Metodología de las Ciencias del Comportamiento, Volumen Especial 2004, 465-469

Usos y abusos de la significación estadística: propuestas de futuro (¿Necesidad de nuevas normativas editoriales?)

Juan Pascual Llobell, María Dolores Frías Navarro y Fernando García Pérez
Universitat de València

Resumen

Históricamente, los científicos sociales y especialmente los psicólogos han confiado en la "comprobación de la significación estadística" como el instrumento por excelencia de análisis de datos. Al comprobar la significación estadística, se especifica una hipótesis nula y se derivan las probabilidades de los datos bajo tal hipótesis. El nivel de significación fijado aporta información sobre la probabilidad de rechazar erróneamente dicha hipótesis. Como afirma Cohen (1990) el esquema de Fisher es tremendamente competitivo y atractivo a la vez, por cuanto "they offered a deterministic scheme, mechanical, and objective, independent of content, and led to clear-cut yes-no decisions" (Pág. 1307). Sin embargo, durante las últimas décadas ha habido un crecimiento exponencial en el número de artículos y publicaciones dedicados a criticar los usos inadecuados de esta estrategia analítica o cuando menos de lo insatisfactoria que resulta para alcanzar el objetivo final de acumulación de conocimiento. El aumento de las críticas ha ido asociado al creciente reconocimiento de las limitaciones asociadas a los tests de significación como único criterio de interpretación de la significación de los resultados. En 1999 el Comité de la American Psychological Association (A.P.A.) sobre Inferencia Estadística recomendaba aplicar un conjunto de reformas a la praxis analítica. El manual de Publicación del APA (2001) se hacía eco de algunas de ellas. Desgraciadamente, el cumplimiento de las mismas que, en parte al menos, supone romper con ciertas tradiciones, no es fácil: probablemente se requieran políticas editoriales más firmes y nuevos currículos académicos.

Abstract

Historically, the social scientists and especially the psychologists have trusted in the "verification of the statistical significance" as the instrument by data analysis excellence. The level of significance contributes information on the probability to reject erroneously null hypothesis. As it affirms Cohen (1990) the plan of Fisher is tremendously competitive and attractive at the same time, "they offered a deterministic scheme, mechanical, and objective, independent of content, and led to clear-cut yes-no decisions" (P. 1307). Nevertheless, during the last decades there has been a growth exponential in the number of articles and publications dedicated to criticize the inadequate uses of this analytic strategy or when less than the unsatisfactory thing that the objective knowledge accumulation.

La explicación científica de los fenómenos psicológicos se sustenta sobre el descubrimiento y confirmación de las relaciones "verdaderas" existentes entre los mismos o con las variables y factores que los explican. Los filósofos no están muy de acuerdo en lo que se entiende por "verdad" o "relación verdadera" (teoría de la correspondencia, teoría de la coherencia o teoría pragmática...), por lo que parece conveniente sustituir el término verdad por el de "validez": afirmamos que una interpretación / explicación de un fenómeno es válida en la misma medida que está confirmada por los datos empíricos o es consistente con otras fuentes de conocimiento, incluidas las investigaciones y teorías previas. Cook y Campbell (1979) fueron los primeros que elaboraron una tipología acerca de las variedades de validez, ampliando desarrollos previos (Campbell y Stanley, 1963), al dimensionarla en cuatro categorías: validez de conclusión estadística, validez de constructo, validez interna y validez externa.

Sin entrar en debates acerca de cuál deberíamos considerar más importante, es cierto que toda investigación que se postule valida, generadora de conocimiento riguroso, ha de cumplir cualesquiera de ella en todas sus vertientes al máximo posible. En esta comunicación, sin embargo, nos limitamos a hablar de la primera.

Validez de la conclusión estadística

La validez de conclusión estadística tiene que ver con el componente de covariación que se establece cuando afirmamos relaciones de causa a efecto o relaciones meramente funcionales entre variables: ¿existe covariación / correlación entre la causa y el efecto? ¿de qué magnitud es? Incorrectamente se puede concluir que la causa y el efecto covarían cuando no es así (error tipo I) o incorrectamente concluir que no covarían cuando de hecho sí lo hacen (error tipo II). Las dos incorrecciones están asociadas a la primera de las preguntas anteriores y hay que establecer las medidas oportunas para que no ocurran. En cuanto a la segunda pregunta, señalemos que una infraestimación o sobreponderación del tamaño del efecto, puede igualmente afectar la interpretación de los datos.

El modelo estándar y habitual en Psicología de tratar el tema de la covariación entre variables ha sido el de la comprobación de la hipótesis de nulidad. Si asumimos, por ejemplo, que la tasa de éxito al aplicar una nueva droga es del 60 % mientras que en el grupo placebo la tasa es sólo del 40%, produciéndose una diferencia del 20%, en la dirección deseada entonces todo hace indicar que la nueva droga es más efectiva contra el cáncer. Sin embargo, también podría pensarse que las dos condiciones son igualmente eficaces (al 50 %) sólo que por azar nuestra muestra-placebo ha obtenido una mortalidad del 60% y la muestra-nueva droga una mortalidad del 40%, produciéndose igualmente un diferencial del 20% entre ambas. El problema de saber si efectivamente (y no solo por azar) la droga es mejor que el placebo lo afrontamos mediante la aplicación de un test de significación, aplicando la siguiente lógica: Si la hipótesis nula es verdadera, se espera que la diferencia promedio entre grupos, extrayendo un numero infinito de muestras será cero, y la diferencia entre dos medias particulares se aproximará razonablemente a cero. Si nuestras muestran producen una diferencia inesperada, en el sentido que no cabe esperarla de dos poblaciones que se suponen idénticas, entonces concluimos que la hipótesis nula no es aceptable. En este caso debemos rechazarla en favor de la hipótesis alternativa.

Como es sabido la posibilidad de rechazar la hipótesis nula depende del tamaño del efecto, del tamaño de la muestra y del nivel alfa. El test de significación estadística, por ejemplo la *t* de Student, está definido por la ratio entre la diferencia de medias (tamaño del efecto) y la dispersión de las diferencias (que depende del tamaño de la muestra). El rango de valores aceptables producidos por el test para rechazar la hipótesis nula está definido por el nivel alfa que se haya elegido. Esta es la lógica que sigue cualquier test de significación estadística. Sin embargo, al combinar la magnitud del efecto con el tamaño de la muestra, el procedimiento puede generar confusión, interpretaciones incorrectas y falacias, de las que por desgracia no resulta fácil deshacerse. (Véase Tabla I).

Tabla I. Falacias respecto de la significación estadística (Tomado de Borenstein, 1997)

1.	Que el efecto sea significativo, no quiere decir que sea importante (p.e. clínicamente importante). Generalmente, lo único que quiere decir es que el efecto es (probablemente) no 0 (*nil*, según Cohen), que puede ser sustancialmente (clínicamente) importante o totalmente irrelevante, aunque real.
2.	Que un estudio compruebe un efecto significativo al .05 y otro al .001, no supone que el segundo es más importante que el primero. Lo sería en el caso que ambos estudios utilizaran muestras de *N* iguales. En caso contrario, nada se concluye al respecto.
3.	Si se comprueba un efecto significativo con p= .03, por ejemplo, en una muestra de niños y el mismo efecto obtiene una probabilidad de .07 en una muestra de jóvenes de 18 años, afirmar que el tratamiento es efectivo en un caso y no en el otro, es cuando menos equivoco. El punto de corte del .05 para indicar significatividad / no significatividad es consuetudinario y nada axiomático. Los valores p se basan en una función continua y, en principio, tan demostrativo es encontrar la combinación .03 y .05 como la combinación. 04 y .06.
4.	Supongamos que un experimento demuestra que cierta droga es estadísticamente no significativa ¿Significa que la droga no es efectiva? Sólo significa que la evidencia empírica obtenida no es suficientemente fuerte para probar que la droga es efectiva. Es perfectamente posible que la droga tenga efectos, pero no se haya conseguido detectarlo (p.e. por un insuficiente número de observaciones).

A pesar de los riesgos anunciados y a pesar de las críticas en contra del uso rutinario de los tests de significación, el procedimiento estándar viene siendo dominante en psicología: la interpretación de los datos se apoya sistemática y rutinariamente sobre los tests de significación y se atiende única y exclusivamente a la interpretación del nivel alfa. Hace ya algunos años, el profesor J. Bernia (1979), expresaba su opinión en torno a la insuficiencia de este modo de proceder al afirmar que "*el análisis estadístico de los datos de un experimento debe comprender, pues, dos niveles: 1º tests de significación (permite afirmar la existencia de una asociación) y 2º. Estimación de la magnitud de dicha asociación.* ***Limitarse al primer nivel del análisis encierra el peligro de ignorar el aspecto más interesante, saber si los efectos son o no relevantes***" (Pág. 254, subrayado nuestro).

Durante varias décadas los metodólogos y estadísticos han advertido de los inconvenientes de hacer inferencias utilizando como único procedimiento la comprobación de la hipótesis nula. Voces como las de Bakan (1966), Carver, (1978) y Morrison y Henkel, (1970), considerando sólo a los pioneros, advirtieron a tiempo de algunos de esos inconvenientes: confusión entre no-rechazo y aceptación, rechazo y significación teórica o desconsideración de los errores tipo II / potencia estadística.

Los hábitos enraizados en la tradición, continuamente alimentada por los manuales de texto y por los docentes, son difíciles de extirpar y todavía hoy existe la creencia encubierta de que el formato de la significación estadística concede validez científica y objetiva a las conclusiones. La mayoría de editores difícilmente aceptan para publicar los artículos que se separan de ese estándar que no va más allá de la obtención del valor *p*. Pero, por suerte, cada vez son más los que se alinean en el lado crítico cuyo máximo exponente, por la dureza de sus postulados, (difíciles de aceptar sin más), podría estar representado por Rothman, editor que fue de *American Journal of Public Health*, quien expresaba así su política editorial:
"*All reference to statistical hypothesis testing and statistical significance should removed from the paper. I ask that you delete p values as well as comments about statistical significance. If you do not agree my standard ....you should feel free to argue the point.....As* ***editor, however, I can hardly be expected to accept papers that the scientific principle that I expose***" (Rothman, 1986, Pág. 559 -el subrayado no existe en el texto original).

En realidad el test de significación estadística, es decir, la comprobación de la hipótesis nula, sólo tiene sentido cuando es razonable suponer que la hipótesis nula es verdadera, pero no ante cualquier situación. Es decir, si el tratamiento no tiene efectos, la única clase de error que podemos cometer es el error tipo I: concluir falsamente que el tratamiento tiene efectos cuando de hecho no los tiene. Pero si el tratamiento tiene efectos el único error que podemos cometer es el error tipo II: concluir que no los tiene al no detectarlo -falta de potencia estadística- cuando de hecho los hay.

Pues bien, ¿hay visos de que la hipótesis de nulidad sea verosímil en psicología? ¿Tiene una cierta probabilidad de ser verdadera? Si definimos la hipótesis nula en términos absolutos, p.e. diferencia de medias igual a 0, cosa usual en psicología (valores *nil* de Cohen), es difícil asumir que los efectos de dos tratamientos sean cuantitativamente iguales si la variable dependiente es medida en una escala continua: siempre cabrá diferenciar sus efectos en algún decimal por lejos que esté de la coma. Si nos atenemos a criterios empíricos, podríamos llegar a la misma conclusión. Hace exactamente una década que Lipsey y Wilson (1993) publicaron una revisión sobre 302 meta-análisis publicados sobre prácticamente cualquier tratamiento psicológico. De ellos, sólo 3 obtenían un tamaño del efecto igual a 0 (Veáse tabla 1 de su estudio original). Es decir solo un 1% de los estudios sugería que la hipótesis nula era verdadera y, además, los tres estudios pertenecían a un mismo tema de estudio. No disponemos de más evidencia empírica que este estudio, por lo que tampoco cabe concluir nada definitivamente, pero es razonable pensar que la hipótesis nula que tanto gusta comprobar a los investigadores es habitualmente falsa. En consecuencia no es al error tipo I al que hay que considerar como prioridad, sino al error tipo II. Y qué consideración se le ha atribuido a éste. Poca, escasa o casi nula.

A juzgar por las revisiones existentes, la existencia de experimentos con baja potencia no es una excepción. A tenor de los trabajos de Seldmeier y Gigerenzer (1989) y Lipsey (1990) entre otros, la baja potencia es muy frecuente siendo la principal causa de existencia de conclusiones nulas falsas y es imposible planificar la potencia de una investigación sin conocer el tamaño del efecto, del que depende principalmente. No obstante cuando los efectos son pequeños, resulta difícil incrementar el tamaño del efecto mediante los métodos usuales (Shadish,Cook y Campbell, 2002), haciendo necesario el recurso al análisis meta-analitico.

Hay un sentimiento generalizado de que las cosas no pueden seguir así. Algunos cortan por lo sano, postulando que los test de significación deben ser abandonados (Schmidt y Hunter, 1997): "*Statistical significance testing retards the growth of scientific knowledge ; it never makes a positive contribution*" (Pág. .37). Otros, quizá la mayoría optan por buscar alternativas complementarias que permitan subsanar los errores de aplicación e interpretación de los tests de significación: estimación del tamaño del efecto y estimación de los intervalos de confianza, son alternativas mayoritariamente propuestas. Dos cuestiones complementarias que se implican mutuamente cabe formularse: cuál es el cambio propuesto y como hacerlo.

Si analizamos el problema desde la óptica de la segunda de las cuestiones, es fácil comprobar que son pocos los agentes que pueden y deben facilitar el cambio, mediante la programación de políticas de futuro adecuadas; básicamente consideramos que hay que tres agentes que pueden contribuir sustancial y eficazmente a que se vayan introduciendo modificaciones en el modo de hacer de los científicos:

1. los editores, exigiendo formatos de publicación que vayan más allá de lo comúnmente establecido
2. los programas docentes que introduzcan y entrenen de verdad en los nuevos hábitos a los investigadores del futuro y
3. la disponibilidad informática de los nuevos procedimientos de análisis

El retraso en la aplicación de cualesquiera de estas políticas mantendrá viva la tradicional manera de afrontar el análisis que cuando menos es insatisfactoria.

No menos de 20 revistas exigen hoy en día en sus prescripciones editoriales que se informe del tamaño del efecto. Entre ellas, *Educational and Psychological Measurement, Journal of Applied Psychology, Journal of Consulting and Clinical Psychology, Journal of Experimental Psychology....* EL comité de estudio del A.P.A. sobre inferencia estadística redactó unas conclusiones (Wilkinson y cols., 1999) que posteriormente ha incorporado en sus recomendaciones el ultimo Manual de Publicación del APA (2001):

"*That it is almost necessary to include some index of effect size or strength of relations in your Results section ... (Es necesario) to provide the reader not only witch information about statistical significance but also with enough information to asses the magnitude of the observed effect or relationship*" (Págs. 25-26).

Se hace necesario informar sobre el tamaño del efecto porque se obliga al investigador a pensar meta-analíticamente, es decir a formular los efectos específicos esperados en función de los efectos anteriormente descubiertos y explícitamente interpretarlos en relación con los efectos relacionados (Cumming y Fich, 2001).

Tabla II. Principios básicos del análisis estadístico de los datos (Informe Wilkinson y cols., 1999)

1.Antes de presentar los resultados, informe sobre las complicaciones, violaciones del protocolo, y sobre otros sucesos no previstos en la recogida de los datos. Estos incluyen los datos perdidos (faltantes), la atrición y la tendencia a no-respuesta. Justificar las técnicas analíticas previstas para neutralizar estos problemas. Describa la no representatividad (de los datos) e informe de los patrones y de la distribución de los datos perdidos y de las consecuencias derivadas.. Documente cómo los análisis aplicados difieren de los previstos antes de descubrir las complicaciones. El uso de las técnicas (estadísticas) que garanticen que los resultados no están contaminados por anomalías de los datos (p.e. *outliers*, puntos de alta influencia, datos perdidos no aleatorios, sesgo de selección, problemas de atrición) debería ser un componente habitual de todos los análisis.
2. La enorme variedad de métodos cuantitativos modernos obliga a los investigadores a la tarea nada trivial de integrar análisis y diseño de investigación. Aunque los diseños y métodos complejos son a veces necesarios para afrontar de manera efectiva las cuestiones a investigar, los acercamientos clásicos simples pueden igualmente aportar una respuesta elegante y suficiente a preguntas importantes. No se debe elegir un método analítico para impresionar a los lectores o para evitar las críticas. Siempre que los supuestos y la fuerza de un método simple sea razonable para los datos y para el problema de investigación, debe utilizarse. El principio (de parsimonia) de OCAM debe aplicarse tanto a la teoría como a los métodos.
3. Hay muy buenos programas de computador para analizar los datos. Mas importante que elegir entre paquetes estadísticos específicos es verificar sus resultados, entender lo que significan y saber cómo lo consiguen. Si usted no puede verificar sus resultados mediante "conjeturas" inteligentes, debería de contrastar los resultados con los ofrecidos por otros programas. A usted no le hará ninguna gracia que su proveedor informe de la existencia de un "huevo" (existencia de un error o defecto en un programa informático) después de que sus datos estén ya en proceso de publicación. No informe de los estadísticos encontrados en una salida de ordenador sin saber cómo han sido calculados o qué significan. No informe de los estadísticos con una precisión mayor de la que es soportada por sus simples datos porque ellos suelen ser impresos con esa precisión por el programa informático. Usar el ordenador es una oportunidad para que usted controle su análisis y su diseño. Si un programa de ordenador no provee el análisis que usted necesita, use otro, en vez de dejar al ordenador que modele su pensamiento.
4. Deberá comprobar que los supuestos exigidos por el análisis se están cumpliendo en los datos. Examine exhaustivamente los residuales. No utilice tests distribucionales e índices acerca de la forma de las distribuciones (curtosis...) como un sustituto del análisis grafico de los residuales.
5.Es difícil imaginar una situación en la que la decisión dicotómica aceptar-rechazar es mejor que informar acerca del valor p existente o, mejor aun, del intervalo de confianza. Nunca use la expresión desafortunada de " aceptar la hipótesis nula". Siempre que se informe del valor p, informe también del tamaño del efecto.
6. Informe siempre del tamaño del efecto de los datos primarios. Si la unidad de medida tiene significado a nivel práctico (p.e. número de cigarros fumados por día) es preferible utilizar una medida no estandarizada del tamaño del efecto (coeficiente de regresión o diferencia de medias) que una medida estandarizada (d o r). Añada algunos comentarios que sitúen el tamaño del efecto en su contexto teórico y práctico.
7. Se estimará el intervalo de confianza para cualquier tamaño del efecto de los efectos principales. Se aportara información sobre los intervalos de los coeficientes de correlacion y para cualquier otro coeficiente de asociación o variación que se analice.
8. Los resultados múltiples requieren un cuidado especial. Hay varias maneras de obtener inferencias razonables cuando nos enfrentamos a la multiplicidad (Bonferroni, tests multivariados, métodos bayesianos...).Es responsabilidad del autor definir y justificar el método aplicado.

El refranero dice que nunca es tarde si la dicha es buena. Aunque tarde, todavía es hora de poder tomar el tren de la reforma. Recientemente Finch, Cummings y Thomason (2001) se preguntaban: "*Why has (methodological) reform proceeded further in some other disciplines, including medicine, than in psychology? The American Medical Association' s ... sets out simple requirements for several important reform practices, including routine use of confidence intervals and evoke of ambiguity in use of statistical terms ... Note how radically different it is from current APA guidelines. What has happened in psychology was not inevitable. We leave to historians and sociologists of science the fascinating and important question of why psychology has persisted for so long with poor statistical practice*" (Págs., 205-206).

Los editores tienen la responsabilidad de ejercer de guardianes promotores de las reformas metodológicas necesarias para hacer progresar adecuadamente la ciencia psicológica, adoptando políticas estrictas que posibiliten el uso adecuado de las prescripciones metodológicas. Recientemente, Thompson (1999) en un esfuerzo de síntesis que le honra ha identificado al menos cinco errores metodológicos que ocurren habitualmente en el ámbito de la Psicología (Thompson se refiere de modo casi exclusivo a la Psicología Educativa, pero sus afirmaciones se pueden generalizar al resto de la Psicología):

"*... de manera harto habitual se cometen cinco errores metodológicos en la investigación educativa: a) uso de métodos "stepwise"; b) no atender en la interpretación de los resultados a la "especificidad de contexto" de los pesos analíticos (beta en la regresión, coeficientes factoriales, función discriminante, función canónica) que son parte de los análisis cuantitativos paramétricos; c) no interpretar los pesos y los coeficientes como parte de la interpretación de los resultados; d) no asumir que la fiabilidad es una propiedad de las puntuaciones y no de los tests; y (e) interpretación incorrecta de la significación estadística y ausencia de información sobre el tamaño de los efectos presentes en cualquier análisis estadístico*" (documento en red).

Probablemente sea el último de los enunciados el que más atención ha recibido y del único por el que nos hemos interesado en esta comunicación. Poner la cuestión sobre el tapete es solo el primer paso para inducir de inmediato lo que parece urgente: **el institucionalizar normativas de publicación actuales, complementarias o alternativas a la comprobación de hipótesis, para un mejor y estricto rigor metodológico.**

Bibliografía

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

Bakan, D. (1996). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437.

Bernia, J. (1979). Psicología experimental I. Valencia: La Nau Llibres.

Borenstein, M. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma and Immunology, 78,* 5-16.

Campbell, D.T. y Stanley, J.C. (1963). Experimental and Quasi-Experimental Desings for Research. En N.L. Gage (Ed). *Handbook of Research on Teaching.* Chicago, IL: Rand McNally

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

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

Cook, T.D. y Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field setting.* Chicago: Rand- NcNally.

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

Finch, S.; Cummings, G. y Thomason, N. (2001) Reporting of statistical inference in the Journal of Applied Psychology: little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

Guenther, R.K. (2002). How probable is the null Hypothesis? *American Psychologist, 45,* 67-68.

Lipsey, M.W. (1990). Design sensitivity: Statistical power for experimental psychology. Thousand Oaks, CA: Sage.

Lipsey, M.W. y Wilson, D.B. (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.

Morrison, D.E. y Henkel, R.E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Rothman, K.J. (1986). *Modern Epidemiology.* Boston: Little, Brown

Schmidt, F.L. y Hunter, J.E. (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow; S.A. Mulaik y J.H. Steiger (eds.), *What if there were no significance tests?* (pp.37-64). Mahwah, NJ: Erlbaum.

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

Shadish, W.R.; Cook, T.D. y Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company. Boston.

Thompson, B. (1999). Five Methodology Errors in Educational Research: A Pantheon of Statistical Significance and Other Faux Pas. En B. Thompson (ed.). *Advances in Social Science Methodology* (pp.23-86).

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