Metodología de las Ciencias del Comportamiento
2002. Volumen Especial

ISSN 1575-9105

Significación y tamaño de efecto: Claves para concluir ¿sobre que?

Trigo, M.E. y López, J.
Universidad de Sevilla

Se exploró el nivel de competencia de un grupo de docentes y estudiantes de Psicología para interpretar de forma conjunta indicadores de tamaño muestral, valor *p* asociado al estadístico de contraste y *r* como índice de tamaño de efecto. Mediante un diseño factorial 2x2x2 de medidas repetidas se crearon ocho situaciones donde dichos indicadores variaban en su grado de coherencia y relevancia para el establecimiento de conclusiones. Los resultaros indicaron que, a pesar de tenerse en cuenta los diversos indicadores, sigue existiendo un sesgo de sobrevaloración del factor significación.

We studied the competence of a group of Psychology teachers and students to evaluate different research outcomes through the effect size, the statistical significance and the sample size used in the study. Subjects had to evaluate eight outcomes, resulting from a 2x2x2 repeated measures design, differing on their coherence and relevance to conclude. The results showed that, even when the three indexes were used, the statistical significance is still judging as the most important.

En los últimos años ha tenido lugar un intenso debate en torno a las pruebas de significación de hipótesis nula que ha puesto de manifiesto la conveniencia de tener en cuenta índices adicionales a la significación estadística para concluir sobre los resultados de un estudio (e.g. Cohen, 1990, 1994; Nickerson, 2000). Se reivindica así la utilización de otros indicadores como el tamaño de efecto y/o la potencia de prueba del estadístico (*APA*, 1994; Wilkinson y *TFSI*, 1999; Vacha-Haase, 2001). Durante estos años se ha publicado abundante material teórico para la reflexión sobre aspectos como la potencia de análisis, así como revisiones de la prácticas metodológicas actualmente utilizadas por los investigadores (Chow, 1998; Harlow, Mulaik y Steiger, 1997; Hubbard y Ryan, 2000; Kirk, 1996, 2001; Keselman *et al.*, 1998; Krueger, 2001). Estos trabajos nos ofrecen un amplio abanico de posibilidades en cuanto a la defensa de unos u otros indicadores, así como en cuanto a la radicalidad de dicha defensa. El objetivo de este estudio es aportar información complementaria sobre cómo utilizan los docentes en el campo de la Psicología estos índices adicionales a la significación estadística para el establecimiento de sus conclusiones y si transmiten estas prácticas a sus alumnos. Más concretamente, se trata de determinar cómo utilizan nuestros alumnos y compañeros de profesión docente los índices de tamaño muestral, nivel de significación y tamaño de efecto a la hora de evaluar la consistencia, validez y relevancia de los resultados obtenidos.

Método

Sujetos

Un total de 13 profesores de la Facultad de Psicología de la Universidad de Sevilla y 56 estudiantes de cuarto curso de esta licenciatura participaron voluntariamente en el estudio.

Materiales

A cada uno de los participantes se les entregó un conjunto de ocho cuestionarios. En cada uno de ellos, tras describir brevemente una determinada situación de estudio con sus correspondientes resultados, debían indicar su grado de acuerdo con cinco afirmaciones sobre los resultados del estudio. El grado de acuerdo debía expresarse a través de una escala tipo likert, donde el nivel 1 indicaba el menor grado de acuerdo y el nivel 5 el máximo. El anexo I describe una de las ocho situaciones de estudio utilizadas y el cuestionario común para todas ellas. Se presentó en todos los casos el mismo estudio bicondicional, con un grupo control y otro experimental, donde se trataba de evaluar el efecto de un tratamiento terapéutico. Las cinco afirmaciones del cuestionario estaban relacionadas con la obtención de resultados estadísticamente significativos (ítem 1), la obtención de índices compatibles entre sí (ítem 2), la posibilidad de concluir de forma clara independientemente del sentido de la conclusión (ítem 3), la posibilidad de concluir en un sentido particular, sobre la existencia de un efecto clínico relevante (ítem 4) y la relación entre la calidad de los resultados y el coste experimental (ítem 5).

Procedimiento

Las ocho situaciones de estudio planteadas (ver Tabla 1) correspondían a las combinaciones cruzadas de un diseño factorial de medidas repetidas 2x2x2. Los factores manipulados fueron el tamaño muestral del estudio, con N=40 y N=14 como valores muestrales grande y pequeño respectivamente; el valor de probabilidad asociado al estadístico de prueba, con p>.10 y p<.01 como niveles de significación liberal y conservador; y el tamaño de efecto, con r=.10 y r=.50 como tamaños pequeño y grande. Todos los sujetos debían responder el mismo cuestionario sobre cada una de las situaciones de estudio, que trataban de plasmar diferentes grados de coherencia y relevancia de los resultados planteados. Así, las situaciones 3 y 8 resultarían muy improbables debido a que contienen elementos contradictorios. Por su parte, las situaciones 4 y 7 representan una falta de consistencia entre los índices de significación estadística y tamaño efecto. Las situaciones 1 y 2 son consistentes en cuanto a significación y tamaño efecto, si bien en la primera de ellas se cuenta con poca potencia, de ahí la dificultad para concluir. Finalmente, en las situaciones 5 y 6 también apuntan ambos índices en la misma dirección, a pesar de la poca potencia que puede darse en la situación 5.

Tabla 1. Situaciones de estudio evaluadas por cada uno de los sujetos.

	p>.10		p<.01	
	n=7	n=20	n=7	n=20
r=.10	(1) Coherente No concluyente	(2) Coherente No relevante	(3) Improbable No coherente	(4) Significación de efectos no relevantes
r=.50	(7) Efectos relevantes no detectados por poca potencia	(8) Improbable No coherente	(5) Efectos relevantes detectados	(6) Ideal Validez interna y externa

En resumen, los sujetos del estudio debían evaluar el grado de coherencia y relevancia y las posibilidades de concluir en cada una de

Correspondencia: María Eva Trigo Sánchez
Universidad de Sevilla
E-mail: trigo@psicoexp.us.es

las situaciones, de forma que pudiésemos analizar las variaciones en sus respuestas en función de los factores manipulados.

Resultados

Se analizó en primer lugar la adecuación de los promedios a lo predicho teóricamente en función de las combinaciones analizadas. En caso de producirse dicha adecuación se procedió a analizar los efectos principales e interactivos de primer y segundo orden correspondientes al AVAR factorial de medidas repetidas. En la descripción de los resultados para cada ítem se destacarán como relevantes aquellos componentes de variación sistemática que cumplan tres criterios: una probabilidad asociada a su correspondiente F inferior a .05, una potencia observada superior a .80, y una interpretación no condicionada por otros componentes de orden superior (por ejemplo, sólo enunciaremos un efecto principal como relevante si dicho factor no está presente en ninguna interacción que lo sea).

Item 1: El efecto de la terapia fue estadísticamente significativo

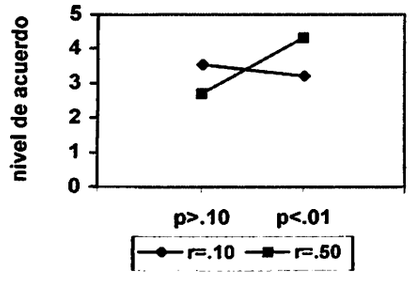
En esta primera cuestión esperábamos encontrar promedios de acuerdo superiores en las situaciones 3, 4, 5 y 6, donde $p<.01$. En la muestra de profesores la adecuación a esta predicción fue casi absoluta. En el AVAR ningún efecto interactivo resultó relevante, pero sí el efecto principal del factor probabilidad (con promedios de 1.21 y 4.66 para las condiciones liberal y conservadora respectivamente).

En la muestra de alumnos también se replicó lo predicho, aunque las diferencias entre los dos grupos de condiciones fueron muy pequeñas. Sólo 10 de los 56 alumnos que compusieron la muestra respondieron con valores altos (4 ó 5) en las condiciones $p<.01$, por lo que decidimos analizar los ítemes posteriores exclusivamente en dichos alumnos. Con esta muestra más reducida los promedios de acuerdo en función de los valores p fueron muy similares a los de los profesores (1.20 versus 4.82). Este fue el único factor que resultó relevante en el AVAR.

Item 2: Los valores de los tres índices aportados (n, p y r) son altamente compatibles entre sí

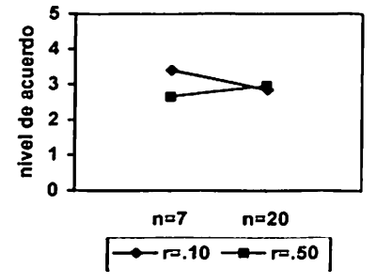
En relación con este ítem sería de esperar el mayor grado de acuerdo en las condiciones 5 y 6 por una parte y 1 y 2 por otra, y el menor grado de acuerdo en las restantes, especialmente 3 y 8. En el grupo de profesores se obtuvieron los mayores promedios en las condiciones 1, 5 y 6. Sin embargo, el promedio de acuerdo en la situación 2, con tamaño muestral grande, probabilidad alta y tamaño de efecto pequeño, fue similar al de la situación 4, donde el efecto resultaba estadísticamente significativo. En el AVAR factorial de medidas repetidas resultó relevante la interacción entre probabilidad y tamaño de efecto, no condicionada por la interacción de segundo orden. Tal y como muestra la Figura 1, cuando el tamaño de efecto es grande, el nivel de acuerdo aumenta al disminuir el nivel de probabilidad (4.32 respecto a 2.71), mientras que con el menor tamaño de efecto la influencia del nivel de probabilidad es menor y en sentido contrario (3.21 respecto a 3.54).

Figura 1. Probabilidad x tamaño de efecto en profesores para el ítem 2.



En cuanto a los alumnos, sólo respondieron adecuadamente ante la situación más clara, la situación 6. En el AVAR resultó relevante la interacción entre el tamaño de efecto y el tamaño muestral (ver Figura 2). Más concretamente, la influencia del tamaño de efecto fue mayor con n=7 (3.40 versus 2.65 para los tamaños de efecto menor y mayor respectivamente) que con n=20 (2.95 versus 2.85). También resultó significativo el efecto principal de la probabilidad asociada al estadístico, con un mayor grado de acuerdo ante la probabilidad menor (3.5 versus 2.42).

Figura 2. Tamaño muestral x tamaño de efecto en alumnos para el ítem 2.

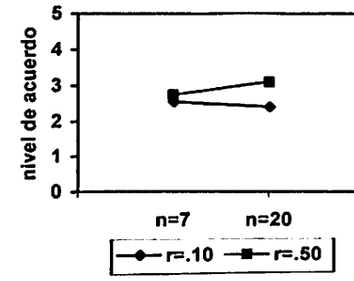


Item 3: Los valores de los tres índices aportados (n, p y r) permiten concluir claramente sobre la existencia o no de un efecto clínico relevante de la terapia

Tanto en las situaciones 5 y 6, donde se puede inferir la presencia de un efecto relevante, como en las condiciones 1 y 2, donde puede inferirse su ausencia, deberían producirse los mayores grados de acuerdo. Los profesores respondieron adecuadamente cuando sí existía el efecto (situaciones 5 y 6), pero no mostraron un alto grado de acuerdo a la hora de establecer conclusiones sobre la ausencia del mismo (situaciones 1 y 2). En el AVAR ninguna interacción resultó relevante, pero sí los tres efectos principales. Más concretamente, los mayores promedios en cuanto al acuerdo con el ítem se dieron en la condición de menor probabilidad (3.64 versus 2.09), mayor tamaño muestral (3.20 versus 2.54) y mayor tamaño de efecto (3.27 versus 2.46).

En la muestra de alumnos se replicaron estos resultados. No obstante, en el AVAR resultó relevante la interacción entre tamaño muestral y tamaño de efecto (ver Figura 3). La influencia del tamaño de efecto fue superior con n=20 (con promedios de 3.10 y 2.40 para el mayor y menor tamaño de efecto respectivamente) que con n=7 (2.75 versus 2.55). También resultó relevante el efecto principal de la probabilidad, con promedios de acuerdo superior en la condición de probabilidad más baja (3.02 respecto a 2.37).

Figura 3. Tamaño muestral x tamaño de efecto en alumnos para el ítem 3.



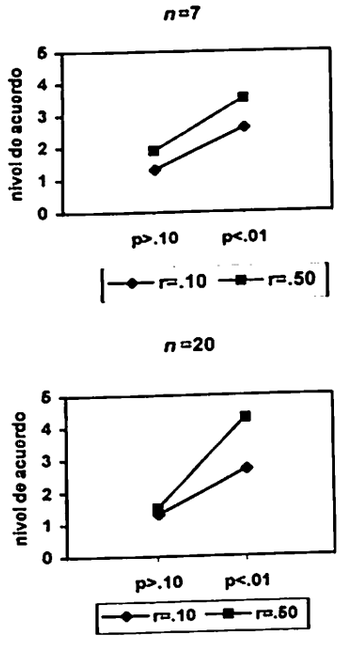
Item 4: La terapia probablemente tiene un efecto clínico relevante

Los mayores promedios de acuerdo son esperables en las situaciones 5 y 6, aunque en este caso habría que añadir también la situación 7, con tamaño de efecto grande pero que no alcanza la significación estadística debido al escaso tamaño muestral. Los promedios menores deberían darse en las situaciones 1 y 2. En la muestra de profesores las respuestas a las situaciones 1 y 2 por un lado y 5 y 6 por el otro se adecuaron a la predicción, pero no ocurrió lo mismo en relación con la situación 7. En el AVAR resultaron relevantes los efectos principales de la probabilidad (3.20 versus 1.98 para el menor y mayor nivel de ésta respectivamente) y el tamaño de efecto (3.30 versus 1.87 para el mayor y menor tamaño de efecto respectivamente).

En la muestra de alumnos los resultados fueron similares en relación con los promedios tanto de las situaciones 5 y 6 como de las situaciones 1 y 2. En el AVAR destaca el efecto interactivo de segundo orden,

siendo la interacción entre probabilidad y tamaño de efecto más fuerte con el mayor tamaño muestral (ver Figura 4). Más concretamente, con el menor tamaño muestral (n=7) las diferencias entre ambos niveles de probabilidad son muy similares para ambos tamaños de efecto (1.90 versus 3.50 para el menor respecto a 1.30 versus 2.60 para el mayor). Con el tamaño muestral mayor (n=20), la diferencia entre ambos niveles de probabilidad es similar a las anteriores para el tamaño de efecto menor (1.30 versus 2.70), pero resulta más pronunciada para el mayor tamaño de efecto (1.50 versus 4.30). La existencia de esta interacción de segundo orden condiciona la interpretación de otros efectos significativos en el AVAR, la interacción entre probabilidad y tamaño de efecto, que depende como acabamos de ver del tamaño muestral, y los efectos principales de ambos factores.

Figura 4. Interacción de segundo orden en alumnos para el ítem 4.



Item 5[1]*: Tasa calidad de resultados / coste experimental*

La situación 5 debería ser juzgada como ideal, al obtenerse un tamaño de efecto suficientemente grande como para que resulte significativo con un tamaño muestral relativamente menor. Lo contrario ocurre con la situación 2. Sin embargo, el comportamiento de los sujetos, profesores y alumnos, no se ajustó a esta predicción, por lo que no tendría mucho sentido llevar a cabo un análisis inferencial.

Discusión y conclusiones

El ítem 1 es una afirmación habitual sobre significación estadística de efectos. Por tanto, era esperable que tanto alumnos como profesores pudiesen adecuar correctamente sus respuestas a la predicción. Esta adecuación serviría como referente para al análisis de los demás ítemes. Sin embargo, una gran cantidad de alumnos (80%) tuvo dificultades para proporcionar la respuesta adecuada en términos de significación estadística del efecto. Entre los posibles factores responsables de este resultado podría mencionarse la reactividad de la situación de prueba experimental, el momento temporal específico de realización de la tarea o el hecho de que la transmisión de conocimiento no sea todo lo fructífera que debiera.

El ítem 2 es un enunciado global sobre la compatibilidad entre los diversos índices aportados. En los profesores, esta compatibilidad se evalúa en función de los índices de tamaño de efecto y significación estadística. Así, un tamaño de efecto grande les resulta muy compatible con la significación estadística y muy poco con la no significación, mientras que un tamaño de efecto pequeño es igualmente compatible con ambas. De acuerdo con la predicción, los índices que se evalúan más compatibles entre sí son un tamaño de efecto grande con una probabilidad pequeña, seguido de un tamaño de efecto pequeño con una probabilidad grande. A diferencia, los alumnos evaluaron la compatibilidad entre los índices aportados en función de la probabilidad por un lado y de la combinación entre tamaño de efecto y tamaño muestral por el otro. Los índices evaluados como más incompatibles entre sí resultan ser un tamaño de efecto grande a partir de un tamaño muestral pequeño, independientemente de la significación estadística. No obstante, el resultado más sorprendente hace referencia a la utilización de un índice único, la significación estadística (efecto principal del factor probabilidad) para evaluar la compatibilidad entre diversos índices. Se muestra así un sesgo hacia la sobrevaloración de la importancia de la significación estadística, al que deberá ponerse remedio a través del proceso de enseñanza-aprendizaje.

Los ítemes 3 y 4 eran similares en el sentido de que ambos plantean concluir sobre el efecto a nivel práctico, el efecto *clínico*, y por tanto, no se ciñen exclusivamente a la significación estadística de dicho efecto. Mientras que con el primero se pretende evaluar la posibilidad o no de concluir claramente sobre la existencia o ausencia de dicho efecto, en el segundo es necesario evaluar exclusivamente la posibilidad de concluir que el efecto clínico es relevante. En los profesores los únicos efectos relevantes en ambos índices tenían la categoría de efectos principales, de forma que tienen en cuenta índices adicionales a la significación estadística para concluir sobre un posible efecto clínico, pero sin tener en cuenta combinaciones concretas de los valores de dichos índices. A diferencia, en los alumnos aparecen como relevantes efectos interactivos de difícil interpretación.

Tanto los profesores como los alumnos mostraron los mayores grados de acuerdo a la hora de concluir sobre la existencia de un efecto clínico relevante cuando el tamaño de efecto era grande y la probabilidad pequeña (5 y 6). Por el contrario, no estuvieron de acuerdo en la posibilidad de concluir que un efecto clínico no es relevante cuando el tamaño de efecto es pequeño y la probabilidad es grande (situaciones 1 y 2 en el ítem 3), ni en afirmar la posible relevancia del efecto clínico cuando el tamaño de efecto es grande aunque no significativo al usar un tamaño muestral pequeño (situación 7 en el ítem 4). Nuevamente se pone de manifiesto con ello la sobrevaloración de la significación estadística para concluir sobre la relevancia o no del efecto encontrado.

Sobre el ítem 5 nos interesa resaltar que sean cuáles sean los factores que expliquen la confusión que presentan las respuestas de los participantes, es infrecuente usar en investigación experimental un indicador económico que plantee la calidad en términos de coste, y por tanto, puede haberse dado una fuerte reactividad a la tarea experimental.

Finalmente, a modo de síntesis del conjunto de resultados obtenidos, hemos de destacar el arduo camino que implicará modificar la práctica habitual de utilizar una decisión dicotómica de rechazar o no *Ho* para extraer conclusiones sobre cualquier estudio, y comenzar a concluir teniendo en cuenta un conjunto de indicadores estadísticos.

1 Fue un ítem que provocó múltiples comentarios, tal vez porque la redacción es poco clara y/o resulta extraño utilizar términos económicos de relación calidad/precio para definir la idoneidad de un estudio.

Referencias

American Psychological Asssociation (1994). *Publication manual of the American Psychological Association (4th ed.).* Washington, D.C.: Author.

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

Cohen, J. (1994). The earth is round (p<0.05). *American Psychologist, 49,* 997-1003.

Chow, S.L. (1998). Précis of Statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21,* 169-239.

Harlow, L., Mulaik, S. y Steiger, J. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hubbard, R. y Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology -And its future prospects. *Educational and Psychological Measurement, 60,*661-681.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C. y Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350-386.

Kirk, R. (1996). Practical significance: A concept who time has come. *Educational and Psychological Measurement, 56,*746-759.

Kirk, R. (2001). Promoting good statistical practices: some suggestions. *Educational and Psychological Measurement, 61,*213-218.

Krueger, J. (2001). Null Hypothesis Significance Testing. On the Survival of a Flawed Method. *American Psychologist, 56,* 16-26.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of and old and continuing controversy. *Psychological Methods, 5,* 241-301.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61,*219-224.

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

Anexo I

Se evaluó el efecto de una terapia para fomentar la conducta de interacción social. Delimitada una población con un nivel de interacción social bajo, se seleccionó una muestra aleatoria de sujetos que fueron asignados por azar a uno de los siguientes grupos:

a_1: control (ausencia de terapia)

a_2: experimental (presencia de terapia)

Una vez aplicado el tratamiento experimental y recogidos los datos sobre interacción social se procedió al análisis de los mismos. Tras comprobar el cumplimiento de los supuestos del modelo de análisis, se aplicó una *t* de Student como prueba de significación estadística con α=.05. Se calculó además el estadístico *r* como índice de tamaño de efecto, considerando los valores de .10, .30 y .50 como tamaños de efecto pequeño, medio y grande respectivamente. La fórmulas correspondientes son:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{S_e\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad r = \sqrt{\frac{t^2}{t^2 + gl_e}}$$

Supongamos que el tamaño muestral por grupo, la probabilidad asociada a la *t* de Student empírica y el tamaño de efecto hubiesen sido:

$$n = 7;\ p > .10;\ r = .10$$

Evalúa las afirmaciones que se hacen a continuación sobre los resultados obtenidos mediante la siguiente escala:

1	2	3	4	5
nada de acuerdo	poco de acuerdo	medianamente de acuerdo	bastante de acuerdo	plenamente de acuerdo

	1	2	3	4	5
El efecto de la terapia fue estadísticamente significativo					
Los valores de los tres índices aportados (*n*, *p* y *r*) son altamente compatibles entre si					
Los valores de los tres índices aportados (*n*, *p* y *r*) permiten concluir claramente sobre la existencia o no de un efecto clínico relevante de la terapia					
La terapia probablemente tiene un efecto clínico relevante					
La relación calidad de resultados/coste experimental es muy elevada					