ERRORS D'INTERPRETACIÓ DELS VALORS *P* EN ESTUDIANTS UNIVERSITARIS DE PSICOLOGIA

ERRORES DE INTERPRETACIÓN DE LOS VALORES *P* EN ESTUDIANTES UNIVERSITARIOS DE PSICOLOGÍA

MISINTERPRETATIONS OF P *VALUES IN PSYCHOLOGY UNIVERSITY STUDENTS*

Laura Badenes-Ribera[*]
Dolores Frías-Navarro[*]
Marcos Pascual-Soler[**]

Doi: 10.7203/anuari.Psicologia.16.2.15

Resum

Antecedents. Informar sobre els valors *p* de les proves de significació estadística és comú en la literatura empírica de Psicologia. No obstant això, les interpretacions incorrectes del valor *p* són abundants i repetitives. Aquests errors afecten les decisions professionals i posen en risc la qualitat de les intervencions psicològiques i l'acumulació d'un coneixement científic vàlid. L'objectiu d'aquest estudi és descriure els errors d'interpretació del valor *p* i la seua interpretació correcta en els estudiants de Psicologia de la Universitat de València. Mètode: La mostra està formada per 63 participants. L'edat mitjana dels participants va ser de 20,05 anys (DT = 2,74). Es va enquestar

[*] Departament de Metodologia de les Ciències del Comportament. Facultat de Psicologia. Universitat de València. Correspondència: <Laura.badenes@uv.es> < M.Dolores.Frias@uv.es>. Fax: 34 96 3864697 Telèfon: 34 963864547.

[**] ESIC Business & Marketing School, València.

NOTA: Estudi subvencionat pel Ministeri d'Economia i Competitivitat, Pla Nacional d'R+D+I (EDU2011-22862). I pel Programa VALr+d per a personal investigador en formació de caràcter predoctoral (ACIF/2013/167). Conselleria d'Educació, Cultura i Esport, Generalitat Valenciana (Espanya).

els participants sobre les seues interpretacions dels valors de *p*. Resultats. Els nostres resultats suggereixen que la majoria dels estudiants universitaris de Psicologia no coneixen la interpretació correcta dels valors *p*. La fal·làcia de la probabilitat inversa presenta més problemes de comprensió. A més, un gran nombre dels estudiants confonen la significació estadística del resultat obtingut amb la seua significació pràctica o clínica. Conclusions principals Aquests resultats destaquen la importància de l'educació estadística.

Paraules clau: Errors d'interpretació, valors p*, test de significació, inferència estadística*.

Resumen

Antecedentes. Informar sobre los valores p *de las pruebas de significación estadística és común en la literatura empírica de psicología. A pesar de esto, las interpretaciones incorrectas del valor* p *són abundantes y repetitivas. Estos errores afectan a las decisiones professionales y ponen en riesgo la calidad de las intervenciones psicológicas y la acumulación de un conocimiento científico válido. El objectivo de este estudio es descrivir los errores de interpretación del valor* p *y su interpretación correcta en los estudiantes de psicología de la Universidad de Valencia. Método: La muestra está formada por 63 participantes. La edad media de los participantes es de 20,05 años (DT = 2,74). Se realizaron encuestas a los participantes sobre sus interpretacions de los valores* p. *Resultados. Nuestros resultados sugieren que la mayoría de estudiantes universitarios de Psicologia no conocen la interpretación correcta de los valores* p. *La fal·làcia de la probabilidad inversa presenta más problemes de comprensión. A su vez, un gran número de estudiantes confunden la significación estadística del resultado obtenido con su significación pràctica o clínica. Conclusiones. Estos resultados destacant la importància de la educación estadística.*

Palabras clave: Errores de interpretación, valores *p*, test de significación, inferencia estadística.

Abstract

Background: Reporting p *values from statistical significance tests is common in psychology's empirical literature. However, the misconceptions about* p *value are abundant and repetitive. These errors affect professional decisions and compromise the quality of psychological interventions and the accumulation of a valid scientific knowledge. The aim of this study is to analyze the errors of interpretation of* p *value and their correct interpretation among college students of psychology at the University of Valencia. We surveyed participants about their interpretations of* p *values. Method: The sample is composed of 63 participants. The mean age of participants was 20.05 years (SD = 2.74). Results:Our findings suggest that college students*

many do not know the correct interpretation of p *values. The fallacy of the inverse probability presents major problems of comprehension. In addition, many students confuse statistical significance of the results obtained with its practical significance or clinical. Main Conclusions: These results highlight the importance of statistical education.*

Key words: Misconception, *p* values, significance testing, statistical inference.

Introducció

La «practica basada en l'evidència» (PBE) exigeix canviar les opinions dels anomenats 'experts' per l'ús del mètode científic i per la valoració de la qualitat de les evidències trobades per abordar els problemes i prendre decisions sobre intervenció, diagnòstic, etiologia o pronòstic (Frías-Navarro, 2011; Frías-Navarro i Pascual-Llobel, 2003; Frías-Navarro, Pascual-Llobel, i García-Pérez, 2000; Pascual-Llobel, Frías-Navarro i Monterde-i-Bort, 2004). La valoració de l'evidència empírica depèn de la qualitat de les anàlisis estadístiques (Faulkner, Fidler i Cumming, 2008).

Dins d'aquest procés de valoració crítica de les proves resulta crucial conèixer i comprendre el procés de contrast d'hipòtesis estadístiques mitjançant l'execució de la prova de significació de la hipòtesi nul·la (Null Hypothesis significance Testing, NHST) sobretot tenint en compte que en l'àmbit de la psicologia el procediment de la NHST és la tècnica per excel·lència. Així, en el 97% dels articles publicats en 10 revistes internacionals de psicologia s'utilitza la prova de significació de la hipòtesi nul·la (Cumming, Fidler, Leonard, Kalinowski, Christiansen *et al*., 2007). Per tant, saber interpretar els valors *p* de probabilitat són competències bàsiques del professional de la psicologia i de totes aquelles disciplines on s'aplica la inferència estadística.

No obstant això, els resultats de les investigacions assenyalen que la interpretació de l'abast dels resultats aportats per la prova de significació estadística no és unànime per part d'investigadors, professionals i estudiants de psicologia (Badenes-Ribera, Frias-Navarro, Monterde-i-Bort, i Pascual-Soler, 2015; Balluerka, Gómez, i Hidalgo, 2005; Haller i Krauss, 2002; Mittag i Thompson, 2000, Palmer i Sesé, 2013).

El valor *p* de probabilitat vinculat al resultat d'una prova estadística (ANOVA, t de Student, khi quadrat, correlació, anàlisi de regressió...) és la probabilitat del resultat observat o un valor més extrem si la hipòtesi nul·la és certa (Gill, 1999; Johnson, 1999). La definició és clara i precisa, però les interpretacions incorrectes d'aquest valor *p* segueixen sent abundants i repetitives (Goodman, 1999, 2008; Wagenmakers, 2007).

En la literatura s'han descrit diverses interpretacions errònies (Carver 1978; Cohen 1994; Dixon, 2003; Falk i Greenbaum 1995; Fidler, 2005; Gill, 1999; Goodman, 2008; Harlow, Mulaik, i Steiger, 1997; Johnson 1999; Kline 2004, 2013; Nickerson 2000). Les interpretacions errònies del valor *p* més comunes són: (1) «Fal·làcia de la probabilitat inversa»; (2) «Fal·làcia de la probabilitat contra l'atzar»; (3) «Fal·làcia de la grandària de l'efecte» i; (4) «Fal·làcia de la significació clínica o pràctica».

Fal·làcia de la probabilitat inversa i fal·làcia de la probabilitat contra l'atzar

La fal·làcia de la «probabilitat inversa» («inverse probability» fallacy) és la falsa creença que el valor *p* fa referència a la probabilitat que la hipòtesi nul·la (H_0) sigui veritable donades certes dades.

La «fal·làcia de les probabilitats contra l'atzar» assenyala que el valor *p* és la probabilitat d'obtenir el resultat per atzar o la probabilitat que el resultat succeeixi com a conseqüència del procés de la selecció de la mostra.

La «fal·làcia de la probabilitat inversa» i la «fal·làcia de les probabilitats contra l'atzar» estan relacionades amb el mateix problema: confondre la probabilitat del resultat, assumint que la hipòtesi nul·la és certa, amb la probabilitat de la hipòtesi nul·la, donades certes dades. Creure que quan p=0,05 la hipòtesi nul·la té un 5% de probabilitat de ser certa o un 95% de ser falsa és un dels errors més greus i comuns.

Tenint en compte el concepte correcte de valor *p*, és clar que aquest valor no representa la probabilitat de la hipòtesi nul·la ja que per calcular el valor *p* es parteix d'un model on s'assumeix des del començament de l'anàlisi que aquesta hipòtesi és certa. Posteriorment, es calcula la probabilitat del resultat obtingut en l'experiment dins de la distribució de la hipòtesi nul·la. En altres paraules, el valor *p* de probabilitat es calcula per a les dades de l'experiment sent la hipòtesi nul·la certa i no es calcula la probabilitat de la hipòtesi nul·la donats els resultats de l'estudi.

En definitiva, les proves de significació estadística no ofereixen informació de la probabilitat condicional de la hipòtesi nul·la donades les dades obtingudes en la investigació (Kirk, 1996; Sharver, 1993). Els resultats d'un determinat estudi poden conduir al rebuig de la hipòtesi nul·la o poden fracassar en el seu rebuig depenent del valor *p* associat al resultat observat, però no poden provar la veritat o falsedat de la hipòtesi nul·la (com tampoc poden fer-ho de la hipòtesi alternativa). Cal tenir molt clar que la hipòtesi nul·la mai no s'accepta; només es manté o es rebutja quan es compara el valor *p* amb el valor d'alfa.

Fal·làcia de la grandària de l'efecte

La «fal·làcia de la grandària» representa una de les crítiques més fortes contra les proves de significació de la hipòtesi nul·la i, en gran mesura, ha provocat el moviment de la reforma estadística que dóna suport a acompanyar els valors *p* amb informació de la grandària de l'efecte i els seus intervals de confiança tal com assenyala el manual de l'American Psychological Association (APA, 2010). Aquesta fal·làcia vincula la significació estadística amb la magnitud de l'efecte trobat. En concret, suposa la falsa creença que valors xicotets de *p* són indicadors d'efectes grans (Gliner, Vaske, i Morgan, 2001; Kline, 2013).

Aquesta fal·làcia podria subjaure la deficiència dels informes científics publicats en revistes d'impacte a l'hora d'informar d'estadístics de la grandària de l'efecte. Els investigadors i els revisors de les revistes es podrien plantejar la qüestió següent: Per què i per a què molestar-se a informar d'una grandària de l'efecte quan es creu que el valor *p* n'és un indicador? (Fidler, 2005; Kirk, 2001).

No obstant això, el valor *p* no informa de la magnitud d'un efecte. La grandària de l'efecte només pot ser coneguda calculant directament el seu valor amb l'estadístic adequat i el seu interval de confiança (Cumming, 2012, Kline, 2004, 2013). Per això, quan s'executa una prova de significació estadística, els resultats dels valors *p* haurien d'anar acompanyats d'un indicador de la grandària de l'efecte, d'aquesta manera permetrien interpretacions més substantives i no únicament probabilístiques reduïdes a mantenir o rebutjar la hipòtesi nul·la (APA, 1996, 2001, 2010; Gliner *et al*., 2001; Wilkinson i the Task Force on Statistical Inference, 1999).

Fal·làcia de la significació clínica o pràctica

La «fal·làcia de la significació clínica o pràctica» vincula la significació estadística amb la importància pràctica del resultat observat. Un resultat estadísticament significatiu no indica que sigui un resultat important. De la mateixa manera que un resultat estadísticament no significatiu podria encara ser important (Gliner, Leech i Morgan, 2002; Kirk, 1996). En aquest sentit, és possible que l'efecte amb significació clínica o importància pràctica no assoleixi la significació estadística i per tant sigui rebutjat. I, al contrari, pot haver efectes amb poca significació clínica o importància pràctica però que hagin arribat a la significació estadística i per això es prenen, sovint, com a significatius o importants.

No obstant això, la significació clínica o pràctica d'un resultat no és una qüestió ni estadística ni de magnitud de l'efecte.

La importància clínica o pràctica d'un resultat està relacionada amb la utilitat de la troballa per al professional de la psicologia i només es pot valorar ate-

nent el context i l'interès de la troballa en una situació clínica concreta. Per tant, la significació clínica d'un resultat és subjectiva i està relacionada amb el judici de l'expert i sotmesa al context de recerca. Per això, el valor de la grandària de l'efecte clínicament important depèn del constructe que s'estigui investigant i de certes variables contextuals, i no és una qüestió ni estadística ni de magnitud de l'efecte (Fredes-Navarro, 2011).

Revisió de la literatura: errors d'interpretació del valor **p** *en estudiants universitaris*

Estudis previs sobre interpretació dels valors *p* de probabilitat de les proves de significació estadística van trobar errors d'interpretació en estudiants universitaris de diferents disciplines que ja havien cursat les assignatures d'estadística. Per exemple, en els estudis de Vallecillos (2002), i Vallecillos i Batanero (1997) i Williams (1988) es va trobar que un gran nombre d'estudiants universitaris van interpretar el valor *p* com la probabilitat de la hipòtesi nul·la.

En l'àmbit de la psicologia, en l'estudi de Haller i Kraus (2002) es va trobar que el 100% dels estudiants universitaris van cometre algun tipus d'error en la interpretació del valor *p*. En concret, el 26% dels estudiants universitaris van creure que una prova d'hipòtesis ofereix la probabilitat de la hipòtesi nul·la, mentre que el 33% van interpretar el valor *p* com la probabilitat de la hipòtesi alternativa. A més, un 15% dels estudiants universitaris va creure que un valor *p* <alfa rebutja absolutament la hipòtesi nul·la i el 13%, que rebutja la hipòtesi alternativa.

En l'estudi de Falk i Greenbaum (1995), es va trobar que el 49,1% dels estudiants universitaris de psicologia van creure que un valor de *p* <alfa significa que la hipòtesi nul·la ha demostrat ser improbable.

En tots aquests supòsits, els estudiants van confondre la probabilitat del resultat, assumint que la hipòtesi nul·la és certa, amb la probabilitat de la hipòtesi nul·la, donades certes dades.

Finalment, com assenyalen Castro-Sotos, Vanhoof, Van den Noortgate i Onghena (2007), no hi ha estudis empírics que analitzin les fal·làcies de la grandària de l'efecte i de la significació clínica o pràctica.

Com es pot observar, aquest tipus d'interpretacions incorrectes són problemes d'interpretació de l'investigador i no del procediment NHST. Darrere d'aquestes interpretacions errònies del valor *p* hi ha unes creences i atribucions sobre el significat dels resultats. Per això, cal comprendre el raonament estadístic o la forma de raonar amb idees estadístiques i donar sentit a la informació estadística que realitzen les persones (Garfield i Gal, 1999, citat en Garfield, 2002).

El raonament estadístic implica fer interpretacions basades en conjunts de dades, representacions gràfiques i resums estadístics combinant idees sobre les dades, l'atzar i el mostreig. Darrere d'aquest raonament hi ha una comprensió conceptual de les idees importants, com la distribució, el centre, la propagació, l'associació, la incertesa, l'atzar i el mostreig (Garfield, 2002).

L'objectiu de la nostra investigació és descriure els errors de raonament estadístic que els estudiants universitaris de psicologia realitzen davant els resultats que aporta una prova d'inferència estadística. La seua visió i interpretació de les evidències són un filtre de qualitat en la seua vida professional que no pot estar sotmès a creences o interpretacions errònies del procediment estadístic que representa l'eina fonamental per obtenir coneixement científic. La competència de la 'lectura crítica' dins del model de la pràctica basada en l'evidència exigeix conèixer i interpretar adequadament la qualitat metodològica de les proves o evidències aportades per la literatura.

En concret, el nostre estudi versa sobre l'anàlisi dels errors d'interpretació del valor *p* vinculats amb la fal·làcia de la probabilitat inversa (analitzada en estudis previs), fal·làcia de la grandària de l'efecte i fal·làcia de la significació clínica o pràctica del resultat, així com la interpretació correcta del valor *p* com a aspectes innovadors del present estudi.

Mètode

Participants

La mostra està formada per 63 estudiants de psicologia de la Universitat de València que ja han cursat les assignatures d'estadística. L'edat mitjana dels participants va ser de 20,05 anys (DT = 2,74). Els homes representen el 20% i les dones el 80%.

Instruments

Es va elaborar un qüestionari estructurat. En primer lloc, el qüestionari inclou preguntes relacionades amb informacions sociodemogràfiques (sexe i edat).

En segon lloc, l'instrument inclou un conjunt de 9 preguntes que analitzen les interpretacions del valor *p* del procediment del contrast d'hipòtesis. En concret, s'analitzen les fal·làcies de probabilitat inversa, de la grandària de l'efecte, de la significació clínica o pràctica i la interpretació correcta del valor *p*, així com la decisió correcta davant un resultat que es considera estadísticament significatiu.

Les preguntes es plantejaven amb l'argument següent:

«*Suposem que un article d'investigació assenyala un valor de p = 0,001 en l'apartat de resultats (alfa = 0,05). Assenyaleu si les afirmacions següents són vertaderes o falses*»

A. Fal·làcia de la probabilitat inversa:

1. S'ha provat que la hipòtesi nul·la és vertadera.
2. S'ha provat que la hipòtesi nul·la és falsa.
3. S'ha determinat la probabilitat de la hipòtesi nul·la (p = 0,001).
4. S'ha deduït la probabilitat de la hipòtesi experimental (p = 0,001).
5. La probabilitat que la hipòtesi nul·la sigui veritable, donades les dades obtingudes, és de 0,01.

B. Fal·làcia de la grandària de l'efecte:

6. El valor $p = 0,001$ confirma de manera directa que la grandària de l'efecte ha estat gran.

C. Fal·làcia de la significació clínica o pràctica:

7. Obtenir un resultat estadísticament significatiu indica que l'efecte detectat és important.

D. Interpretació correcta i decisió adoptada:

8. Es coneix la probabilitat del resultat de la prova estadística, assumint que la hipòtesi nul·la és certa.
9. Atès que p=0,001 aleshores el resultat obtingut permet concloure que les diferències no es deuen a l'atzar.

Procediment

Aquest treball s'emmarca dins de la línia d'investigació sobre cognició i educació estadística que el nostre equip d'investigació desenvolupa des de fa anys al Departament de Metodologia de les Ciències del Comportament de la Universitat de València (Spain) (REME).

La participació en l'estudi va ser voluntària. Als participants se'ls assegurà l'anonimat en l'emplenament dels qüestionaris de llapis i paper. L'emplenament es va dur a terme en hores de classe. La recollida de dades va tenir lloc a l'octubre de 2014, a l'inici del curs acadèmic a Espanya.

Anàlisi d'estadístiques

L'anàlisi de les dades es va realitzar mitjançant el programa estadístic IBM SPSS v. 20 per a Windows. Les anàlisis detallen les freqüències i percentatges

d'acord amb cadascuna de les afirmacions relatives al procés d'interpretació dels resultats d'una investigació. A més, aquest tipus d'anàlisi permet la comparació amb els resultats d'altres investigadors que també han analitzat les diferents fal·làcies amb freqüències i percentatges.

Resultats

La taula 1 mostra el percentatge de respostes dels participants que donen suport a les afirmacions falses sobre el valor *p* relacionades amb la fal·làcia de la probabilitat inversa. S'observa que gran part dels participants perceben com a veritable alguna de les afirmacions falses sobre el valor *p*.

TAULA 1
Fal·làcia de la probabilitat inversa (%)

ÍTEM	%
1. S'ha provat que la hipòtesi nul·la és vertadera	25.4
2. S'ha provat que la hipòtesi nul·la és falsa	55.6
3. S'ha determinat la probabilitat de la hipòtesi nul·la (p = 0.001)	64.5
4. S'ha deduït la probabilitat de la hipòtesi experimental (p = 0.001)	29
5. La probabilitat que la hipòtesi nul·la sigui veritable, donades les dades obtingudes, és de 0.01	50.8
% Participants que han valorat correctament les 5 afirmacions com falses	3.2

Les afirmacions falses que més suport han rebut per part dels estudiants són «s'ha provat que la hipòtesi nul·la és falsa», «s'ha determinat la probabilitat de la hipòtesi nul·la (p = 0,001)» i «la probabilitat que la hipòtesi nul·la sigui veritable, donades les dades obtingudes, és de 0,01». Només el 3,2% dels participants han valorat correctament les 5 afirmacions com a falses.

Fal·làcia de la grandària de l'efecte i fal·làcia de la significació clínica o pràctica

La taula 2 mostra el percentatge de respostes dels participants que donen suport a les afirmacions falses del valor de *p* vinculades a la grandària de l'efecte. S'observa que el 50% dels estudiants valoren les dues afirmacions com a falses, sent aquestes respostes correctes.

TAULA 2
Fal·làcia de la grandària del efecte i de la significació clínica/pràctica (%)

ÍTEM	%
6. El valor $p = 0.001$ confirma de manera directa que la grandària de l'efecte ha estat gran	17.5
7. Obtenir un resultat estadísticament significatiu indica que l'efecte detectat és important	41.3
% Participants que han valorat correctament les 2 afirmacions com falses	50.8

L'afirmació falsa que més suport ha rebut és «obtenir un resultat estadísticament significatiu indica que l'efecte detectat és important». En conseqüència, els participants confonen la significació estadística dels resultats obtinguts amb la seua significació pràctica o clínica.

Interpretació correcta del valor p *i decisió adoptada*

La taula 3 mostra el percentatge de participants que donen suport a les afirmacions correctes sobre el valor *p* de probabilitat. S'observa que els participants presenten més problemes amb la interpretació estadística del valor *p*. La interpretació millora quan es fa en termes de probabilitat comparada amb la interpretació del valor *p* com a conclusió estadística.

TAULA 3
Interpretació correcta del valor p i decisió adoptada (%)

ÍTEM	%
6. Es coneix la probabilitat del resultat de la prova estadística, assumint que la hipòtesi nul·la és certa	61.3
9. Atès que p=0.001 aleshores el resultat obtingut permet concloure que les diferències no es deuen a l'atzar	50
% Participants que han valorat correctament les 2 afirmacions com vertaderes	27

Només un petit percentatge dels participants valoren correctament les dues afirmacions com a vertaderes.

Discussió

Els resultats indiquen que la comprensió i l'aplicació correcta de molts conceptes estadístics continua sent problemàtica. Confondre el nivell de significació d'alfa amb la probabilitat que la hipòtesi nul·la sigui certa, interpretar un resultat estadísticament significatiu com un resultat important o útil són inter-

pretacions errònies o falses creences que continuen existint entre els estudiants universitaris de psicologia. Aquestes dades són consistents amb previs estudis amb estudiants universitaris de psicologia (Falk i Greenbaum 1995; Haller i Kraus, 2002) i d'altres disciplines (Vallecillos, 2002; Vallecillos i Batanero, 1997; Williams, 1988).

La «fal·làcia de la probabilitat inversa» és la que amb més freqüència s'observa. El valor *p* no representa la probabilitat de la hipòtesi nul·la, ja que per calcular el valor *p* s'assumeix des del començament de l'anàlisi que aquesta hipòtesi és certa. És a dir, el valor *p* de probabilitat es calcula per a les dades de l'experiment sent la hipòtesi nul·la certa ($Pr(Dades|H_0)$) i no es calcula la probabilitat de la hipòtesi nul·la donats els resultats de l'estudi ($Pr(H_0|Dades)$). Per aquesta raó, com assenyalen Palmer i Sése (2013), mai no hauria d'expressar que «s'accepta» la hipòtesi nul·la quan el valor *p* és menor que el d'alfa, la hipòtesi nul·la es rebutja o no es rebutja (p. 51). No rebutjar la hipòtesi nul·la no implica la veracitat de la hipòtesi nul·la (Finch, Cumming, i Thomason, 2001; Monterde-i-Bort, Frias-Navarro, i Pascual-Llobel, 2010; Monterde-i-Bort, Pascual-Llobel, i Frias-Navarro, 2006; Morgan, 2003).

A més, el valor *p* no ofereix informació de la magnitud de l'efecte o importància del resultat (Gliner *et al*., 2002; Grant, 1962; Rosenthal, 1993; Shaver, 1993). No obstant això, un gran nombre dels estudiants universitaris de psicologia confonen la significació estadística dels resultats obtinguts amb la seua significació pràctica. Quant a això, la presentació de molts asteriscos al costat del valor *p* de probabilitat o valors *p* molt xicotets només assenyalen que en aquest disseny la hipòtesi nul·la és poc plausible, però d'ací no es pot inferir que l'efecte trobat és important, que la relació entre les variables és forta o que hi ha una rellevància substantiva (Frías-Navarro, 2011; Gliner, *et al*., 2001).

En definitiva, rebutjar la hipòtesi nul·la no indica la importància dels resultats. Cal no confondre la significació estadística dels resultats amb la significació clínica o pràctica (Palmer i Sesé, 2013; Thompson, 1996). En aquest sentit, per a distingir entre la importància o significació pràctica de les troballes i la seua significació estadística es recomana utilitzar el terme «estadísticament significatiu» per descriure els resultats que rebutgen la hipòtesi nul·la, és a dir, els resultats vinculats amb un valor $p <$ alfa (APA, 2010; Cumming, 2012; Gliner *et al*., 2001; Kline, 2013; Monterde-i-Bort *et al*., 2006, 2010; Thompson, 1996) o eliminar la paraula «significatiu» de l'informe dels resultats de valors de $p <$ alfa (Batanero, Tauber, i Sanchez, 2004). Altres autors i la mateixa APA (2010) recomanen acompanyar els valors *p* de les proves de significació estadística amb l'estimació de la grandària de l'efecte i els seus intervals de confiança (Cumming, 2014, 2013; Cumming, Fidler, Kalinowski, i Lai, 2012; Frías-Navarro *et al*., 2000; García-García, Ortega-Campos i De la Fuente, 2011; Gliner

et al., 2001; Monterde-i-Bort *et al*., 2006, 2010; Palmer i Sése, 2013; Pascual-Llobel *et al*., 2004; Sackett, Rosenberg, Gray i Richardson, 1996; Sackett, Richardson, Rosenberg, i Haynes, 1997). No obstant això, altres autors advoquen per reemplaçar el procediment de la NHST per altres alternatives d'anàlisi com els intervals de confiança (Fidler i Lotus, 2009; Gardner, i Altman, 1986) o les tècniques bayesianes (Kruschke, 2011; Masson, 2011).

El present treball aporta una evidència més en la línia de la necessitat de l'educació estadística atesos els problemes que envolten la interpretació adequada dels resultats obtinguts amb el procediment de significació de la hipòtesi nul·la per garantir una formació de qualitat als futurs professionals de la psicologia (Cumming, 2012; Gliner *et al*., 2002; Kline, 2013; Haller i Kraus, 2002) i de les altres disciplines que utilitzen el procediment NHST.

Els resultats d'aquesta investigació descriuen el tipus de fal·làcies que els estudiants de psicologia realitzen sobre la interpretació del valor *p* de probabilitat que acompanya una prova d'inferència estadística. Aquesta informació és fonamental per abordar i planificar estratègies d'educació estadística dirigides a intervenir sobre les interpretacions incorrectes. La investigació futura sobre aquest camp d'estudi ha d'anar dirigida ara a la intervenció sobre les fal·làcies o errors d'interpretació vinculats a la interpretació del valor *p* de probabilitat.

El model de la pràctica basada en l'evidència exigeix disposar d'un coneixement adequat dels fonaments de la metodologia d'investigació per poder valorar de forma crítica les proves o evidència que els estudis detallen en els seus informes. En aquest sentit, i com assenyala Kirk (2001), per promoure unes bones pràctiques estadístiques és necessari un enfocament multifacètic. Un enfocament que impliqui autors dels llibres de text, els professors que imparteixen docència en programes de graus i postgraus, els autors de paquets estadístics de programari, editors de revistes i manuals de publicació. Els llibres de text han de presentar els conceptes estadístics correctament i els professors han d'estar preparats per ensenyar els conceptes correctament (p. 216). Per exemple, com assenyalen Gliner *et al*. (2002), els autors de llibres de text haurien d'incloure una secció sobre l'actual debat i crítiques del procediment NHST sobre si les proves de significació estadística són el millor mètode per avançar en l'acumulació del coneixement científic i vàlid. També hauria d'afegir informació sobre com calcular i informar sobre la grandària de l'efecte i els seus intervals de confiança tant en els resultats estadísticament significatius com en els resultats no estadísticament significatius. I, finalment, l'autor hauria de posar exemples per decidir si el resultat té importància pràctica o clínica. En el primer llibre de la col·lecció Reforma Estadística editat a Espanya es detallen totes aquestes qüestions en diferents capítols (Frías-Navarro, 2011).

Limitacions

Els resultats de l'estudi s'han d'interpretar amb certes limitacions. El procediment de mostreig (mostra de conveniència) limita la validesa externa dels nostres resultats. Seria important replicar l'estudi amb un major nombre d'estudiants universitaris de psicologia i d'altres disciplines (per millorar la generalització). No obstant això, els resultats del present estudi van en la línia dels resultats d'estudis previs en mostres d'estudiants universitaris de diferents disciplines (Vallecillos, 2002, Vallecillos i Batanero, 1997; Williams, 1988), d'estudiants universitaris de psicologia (Falk i Greenbaum 1995; Haller i Kraus, 2002), de professors universitaris de psicologia (Badenes-Ribera *et al*., 2015; Haller i Kraus, 2002; Monterde-i-Bort, 2006, 2010; Oakes, 1986), en mostres de membres de l'American Educational Research Association (AERA) (Gordon, 2001; Mittag, i Thompson, 2000) i en professionals de l'estadística (Lecoutre, Poitevineau, i Lecoutre, 2003). Tot això ens indica la necessitat de formar adequadament els professionals de la psicologia en metodologia d'investigació per produir un coneixement científic i vàlid i millorar la pràctica professional. La PBE requereix professionals que valorin críticament les troballes dels estudis o investigacions psicològiques i, per això, és necessària una formació en conceptes estadístics, en metodologia de dissenys d'investigació i en resultats de proves d'inferència estadística.

Referències

American Psychological Association (1996). *Task Force on Statistical Inference Report*. Washington, DC: American Psychological Association.

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

American Psychological Association (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Badenes-Ribera, L.; Frias-Navarro, D.; Monterde-i-Bort, H. i Pascual-Soler, M. (*in* press). Interpretation of the p value. A national survey study in academic psychologists from Spain. *Psicothema*, *27*. doi: 10.7334/psicothema 2014.283.

Balluerka, N.; Gómez, J. i Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 55-70.

Batanero, C.; Tauber, L. M. i Sanchez, V. (2004). Students' reasoning about the normal distribution. En D. Ben-Zvi i J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257-276). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Castro-Sotos A. E.; Vanhoof, S.; Van den Noortgate, W. i Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review 2*, 98-113. doi: 10.1016/j.edurev.2007.04.001.

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

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge

Cumming, G. (2013). The new statistics: A how to guide. *Australian Psychologist*, *48*, 161-170. doi: 10.1111/ap.12018.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29. doi: 10.1177/0956797613504966.

Cumming, G.; Fidler, F.; Leonard, M.; Kalinowski, P.; Christiansen, A.; Kleining, A.; ... i Wilson, S. (2007). Statistical reform in Psychology: Is anything changing? *Psychological Science*, *18*, 230-232. doi: 10.1111/j.1467-9280.2007.01881.x.

Cumming, G.; Fidler, F.; Kalinowski, P. i Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*, *1*38-146. doi: 10.1111/j.1742-9536.2011.00037.x.

Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 133-149.

Falk, R. i Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75-98. doi: 10.1177/0959354395051004.

Faulkner, C., Fidler, F., i Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behavior Research and Therapy*, *46*, 270-281. doi: 10.1016/j.brat.2007.12.001.

Fidler, F. (2005). *From statistical significance to effect estimation: statistical reform in psychology, medicine and ecology*. PhD Thesis History and Philosophy of Science. Melbourne, Australia. Department of History and Philosophy of Science. University of Melbourne.

Fidler, F. i Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology*, *217*, *27-37*. doi: 10.1027/0044-3409.217.1.27.

Finch, S.; Cumming, G. i Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181-210. doi: 10.1177/00131640121971167.

Frías-Navarro, D. (2011*).Técnica estadística y diseño de investigación*. València: Palmero Ediciones.

Frías-Navarro, D.; Monterde-i-Bort, H.; Pascual-Soler, M.; Badenes-Ribera, L. i Pascual-Llobel, J. (2014). *Improvement in statistical practice: results from a Spanish survey (N=744)*. VI European Congres of Methodology. Universiteit Utrecht. The Netherlands July 23-25.

Frías-Navarro, D. i Pascual-Llobell, J. (2003). Psicología clínica basada en pruebas: efecto del tratamiento. *Papeles del Psicólogo*, *85*, 11-18. Disponible en <http://www.papelesdelpsicologo.es/vernumero.asp?ID=1074>.

Frias-Navarro, D.; Pascual-Llobel, J. i Garcia-Perez, F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema*, *12*, 236-240.

García-García, J.; Ortega-Campos, E. i De la Fuente-Sánchez, L. (2011). The use of the effect size in JCR Spanish Journals of Psychology: From theory to fact. *The Spanish Journal of Psychology*, *14*, 1050-1055. doi: 10.5209/rev_SJOP.2011.v14.n2.49.

Gardner, M. J. i Altman, D. G. (1986). Confidence intervals rather than p-values: estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750. doi: 10.1136/bmj.292.6522.746.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of statistic education*, *10*. Disponible en <www.amstat.org/publications/jse/v10n3/garfield.html>.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*, 647-674. doi: 10.1177/106591299905200309.

Gliner, J. A.; Vaske, J. J. i Morgan, G. A. (2001). Null hypothesis significance testing: Effect size matters. *Human Dimensions of Wildlife*, *6*, 291-301. doi: 1087-1209/01.

Gliner, J. A.; Leech, N. L. i Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say?. *The Journal of Experimental Education*, *71*, 83-92. doi: 10.1080/00220970209602058.

Goodman, S. (1999). Toward evidence-based medical statistics 1: The p value fallacy. *Ann Intern Med*. *130*, 995-1004.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin Hematol 45*, 135-140. doi: 10.1053/j.seminhematol.2008.04.003.

Gordon, H. R. D. (2001). American vocational education research association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Educational Research*, *26*, 1-18.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics on investigating theoretical models. *Psychological Reviews*, *69*, 54-61.

Haller, H. i Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research Online* [On-line serial], 7, 120. Disponible en <http://www.metheval.uni-jena.de/lehre/0405-ws/evaluationuebung/haller.pdf>.

Harlow, L. L.; Mulaik, S. A. i Steiger, J. H. (Eds.) (1997). *What if there were no significance tests*? London: Lawrence Erlbaum Associates, Publishers.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management*, *63*, 763-772.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759. doi: 10.1177/0013164496056005002.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213-218. doi: 10.1177/00131640121971185.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline, R. B. (2013). *Beyond significance testing: Statistic reform in the behavioral sciences*. Washington, DC: American Psychological Association.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299-312. doi: 10.1177/1745691611406925.

Lecoutre, M. P.; Poitevineau, J. i Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Tests. *International Journal of Psychology*, *38*, 37-45. doi: 10.1080/00207590244000250.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavioral Research*, *43*, 679-690. doi: 10.3758/s13428-010-0049-5.

Mittag, K. C. i Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance test and others statistical issues. *Educational Researcher*, *29*, 14-20.

Monterde-i-Bort, H.; Frías-Navarro, D. i Pascual-Llobel, J. (2010). Uses and abuses of statistical significance tests and other statistical resources: A comparative study. *European Journal of Psychology of Education*, *25*, 429-447. doi: 10.1007/s10212-010-0021-x.

Monterde-i-Bort, H.; Pascual-Llobel, J. i Frias-Navarro, D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema*, *18*, 848-856.

Morgan, P. (2003). Null hypothesis significance testing: Philosophical and practical considerations of a statistical controversy. *Exceptionality: A Special Education Journal*, *11*, 209-221. doi: 10.1207/S15327035EX1104_2.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301. doi: 10.1037//1082-989X.S.2.241.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Chicester: John Wiley & Sons.

Palmer, A. i Sesé, A. (2013). Recommendations for the use of statistics in Clinical and Health Psychology. *Clínica y Salud*, *24*, 47-54. doi: http://dx.doi.org/10.5093/cl2013a6.

Pascual-Llobel, J.; Frias-Navarro, D. i Monterde-i-Bort, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica clínica basada en la evidencia. *Papeles del Psicólogo*, *87*, 1-8.

Rosenthal, R. (1993). Cumulating evidence. En G. Keren y C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (p. 519-559). Hillsdale, NJ: Erlbaum.

Sackett, D. L.; Richardson, W. S.; Rosenberg, W. M. C. i Haynes, R. (1997). *Evidence based medicine*. *How to practice & teach EBM*. New York: Churchill Livingstone.

Sackett, D. L.; Rosenberg, W. M. C.; Gray, J. A. M. i Richardson, W. S. (1996). Evidence based medicine. What it is and what it isn't. *British Medical Journal*, *312*, 71-72. doi: 10.1136/bmj.312.7023.71.

Shaver, J. P. (1993). What statistical significance testing is, and what is not. *The Journal of Experimental Education*, *61*, 293-316.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30. doi: 10.3102/0013189X025002026.

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypotheses testing by university students. *Themes in Education*, *3*, 183-198.

Vallecillos, A. i Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathematiques*, *17*, 29-48.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wilkinson, L. i the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist*, *54*, 594-604. doi: 10.1037/0003-066X.54.8.594.

Williams, A. M. (1998). Students' understanding of the significance level concept. En L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W. Wong (Eds.), *Proceedings of the fifth international conference on teaching statistics* (pp. 743-749). Voorburg, The Netherlands: International Statistical Institute.