Psicothema
2000. Vol. 12, Supl. nº 2, pp. 549-552

ISSN 0214 - 9915 CODEN PSOTEG

Contraste de hipótesis e investigación psicológica española: análisis y propuestas

Antonio Valera Espín, Julio Sánchez Meca y Fulgencio Marín Martínez
Universidad de Murcia

El propósito de este trabajo es analizar la aplicación del procedimiento estadístico más común para estudiar relaciones entre variables y fenómenos empíricos en Psicología: La prueba del contraste de la hipótesis nula. Para comprobar si su uso es correcto en la investigación psicológica española se realizó un estudio de la potencia de las investigaciones publicadas en revistas españolas. El análisis de los 169 experimentos identificados, con un total de 5.480 pruebas de significación estadísticas, arrojó valores de potencia de 0.18, 0.58, 0.83 y 0.59, para tamaños del efecto bajo, medio, alto y estimado, respectivamente. Estos valores descendieron drásticamente en un 20% aproximadamente cuando se repitieron los cálculos controlando la inflacción de la tasa de error Tipo I mediante el ajuste de Bonferroni. Estos resultados, por lo demás similares a los encontrados en los estudios de potencia internacionales, deben hacer reflexionar a la comunidad científica sobre la necesidad de dedicar especial atención al control de la potencia cuando se diseña una investigación. Además, se discuten diversas propuestas complementarias al uso de las pruebas de significación para enriquecer la información aportada en las investigaciones.

Hypothesis testing and Spanish psychological research: Analyses and proposals. The purpose of this paper was to analyse the application of the most common statistical procedure for studying relationships among variables and empirical phenomena in psychology: The null hypothesis statistical test. In order to determine whether its use in psychological Spanish research is adequate, we carried out a power study of the papers published in Spanish journals. The analysis of the 169 experiments selected, with a total of 5,480 statistical tests, showed power values of 0.18, 0.58, 0.83, and 0.59 to low, medium, high, and estimated effect sizes, respectively. These values drastically decreased in about a 20% when the calculations were repeated controlling the Type I error inflation through Bonferroni adjustment. The results were very similar to those obtained in other international power studies and lead us to think about the need for a special attention for controlling the statistical power in designing a research. On the other hand, we discuss several complementary proposals to the use of significance tests that may improve the information obtained.

Prácticamente desde que hicieron su aparición, las pruebas de contraste de la hipótesis nula han estado marcadas por las diferencias de criterio de los autores que se han ocupado de ellas. La polémica histórica la protagonizaron el propio Fisher que se enfrentó teóricamente a los autores que intentaron modificar su modelo, J. Neyman y E. Pearson.

La lógica de este nuevo modelo propuesto por Neyman y Pearson se ajusta a las ideas siguientes:

1º Se postula un modelo que se apoya en una serie de supuestos.

2º Dentro del modelo se formulan dos hipótesis en términos probabilísticos. La hipótesis nula (H_0) establece que no hay relación entre las variables y que si se replica el experimento indefinidamente el resultado o estadístico seguirá una distribución de probabilidad hipotética. La hipótesis alternativa afirma que existe relación entre las variables y que el estadístico sigue una distribución de probabilidad diferente. Sobre la distribución de probabilidad de la primera hipótesis se establece una región de rechazo de la misma.

3º Se aplica la prueba estadística con los datos y se obtiene la probabilidad de obtener los mismos suponiendo una H_0 verdadera.

Así pues, utilizar una prueba de significación como mecanismo de elección no lleva siempre a una solución correcta, ya que cabe la posibilidad de error, un rechazo equivocado o una aceptación errónea, y dos posibilidades de acierto, de los cuales la potencia representa la probabilidad de rechazar una hipótesis nula falsa.

Fisher nunca estuvo de acuerdo con las modificaciones de su teoría, no aceptó el modelo de dos hipótesis, pero el enfoque que se ha impuesto es una solución mixta (Gigerenzer, 1993), una teoría híbrida que combina los dos modelos y que es en buena parte la responsable de la polémica y críticas que han ido surgiendo a lo largo de la historia de estos procedimientos. De entre todas las críticas, el principal problema a que se suele hacer mención en la aplicación de las pruebas de significación es la no consideración de la potencia estadística en la planificación. Este problema afecta a la validez de la conclusión estadística, que como es sabido, se

Correspondencia: Julio Sánchez Meca
Facultad de Psicología
Universidad de Murcia
30100 Murcia (Spain)
E-mail: jsmeca@fcu.um.es

ocupa de evaluar si la prueba estadística será capaz de detectar relaciones entre las variables intervinientes, en el supuesto de que tales relaciones existan. No tener en cuenta la potencia estadística conlleva una serie de peligros, el mayor de los cuales es una alta probabilidad de resultados no significativos ambiguos, puesto que bien podría ser cierta la hipótesis nula, o bien podría suceder que la potencia sea tan baja que descubrir el efecto resulte una cuestión de mera suerte.

Desde que J. Cohen (1962) realizase el trabajo pionero, numerosos estudios de potencia ya han demostrado el absoluto olvido de la potencia en las aplicaciones de las pruebas de significación en el ámbito de la Psicología (Acklin et al., 1992; Brown y Hale, 1992; Chase y Chase, 1976; Clark-Carter, 1997; Frías et al., 1994; Kazdin y Bass, 1989; Lipsey et al., 1985; Pascual et al., 1994; Rossi, 1990; Sánchez et al., 1992; Sedlmeier y Gigerenzer, 1989; Valera et al., 1998). Concretando, existen problemas en el diseño y aplicación de las pruebas de significación porque no se toma en consideración el tamaño del efecto y no se controla la potencia estadística del contraste.

Por esta razón el objetivo de nuestro estudio ha sido analizar la potencia estadística de la investigación publicada en revistas españolas. Además pretendemos comparar los resultados de nuestro estudio con los de otros estudios españoles y extranjeros. El primero de estos objetivos, planteado en forma de preguntas, quedaría así: (a) ¿En qué grado las investigaciones que aplican pruebas de significación alcanzan a priori los mínimos de potencia estadística convenientes?; (b) ¿Qué niveles suelen lograr las investigaciones incluidas en el estudio en cuanto a potencia a partir del tamaño de efecto estimado con los resultados?; (c) ¿Son adecuados los tamaños muestrales empleados habitualmente? y (d) ¿Cuál suele ser la magnitud de los tamaños del efecto obtenidos?

Método

Para la selección de los estudios nos centramos en revistas españolas que aplican pruebas estadísticas inferenciales y que han sido publicadas en España en el año 1991. En total se revisaron 18 revistas que incluían 573 artículos.

Las variables codificadas en el estudio con el propósito de que nos sirvieran para un análisis más pormenorizado fueron: la revista, la disciplina y el tipo de diseño.

Para calcular la potencia de cada estudio utilizamos como unidad de análisis el experimento (utilizamos también como unidad de análisis el artículo y el tipo de prueba estadística, pero los resultados de los análisis fueron similares). Como tamaños del efecto empleamos los convencionales propuestos por Cohen (1988) -bajo, medio y alto- y el tamaño del efecto estimado. Fijamos como nivel de significación el 5%, para contraste bilateral. Controlamos la tasa de error Tipo I con el procedimiento Bonferroni no ordenado. Y además de todos los cálculos de potencia para cada tamaño del efecto, en cada experimento se obtuvo el tamaño muestral necesario para garantizar una potencia adecuada de .80.

Resultados

De los 573 artículos revisados, 243 eran estudios empíricos. De estos 243 estudios se eliminaron 74 experimentos en los que no era posible el cálculo de la potencia (estudios descriptivos, de caso único o en los que no se aportaban los datos para realizar los cálculos), quedando un total de 169 estudios analizados.

El número total de pruebas estadísticas que se incluyeron en el estudio de potencia fue de 5.480. Destaca el empleo en los artículos revisados de la prueba de significación de un coeficiente de correlación (el 56.37%), seguido de la prueba T para la comparación de medias (22.15%) y modelos de ANOVA (10.2%).

La Tabla 1 presenta los resultados del análisis de la potencia estadística. Las estimaciones de potencia obtenidas para cada tamaño del efecto utilizando la potencia media de cada experimento fueron las siguientes: 0.18 para tamaño del efecto bajo, 0.58 para tamaño del efecto medio, y 0.83 para tamaño del efecto alto. En la última fila de la Tabla 1, en la que presentamos los intervalos confidenciales en torno a la media al 95%, puede verse que sólo para el tamaño del efecto alto se supera el deseable valor de potencia de 0.80 (0.81 — 0.86).

Controlando la tasa de error Tipo I con el procedimiento Bonferroni, los valores de potencia medios sufren un descenso drástico. De 0.18 a 0.07 para tamaño del efecto bajo, de 0.58 a 0.38 para tamaño del efecto medio y de 0.84 a 0.67 para tamaño del efecto alto.

Agrupando la información de acuerdo con las revistas analizadas se comprueba que el patrón de resultados es muy similar (ver Tabla 2). Destacan por sus mayores niveles de potencia logrados, *Boletín de Psicología* y *Revista de Psicología del Trabajo y las Organizaciones*. Se entienden sus mejores niveles de potencia porque

Tabla 1
Potencia calculada a partir de las medias de cada experimento

	Efecto Bajo	Efecto Medio	Efecto alto	Efecto Estimado
N	169	169	169	154
Mínimo	.006	.149	.315	.05
Máximo	1	1	1	1
Media	.18	.58	.84	.59
Mediana	.13	.54	.89	.59
Desv. Típica	.15	.26	.17	.25
Cuartil 1	.088	.36	.75	.41
Cuartil 2	.20	.85	.99	.77
I.C. 95%	.157——-.203	.541——-.620	.810——-.862	.551——-.629

Tabla 2
Valores de potencia medios por revistas

Revista	Nº Experimentos	Efecto Bajo	Efecto Medio	Efecto Alto	Efecto Estimado
Anales de Psicología	2	.126	.538	.785	.701
A. y Modificación de Conducta	25	.174	.663	.88	.518
Anuario de Psicología	10	.108	.414	.729	.583
Boletín de Psicología	7	.308	.631	.885	.618
Cognitiva	11	.13	.51	.754	.675
Comunicación, L. y Educación	2	.18	.676	.91	.634
Estudios de Psicología	18	.154	.494	.853	.65
Infancia y Aprendizaje	13	.186	.653	.88	.566
Investigaciones Psicológicas	1	.06	.192	.599	1.00
Psicológica	19	.205	.603	.868	.598
Psicothema	10	.149	.549	.795	.436
Revista de Historia de la Psicología	0	—	—	—	—
Revista E. de Terapia del Compto.	2	.085	.278	.582	—
Revista de Psicología de la Salud	3	.24	.528	.717	.443
Revista de Psicología Social Aplicada	2	.187	.743	.917	.924
Revista de Psicología de la Educación	9	.23	.678	.881	.645
R. de Psicología Social y Aplicada	29	.183	.549	.812	.563
R. de P. del Trabajo y las Organiz.	6	.255	.823	.956	.704

ambas revistas utilizan por término medio más unidades de observación que el resto. Más de 300 en *Boletín de Psicología* y casi 250 en *Revista de Psicología del Trabajo y las Organizaciones*.

Agrupando la información según la disciplina (ver Tabla 3), vemos que es el área de *Metodología* la que alcanza los mejores niveles de potencia, aunque sólo 4 estudios pudieron catalogarse como metodológicos y además son estudios correlacionales (psicométricos) que emplean un número de sujetos alto.

En cuanto al tipo de diseño (ver Tabla 4), los estudios correlacionales consiguen los niveles de potencia más adecuados para tamaños del efecto estimados. Las investigaciones por encuesta, aunque consiguen niveles de potencia muy altos para los tamaños del efecto convencionales, dado el alto número de sujetos que suelen emplear, no alcanzan una potencia media adecuada para tamaños del efecto estimados. Hay que tener en cuenta que sólo dos investigaciones por encuesta se pudieron incluir en el estudio y que los tamaños del efecto estimados en estas investigaciones suelen ser muy bajos porque al fin y al cabo se trata de opiniones, hay mucha variabilidad.

Otro de los objetivos de este estudio era comprobar si los tamaños muestrales resultan adecuados. Comprobamos que por término medio se emplean 102 unidades de observación por experimento, mientras que deberían utilizarse 227. Considerando la mediana como estadístico más robusto, se emplean 53 sujetos por experimento, mientras que deberían emplearse 103. Es decir, deberían doblarse los tamaños muestrales habitualmente empleados para conseguir una potencia adecuada.

Tabla 3
Valores de potencia medios según el área de conocimiento

Disciplina	Experimentos	Efecto Bajo	Efecto Medio	Efecto Alto	Efecto Estimado
Psicología Básica	51	.16	.52	.80	.55
Personalidad, Eval. y Trat. Psicológico	46	.17	.60	.84	.53
Psicología Evolutiva	36	.18	.62	.85	.60
Metodología	4	.50	.99	1.00	.88
Psicobiología	6	.14	.48	.77	.40
Psicología Social	26	.18	.57	.87	.73

Tabla 4
Valores de potencia medios según el tipo de diseño

Tipo de diseño	Experimentos	Efecto Bajo	Efecto Medio	Efecto Alto	Efecto Estimado
Estudio Correlacional	35	.17	.54	.84	.72
Estudio Cuasi-experimental	65	.20	.64	.85	.60
Estudio Experimental	67	.15	.54	.81	.50
Encuesta por Muestreo	2	.53	.99	1.00	.63

Tabla 5
Resultados de potencia en ciencias sociales, en publicaciones anglosajonas, españolas y en nuestro medio

	Efecto Bajo	Efecto Medio	Efecto Alto	Efecto Estimado
Ciencias Sociales	.24	.64	.85	
Psicología (Ingleses)	.19	.58	.84	.77
Psicología (Españoles)	.16	.55	.82	.61
Estudio de Potencia	.18	.58	.84	.59

Se realizó también un estudio de los tamaños del efecto para comprobar de qué magnitud suelen ser en las investigaciones psicológicas españolas. Para realizarlo convertimos todos los tamaños del efecto estimados a partir de los datos a coeficientes de correlación. Nuestros resultados apoyan la idea de que los tamaños del efecto en Psicología suelen ser de tipo medio. Adoptando la convención de Cohen (1988) que considera como medio un coeficiente de correlación de 0.30, se ha comprobado que la media de los tamaños del efecto encontrados en nuestro estudio es 0.31 y la mediana 0.27.

Por último, en la Tabla 5 se presenta la comparación de este trabajo con otros estudios de potencia, demostrando que las diferencias en cuanto a resultados de las distintas investigaciones de este tipo no son en absoluto notables.

Conclusiones

De acuerdo con los resultados de este trabajo, puede concluirse que el grado de potencia estadística que se alcanza en general en las investigaciones españolas que aplican pruebas de contraste está a un nivel similar que en las investigaciones internacionales, lo cual no debe tomarse como motivo de alegría puesto que tampoco llegan al nivel que se considera deseable. Por otra parte, se confirma que los tamaños del efecto que suelen encontrarse en Psicología son de tipo medio. En cuanto a la adecuación en general de los tamaños muestrales, puede decirse que habría que doblar el número de sujetos que suelen utilizarse en las investigaciones psicológicas para alcanzar los niveles de corrección debidos.

Por lo tanto, puede afirmarse que existen fallos de diseño y de aplicación, además de incomprensiones y malinterpretaciones, de las pruebas de significación en las investigaciones.

Una buena parte de los problemas en la aplicación de las pruebas de significación se deben a las dificultades que entraña el cálculo de la potencia. Pero ya existen instrumentos que facilitan la tarea. La comunidad científica debe comenzar a utilizar las alternativas que complementan y mejoran la aplicación del contraste de hipótesis estadísticas. Las propuestas pueden clasificarse en dos grandes apartados: Las que pretenden mejorar el modelo de las pruebas de significación y las que rompen por completo con el mismo por considerarlo carente de sentido científico. Presentamos a continuación algunas alternativas que pueden mejorar estos procedimientos.

Los *intervalos de confianza* ofrecen más información que las pruebas de significación porque, además de permitirnos la aceptación o rechazo de una hipótesis estadística, nos indican de qué magnitud es el efecto y acotan la precisión de la estimación.

La principal propuesta como complemento a las pruebas de significación es el *tamaño del efecto*. Nos indica el grado en que la hipótesis nula es falsa, el grado en que el fenómeno está presente en la población. Los tamaños del efecto resultan además fundamentales en el análisis de la potencia, porque es uno de los parámetros de los que la potencia es función.

El BESD (*Binomial Effect Size Display*). Recientemente Kirk (1996) considera que incluso un tamaño del efecto pequeño puede tener una significación práctica. Piensa que todos los investigadores al presentar los resultados de sus trabajos deben emitir un juicio sobre la utilidad del efecto encontrado, aunque éste pueda parecer pequeño. En este sentido, puede ayudar la presentación binomial del tamaño del efecto propuesta por Rosenthal (1991), que consiste en transponer a una tabla 2x2, de acuerdo con el tamaño

del efecto, los porcentajes de éxito y los de fracaso después de la aplicación de un tratamiento.

Por otra parte, Rosenthal y Rubin (1994; véase también Hallahan y Rosenthal, 1996; Rosnow y Rosenthal, 1996) han propuesto un nuevo estadístico, el *valor contranulo* (counternull value), que demuestra la relatividad de los resultados de una prueba de significación y que se define como la magnitud no nula del tamaño del efecto que está apoyada por exactamente la misma cantidad de evidencia que el valor nulo del tamaño del efecto.

El *indicador del tamaño del efecto en lenguaje común*. Al igual que el valor contranulo, este índice, propuesto por McGraw y Wong (1992), es un complemento del tamaño del efecto. Permite relativizar la importancia de un tamaño del efecto. 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.

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.

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

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

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. Seoane (Coords.), *III Simposium de Metodología de las Ciencias Sociales y del Comportamiento* (pp. 1.057-1.063). Santiago de Compostela: Servicio de Publicaciones.

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). Hillsdale, NJ: Erlbaum.

Hallahan, M. y Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behavioral Resesearch & 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 Psychology, 57*, 138-147.

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

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.

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

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 las Ciencias Sociales y del Comportamiento* (pp. 1.093-1.101). Santiago de Compostela: Servicio de Publicaciones.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. *American Psychologist, 46*, 1.086-1.087.

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

Rosnow, R.L. y Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331-340.

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

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? *Psychological Bulletin, 105*, 309-316.

Valera, A.; Sánchez, J.; Marín, F. y Velandrino, A.P. (1998). Potencia Estadística de la Revista de Psicología General y Aplicada (1990-1992). *Revista de Psicología General y Aplicada. 51*(2).