Propuestas alternativas o complementarias al contraste de hipótesis

José Luis Palacios Gómez (Universidad Complutense de Madrid)

Julio de 2007

En el texto de Nickerson *Null hypothesis significance testing: a review of an old and continuing controversy* (2000: 277ss) se recogen una serie de propuestas alternativas o complementarias al contraste de hipótesis. Fundamentalmente son ocho:

a) Proporcionar alguna indicación de la variabilidad o la precisión de las medidas
b) Informar de los intervalos de confianza alrededor de las estimaciones puntuales
c) Informar del tamaño del efecto
d) Utilizar la potencia de la prueba
e) Usar pruebas de "tres-resultados"
f) Usar estimación de parámetros y técnicas de ajuste de modelos
g) Demostrar la replicabilidad de los resultados
h) Utilizar métodos de inferencia bayesiana

a) La variabilidad debida al error de medida es un indicador que suele faltar en los informes de los experimentos psicológicos (especialmente en las representaciones gráficas de los mismos). Medidas de la variabilidad como la desviación típica son simplemente descriptivas de los datos que se usan pero dicen poco respecto de las poblaciones de las que se extraen las muestras. Pero existe una relación entre el error típico de la media y la desviación típica de la población de medias muestrales, de modo que se conoce que aproximadamente dos tercios de las medias de muestras aleatorias replicadas caen dentro de un rango de aproximadamente 1,4 veces el error estándar a cada lado de la media muestral (el error típico de la media es un estimador de la desviación típica de las medias muestrales alrededor de la media poblacional).

b) Los intervalos de confianza (como una forma de indicador de variabilidad) se han defendido enfáticamente como una de las principales alternativas (o al menos como complemento) al contraste de hipótesis. En algunos campos de investigación, y particularmente en los de investigación biomédica, parece haber un amplio consenso sobre que los intervalos de confianza constituyen una mucho mejor aproximación a la presentación e interpretación de los resultados que los test de significación (Altman et al., 2001: 10), argumentándose que "presentar solamente valores p puede hacer que se les dé más mérito que el que tienen realmente. Particularmente hay una tendencia a igualar significación estadística con importancia médica o relevancia biológica. Pero diferencias pequeñas (...) pueden ser estadísticamente significativas con muestras amplias, mientras que efectos (clínicos) importantes pueden ser estadísticamente no significativos sólo porque el número de sujetos estudiados es pequeño" (Altman et. al., 2001: 17).

Los intervalos de confianza se usan comúnmente para medias o para diferencias de medias, pero pueden utilizarse para calcular otros estadísticos (por ejemplo, intervalos de confianza alrededor de medias para diseños entre-sujetos, aunque esto es mucho menos frecuente). Una de las razones principales que se esgrimen para utilizar los intervalos de confianza en vez del test de hipótesis es que resultan más informativos que éste. Si se obtiene una estimación del tamaño del efecto, se calcula un intervalo de confianza (generalmente al 99%, 95% o 90%) para esa estimación, lo cual, se dice, es más preciso y menos ambiguo que un test de significación, aunque algunos autores han advertido sobre el riesgo de cometer los mismos errores de mala interpretación y uso inadecuado que aquejan al contraste de hipótesis (Abelson, 1997; Hayes, 1988). Así, por ejemplo, para trabajar con intervalos de confianza hay que asumir ciertos requisitos estadísticos (como en el contraste de hipótesis) relativos a la aleatoriedad y al tamaño muestral que si no se cumplen hacen muy poco preciso (o incluso erróneo) el uso de intervalos. Esto puede ser cierto, pero al menos en primera instancia el intervalo de confianza es más intuitivamente interpretable que el contraste de hipótesis, ya que un intervalo de confianza

de, por ejemplo, 95%, indica que si un investigador repitiese su estudio en las mismas condiciones pero con distintas muestras aleatorias, noventa y cinco de cada cien veces obtendría intervalos que contendrían el verdadero parámetro poblacional y cinco veces obtendría intervalos que no lo contendrían. En otras palabras, se puede tener gran confianza en que el intervalo resultante abarca el valor verdadero (en la población), pues dicho intervalo se ha obtenido por un método que casi siempre acierta. Esto no equivale a decir que hay una probabilidad de 95% de que el verdadero valor se encuentre dentro del intervalo, error de interpretación que es bastante común y que se suele citar como una de los más frecuentes entre los malos usos o incorrectas comprensiones de los intervalos de confianza.

Los intervalos de confianza han sido defendidos muy frecuentemente como la mejor alternativa al contraste de hipótesis, pero su escasa aparición en los artículos de las revistas científicas de ciencias humanas y sociales, como señala Kirk (1996), parece que tiene que ver tanto con lo "embarazoso" de la relativa amplitud del intervalo (sobre todo cuando se busca un alto porcentaje de confianza) como con que "lo que no está en SPSS no existe" (Sánchez-Bruno y Borges del Rosal (2005: 148), en referencia al cálculo de intervalos cuando el tamaño del efecto se mide con coeficientes de correlación. Abundando en esta cuestión, Reichardt y Gollob (1997) han apuntado varias fuentes potenciales de resistencia al uso de intervalos de confianza: a) tradición de usar contraste de hipótesis; b) falta de reconocimiento de las condiciones en que los intervalos de confianza son preferibles a las pruebas de significación estadística; c) relativa parvedad de los programas de ordenador y de fórmulas para calcular intervalos de confianza; d) frecuente pequeño tamaño de los estimadores de los parámetros –los intervalos revelan esto, mientras que las pruebas de significación no; e) decepcionante –por excesiva- amplitud de los intervalos; f) incertidumbre a la hora de elegir entre varios intervalos de los posibles y defender la elección hecha; g) errores en la crítica de las pruebas de significación, que dañan la credibilidad de los intervalos frente a aquéllas; h) rechazo a abandonar los test de significación porque se aprecia que siguen siendo útiles en la investigación inferencial.

c) Muchos investigadores recomiendan informar del tamaño del efecto junto con o en vez de los resultados de las pruebas de significación estadística (la propia *APA* recomienda tal cosa). Aunque el término "tamaño del efecto" parece referirse a los resultados de los experimentos, de hecho puede aplicarse a la magnitud de cualquier medida, tal como la diferencia entre medias o el grado de asociación entre dos variables, así como al impacto teórico o práctico de algún hallazgo científicos (no necesariamente de una intervención sobre cualquier ámbito, aunque también). Respecto del tamaño del efecto, Carver (1993: 291) ha señalado que es preferible informar sobre tamaño del efecto y error muestral que de un estadístico tal como *t*, que combina ambos (como hace efectivamente el contraste de hipótesis): "La significación estadística no nos dice nada directamente relevante sobre si los resultados que hemos hallado son grandes o pequeños, ni sobre si el error muestral es grande o pequeño. Podemos eliminar este problema informando sobre ambos separadamente". Sin embargo, informar sobre el tamaño del efecto no está completamente libre de problemas: por ejemplo, no siempre se sabe cuál de los varios posibles estimadores del tamaño del efecto es el más apropiado en un contexto determinado, ni si se puede generalizar sin más el tamaño del efecto a otras poblaciones diferentes en algo a la estudiada. Tampoco es fácil determinar si un efecto grande es mayor o menor garantía de una importancia teórica o práctica que una *p* pequeña en una prueba de hipótesis. Algunos autores han señalado que el tamaño del efecto conviene complementarlo con la prueba de hipótesis (Hagen, 1998, por ejemplo) o que es recomendable determinar primero si un efecto es estadísticamente improbable y luego, si lo es, informar de su tamaño (Robinson y Levin, 1997); también se puede determinar previamente qué tamaño mínimo del efecto puede aceptarse para el asunto investigado y luego hacer una prueba de hipótesis para saber si ese tamaño es además estadísticamente significativo (Fowler, 1985). Por otro lado, también hay quien afirma que el tamaño del efecto debería calcularse independientemente de que *p* señale significación estadística (Rosnow y Rosenthal, 1989). En sentido contrario, otros autores (como Chow, 1996) han señalado que informar sobre el tamaño del efecto puede ser importante cuando un investigador está interesado en un efecto experimental *per se* y

el propósito del experimento es determinar si ese efecto es de suficiente tamaño como para mostrar importancia práctica (experimento utilitario), pero que no es relevante cuando el propósito del experimento es probar las implicaciones de una teoría explicativa (experimento de corroboración de teoría). En este último caso, Chow dice que la única consideración relevante es si el resultado obtenido es consistente con lo que predice la teoría, dentro del criterio representado por el nivel alpha, y que dar énfasis al tamaño del efecto en este contexto puede ser un error. Chow también apunta que obtener un resultado que es consistente en este sentido estadístico con las implicaciones de una teoría ciertamente no prueba que la teoría sea cierta, pero arguye que ello constituye una apoyatura a la sostenibilidad de la misma.

d) Parece estar ampliamente admitido que cuando se quiere concluir, a partir del fracaso de una prueba para establecer la significación estadística, que no hay un efecto o que cualquier efecto que hubiese es tan pequeño que resulta despreciable, se debería hacer si es posible una prueba de potencia y solamente si la potencia de la prueba resultase alta se podría considerar cierta la hipótesis nula. Se admite también que una razón hipotética para explicar la falta de atención a la potencia por parte de los investigadores es precisamente la dificultad para realizar tal prueba. Gran parte de la discusión sobre las pruebas de potencia en la literatura psicológica se ha centrado sobre la cuestión de cuánto de amplia tiene que ser una muestra para que un determinado efecto resulte estadísticamente significativo y sobre el problema que tiene el investigador para determinar una muestra suficientemente amplia para detectar un efecto si efectivamente lo hay. Sin embargo, algunos autores han señalado que es importante considerar la potencia en los experimentos no solamente para detectar los efectos que se buscan, sino también para evitar detectar efectos tan pequeños que resulten despreciables. Si un investigador está interesado en un efecto solamente si éste es suficientemente grande y robusto –detectable incluso con una muestra pequeña- entonces realizar una prueba de potencia no es propiamente un requisito e incluso puede ser una técnica indeseable: la importancia de la prueba de potencia depende de los propósitos del

investigador y sólo deben realizarse con criterio, no automáticamente. Shaver (1993: 309) ha criticado la utilización de las pruebas de potencia sobre la base de que su propósito principal es determinar el tamaño de la muestra que se necesita para arrojar un resultado estadísticamente significativo de un efecto de una cierta magnitud y que no tienen sentido si la significación estadística es de poca entidad: “Si los tamaños del efecto son importantes porque la significación estadística no es un indicador adecuado de la magnitud del resultado, ¿por qué hacer el juego de ajustar las especificaciones de la investigación de tal manera que se obtenga un resultado estadísticamente significativo si se ha obtenido un tamaño del efecto especificado de antemano?”. Como todas las herramientas para hacer análisis estadísticos, el test de potencia tiene sus limitaciones y la probabilidad de que un experimento arroje un resultado estadísticamente significativo no depende solamente del tamaño del efecto y del tamaño muestral sino también de la variabilidad interna de los datos; ésta puede ser influenciada por muchos factores, entre los cuales cabe destacar el efecto de las variables intervinientes no controladas, y los test de potencia no pueden captar este hecho.

e) Algunos investigadores han propuesto substituir las tradicionales pruebas de significación de una y dos colas por las pruebas de “tres resultados”. Una prueba de una cola no permite determinar que un efecto es estadísticamente significativo in la dirección opuesta a la hipotetizada y que una prueba de dos colas no justifica sacar conclusiones sobre la dirección del efecto. La prueba de “tres resultados” puede entenderse como una combinación de las de una y dos colas, con la cual se puede decidir si la media 1 es menor que la media 2, que la media 1 es mayor que la media 2 o que la dirección de la diferencia de medias (si la hay) entre tres tamaños es indeterminada. Harris (1997) argumenta que muchos de los errores de interpretación de las pruebas de hipótesis se deben a que parecen conducir a una decisión entre dos hipótesis, algo que podría evitarse utilizando la prueba de “tres resultados” (alejaría, por ejemplo, el riesgo de interpretar el no rechazo de la H_0 como una aceptación de la misma). La prueba de “tres resultados” proporciona un análisis más preciso de los resultados que las

pruebas tradicionales, pero no evita muchos de los problemas que están asociados a éstas y no han obtenido mayor popularidad entre los investigadores.

f) La estimación de parámetros y los ajustes de modelos son otras alternativas a las pruebas de hipótesis. Cuando se utilizan estas técnicas, uno no se pregunta si dos muestras difieren estadísticamente en algún modo específico o más que de ese modo específico, sino más bien se intenta estimar el valor de algún parámetro de una población o determinar cuánto de bien un modelo predice el valor de una o más variables experimentales. Pero con la primera técnica lo que se logra es establecer el grado de objetividad de una medida y no decidir si se acepta o se rechaza una teoría para trabajar con ella o para mejorarla. Con respecto a la segunda técnica, se ha defendido (Granaas, 1998) que el ajuste de modelos no sólo es más potente que las pruebas de hipótesis sino también que es una técnica más fácil de aprender y utilizar: un determinado modelo se mantiene mientras que se ajusta bien a los datos y cuando deja de hacerlo se sustituye por otro. A pesar de que la estimación de parámetros y el ajuste de modelos se contemplan como alternativas a las pruebas de hipótesis, lo cierto es que algunas formas de pruebas de hipótesis juegan un importante papel en algunas medidas de bondad del ajuste. Además, la bondad del ajuste no es el único criterio para evaluar un modelo y va contra el principio de parsimonia aumentar el número de parámetros de un modelo para lograr un ajuste óptimo.

g) La “replicabilidad” se considera usualmente como una condición *sine qua non* para considerar científico un hallazgo experimental. Sin embargo, la reproducción exacta de todas las características y condiciones de un experimento es muy infrecuente. A pesar de ello, la replicación de resultados se ha considerado como una buena, sino la mejor, alternativa a las pruebas de hipótesis. Autores como Carver (1993) defienden que los mejores artículos de la literatura científica son aquellos que incluyen no una prueba de significación estadística, sino los que la remplazan por una replicación de resultados, aunque otros, como Lykken (1968) han visto la

replicación como un ideal impracticable. Por otro lado, Llobel et al. (2000) han señalado que el hecho de obtener un valor de p bajo o muy bajo (por debajo del nivel alpha estándar) no es un predictor concluyente de la replicabilidad de los datos, pues aunque la probabilidad de cometer un error de Tipo I sea muy pequeña no implica que precisamente en esa ocasión no se haya cometido dicho error. Así pues, la replicación entendida como obtener de nuevo un resultado estadísticamente significativo es también una importante salvaguarda contra el error de Tipo I: obtener una significación para un nivel de 0,05 no evita la posibilidad de que haya intervenido en ello el mero azar, pero volver a obtener el mismo resultado en una replicación aleja obviamente ese riesgo. Evidentemente, a más replicaciones con resultados de significación estadística semejantes, menor probabilidad de cometer un error de Tipo I.

h) Muchos autores han defendido que se remplace o que se complemente la prueba de significación con la utilización de razones de probabilidad y la estimación de probabilidades posteriores con la aplicación de la regla de Bayes. En teoría esta posibilidad tiene mucho de recomendable, ya que la evaluación de hipótesis bayesiana proporciona una prueba para fortalecer la hipótesis nula o su alternativa. Así, dado un conjunto de modelos competidores sobre algún proceso de interés, si se puede especificar para cada modelo una probabilidad a priori, y siendo cierta la probabilidad del resultado condicional, entonces se puede usar la regla de Bayes para calcular la probabilidad de ocurrencia de cada modelo y determinar cuál es el que tiene mayor probabilidad de ocurrencia a posteriori. Otra gran ventaja del método bayesiano de evaluación de datos es que constituye un procedimiento para acumular los efectos de evidencia (prueba) en determinados estudios a lo largo del tiempo (las probabilidades posteriores de un estudio pueden convertirse en las a priori del siguiente, y etcétera). El problema al respecto reside en que gran parte de las investigaciones no se hacen bajo la forma de una especie de serie temporal, sino aisladamente, de manera que muy frecuentemente no se puede seguir esa dinámica de pruebas bayesianas encadenadas: si no se pueden establecer probabilidades a priori que descansen sobre experiencia acumulada, la

probabilidad a priori es más bien subjetiva y el método pierde gran parte de su valor. Por ello, algunos autores como Wilson, Miller y Lowler (1967: 191) defienden el uso del método bayesiano cuando se dispone de la información apriorística necesaria y la prueba de hipótesis cuando no se dispone de ella: “Si tenemos una alternativa claramente definida, y podemos decir que la ´realidad´ debe ser una u otra, entonces podemos justificar una razón de probabilidad o algún procedimiento similar. Si no tenemos tales modelos alternativos, no podemos inventarlos para evitar errores teoréticos”. Otros autores han puesto reparos de orden filosófico al uso del método bayesiano (como el propio Fisher), esencialmente sobre la base de que no se puede construir ciencia a partir de creencias o convencimientos subjetivos. También se ha argumentado que el método bayesiano y el contraste de hipótesis son mutuamente excluyentes para evaluar datos, aunque Nickerson (2000: 285) opina que esto no es necesariamente cierto puesto que ambos métodos tienen sus fortalezas y debilidades, de manera que en situaciones distintas uno u otro puede resultar el mejor y que en todo caso pueden combinarse para lograr el mejor rendimiento posible.

De todo lo anterior cabe concluir que las pruebas de significación estadística ciertamente no son una panacea que constituya la referencia única, ni siquiera fundamental, para conceder valor (representatividad) a los resultados de una investigación. Su uso indiscriminado, su utilización en exclusiva para otorgar relevancia general a los hallazgos de los estudios científicos sobre las más diversas cuestiones, y seguramente también las interpretaciones erróneas que se hacen de su fundamentación teórica y de su función y significado práctico, son cuestiones que muy probablemente han originado muchas de las críticas y rechazos que han suscitado entre la comunidad científica (particularmente en las ciencias humanas y de la salud). Algunas de estas críticas están motivadas porque se les quiere dar funciones que no tienen (como señala Chow en varias ocasiones en las obras que hemos consultado, la utilidad más relevante que tienen las pruebas de hipótesis es la de conseguir un razonable nivel de seguridad respecto de que los resultados de una investigación con muestras no son mera casualidad), pero otras parecen estar bien razonadas: muestras

demasiado pequeñas pueden hacerlas poco útiles a efectos de “clínica”, muestras demasiado grandes tienden a producir significatividad “artificial”, muestras no aleatorias las vuelven inútiles a efectos de generalización, etc. En nuestra opinión, todo parece indicar que las pruebas de hipótesis son ciertamente una herramienta útil en la investigación con componente inferencial, pero que conviene que se complemente con otras herramientas (algunas o todas de las descritas más arriba) para dar mayor robustez a la inferencia. No parece que estas herramientas, las pruebas de significación incluidas, sean excluyentes, sino más bien que su articulación o combinación resulta conveniente siempre que se pueda llevar a cabo: su uso conjunto arrojará más luz sobre los resultados de la investigación y hará más confiable la inducción a las poblaciones de las que se extraen los sujetos o grupos “experimentales”. Es cierto que puede darse el caso de que dos o más técnicas inferenciales usadas conjuntamente pueden arrojar resultados contradictorios (por ejemplo, que una diferencia entre medias sea estadísticamente significativa para un nivel alpha estándar pero que caiga fuera del intervalo de confianza correspondiente), y que tal circunstancia puede generar confusión e incertidumbre en el investigador, pero también lo es que esta situación es mejor, en términos científicos, que tener certezas infundadas.

REFERENCIAS

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psichological Science, 8*, 12-15, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 278.

Altman, D.G., Machin, D., Bryant, T.N. y Gardner, M.J. (2001). *Statistics with confidence*. Bristol: British Medical Journal Books.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 281.

Chow, S.L. (1996). *Statistical Significance. Rationale, validity and utility*. London: Sage Publications.

Fowler, R.L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology, 70*, 215-218, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 281.

Granaas, M.M. (1998). Model fitting: A better approach. *American Psychologist*, 53, 800-801, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 283.

Hagen, R.L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist*, 52, 15-24, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 281.

Harris, R.J. (1997). Significance tests have their place. *Psychological Science*, 6, 8-11, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 283.

Hayes, A.F. (1988). Reconnecting data análisis and research design: Who needs a confidence interval?. *Behavioral and Brain Science*, 21, 203-204, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 278.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 279.

Llobell, J.P., García Pérez, J.F. y Frías, M.D. (2000). Significación estadística, importancia del efecto y replicabilidad de los datos. *Psicothema, 12*, supl. nº2, 408-412.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psichological Bulletin*, 70, 151-159, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 283.

Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 241-301.

Reichardt, C.S. y Gollob, H.F. (1997). When confidence intervals should be used instead statistical test, and vice versa. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.) ((pp. 259-284). *What if there were no significance test?*. Hillsdale (NJ): Erlbaum, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 279.

Robinson, D. y Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 269 (5), 21-26, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 281.

Rosnow, R.L y Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 281.

Sánchez-Bruno, A. y Borges del Rosal, A. (2005). Transformación Z de Fisher para la determinación de intervalos de confianza del coeficiente de correlación de Pearson. *Psicothema, 17* (1), 148-153.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 282.

Wilson, W., Miller, H.L. y Lower, J.S. (1967). Much ado about the null hypothesis. *Psychological Bulletin*, 68, 188-196, citado por Nickerson R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 2*, 285.