anales de psicología, 1992, 8(1-2), 33-38

Software para el cálculo de la potencia: un estudio comparativo

Fulgencio Marín Martínez [1]
Antonio P. Velandrino Nicolás
Antonio Valera Espín
Julio Sánchez Meca
Universidad de Murcia

Resumen. Se comparan tres paquetes de software estadístico específicamente diseñados para el cálculo de la potencia estadística: STAT-POWER (Bavry, 1991), POWER (Borenstein y Cohen, 1988) y DESIGN (Dallal, 1988). Los criterios de comparación son la intera ctividad, amplitud y precisión. Ninguno de los tres supera a los demás: STAT-POWER es el más exhaustivo, mientras que POWER es el más interactivo y puede utilizarse tanto para la investigación como para la docencia. Finalmente, se discuten las ventajas y limitaciones de cada programa.
Palabras clave: Software estadístico; potencia; tamaño del efecto; tamaño muestral

Abstract. Three statistical packages specifically designed for calculating statistical power are compared: STAT-POWER (Bavry, 1991), POWER (Borenstein & Cohen, 1988), and DESIGN (Dallal, 1988). The comparison criteria are interactivity, completeness, and accuracy. None of them does a better performance: STAT-POWER is the most complete, while POWER is the most interactive, and can be used both for research and for teaching. Finally, the potentials and limitations of each program are discused.
Key words: Statistical software; power; effect size; sample size

Introducción

Hace treinta años que Jacob Cohen publicó el primer estudio sobre la potencia estadística de la investigación en Psicología (Cohen, 1962), advirtiéndonos de las consecuencias metodológicas que se derivan del desarrollo del contraste de hipótesis sin un mínimo de potencia para detectar los efectos de interés en un campo de estudio. Sin embargo, la investigación actual en ciencias sociales y del comportamiento no parece haberse hecho eco de estas consideraciones. Así lo ponen de manifiesto recientes estudios de potencia (Kazdin y Bass, 1989; Rossi, 1990; Sedlmeier y Gigerenzer, 1989), en los que los resultados arrojan una potencia media en torno al 50% para tamaños del efecto de una magnitud intermedia.

El sugerente y un tanto sarcástico interrogante "¿Tienen los estudios de potencia algún efecto sobre la potencia de los estudios?" con el que Sedlmeier y Gigerenzer (1989) titulan uno de sus más recientes artículos, plantea la falta de receptividad en la comunidad científica a las sugerencias de que se cuantifique y controle la potencia de la investigación. De hecho, en la actualidad ni siquiera los consejos editoriales de las revistas de mayor impacto en psicología adoptan como criterio de valoración de un trabajo empírico inferencial su potencia estadística.

Por otro lado, la inexistencia hasta hace bien poco de un instrumental que facilitara el cálculo de la potencia, representaba una seria limitación de cara a la concienciación del investigador social para el control de la potencia de sus estudios. Las primeras tablas en las que se ponían en relación los parámetros $1 - \beta$ (potencia), tamaño del efecto (TE) y tamaño muestral (n) para los niveles de significación estándar (α = .05, .01 ó .10), incluyendo gran parte de los contrastes de hipótesis más usuales en la investigación social, fueron elaboradas por Cohen (1969), quien las amplió en las sucesivas ediciones de dicho manual hasta las más recientes (Cohen, 1977, 1988). Otras aportaciones en esta línea fueron las de Kraemer y Thiemann (1987).

Sin embargo, la década de los 80 irrumpiría con una amplia difusión de los microordenadores, paralela al desarrollo de programas de estadística compatibles con el hardware de un microordenador.

[1] *Dirección:* Fulgencio Marín Martínez. Dpto. Metodología y Análisis del Comportamiento. Facultad de Psicología. Apdo. 4021. 30080 Murcia (Spain).
 ISSN: 0212-09728.

El investigador en ciencias sociales se hizo rápidamente usuario del software disponible para el análisis estadístico de sus datos, demandando programas que le permitan desarrollar los cálculos de la forma más rápida, directa, sencilla e interactiva. Pudo con ello prescindir del manejo manual de tablas estadísticas, más incómodo, y que en muchas ocasiones obligaba a interpolar, con la consiguiente pérdida de precisión en los resultados. Como consecuencia de esta demanda, han surgido numerosos programas de software específicamente diseñados para el cálculo de la potencia estadística (cf. por ej., Goldstein, 1989).

Estudio comparativo.

El objetivo de este trabajo es el de valorar las aportaciones de tres paquetes estadísticos especialmente diseñados para el cálculo de la potencia y que parecen estar experimentando una gran difusión: STAT-POWER (Bavry, 1991), POWER (Borenstein y Cohen, 1988) y DESIGN (Dallal, 1988). Precisamente porque en el contexto metodológico actual se prescinde del control de la potencia, estimamos de gran interés informar al investigador de las posibilidades informáticas a este respecto. Asimismo, creemos que el software presentado supone un estímulo para que los investigadores comprueben la potencia de sus estudios y consideren la necesidad de controlarla a priori. Todo ello gracias a la facilidad con la que estos programas pueden ser utilizados para gran parte de las pruebas estadísticas más frecuentes, simplemente informando de una serie de parámetros que podrán variar en función de la prueba estadística (tamaño muestral, medias y desviaciones típicas, el nivel de significación, coeficientes de correlación, varianza de error en un ANOVA, etc).

Estos programas pueden instalarse en microordenadores PC/XT/AT/386/486 y compatibles que soporten el sistema MS-DOS. No precisan de coprocesador matemático y el más extenso puede ocupar un máximo de un Megabyte de memoria. Por consiguiente, están al alcance de cualquier usuario particular de PCs.

De entre los criterios que pueden utilizarse para comparar los tres paquetes, hemos escogido los de interactividad, amplitud y precisión. Otros criterios como los de velocidad, conectividad o manipulación de datos se han excluído, por no existir una diferenciación importante entre los programas.

1. Interactividad

De los tres paquetes mencionados, el de Bavry (1991) y el de Borenstein y Cohen (1988) funcionan mediante un sistema de menús que los hace muy interactivos. De entrada, el programa pide al usuario que seleccione la prueba estadística por la que está interesado, presentando una gama más amplia de alternativas el programa de Bavry en comparación con el de Borenstein y Cohen. El programa de Dallal (1988), aún no siendo tan interactivo, ya que no dispone de un sistema de menús, resulta también fácil de manejar por medio de una secuencia de comandos.

Una vez elegida la prueba estadística, y con la excepción de DESIGN que se centra exclusivamente en el cálculo de tamaños muestrales, los otros dos programas dan la opción de calcular la potencia del contraste de hipótesis ya desarrollado, o bien del tamaño muestral necesario para que el diseño de una determinada investigación garantice el nivel de potencia deseado.

En cuanto al cálculo de la potencia, el investigador se interesará fundamentalmente por la potencia para detectar el tamaño del efecto (TE) real del que dan evidencia los datos muestrales que lo estiman, aunque también puede interesarle la potencia con la que su investigación detectaría una determinada magnitud en el efecto objeto de estudio.

Si el objetivo es buscar el tamaño muestral asociado a una determinada potencia, se impone informar de la magnitud del efecto que queremos poner a prueba, para lo que pueden sernos de utilidad los criterios de Cohen que cuantifican lo que sería un TE alto, medio o bajo en ciencias sociales y del comportamiento (Cohen, 1969).

El único de los tres programas que permite introducir directamente el valor del TE en unidades

estándar (*d, f, h,* etc, dependiendo del tipo de contraste de hipótesis) es el de Borenstein y Cohen. Puesto que todos estos índices pueden ser fácilmente transformados para hacerlos directamente comparables (véase Sánchez Meca y Ato, 1989), por ejemplo en un coeficiente de correlación de Pearson o en una puntuación z, se conocen sus valores correspondientes a los criterios de Cohen de un TE bajo, medio y alto. Asimismo, también podría buscarse su equivalencia bajo cualesquiera otros criterios.

Por consiguiente, disponer de una métrica común que permita valorar de una forma estándar e independiente del contraste de hipótesis empleado, la relación o el efecto planteados en una investigación, facilita al investigador establecer la magnitud del efecto por la que está interesado. Su estudio deberá presentar una potencia elevada en relación a dicha magnitud.

Con los programas de Bavry (1991) y Dallal (1988), definir a priori un determinado TE puede resultar sencillo en contrastes como los de un coeficiente de correlación de Pearson, donde la propia correlación es un índice del TE. Sin embargo, en otros contrastes como el ANOVA, se requeriría buscar una combinación de medias y desviaciones típicas equivalente al TE deseado, lo que puede llegar a resultar muy engorroso conforme el contraste es más complejo.

Tras la selección de la prueba estadística, el programa de Borenstein y Cohen presenta un segundo menú que además de contemplar las posibilidades del cálculo del tamaño muestral o de la potencia, incluye otras opciones tales como la elaboración de gráficos o tablas, e incluso el desarrollo de una simulación Monte Carlo. Ambas opciones se presentan con un carácter más didáctico que operativo, dando cuenta de las interrelaciones entre la potencia y el resto de parámetros que intervienen en el diseño de un contraste de hipótesis.

Asimismo, el programa DESIGN construye gráficos para ilustrar la relación entre el tamaño muestral y la potencia, partiendo de los datos que el propio usuario introduce al plantear un determinado contraste, y manteniendo constantes el TE y el nivel de probabilidad α.

De especial interés es el carácter didáctico del programa de Borenstein y Cohen, en consonancia con el notable y persistete esfuerzo de Cohen, desde 1962 hasta sus más recientes publicaciones (1988, 1990, 1992), por concienciar a los investigadores sociales de la necesidad de controlar la potencia de los estudios, a fin de garantizar el preciso rigor metodológico.

A partir de los datos concretos que de un contraste de hipótesis son introducidos por el usuario, el programa elabora una serie de gráficos. Además de representar la magnitud de cada uno de los cuatro parámetros (tamaño muestral, tamaño del efecto, nivel de significación y potencia), estos gráficos permiten que el propio usuario altere el valor de uno o varios de dichos parámetros, comprobando en qué medida se modifican los restantes.

En concreto, se puede visualizar gráficamente cómo cambia la potencia de un contraste modificando el TE (con α y n constantes), o bien el tamaño muestral (con TE y α constantes), o la probabilidad α (con TE y n constantes).

Por otro lado, se pueden desarrollar simulaciones Monte Carlo para gran parte de los contrastes, consistentes en plantear un proceso indefinido de muestreo aleatorio, en el que se aprecian las diferencias entre los valores que para la población define el usuario (medias, proporciones, etc) y los correspondientes a cada "extracción" muestral. Sirviéndose de unos gráficos muy didácticos, el programa nos informa sucesivamente de los resultados para una determinada muestra (TE estimado), los parámetros de la población (TE real), si son o no significativos los resultados muestrales y la potencia del contraste.

Este proceso resulta especialmente ilustrativo de las consecuencias de asumir una determinada potencia. Dado que el usuario puede solicitar un número muy elevado de muestreos, observará la concordancia entre la potencia del estudio y el porcentaje de muestras en las que se rechaza la hipótesis nula, bajo las condiciones del TE, nivel α y tamaño muestral por él mismo definidas.

2. Amplitud

Utilizamos el término amplitud para indicar el abanico de pruebas estadísticas concretas que pueden ser analizadas con el software revisado para obtener valores de potencia. Como puede observarse en la Tabla I, el programa de potencia más completo en este sentido sólo es capaz de analizar 14 pruebas diferentes de forma directa (Bavry, 1991). Si bien es posible obtener la potencia para los procedimientos estadísticos más utilizados en las investigaciones psicológicas disponiendo de los tres programas, lo cierto es que, dada la enorme variedad de pruebas de contraste existentes, estos paquetes resultan limitados. Sólo uno de ellos (Bavry, 1991) permite, por ejemplo, el cálculo de la potencia para una prueba multivariante, pese a la gran popularidad de que gozan cada vez más estos procedimientos. Tampoco dedican ningún módulo a las comparaciones múltiples o a las pruebas post-hoc.

El control de la potencia en un diseño de investigación puede resultar una tarea difícil si tenemos en cuenta que para cada TE, nivel de significación y tamaño muestral diferentes, la distribución de probabilidad de la hipótesis alternativa sigue una distribución no central. Con la mayoría de las pruebas estadísticas, calcular la potencia utilizando las fórmulas al uso y una simple calculadora de bolsillo es un proceso lento y tedioso. Es por ello que, sin menoscabo de la utilidad de las tablas publicadas (Cohen, 1988; Kraemer y Thiemann, 1987), consideramos importante que los instrumentos informáticos dedicados a la potencia abarquen un amplio número de pruebas estadísticas. Además, otra ventaja con respecto a las tablas en la planificación de la investigación es que el usuario puede elegir con rapidez y facilidad entre diferentes diseños.

De los tres programas analizados, ya hemos visto que el confeccionado por Bavry permite el cálculo de la potencia para un mayor número de pruebas de contraste. Además, este instrumento incluye un módulo que calcula los percentiles para un buen número de distribuciones de probabilidad centrales y no centrales, lo que permite el cálculo de la potencia de prácticamente todas las pruebas estadísticas, siempre y cuando el usuario domine los conceptos y el proceso de cálculo.

Prueba estadística	Bavry	Borenstein & Cohen	Design
Prueba *T* dos muestras independientes	SI	SI	SI
Prueba *T* dos muestras relacionadas	SI	NO	SI
Prueba *T* para una muestra	SI	NO	SI
Significación de un coeficiente de correlación	SI	SI	SI
Diferencias entre coeficientes de correlación independientes	SI	NO	NO
Contraste de una proporción	SI	NO	NO
Diferencias entre proporciones independientes	SI	SI	SI
Chi cuadrado y tablas de contingencia	SI	SI	NO
Análisis de varianza simple	SI	SI	SI
Análisis de varianza de medidas repetidas	SI	NO	NO
Análisis de varianza factorial	SI	SI	SI
Análisis de varianza mixto	SI	NO	NO
Análisis de varianza de medidas repetidas multivariado	SI	SI	NO
Regresión y correlación multiple	SI	SI	NO

Tabla I: Relación de pruebas estadísticas que pueden ser analizadas por cada uno de los programas revisados.

Tanto el programa de Bavry como el de Borenstein y Cohen ofrecen la posibilidad de calcular, bien la potencia estadística –en función de α, n.y el TE–, o bien del tamaño muestral -dados α, el TE y la potencia que el investigador considera deseable. La obtención del valor de potencia estadística es útil a posteriori, es decir, cuando el investigador ha concluido el trabajo y quiere comprobar si ha conseguido un nivel de potencia adecuado. La utilidad del cálculo del tamaño muestral viene dada a priori, cuando el investigador está planificando el diseño de investigación y se dispone a buscar

un tamaño muestral que asegure determinado nivel de potencia. El programa de Dallal (1988) sólo permite cálculos de tamaño muestral, al estar pensado exclusivamente para la planificación de la investigación, tal y como indica su nombre. Sin embargo, es posible el cálculo de la potencia realizando un proceso iterativo de aproximaciones sucesivas al valor buscado. El proceso finalizará en la medida en que más nos acerquemos al tamaño muestral definido de antemano.

3. Precisión

En la Tabla II podemos comprobar que, en general, no existen apenas diferencias entre los valores de potencia obtenidos con los tres programas analizados, manteniendo constantes el tamaño muestral, el tamaño del efecto y el nivel de significación. Era de esperar esta fuerte concordancia entre los tres paquetes si tenemos en cuenta que la precisión informática al aplicar cualquier fórmula es muy alta. Incluso utilizando fórmulas aproximadas de la función de densidad de las distribuciones no centrales, se alcanzan valores de potencia precisos. De hecho, puede asegurarse que las diferencias encontradas se deben a errores de redondeo (excepto en el caso del contraste de dos proporciones para muestras independientes, donde DESIGN no ofrece valores muy cercanos a los otros dos programas).

Tamaño	Prueba T 2 muestras Efecto d = .5, NS=.05			Significación de coeficiente Efecto R = .3, NS=.05		
Muestral	Bavry	B. & C.	DESIGN	Bavry	B. & C.	DESIGN
10	.185	.18	.185	.123	.14	.133
20	.338	.34	.338	.244	.26	.254
30	.478	.48	.478	.362	.37	.370
40	.598	.60	.598	.472	.48	.477
50	.697	.70	.697	.570	.57	.572
60	.775	.77	.775	.654	.65	.654
70	.836	.84	.836	.725	.72	.723
80	.882	.88	.882	.783	.78	.780
90	.916	.92	.916	.831	.83	.827
Tamaño	Contraste 2 proporciones Efecto h = .5, NS=.05			ANOVA un sentido 3 grupos Efecto f = .25, NS=.05		
Muestral	Bavry	B. & C.	DESIGN	Bavry	B. & C.	DESIGN
10	.201	.20	.054	.196	.199	.196
20	.354	.35	.170	.377	.386	.377
30	.492	.49	.303	.543	.553	.543
40	.611	.61	.433	.679	.688	.679
50	.707	.71	.550	.783	.789	.783
60	.784	.78	.650	.857	.862	.857
70	.842	.84	.733	.909	.912	.909
80	.887	.89	.799	.943	.945	.943
90	.919	.92	.851	.965	.966	.965

Tabla II. Tabla comparativa de valores de potencia en los tres programas analizados.

La mayor precisión es una de las ventajas básicas de los programas de software sobre las tablas. Las tablas de potencia no pueden ofrecer valores exactos en todos los casos porque se necesitaría una tabla diferente para cada prueba estadística en función de la variación del nivel de significación, del tamaño muestral, del TE y de los grados de libertad de la distribución considerada. Como se comprenderá, el número de tablas que puede obtenerse con tantos parámetros en juego es prohibitivo. No obstante, hemos comprobado que las tablas de Cohen (1988), y los procedimientos ideados por éste para utilizarla en casos especiales, consiguen un buen ajuste a los valores de potencia calculados con los instrumentos informáticos analizados aquí.

Conclusión

La complejidad que supone el cálculo de la potencia estadística para un investigador queda sustancialmente reducida gracias a la existencia de programas informáticos diseñados a tal efecto. Dada la importancia que el control de la potencia tiene para la validez de la conclusión estadística, los investigadores que utilizan pruebas de contraste en el análisis de sus datos deberían dedicar parte de sus esfuerzos a la planificación a buscar de manera rigurosa el tamaño muestral más adecuado.

Los programas aquí revisados podrían ser en este sentido de gran utilidad. Todos ellos son muy interactivos, de fácil manejo, precisos, rápidos en los cálculos, y con exigencias de *hardware* mínimas; aunque todavía no contemplen la totalidad de pruebas estadísticas de contraste existentes. En cuanto a sus diferencias, destacan la mayor amplitud de STAT-POWER (Bavry, 1991) en relación al número de pruebas que abarca, así como la inclusión de índices estándar para el TE en el programa de Borenstein y Cohen (1988).

Por último, el carácter didáctico del programa de Borenstein y Cohen resulta muy acertado para concienciar, ya desde los primeros niveles académicos en los estudios de licenciatura, de la importancia y relevancia conceptual del control de la potencia en la investigación.

Referencias

Bavry, J.L. (1991). *STAT-POWER: Statistical Design Analysis System* (2nd ed.). Chicago, IL: Scientific Software, Inc.

Borenstein, M. y Cohen, J. (1988). *Statistical Power Analysis: A Computer Program.* Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (rev. ed.). New York: Academic Press.

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, 45(12),* 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155–9.

Dallal, G.E. (1988). *DESIGN: A Suplementary Module for SYSTAT* (Vers. 2.0). Evanston, IL: SYSTAT, Inc.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *American Statistician, 43,* 253-260.

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.

Kraemer, H.C. y Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research.* Newbury Park, CA: Sage.

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

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

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

(Original recibido: 15-6-1992)
(Original aceptado: 27-7-1992)