anales de psicología
1997, vol . 13, nº 1, 85-90

Murcia (España). ISSN: 0212-9728

Pruebas de significación y magnitud del efecto: Reflexiones y propuestas

Antonio Valera Espín y Julio Sánchez Meca*

Universidad de Murcia

Resumen: Se describen diversas propuestas que permiten complementar la información que aporta un contraste de hipótesis estadísticas. El uso de estas propuestas reduce en buena medida una de las críticas más duras que las pruebas de significación han recibido: Las pruebas de significación no ofrecen información sobre la magnitud de la relación entre las variables implicadas. Las propuestas que se discuten en este trabajo son: los intervalos de confianza, el tamaño del efecto, la presentación binomial del tamaño del efecto, el valor contranulo y el indicador en lenguaje común del tamaño del efecto.
Palabras clave: Pruebas de significación, contraste de hipótesis estadísticas, intervalos de confianza, tamaño del efecto.

Title: Significance tests and effect magnitude: Reflections and proposals.
Abstract: Several proposals that enable to complement the information offered in statistical hypothesis testing are described. Using these proposals reduce the most hard critic that significance tests have suffered: Significance tests do not offer information about the magnitude of the relationship among the involved variables. The proposals that are discussed in this paper are: Confidence intervals, effect size, binomial effect size display, counternull value and common language effect size indicator.
Key-words: Significance tests, statistical hypothesis testing, confidence intervals, effect size.

1. **Introducción**

Desde que hicieron su aparición, las pruebas de significación han estado marcadas por las diferencias de criterio de los autores que se han ocupado de su desarrollo teórico. Hoy día continúan siendo técnicas controvertidas, en absoluto exentas de modificaciones, críticas, ampliaciones o propuestas alternativas. La polémica histórica la protagonizó ya quien puede considerarse el pionero, Fisher, que se enfrentó teóricamente con dos autores que intentaron mejorar su trabajo, Jerzy Neyman y Egon Pearson. La principal diferencia entre estas dos teorías se debe a que en el modelo fisheriano no se acepta la idea de dos hipótesis estadísticas contrapuestas. A pesar de la controversia, la teoría que se ha impuesto ha sido una solución mixta entre las dos teorías (Gigerenzer, 1993).

La teoría híbrida que combina los dos enfoques es en buena parte responsable de muchas de las críticas que estos procedimientos han recibido. En 1970,

Morrison y Henkel editaron *The significance test controveny,* que incluía contribuciones críticas a las pruebas de significación de la hipótesis nula de, entre otros, Bakan (1966), Lykken (1968), Meehl (1967) y Rozeboom (1960). Más recientemente, metodólogos tan eminentes como Carver (1993), Cohen (1990,1994), Gigerenzer (1993), Kirk (1996), Loftus (1993a, 1993b, 1993c), Meehl (1978, 1986, 1990a, 1990b), Oakes (1986), Rosenthal (1990), Rosnow y Rosenthal (1989), Rosnow y Rosenthal (1989), Schmidt (1996), también cuestionan los planteamientos teóricos y la aplicación de las pruebas de significación de la hipótesis nula en las ciencias del comportamiento. Las críticas, tan numerosas que sería difícil considerarlas exhaustivamente, fluctúan desde las que sólo se lamentan del uso inapropiado de unas técnicas potentes y válidas, a las que consideran tales técnicas como poco interesantes e incluso erróneas.

De todas las numerosas críticas al contraste de hipótesis estadísticas, la más destacable es que los investigadores al utilizarlas no suelen tener en cuenta la potencia estadística en la planificación de sus investigaciones. Numerosos estudios de potencia ya han demostrado el absoluto olvido de la potencia en las aplicaciones de las pruebas de significación (Acklin, McDowell y Omdoff, 1992; Brown y Hale, 1992; Clark-Carter, 1997; Cohen, 1962; Chase y Chase, 1976; Frías, García *y* Pascual, 1994; Kazdin *y* Bass, 1989; Lipsey, Crosse, Dunkle, Pollard y Stobart, 1985; Pascual, Frías y García, 1994; Rossi, 1990; Sánchez, Valera, Velandrino y Marín, 1992; Sedlmeier y Gigerenzer, 1989; Valera, 1998; Valera, Sánchez, Marín y Velandrino, en prensa).

Pero al aplicar un contraste de hipótesis estadísticas, además de considerar el tamaño muestral necesario que permita garantizar una potencia adecuada en la planificación, es importante que se utilicen otros procedimientos estadísticos complementarios que ayudan a extraer el sentido de los datos y que han sido propuestos recientemente por diversos autores.

En este trabajo se describen las principales alternativas propuestas para mejorar o complementar la información que aporta la aplicación de una prueba de signi-

(*)**Dirección para correspondencia**: Julio Sánchez Meca. Departamento de Psicología Básica y Metodología. Universidad de Murcia, Campus de Espinardo (edificio "Luis Vives"). Aptdo. 4021 30080 Murcia (España). E-mail: jsmeca@um.es

ficación. No se tratan aquí aquellas alternativas que no siguen el modelo de la inferencia estadística clásica (métodos basados en el remuestreo y las pruebas de permutación, la perspectiva bayesiana, el modelado estadístico, etc.). En concreto, se explica un procedimiento basado en la misma lógica que el contraste de hipótesis, el intervalo de confianza, y un estadístico que ha demostrado su utilidad para estimar la magnitud de la relación existente entre las variables implicadas, el tamaño del efecto; en cuanto a la presentación binomial del tamaño del efecto, el indicador en lenguaje común del tamaño del efecto y el valor contranulo, son tres estadísticos que nos permiten enriquecer la información que aporta el propio tamaño del efecto.

2. Intervalos de Confianza

En general, se asegura que calcular intervalos de confianza en tomo a las estimaciones es un útil complemento, o incluso un buen sustituto, a las pruebas de significación (Bakan, 1966; Cohen, 1990; Loftus, 1991, 1993a, 1993b, 1993c, 1995, 1996; Loftus y Masson, 1994).

La obtención de intervalos de confianza en el análisis de los datos de las investigaciones psicológicas puede considerarse una buena solución al conjunto de críticas que se achacan a las pruebas de significación de la hipótesis nula. El principal argumento a favor de que las pruebas de significación deban sustituirse por intervalos confidenciales sostiene que, mientras las pruebas de hipótesis sólo responden a la cuestión de si unos estadísticos que representan parámetros difieren altamente, los intervalos confidenciales, además de esta información, estiman los parámetros.

Los intervalos de confianza no suponen un problema de interpretación lógica. Un intervalo de confianza no dice nada de la probabilidad de los datos dada una hipótesis (como sucede en el contraste de hipótesis). Simplemente se comprueba la confianza de acuerdo con una distribución de probabilidad de que el verdadero valor poblacional se encuentre comprendido dentro de un rango de estimaciones. Es un procedimiento altamente comprensivo, incluso para inexpertos en análisis de datos. La simplicidad interpretativa caracteriza a esta técnica; no caben equívocos entre diferentes tipos de error, ni confusión entre probabilidades. La aplicación es también fácil; para calcular un intervalo de confianza sólo necesitamos la estimación, la puntuación en la distribución de probabilidad que le corresponde al nivel de confianza deseado y el error típico del estadístico. Por ejemplo, la fórmula a aplicar para calcular un intervalo de confianza en tomo a la diferencia de medias independientes Δ es, suponiendo desconocidas las varianzas poblacionales:

$$p(d - \left|_{a/2t_v}\right|\hat{\sigma}_d \leq \Delta \leq d + \left|_{a/2t_v}\right|\hat{\sigma}_d) = 1 - \alpha$$

siendo $\Delta = \mu_1 - \mu_2$ el parámetro a estimar, $d = \hat{Y}_1 - \hat{Y}_2$ el estimador $a/2t_v$, la puntuación de la distribución que le corresponde al nivel de confianza ($a/2$) especificado para un contraste bilateral, $\hat{\sigma}_d$ el error típico de la diferencia de medias y los grados de libertad $v = n_1 + n_2 - 2$.

Por último, los intervalos de confianza ofrecen más información que las pruebas de significación. Por una parte, cuando un intervalo confidencial para una diferencia no incluye el cero, la hipótesis de no diferencia es rechazada. Llegamos, por lo tanto, con este procedimiento a la misma conclusión que con una prueba clásica de significación; pero además, ofrece más información indicando la dirección y la magnitud de la diferencia.

Un ejemplo ayudará a comprender la propuesta de los intervalos de confianza. Supóngase que después de aplicar un tratamiento la media del grupo experimental es 8.33, mientras que la del grupo control es 6.7. El número de sujetos para cada grupo es 30, y las desviaciones típicas son 3.845 para el grupo experimental y 3.007 para el grupo control. Si el investigador aplica un contraste de hipótesis sirviéndose de una prueba *T* para muestras independientes los resultados no resultan significativos *(T(58)=-1.833; p=.072).* El intervalo de confianza nos indica, además de permitimos tomar una decisión similar al contraste puesto que el cero está incluido en el intervalo, la precisión de la estimación: *p(-0.149 ≤ Δ ≤ 3.415)=1-α.*

Frick (1996), por su parte, considera que el uso de intervalos confidenciales no soluciona los problemas que se le achacan a las pruebas de significación. Asegura que estas funciones no evitan en absoluto la lógica de las pruebas de la hipótesis nula (por el contrario, se basan en ella) y tacha de ilógicas las ideas de los autores que critican estas pruebas y, al mismo tiempo, defienden el uso de los intervalos de confianza. Cortina y Dunlap (1997) opinan que los intervalos de confianza también resultan ser procedimientos imperfectos: Sólo si el valor de *a* es 0 es posible un intervalo con un nivel de confianza del 100%, y en tales casos el rango iría de menos infinito a más infinito (o de -1 a 1 en correlaciones), por lo que la información no sería útil.

Entre sus defensores, destaca Loftus (1991, 1993a, 1993b, 1993c, 1995), que apuesta por técnicas gráficas que presenten las medias muestrales indicando sus intervalos de confianza. De esta forma, tendremos información de la mejor estimación del patrón subyacente a las medias de la población y, al mismo tiempo, el grado en que podemos tomar en serio el patrón de medias obtenido al considerar los errores de muestreo.

En relación con los gráficos que reflejan los errores típicos, Dunlap y May (1989) muestran cómo pueden ser un sustituto de las pruebas de contraste, puesto que permiten inferir la significación. Si los errores típicos respectivos de dos medias representadas gráficamente se tocan o se solapan, significa que las dos medias no difieren significativamente. Si, por el contrario, las medias difieren tres o más veces la longitud del error típico, entonces difieren significativamente por lo menos a un nivel de a =.05

3. **Tamaños del efecto**

La principal propuesta como complemento a las pruebas de significación consiste en aportar tamaños del efecto como resultados de los análisis en una investigación. Como a los intervalos de confianza, al tamaño del efecto se le otorga más valor informativo que a los contrastes de hipótesis y, sin duda, constituye un buen complemento de los mismos. Así, el principio de bondad-suficiente de Serlin y Lapsley (1985) integra el tamaño del efecto en la metodología de las pruebas de significación.

La utilidad de los tamaños del efecto ha quedado demostrada gracias a la aparición del meta-análisis. De hecho, el desarrollo de este estadístico podemos atribuírselo sobre todo a los teóricos y aplicadores de esta técnica, el meta-análisis, muy útil a la hora de resumir, estructurar o acumular el conocimiento científico obtenido en investigaciones empíricas que emplean metodología estadística (Cooper y Hedges, 1994; Marín, 1996; Rosenthal, 1991; Sánchez y Ato, 1989; Sánchez, Marín y Valera, 1992).

Un tamaño del efecto responde a preguntas tales como: ¿Cuál ha sido la magnitud del efecto de un tratamiento?; ¿en qué cantidad se manifiesta una diferencia entre estadísticos?; ¿cómo de fuerte es una relación entre variables? Cohen (1988) define este estadístico como el grado en que el fenómeno está en la población o, en el contexto de una prueba de significación, el grado en que la hipótesis nula es falsa.

Al tratarse de una estimación del tamaño del efecto en la población, puede calcularse un intervalo de confianza en tomo suyo. De este modo, todo lo dicho anteriormente para los intervalos confidenciales es directamente aplicable a los tamaños del efecto. El intervalo confidencial asociado a un tamaño del efecto nos indica el rango dentro del cual es probable que se encuentre el efecto real en la población.

Una ventaja importante de los tamaños del efecto es que, al ser transformaciones a una escala común, los resultados de diferentes estudios o experimentos son directamente comparables. Es precisamente esta característica lo que les hace imprescindibles al realizar un estudio meta-analítico. Aunque para cada prueba estadística existe un tamaño del efecto diferente (e incluso varios modos de calcular la magnitud de un efecto para una misma prueba), es posible la transformación entre los diferentes tipos. Así, es común en los estudios meta-analíticos que todos los tamaños del efecto diferentes *(d* de Cohen, *f, g, h, q, W,* etc.) se conviertan, por ejemplo a coeficientes de correlación producto-momento (Rosenthal, 1991).

Obviamente, no existe contradicción entre pruebas de significación y tamaños del efecto. Si se desea tener en cuenta la potencia estadística al aplicar una prueba de significación, un paso fundamental es proponer un tamaño del efecto. Tanto en los cálculos de potencia a priori como en los cálculos a posteriori, se requiere de una estimación de la magnitud del efecto. Pero el principal problema al planificar una prueba de significación está en determinar cuánto vale el tamaño del efecto que se pretende investigar. Lipsey (1990) denomina al tamaño del efecto el parámetro problemático, precisamente por la dificultad que supone establecerlo a priori. Sin embargo, este autor ofrece una solución que día a día va resultando más factible: Utilizar los resultados de los meta-análisis ya realizados. Vista la solución que aporta, resultaría conveniente que al realizar un estudio empírico la revisión subjetiva sobre el tema se completara con un meta-análisis (a no ser que ya esté hecho o que el estudio sea el primero sobre el tópico). Lipsey propone que para determinar un tamaño del efecto en la planificación de una investigación es muy útil considerar la distribución de los tamaños del efecto de las investigaciones anteriores similares. Sugiere un procedimiento para sustituir la convención de tamaños del efecto propuesta por Cohen (1988). Si existe un meta-análisis sobre el tema puede utilizarse como tamaño del efecto pequeño la media o mediana de la distribución de puntuaciones de tamaños del efecto que quedan por debajo del percentil 33; como tamaño del efecto medio la media o mediana del 34% central de las puntuaciones de tamaños del efecto; y como tamaño del efecto alto la media o mediana de las magnitudes del efecto que estén por encima del percentil 67.

En el ejemplo que propusimos para describir los intervalos de confianza, el tamaño del efecto que resulta es: $d = (\hat{Y}_1 - \hat{Y}_2)/s$ =(8.333-6.7)/3.43=0.47, un valor cercano a lo que Cohen considera como medio en su clasificación, lo que demuestra que el investigador ha obtenido un efecto a pesar de que con la aplicación del contraste de hipótesis no alcanzó la significación estadística.

4. **El valor contranulo**

Rosenthal y Rubin (1994; véase también Hallahan y Rosenthal, 1996; Rosnow y Rosenthal, 1996) han propuesto un nuevo estadístico que demuestra la relatividad de los resultados de una prueba de significación.

Lo denominan el valor contranulo *(counternull value)* de un tamaño del efecto obtenido y lo definen como la magnitud no nula del tamaño del efecto que está apoyada por exactamente la misma cantidad de evidencia que un tamaño del efecto nulo. Para comprenderlo más claramente consideremos el valor contranulo como la hipótesis nula (H_o) y su valor de probabilidad resultante será el mismo que el obtenido para la H_o real, es decir, si la H_o *a* comprobar fuese el valor contranulo, el valor de probabilidad que obtendríamos con nuestros datos sería el mismo que se ha hallado para la prueba con la H_o empleada realmente.

La fórmula del valor contranulo es: $TE_{contranulo} = 2TE_{obtenido} - TE_{nulo}$, siendo $TE_{contranulo}$ el tamaño del efecto cotranulo, $TE_{obtenido}$ el tamaño del efecto obtenido con los datos empíricos y TE_{nulo} el tamaño del efecto que se plantea habitualmente en la hipótesis nula. Si el tamaño del efecto nulo es igual a cero, como ocurre en muchas ocasiones, el valor contranulo será igual a dos veces el tamaño del efecto obtenido. Cuando se trabaja con coeficientes de correlación, la fórmula del valor contranulo es: $r_{contranulo} = \sqrt{(4r^2)/(1+3r^2)}$ (Rosnow y Rosenthal, 1996).

La utilidad del valor contranulo está en que ayuda a evitar pensar que, ya que la H_o no fue rechazada, la mejor estimación debe ser el valor nulo (generalmente cero). Este nuevo estadístico nos dice que la evidencia en favor de la conclusión de que la mejor estimación del tamaño del efecto es el valor nulo es exactamente la misma que para la conclusión de que el tamaño del efecto es el doble del obtenido.

El ejemplo propuesto anteriormente puede ayudar a comprender mejor este estadístico. El resultado del contraste de hipótesis obligaba a aceptar la hipótesis nula, sin embargo el tamaño del efecto contranulo es $TE_{contranulo} = 2TE_{obtenido} - TE_{nulo} = 2(0.47) - 0 = 0.94$, lo que indica que existe evidencia para pensar que una buena estimación del tamaño del efecto puede ser hasta del doble del obtenido lo que probablemente permitiría rechazar la H_o puesto que alcanza un valor muy alto.

5. **Presentación binomial del tamaño del efecto**

Kirk (1996) entiende que, en ocasiones, un resultado que en un contraste de hipótesis ha sido "no significativo" puede tener, sin embargo, una significación práctica. Una estimación puntual o un intervalo de confianza (la consideración de la magnitud del efecto, en definitiva) pueden usarse para decidir si los resultados son realmente pobres o, por el contrario, útiles e importantes. Considera obligatorio que los investigadores emitan un juicio sobre la utilidad del resultado obtenido, independientemente de si la magnitud del efecto pueda considerarse baja, moderada o alta de acuerdo con la convención estadística al respecto. Para que la significación práctica sea un concepto útil, su determinación no debe estar ritualizada por arbitrariedades metodológicas: No rechazar una H_o no implica necesariamente que no exista efecto y, por otro lado, un resultado estadísticamente significativo en base a un valor de probabilidad tampoco implica que la magnitud del efecto tenga una importancia práctica. Considerando un caso extremo, es evidente que la magnitud del efecto de un tratamiento que convencionalmente pueda considerarse baja, puede resultar muy importante si permite salvar algunas vidas (no importa que sean pocas).

Rosnow y Rosenthal (1996) y Rosenthal y Rubin (1982) han confeccionado un procedimiento que puede ayudar a tomar una decisión al investigador sobre la utilidad práctica de un tratamiento. Esta forma de considerar la magnitud de un efecto la denominan *`BESD" (Binomial Effect SiZe Display) y* consiste en transponer la correlación entre las variables en juego a una tabla 2x2 que incluya las proporciones de éxito (y de fracaso) del tratamiento. En el ejemplo que venimos utilizando, es posible convertir el tamaño del efecto, d, a coeficiente de correlación de Pearson, *r y* a partir de este último índice confeccionar una tabla de proporciones de éxito y fracaso. Así, volviendo al ejemplo propuesto, puede convertirse el tamaño del efecto a correlación de Pearson mediante $r = d/\sqrt{d^2+4} = 0.47/\sqrt{0.47^2+4} = .229$ y confeccionar una tabla como la siguiente:

	% éxito	% fracaso
Tratamiento	61.4	38.6
Control	38.6	61.4

La conversión de *r* a *BESD* consiste en calcular la razón de éxito en la condición tratamiento de acuerdo con *.50 + r/2* (y multiplicar por cien para obtener porcentajes) y la razón de fracaso del tratamiento según .50 - *r/2*. La presentación resultante ofrece una información valiosa para la consideración de la utilidad práctica del tratamiento (en otros casos puede tener el sentido de proporciones de curación, proporciones de mejora, proporciones de supervivencia, etc.). En el ejemplo, puede interpretarse la tabla entendiendo que el grupo tratado consigue un 22.8% más de eficacia que el grupo control. Esta diferencia, dependiendo del contexto, puede resultar de gran importancia práctica o real.

6. **El índice universal del tamaño del efecto**

McGraw y Wong (1992) proponen otro estadístico que ayuda a la interpretación de los resultados cuando se

comparan medias, se trata de un índice de la magnitud del efecto que podemos considerar de aplicación universal, el índice *CL (Common Language Effect SiZe Indicator).* Consiste en calcular la probabilidad de obtener una diferencia entre puntuaciones mayor que cero en la distribución de las diferencias. Así, si se comparan dos medias, *CL* es la probabilidad de obtener una puntuación de diferencias entre ellas mayor que cero en una distribución normal cuya media es la diferencia entre las dos medias muestrales. Para su cálculo debe utilizarse la distribución normal tipificada buscando el área de probabilidad por encima del valor $z = \left[0 - (\hat{Y}_1 - \hat{Y}_2)\right] / \left(\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}\right)$, donde $\hat{Y}_1$ y $\hat{Y}_2$ son las medias de las dos muestras y $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ son las varianzas de los datos de cada muestra.

En el ejemplo que ya proponíamos anteriormente, la aplicación de la fórmula anterior nos ofrece como resultado un valor de z de -0.3345 que se corresponde con una probabilidad de obtener una diferencia entre puntuaciones mayor que cero de .6293.

7. **Conclusiones**

Todas estas alternativas basadas en tamaños del efecto e intervalos de confianza pretenden que los investigadores se centren más en la significación práctica de los resultados de sus estudios que en conseguir la anhelada significación estadística. De hecho, se ha comprobado en el ejemplo que, aunque no resultó estadísticamente significativo el resultado al aplicar el contraste de hipótesis, los demás procedimientos apuntan hacia la existencia de un efecto del tratamiento.

En cualquier caso, no debemos considerar las pruebas estadísticas como el instrumento único y perfecto para corroborar teorías psicológicas o evidenciar hechos científicos. Una prueba estadística debe entenderse como una humilde herramienta para ayudamos en la comprensión, organización o estructuración de los datos de la realidad. Como nos recuerdan Gigerenzer, Swijtink, Porter, Daston, Beatty, y Krüger (1989), grandes maestros de la Psicología, tales como Freud, Piaget o Skinner, entre otros, no se sirvieron de la estadística para construir sus teorías. Cohen (1965) escribe: "Statistical analysis is a tool, not a ritualistic religion". Es evidente que para obtener conclusiones absolutas debemos basamos en pruebas también absolutas.

De acuerdo con Frick (1995, 1996), el uso del contraste de hipótesis puede tener poco valor cuando se pretende obtener predicciones cuantitativas o cuando se quiere comprobar aplicaciones prácticas (en tales casos las estimaciones del tamaño del efecto pueden ofrecer información más valiosa). Sin embargo, la aplicación de las pruebas de significación puede resultar ideal cuando lo que se pretende es comprobar leyes en sentido ordinal o cualitativo, es decir, cuando el tamaño del efecto interviniente no es trascendental, puesto que la alta variabilidad existente entre las observaciones, por ejemplo, confiere una importancia relativa. Además, aunque no es un método apropiado para apoyar creencias, sí puede emplearse como procedimiento para establecer una evidencia suficiente.

En cualquier caso, se hace necesario evitar el uso indiscriminado de las pruebas de significación y, cuando éstas se utilicen, deberían complementarse acompañándolas con procedimientos estadísticos que informen acerca del grado, dirección e importancia real de los resultados obtenidos con las pruebas de significación. Estos procedimientos estadísticos deben pasar por la aplicación de índices del tamaño del efecto, ya que son capaces de proporcionar información acerca de la relevancia práctica, clínica o social de un resultado empírico en el ámbito de las ciencias sociales en general, y de las ciencias del comportamiento en particular.

Referencias

Acklin, M.W.; McDowell, C.J. II y Orndoff, S. (1992). Statistical power and the Rorschach: 1975-1991. *Journal of Personality Assessment, 59,* 366-379.

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

Brown, J. y Hale, M.S. (1992). The power of statistical studies in consultation-liaison psychiatry. *Pychosomatics, 33*, 437-443.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287-292.

Chase, L.J. y Chase, R.B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61,* 234-237.

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

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

Cohen, J. (1965). Some Statistical issues in psychological research. En B.B. Wolman (ed.), *Handbook of Clinical Psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.

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

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

Cooper, H.M. *y* Hedges, L.V. (1994). *The Handbook of Research Synthesis.* New York Sage.

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

Dunlap, W.P. y May, J.G. (1989). Judging statistical significance by inspection of standard error bars. *Bulletin of the Pychonomic Society, 27*, 67-68.

Frías, D.; García, J.F. y Pascual, J. (1994). Estudio de la Potencia de los Trabajos Publicados en "Psicológica". Estimación del Número de Sujetos Fijando Alfa y Beta. En C. Arce *y J.* Seoa-

ne (Coords.), *III Simposium de Metodología de las Ciencias Sociales y del Comportamiento (pp.* 1057-1063). Santiago de Compostela: Servicio de Publicaciones e Intercambio Científico de la Universidad.

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

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

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. En G. Keren y C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues (pp.* 311-339. Erlbaum.

Gigerenzer, G.; Swijtink, Z.; Porter, T.; Daston, L.; Beatty, J. y Kruger, L. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life.* Cambridge, England: Cambridge University Press.

Hallaban, M. y Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behavioral Research & Therapy, 34,* 489-499.

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

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

Lipsey, MOW. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. London: Sage.

Lipsey, M.W.; Crosse, S.; Dunkle, J.; Pollard, J. y Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. *New Directions for Program Evaluation, 27*, 7-28.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 1-5.

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, & Computers, 25,* 250-256.

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

Loftus, G.R. (1993c). *Why psychology will never be a real science until we change the way we analyze data.* Paper presented at the 102nd annual convention of the American Psychological Association. Los Angeles, CA, August.

Loftus, G.R. (1995). Data analysis as insight: Reply to Morrison and Weaver. *Behavior Research Methods, Instruments, & Computers, 27,* 57-59.

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

Loftus, G.R. y Masson, M.E. (1994). Using confidence intervals in within-subject designs. *Pychonomic Bulletin Review, 1*, 476-490.

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*,151-159.

Marín, F. (1996). *Enfoques meta-analíticos: Un estudio comparativo mediante simulación Monte Carlo.* Tesis Doctoral no publicada, Universidad de Murcia.

McGraw, K.O. y Wong, S.P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361-365.

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

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology, 46,* 806-834.

Meehl, P.E. (1986). What social scientists don't understand. En D. Fiske *y* R. Shweder (Eds.), *Metatheoy in Social Science: Pluralisms and Subjectivities* (pp. 315-329). Chicago: University of Chicago Press.

Meehl, P.E. (1990a). Appraising and amen theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Meehl, P.E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244.

Morrison, D.E. *y* Henkel, R.E. (1970). *The Significance Test Controversy.* London: Butterworths.

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* Chichester. Wiley.

Pascual, J.; Frías, M.D. y García, J.F. (1994). Análisis comparativo del tamaño del efecto y la potencia en función de la naturaleza del trabajo en la revista "Anuario de Psicología". En C. Arce *y J.* Seoane (Coords.), *III Simposium de Metodología de lar Ciencias Sociales y del Comportamiento* (pp. 1093-1101). Santiago de Compostela: Servicio de Publicaciones e Intercambio Científico de la Universidad.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior of Personality, 5,*1-30.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R. y Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

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

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

Rosnow, R.L. y Rosenthal, R. (1996). Computing contrasts, effect sizes, and countemulls on other people's published data: General procedures for research consumers. *Pychological Methods,* 1, (en prensa).

Rossi, J.S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting & Clinical Psychology, 5*, 646-656.

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

Sánchez, J. y Ato, M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación. En J. Arnau y H. Carpintero (Coords.), *Tratado de Psicología General,* Vol. I (pp. 617-669). Madrid: Alhambra.

Sánchez, J.; Marín, F. y Valera, A. (julio, 1992). *Averaging dependent effect sizes: A problem in meta-analysis* Poster presentado al XXV International Congress of Psychology, Bruselas.

Sánchez, J.; Valera, A.; Velandrino, A.P. y Marín, F. (1992). Un estudio de la potencia estadística en Anales de Psicología. *Anales de Psicología, 8,*19-32.

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

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

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

Valera, A. (1998). *Contraste de hipótesis en investigación psicológica en España: Un estudio de potencia.* Tesis Doctoral no publicada, Universidad de Murcia.

Valera, A.; Sánchez, J.; Marín, F. y Velandrino, A. (en prensa). Potencia estadística de la investigación publicada en la Revista de Psicología General y Aplicada. *Revista de Psicología General y Aplicada.*

Artículo recibido: 20-2-98, aceptado: 27-3-98