Morrison y Henkel, 1970) diseñados para recopilar ensayos sobre el proceso de decisión estadística y revisar los principios de las pruebas de comprobación de hipótesis. En todos ellos, de manera más o menos explícita se comentan las distintas tradiciones desde la que cabe afrontar el problema: por ejemplo Huberty (1993) ha valorado el impacto reflejado en los manuales de estadística de la controversia entre Fisher y Neyman-Pearson. Bastante menos se ha escrito sobre la alternativa bayesiana que, sin duda ha tenido menos impacto histórico en Psicologia, pero a la que no se puede olvidar en la actualidad como una via de gran futuro (Dixon, 1998; Dixon y O'Reilly, 1999)

Recientemente, la *American Psychological Association* (A.P.A.) constituyó en Marzo de 1996 un grupo de trabajo sobre Inferencia Estadística con el mandato prioritario de clarificar los temas importantes que están relacionados con la práctica estadística contemporánea en Psicología. Una de las cuestiones propuestas a la comisión con carácter de urgencia giraba en torno a las pruebas de significación estadística y al uso que de las mismas se hace en la praxis habitual de los científicos, praxis guiada, tutelada y obligada por las normas de publicación impuestas por los comités editoriales de las revistas y por la propia APA. El grupo de

trabajo inicial estaba formado por profesionales sobresalientes como Robert Rosenthal, Robert Abelson y Jacob Cohen (directores del grupo) junto con Mark Applebaum, Leona Aiken, Gwyneth Boodoo, David Kenny, Helena Kraemer, Donald Rubin, Bruce Thompson, Howard Wainer y Lee Wilkinson. Como supervisores fueron nominados Lee Cronbach, Paul Meehl, Frederick Mosteller y John Tukey. Su primer trabajo elaborado en la reunión de Diciembre de 1996 señala que de ninguna manera procede rechazar la práctica de la comprobación de la hipótesis nula y la obtención del valor *p*. Pero claramente se sugeriría la necesidad de complementar la presentación, análisis e interpretación de los datos con otros estadísticos, por ejemplo, la estimación del tamaño del efecto o de los intervalos de confianza.

Anteriormente en su cuarta edición del manual de publicación de la *American Psychological Association* (1994) se incluían ciertas recomendaciones sobre el estilo de los informes de investigación y se estimulaba a los investigadores a proporcionar información expresa sobre el tamaño del efecto además de los valores de probabilidad asociados a las pruebas de significación estadística, haciendo más fácil de esta manera la posterior integración de los resultados empíricos (vía meta-análisis, por ejemplo) y posibilitando una interpretación más adecuada de los resultados desde su importancia sustantiva, clínica o aplicada.

Influenciados sin duda por ese mismo espíritu, consejos editoriales de algunas revistas científicas empezaron a recomendar a los autores la importancia de informar y en su caso interpretar acerca de las medidas de magnitud del efecto. Por ejemplo, han adoptado dicho criterio la revista *Memory and Cognition* (Loftus, 1993b), la revista *Educational and Psychological Measurement* (Thompson, 1994a), y más recientemente el *Journal of Experimental Education* (Heldref Foundation, 1997), el *Journal of Consulting and Clinical Psychology* (Kendall, 1997), el *Journal of Applied Psychology* (Murphy, 1997), el *Journal of Learning Disabilities* (Hresko, 2000), el *Language Learning* (Ellis, 2000) y el *Research in the Schools* (McLean y Kaufman, 2000).

Las reuniones científicas tampoco se han mantenido ajenas al problema, haciéndose eco de la polémica y dedicando sesiones al debate de la controversia, (planteándose en algunos casos su abandono, véase por ejemplo el trabajo de Schmidt, 1996). En las reuniones anuales de la A.P.A. y de la *American Psychological Society* (A.P.S.) celebradas en 1996 se propuso a debate la siguiente cuestión: "*should significance tests be banned*". En 1997 esta misma pregunta fue recogida por Hunter (1997). Y el mismo Jacob Cohen y Bruce Thompson fueron invitados a participar ese mismo año en el Congreso que se realizó en Chicago

promovido por la *American Psychological Association* con dos trabajos cuyos títulos fueron "*Much ado about nothing*" (Cohen, 1997) y "*If statistical significance tests are broken/misused, what practices should supplement or replace them*" (Thompson, 1997). Desgraciadamente Cohen, profesor emérito de la Universidad de Nueva York, no podrá proseguir su trabajo, falleció el 20 de Enero de 1998. El interrogante planteado es uno de los temas de mayor actualidad.

En el número de Julio de 1998 de la revista *American Psychologist* se retoma de nuevo el tema mediante un conjunto de trabajos que critican y valoran la defensa que Hagen (1997) realizó en esta misma publicación sobre las pruebas de la hipótesis nula. El debate y la actualidad del tema siguen vigentes a ida de hoy y parece destinado a aparecer de manera virulenta y con carácter cíclico de tanto en tanto probablemente porque no ha sido resuelto de manera satisfactoria o porque todavía queda por descubrir la mejor opción posible.

Podemos agrupar en tres grandes áreas o categorías las principales críticas (Judd y McClelland, 1989; Hagen, 1997; Kirk, 1996):

1. Las pruebas de contraste estadístico de hipótesis no nos dicen lo que realmente queremos saber.
2. La hipótesis de nulidad siempre es falsa en sentido estricto y, en consecuencia, su comprobación resulta trivial.
3. El nivel inflexible de error de Tipo I, fijado a priori, puede conducir a incertidumbre en la interpretación de los resultados.

Las pruebas de contraste estadístico de hipótesis NO nos dicen lo que realmente queremos saber

La respuesta que el razonamiento científico busca tiene que ver con esta pregunta: "*dado que hemos encontramos estos datos, ¿cuál es la probabilidad de la hipótesis nula?*" Es decir, ($P(H_0|Datos)$), pero las pruebas de significación estadística responden a otra cuestión bien diferente que es la siguiente, "*dado que la hipótesis(nula) es verdadera, cuál es la probabilidad de estos datos o datos más extremos*" ($P(Datos|H_0)$). Por supuesto obtener un valor de ($P(Datos|H_0)$) pequeño no implica que ($P(H_0|Datos)$) también lo sea, porque ambos tipos de probabilidad condicional son inconfundibles.

Cuando los investigadores pasan de comprobar que "*el valor de p asociado a la prueba estadística es menor de 0.05*" a concluir que "*la hipótesis de nulidad probablemente es falsa*", están dando un paso en falso que agrava aún más si cabe el problema. Afirmar la probabilidad, baja o alta, de unos datos, no informa directamente nada acerca de la probabilidad, alta o baja, de la hipótesis de referencia. Otro tanto cabe decir cuando del no rechazo de la hipótesis nula ($p > 0.05$) se concluye

que la hipótesis de nulidad es cierta.

Otros razonamientos igualmente incorrectos afirman que el complemento de *p* es la probabilidad de encontrar un nuevo resultado estadísticamente significativo caso de volver a replicar eela investigación. Todas estas interpretaciones constituyen un error que Falk y Greenbaum (1995) han rotulado como "*the illusion of attaining probability*".

En definitiva, las interpretaciones inadecuadas usuales de las pruebas de significación estadística (el valor *p*) han fundamentado en parte las duras críticas que esta técnica ha recibido (Manzano, 1997). Según Carver (1978) hay por lo menos tres errores comunes (no son los únicos) que una vez identificados conviene corregir:

1. interpretar el valor de *p* como la probabilidad de que los resultados sean debidos al azar
2. interpretar *p* como la probabilidad de poderlos replicar y
3. creer que los resultados están directamente relacionados con la probabilidad de verdad de la hipótesis de investigación

Chow (1996) ha destacado de manera clara el sentido de la probabilidad (valor *p*) cuyo valor estimamos en nuestros análisis y señala dos cuestiones:

1. "*la probabilidad asociada no se refiere a la probabilidad de un valor particular del test estadístico; se refiere a la probabilidad de un valor particular del test estadístico (por ejemplo t = 2.05) más las probabilidades de todos los posibles valores más extremos (por ejemplo t = 2.56, 3, 29 etc.)*" (página 38-39). El valor de *p* es, por tanto, una probabilidad acumulada.
2. "*al mismo tiempo, p es una probabilidad condicional. A saber, es una probabilidad acumulada contingente a la aceptación de* H_0. *Es decir, tratar p como la probabilidad de que los resultados sean producidos por azar es poner el carro antes que el caballo*" (página 39). El valor de *p* es, por tanto, una probabilidad condicional.

La hipótesis de nulidad siempre es falsa en sentido estricto y su comprobación resulta trivial

Otras críticas plantean la trivialidad, inclusive la circularidad, del contraste de hipótesis. Como Kish (1975) señala, los tests de significación estadística son particularmente inefectivos tal como, por lo general, son aplicados en la investigación social, es decir, como comprobación denunciadas en terminos de diferencias cero o relaciones nulas.

Virtualmente la hipótesis de nulidad, (entendida como la acabamos de definir o como Cohen (1994) la llamaba "nil hypothesis") siempre es falsa (con un tamaño de muestra suficiente, cualquier diferencia por nimia y trivial que sea puede generar diferencias estadísticamente significativas),

de modo que su rechazo aporta escasa información y si acaso sólo nos informa que el diseño de investigación ha tenido suficiente potencia estadística para detectar un efecto que puede ser grande o pequeño, útil o inútil. Un resultado improbable, (conclusión obtenida), no es necesariamente un resultado importante (conclusión deseada). Cuando concluimos que un resultado es significativo queremos decir que lo es desde un punto de vista estadístico pero puede que substantivamente no sea relevante ni meritorio.

El nivel inflexible de error de Tipo I, fijado a priori, puede generar incertidumbre en la interpretación de los resultados

Cuando el investigador fija a priori el riesgo que asume de error de Tipo I (normalmente en 0.05) establece un valor dicotómico de decisión estadística –estadísticamente significativo/estadísticamente no significativo– que en ocasiones conduce a problemas de incertidumbre en la interpretación de los resultados. Recordemos aquí la alta sensibilidad del valor de *p* respecto al tamaño muestral y por lo tanto no puede ser considerado como una propiedad cuantitativa intrínseca del fenómeno estudiado (Abelson, 1997a) dada su dependencia directa con el número de observaciones incluidas en el diseño de investigación. Por ejemplo puede ocurrir que dos investigaciones que obtienen el mismo efecto del tratamiento interpreten los resultados de forma opuesta. En uno de los estudios se obtiene un valor de $p = 0.049$ y por lo tanto las diferencias encontradas se interpretan como estadísticamente significativas mientras que en el segundo estudio las diferencias no se consideran significativas porque el valor de probabilidad es de 0.051 cuando realmente las diferencias estaban en el mayor tamaño de la muestra de la primera investigación. Como Rosnow y Rosenthal (1989) señalan "*seguramente, Dios ama al 0.06 tanto como al 0.05*" (página 1277).

Recuerde el lector, a modo de conclusión, que el test de significación estadística y el valor *p* asociado al mismo *no informa*:

1. De la probabilidad de la hipótesis nula.
2. De la probabilidad de la hipótesis alternativa.
3. De la posibilidad de replicar los datos en investigaciones posteriores.
4. El rechazo de la hipótesis nula no añade valor informativo.

La prueba de significación de la hipótesis nula. alternativas de análisis

Pero, aunque las críticas sean sensatas y razonadas, ni son toda la realidad, ni han deslegitimado el proceso estándar de inferencia estadística. En los últimos años también se han levantado voces a favor del uso

de la significación estadística y del contraste de hipótesis (Abelson, 1997a, 1997b; Cortina y Dunlap, 1997; Fritz, 1995, 1996; Greenwald, Gonzalez, Harris y Guthrie, 1996; Hagen, 1997; Harris, 1993; Levin, 1993) y se han propuesto vias alternativas de análisis que amplien y mejoren la información aportada por pruebas estándar.

Destacan la estimación de los intervalos de confianza (Bakan, 1966; Cohen 1990; Loftus, 1991, 1993a; Hunter, 1997; Thompson, 1997) y del tamaño del efecto (Cohen 1990; Schmidt 1996), análisis bayesianos (Lindley, 1965), estudios de reclinación y meta-análisis (Greenwald y cols., 1996, Pollard y Richardson, 1987; Schmidt 1992) y planteamientos de 'non nil hypothesis' en términos de Cohen, o de 'good enough hypothesis' siguiendo los trabajos de Serlin y Lapsley, (1985, 1993) y Rouanet (1996). Desde esta perspectiva recientemente Murphy y Myors (1999) presentan una alternativa de análisis con hipótesis de efectos mínimos. También se ha incrementado la consideración de la validez de conclusión estadística de la investigación, destacando la importancia de la potencia estadística (Cohen, 1988; Lipsey y Wilson, 1993; Schmidt, 1992).

La información que la técnica del intervalo de confianza proporciona al investigador es semejante a la ofrecida por la prueba de contraste de hipótesis sólo que define la amplitud de valores que indican entre qué cantidades se encuentran las diferencias estadísticamente significativas. No informa sobre la probabilidad de los datos dada una hipótesis sino que de acuerdo con una distribución de probabilidad comprueba la confianza de que el verdadero parámetro poblacional se encuentre comprendido dentro de una amplitud de estimaciones (Valera y Sánchez, 1997) y si no se incluye el cero, la hipótesis de no diferencia es rechazada.

Sin embargo, el intervalo de confianza tampoco soluciona todas las críticas anteriores ya que como Fritz (1996) indica, no superan la lógica de las pruebas de la hipótesis nula. En última instancia, la prueba de la hipótesis nula y el cálculo de los intervalos de confianza tienen los mismos fundamentos y son estadísticamente indistinguibles (p. e.: Chow, 1996, página 21). "*Los intervalos de confianza y el contraste de hipótesis nula están basados sobre la misma información. Por un ejemplo, un intervalo del 95% acerca de la diferencia entre dos medias y un test de significación al 0.05, ambos utilizan la diferencia entre las medias muestrales, el valor t correspondiente a la distribución según los grados de libertad, la varianza muestral y el tamaño de la muestra. Los dos métodos simplemente presentan la información de distinta manera. El resultado final es un énfasis sobre la estimación de parámetros o un énfasis sobre el error muestral, siendo así que en cada caso se dispone de unas ventajas y de unos inconvenientes*" (Cortina y Dunlap, 1997, página 170).

Otra alternativa viable al procedimiento clásico defiende la necesidad de formular la hipótesis nula no en términos de valor cero si no en valores distintos de cero, por mínimos que sean. Anteriormente comentamos que en la mayoría de los casos la hipótesis nula se formula como hipótesis *nil* o vacía, o hipótesis de efecto cero. Sin embargo, la formulación de la hipótesis nula no tiene por qué ser expresada necesariamente en esos términos.

Nos podemos hacer dos preguntas a la hora de formular la hipótesis de investigación: ¿El tratamiento tiene efectos? (La hipótesis nula se formularía en términos de no-efectos, valor 0) ¿El tratamiento tiene un efecto de una cuantía determinada o por encima de un valor determinado? (La hipótesis nula no se formula necesariamente en términos de valor 0). En este ultimo caso nos preguntamos si el tratamiento tiene efectos suficientemente grandes como para ser estadísticamente significativos (para la vida real, de acuerdo con un criterio teórico...) En realidad, nos estamos refiriendo a la posibilidad de aplicar un test de *efectos mínimos*. Pero, ¿es esto posible?

La hipótesis a testar en este caso es que el tratamiento o la intervención tiene un efecto igual a o menor que un valor mínimo fijado a priori (de tamaño x). Si, por ejemplo, el investigador decide que el tratamiento que explica el 1% de la varianza tiene un efecto excesivamente pequeño como para que sea útil o practico su uso, entonces su tarea consistirá en desarrollar un test estadístico que determine cuán grande ha de ser un tamaño del efecto en una muestra particular para que explique el 1% o más de la varianza. Ese test será una razón *F*, determinando el valor necesario para rechazar la hipótesis que el efecto en la población es igual o menor que el fijado.

Recordemos como se procede a la hora de comparar la *F* empírica con la teórica. Las tablas de la *F* que aparecen en los manuales están basadas sobre la distribución centrada de la *F*, es decir, la distribución de la *F* que esperaríamos en caso de efectos cero en la población. El test de efectos mínimos se basa en la distribución *F* no centrada. La forma de la distribución esta determinada por sus grados de libertad y por el parámetro lambda que es esencialmente dependiente del tamaño del efecto y del tamaño de la muestra. Una buena estimación en el modelo lineal general de la lambda viene dado por la razón N2xVE/(1-VE) donde N2 se refiere a los grados de libertad y VE es la varianza explicada de la variable dependiente. Cuanto más grande sea el valor lambda, permaneciendo constantes el resto de elementos, mayor será la *F* necesaria para rechazar la hipótesis nula.

Para desarrollar las tablas con las que testar la hipótesis nula de

efectos mínimos, el investigador deberá a) desarrollar la definición operacional de tamaño mínimo; b) calcular el tamaño de no centralidad para tal efecto mínimo y c) tabular la correspondiente distribución de *F* no centrada. En este planteamiento se nos presentan dos problemas. Primero cómo definir de manera sensata el "efecto mínimo". Segundo, cómo generar las tablas de F no centrada.

La definición de efecto mínimo requiere un juicio de valor que escapa al análisis y es previo. Es un problema dependiente de la teoría, de la praxis o del consenso. Dependerá del área a investigar y de otras concomitancias. También dependerá de criterios estadísticos porque conforme se defina un tamaño del efecto mínimo mayor, la potencia del modelo general lineal para detectar tal efecto decrece, es decir aumenta por tanto la probabilidad de cometer error Tipo II. Antes de llevar a cabo la investigación habrá que delimitar cual es la relación de riesgo que se acepta respecto de los errores de Tipo I y II. La relación correcta se puede suponer que es de uno a cuatro como afirma Cohen (1988) o se puede suponer mayor o menor según las circunstancias concretas.

Una de estas circunstancias puede venir determinada por el área de investigación. Por ejemplo, los test de habilidades cognitivas, explican el 20 o el 25% de la varianza del criterio, sea éste por ejemplo, el éxito académico o el rendimiento escolar. Formular la hipótesis nula en términos de 0 es fácil de rechazar si el efecto verdadero es de 20%. Bastará una muestra muy pequeña de sujetos. Si construimos un nuevo test, éste debería explicar como mínimo el 10% de la varianza. Incluso podría definir la hipótesis nula en términos de como mínimo lo que explican los otros ya existentes en el mercado (20%) y como alternativa, más que los otros tests. Este planteamiento tiene una ventaja, nos obliga a la revisión de las investigaciones anteriores para fijar un tamaño de efectos mínimos, a partir de los cuales generamos las hipótesis nula y alternativa.

Otras veces, la determinación del tamaño de efecto mínimo puede que exija un análisis de utilidad o de costos y de relevancia social o clínica o cualquier otro criterio que se estime oportuno. Siempre y en cualquier caso, habrá que justificarlo y definirlo y no dar por supuesto como un "prejuicio" dogmáticamente establecido que ese valor ha de ser siempre el valor cero.

Existen aproximaciones estadísticas para generar tablas como las de Tiku y Yip (1978), o programas informáticos para computar las probabilidades como los de Narula y Weistroffer (1986) a los que nos remitimos. Recientemente Murphy y Myors (1999) han desarrollado unas tablas de carácter general para aplicar los tests de efectos mínimos al Modelo Lineal General.

La alternativa conocida como "test de hipótesis de efectos mínimos" parece ser una vía interesante y de futuro para mejorar muchas de las investigaciones que se realizan tanto en el área de la psicología aplicada como de la ciencia básica; en ambas situaciones la reformulación de la hipótesis nula en los términos descritos puede generar una investigación más precisa que sirva a los intereses de predecir con mayor exactitud las derivaciones teóricas y a la mejor elección de los tratamientos psicológicos más eficientes. Es una buena manera de aplicar la sugerencia recogida en la última edición del Manual del A.P.A. (1994) que obliga a "*informar e interpretar el tamaño del efecto poniéndolo en relación con los resultados previos ya conocidos*".

CONCLUSIONES

Actualmente no existe un único método mejor para analizar los datos de la investigación. El procedimiento conocido como "comprobación de la significación estadística de la hipótesis nula" es uno más dentro de un arsenal de técnicas que deben ser combinadas para ayudar a lo que de verdad interesa al científico y es el objetivo final de todo proceso investigador: la comprobación de teorías donde la hipótesis sustantiva ha de guiar la elección del procedimiento más válido de análisis. La ubicación de NHST dentro del proceso del diseño de la investigación necesita cuanto menos de nuevos ajustes y complementos que aporten información adicional como el tamaño del efecto o tengan en cuenta otros aspectos de la comprobación estadística como la potencia de la prueba estadística a efectos de proporcionar una mayor comprensión del resultado empírico obtenido y haciendo más fácil, por ejemplo, los procesos de integración posterior de la información vía métodos estadísticos, por ejemplo, el meta-análisis. En un futuro próximo plantear las hipótesis nulas con un mayor nivel de precisión en términos de efectos mínimos también ayudará a valorar de manera más óptima la eficacia de los tratamientos psicológicos dentro de la Psicología Aplicada y a afinar mejor nuestras hipótesis experimentales derivadas de la teoría en aras de un conocimiento más fino y acumulativo.

BIBLIOGRAFÍA

Abelson, R. P. (1997a). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). En L. L. Harlow, S. A. Mulaik y J. H. Steiger (Eds.), *What if there were no*

significance tests? Mahwah, NJ: Lawrence Erlbaum Associates.
Abelson, R. P. (1997b). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.
American Psychological Association (A.P.A.) (1994) *Publications manual of the American Psychological Association* (4th ed.). Washington, DC: Author.
Bakan, D. (1966). The effect of significance testing in psychological research. *Psychological Bulletin*, *66*, 423-437.
Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.
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.
Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.
Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *70*, 107-115.
Carlton y Strawdwemann 1996
Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.
Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.
Chow, S. L. (1996). *Statistical significance. Rationale, validity and utility*. London, UK: Sage Publications.
Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
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.
Cohen, J. (1997, August). *Much ado about nothing*. Lecture presented at the annual meeting of the American Psychological Association, Chicago.
Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.
Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116-127. (Traducido en la revista *Escritos de Psicología*, I, 1997).
Dixon, P. (1998). Why scientists value p values. *Psychonomic Bulletin Research*, *5*, 390-396.
Dixon, P. & O'Reilly, T. (1999). Scientific versus statistical inference.

Canadian Journal of Experimental Psychology, 53, 133-149.
Ellis, N. (2000). Editorial. *Language Learning, 50*, (3).
Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology, 2*, 75-98.
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburg: Oliver and Boyd. (Reeditada en 1990 por la editorial Oxford University Press).
Fisher, R. A., & MacKenzie, W. A. (1923). Studies in crop variation: 2. The manurial response of different potato varieties. *Journal of Agricultura Science*, 13, 311-320.
Folger, R. (1989). Significance tests and the duplicity of binary decisions. *Psychological Bulletin, 106*, 155-160.
Fritz, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23*, 132-138.
Fritz, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.
Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*, 54-61.
Greenwald, A. G. (1993). Consequences of prejudice against the null hypothesis. En G. Keren y C. Lewis, (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect size and *p*-values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175-183.
Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.
Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significances tests?* Mahwah, NJ: Lawrence Erlbaum Associates.
Harris, R. J. (1993). Multivariate analysis of variance. En L. K. Edwards (Ed), *Applied analysis of variance in behavioral science*. New York, NY: Marcel Dekker.
Heldref Foundation (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.
Hresko, W. (2000). Editorial policy. *Journal of Learning Disabilities, 33*, 214-215.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, *61*, 317-333.

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

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego, CA: Harcourt, Brace and Jovanovich.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *65*, 3-5.

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

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

Kish, L. (1975). Representation, randomization and control. En H. M. Blalock et al. (Eds.), *Quantitative Sociology*. New York, NY: Academic Press. (Traducción recogida en 1980 en F. Alvira, M. D. Avia, R. Calvo y J. F. Morales, *Los dos métodos de las Ciencias Sociales*. Madrid: Centro de Investigaciones Sociológicas)

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

Lindley, D. V. (1965). *Introduction to probability and statistics from bayesian viewpoint. Part 2: Inference*. Cambridge, Mass.: Cambridge University Press.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational and behavioral treatment. *American Psychologist*, *12*, 1181-1209.

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

Loftus, G. R. (1993a). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments and Computers*, *25*, 250-256.

Loftus, G. R. (1993b). Editorial comment. *Memory & Cognition*, *21*, 1-3.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 131-139.

Manzano, V. (1997). Usos y abusos del error de Tipo I. *Psicológica. Revista de Metodología*. *18*, 153-169.

McLean, J. E., & Kaufman, A. S. (2000). Editorial: Statistical significance testing and Research in the Schools. *Research in the Schools*, *7*, (2).

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

Meehl, P. E. (1978). Theoretical risk and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting*

and Clinical Psychology, *46*, 806-834.
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195-244.
Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy: a reader*. Chicago: Aldine.
Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.
Murphy, K. R. & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimun-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234-2484.
Narula, S. & Weistroffer, H. (1996). Computation of probability and non centrality parameter of a noncentral F distribution. *Communications in Statistics*, *B20*, 871-878.
Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641-650.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: John Wiley & Sons.
Pollard, P. (1993). How significant is 'significance'. En G. Keren y C. Lewis, (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.
Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*, 159-163.
Rosenthal, R. (1990a). How are we doing in soft psychology. *American Psychologist*, *45*, 775-776.
Rosenthal, R. (1990b). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*, 1-30.
Rosenthal, R., & Rubin, D. B. (1985). Statistical analysis: summarizing evidence versus establishing facts. *Psychological Bulletin*, *97*, 527-529.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284.
Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, 119, 149-158
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.
Schmidt, F. L. (1992). What do data really mean. Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.
Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.
Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research:

the good-enough principle. *American Psychologist*, *40*, 73-83.
Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good enough principle. En G. Keren y C. Lewis, (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.
Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research*, *40*, 371-393.
Signorelli, A. (1974). Statistics: Tool of master of the psychologist? *American Psychologist*, *11*, 221-223.
Thompson, B. (1994a). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.
Thompson, B. (1997, August). *If statistical significance tests are broken/ misused, what practices should supplement or replace them?*. Address presented at the annual meeting of the American Psychological Association, Chicago.
Tiku, M. L., & Yip, D. Y. N. (1978). A four-moment aproximation based on the F distribution. *Australian Journal of Statistics*, *20*, 257-261.
Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *1*, 100-116.
Valera, A., & Sánchez, J. (1997). Pruebas de significación y magnitud del efecto: reflexiones y propuestas. *Anales de Psicología*, *13*, 85-90.