Statistical significance and other complementary measures for the interpretation of the research results

Significación estadística y otras medidas complementarias para la interpretación de los resultados de investigación

Mildrey Torres Martínez, Magaly Herrera Villafranca and Yaneilys García Ávila
Instituto de Ciencia Animal, Apartado Postal 24, San José de las Lajas, Mayabeque, Cuba
Email: femtorresm@ica.edu.cu
Mildrey Torres Martínez: https://orcid.org/0000-0001-7942-0195
Magaly Herrera Villafranca: https://orcid.org/0000-0002-2641-1815
Yaneilys García Avila: https://orcid.org/0000-0003-0126-6233

The contrast hypothesis constitutes the most used method in the scientific research to estimate the statistical significance of any find. However, nowadays its use is questionable because it did not have other statistical criteria that make possible the credibility and reproducibility of the studies. From this condition, this study review how the use of the null hypothesis significance testing has been and the recommendations made regarding the application of other complementary statistical criteria for the interpretation of the results. It is described the main controversy of only use the probability value to reject or accept a hypothesis. The interpretation of a non significant value, as prove of effect absence or a significant value as existence of it, is a frequent mistake in scientific researchers, according to the reviewed literature. It is suggested to make a rigorous assessment of the obtained data in a research and include in the study reports other statistical tests, as the test power and the effect size of the intercession, to offer a complete interpretation and increase the results quality. Specifically, it is recommended to the editors of scientific journals to consider the report of the mentioned statisticians in the papers who required, as part of the criteria to take into account for their evaluation.

Key words: *null hypothesis significance testing, probability value, statistical power, effect size*

El contraste de hipótesis constituye el método que más se emplea en la investigación científica para estimar la significación estadística de cualquier hallazgo. Sin embargo, en la actualidad su utilización es cuestionable porque le falta integrar otros criterios estadísticos que posibiliten la credibilidad y reproducibilidad de los estudios. A partir de esta condición, este trabajo reseña cómo ha sido la utilización de la prueba de significación de la hipótesis nula y las recomendaciones que se le han hecho en cuanto a la aplicación de otros criterios estadísticos complementarios para la interpretación de los resultados. Se describe aquí la polémica fundamental de utilizar solamente el valor de probabilidad para rechazar o aceptar una hipótesis. La interpretación de un valor no signi¬ficativo, como una prueba de ausencia de efecto o de un valor significativo como existencia del mismo, es un error frecuente en investigaciones científicas, según refiere la literatura revisada. Se sugiere realizar una valoración rigurosa de los datos obtenidos en una investigación e incluir en los informes de trabajo otras pruebas estadísticas, como la potencia de la dócima y el tamaño del efecto de la intervención, para ofrecer una interpretación más completa e incrementar la calidad de los resultados. Específicamente, se recomienda a los editores de revistas científicas que se considere el informe de dichos estadísticos en los trabajos que así lo requieran, como parte de los criterios a tener en cuenta para su evaluación.

Palabras clave: *prueba de significación de la hipótesis nula, valor de probabilidad, potencia estadística, tamaño del efecto*

In the context of the research activity, the null hypothesis significance testing (NHST) is the inductive inferential method most used in the reports (Antúnez *et al.* 2021). However, the criticisms to the use of this test are so numerous that will be difficult to exhaustively deal with them in only one paper. It has provided evidence from the ones are focus on the incorrect use in the research reports up to those who question their scientific use and propose its abandon (Díaz-Batanero *et al.* 2019).

Thorough years, the controversy about the (NHST) has been so intensive, that some scientific and professional associations, as the American Psychological Association, American Education Research Association and the American Statistical Association, recommend to make changes in the editorial policy of the scientific journals with respect to the test use and to the favorable use of other criteria

En el contexto de la actividad investigativa, la prueba de significación de la hipótesis nula (PSHN) es el método inferencial inductivo más utilizado por excelencia en los informes (Antúnez *et al.* 2021). Sin embargo, las críticas al uso de esta prueba son tan numerosas que sería difícil abordarlas de manera exhaustiva en un solo trabajo. Se han documentado desde las que se centran en su incorrecta utilización en los informes de investigación hasta las que cuestionan su utilidad científica y proponen su abandono (Díaz-Batanero *et al.* 2019).

Durante años, la polémica sobre la PSHN ha sido tan intensa, que algunas asociaciones científicas y profesionales, como la American Psychological Association, la American Education Research Association y la American Statistical Association, recomiendan realizar cambios en las políticas editoriales de las revistas científicas con respecto a la utilización de la prueba y al uso favorable de otros criterios que permitan discutir más

that allow to discuss more the founded results (Frías *et al.* 2002).

The proposed changes are not alternatives to the classic statistical inference model, are a way of compensate some of the limitations of the NHST. These recommendations make reference mainly to two aspects: the need of taking into account the test power in the studies and to include estimations of the effect size (ES) (Hickey *et al.* 2018).

This study expects to review the use of the NHST and the recommendations for the application of other complementary statistical measurements in the results interpretation.

All scientific research has as objective to look for the explication of phenomenon, to make theories on their performances and with this, can derive estimations on the reality. However, to prove theories or estimate effects of a treatment, the researchers have to make a process of hypothesis verification, in which the scientific hypothesis is translated to the statistics (Kuffner and Walker 2019). According to Frías *et al.* (2002), the statistics technique of the hypothesis contrast and the research design has been mutually needed during decades.

The null hypothesis significance testing: p value. The methodological proposal of the NHST was developed between 1915 and 1933, as a result of the analysis of two thought schools: that of Ronald Fisher (1890-1962) and which was represented by Jerzy Neyman (1894-1981) and Egon Pearson (1895-1980). The main difference between these two theories do not lies in the calculations, but in the conceptions and in the underlying reasoning (Bono and Arnau 1995).

Fisher (1925) only defined a null hypothesis (H_0) and from this one, based on the sampling distribution of the statistical test, he estimated the probability of the data sample to decide it reject or not. In a general way, the decision rule was based on a probability value (p) from which it was accepted or not H_0, if the calculated p value was lower than 0.05. It should highlight that, although Fisher (1935, 1950, and 1955) gave priority to a significance level of 0.05, never prescribed that said level should keep fixed, so it depends on the characteristics of the research.

Neyman and Pearson (1928) proposed the addition of an alternative hypothesis (H_1) in comparison with H_0, which lead to the definition of two regions: reject and acceptation. From these contributions, the decision process can lead to two potential errors, type I, defined as the probability (α) of reject H_0 when it is true and the error type II, understandable as the probability (β) of accepting H_0 being false, which mean that there is not effect of treatments when in actual fact there is. The control of the last error allows increase the probability of find the true positives and correctly reject H_0, with a degree of certainty named: power (1- β) (Cochran and Cox 1999).

los resultados encontrados (Frías *et al.* 2002).

Los cambios propuestos no suponen alternativas al modelo de inferencia estadística clásica, sino una forma de compensar alguna de las limitaciones de la PSHN. Estas recomendaciones hacen referencia fundamentalmente a dos aspectos: la necesidad de tener en cuenta la potencia de la dócima en los estudios y de incluir estimaciones del tamaño del efecto (TE) (Hickey *et al.* 2018).

Por lo anterior, el siguiente trabajo pretende reseñar cómo ha sido la utilización de la PSHN y las recomendaciones para la aplicación de otras medidas estadísticas complementarias en la interpretación de los resultados.

Toda investigación científica tiene como objetivo buscar la explicación de los fenómenos, elaborar teorías sobre sus comportamientos y con ello, poder derivar estimaciones sobre la realidad. Sin embargo, para comprobar teorías o estimar efectos de un tratamiento, los investigadores tienen que realizar un proceso de comprobación de hipótesis, en el que se traduce la hipótesis científica a la estadística (Kuffner y Walker 2019). Según Frías *et al.* (2002), la técnica estadística del contraste de hipótesis y el diseño de la investigación se han necesitado mutuamente durante décadas.

La prueba de significación estadística de la hipótesis nula: el valor p. La propuesta metodológica de la PSHN se desarrolló entre 1915 y 1933, como un resultado del análisis de dos escuelas de pensamiento: la de Ronald Fisher (1890-1962) y la representada por Jerzy Neyman (1894-1981) y Egon Pearson (1895-1980). La principal diferencia entre estas dos teorías no radica en los cálculos, sino en las concepciones y en el razonamiento subyacente (Bono y Arnau 1995).

Fisher (1925) definió únicamente una hipótesis nula (H_0) y a partir de ella, sobre la base de la distribución muestral del estadístico de prueba, estimó la probabilidad de una muestra de datos para decidir su rechazo o no. De forma general, la regla de decisión se basó en un valor de probabilidad (p) a partir del cual se aceptaba o no H_0, si el valor p calculado era menor que 0.05. Se debe señalar que, aunque Fisher (1935, 1950, 1955) daba prioridad a un nivel de significación de 0.05, nunca prescribió que dicho nivel se debiera mantener fijo, sino que dependía de las características de la investigación.

Neyman y Pearson (1928) propusieron la adición de una hipótesis alternativa (H_1) en contraposición con H_0, lo que condujo a la definición de dos regiones: rechazo y aceptación. A partir de estos aportes, el proceso de decisión puede conducir a dos potenciales errores, el de tipo I, definido como la probabilidad (α) de rechazar H_0 cuando es cierta y el error de tipo II, entendido como la probabilidad (β) de aceptar H_0 siendo falsa, lo que significa que no existen efectos de tratamiento cuando en realidad sí los hay. El control de este último error permite aumentar la probabilidad de encontrar los verdaderos positivos y rechazar correctamente H_0, con un grado de certeza denominado: potencia (1- β) (Cochran y Cox 1999).

During the course of years, the p value of Fisher was turn in a way of adequately estimating the result of the intervention group, when assuming that H_0 is correct. The period 1940-1960 was known as the inference revolution and the statistical manuals of the period showed the hybrid model of the NHST between the approach of Fisher and Neyman-Pearson. This period was characterized by the exponential increase of the application of the HNST procedure by the researchers, in which the inference of the sample to the population was considered the crucial point of the studies (Bono and Arnau 1995).

The hypothesis contrast acquired great importance in the seventies and eighties. Many journals took as criterion the obtaining of results statistically significant to accept articles (publication bias) (Cohen 1994). The Journal of Experimental Psychology, for example, considered among their editorial rules to accept only those articles with significant results at level of 0.05. Those statistically significant at 0.01deserves a priority place in the journal. However, some of these results represented little practical interest and in the most of papers it was not considered the risk level which the research was able to accept when interpreted the results of the statistical test (Cohen 1992). In addition, researchers of the social and psychological sciences, not only need to know if the effect of treatment was significant or not, so they wish to obtain the true value of the effect (Rothman 1978).

From the above, to the nineties there were important statistical elements for a total interpretation of the results. Authors like Schmidt (1996) exhort to focus on the ES estimation for the final discussion of the finds. Wilkinson *et al.* (1999) recommended informing this statistical close with the probability value. Also, in the fourth edition of the publication manual of the American Psychological Association (1994), certain recommendation on the style of the research reports are performed and is emphasize in that the researchers should provide the probability values contributed by the statistical significance tests close to the values of the effect size (ES) and the statistical power as reliability precaution of the result. In this sense, a link between the significant, important and valid is established.

Despite these recommendations, there are still many researchers which are publishing that even take into account them. However, for the beginning of the new century is show a tendency in favor of not only inform the hypothesis contrast, as only one element to find or not significance differences, so it should be enclose of other complementary measures that allow a practical and precise scientific discussion (Marín and Paredes 2020).

Serdar *et al.* (2021) state that the controversy of using the NHST as instrument valid for the scientific progress is still keeping, which is show in meetings and

Con el transcurso de los años, el valor p de Fisher se convirtió en una forma de estimar adecuadamente el resultado del grupo intervención, al asumir que la H_0 es correcta. El período entre 1940-1960 se conoció como la revolución de la inferencia y los manuales de estadística de la época presentaron el modelo híbrido de la PSHN entre los enfoques de Fisher y Neyman-Pearson. Esta etapa se caracterizó por el incremento exponencial de la aplicación del procedimiento PSHN por parte de los investigadores, en el que la inferencia de la muestra a la población se consideró el punto crucial de los estudios (Bono y Arnau 1995).

El contraste de hipótesis adquirió gran importancia en la década de los 70 y 80. Numerosas revistas tomaron como criterio la obtención de resultados estadísticamente significativos para aceptar artículos (sesgo de publicación) (Cohen 1994). La Journal of Experimental Psychology, por ejemplo, contemplaba entre sus normas editoriales aceptar sólo aquellos manuscritos con resultados significativos al nivel de 0.05. Los estadísticamente significativos al 0.01 merecían un lugar priorizado en la revista. Sin embargo, algunos de estos resultados presentaban poco interés práctico y en la mayoría de los trabajos no se contemplaba el nivel de riesgo que el investigador estaba dispuesto a aceptar al interpretar los resultados de una prueba estadística (Cohen 1992). Además, investigaciones de las ciencias sociales y sicológicas, no sólo necesitaban conocer si el efecto del tratamiento fue significativo o no, sino que deseaban obtener en magnitud el verdadero valor de dicho efecto (Rothman 1978).

A partir de lo anterior, para la década de los 90 se dieron elementos estadísticos importantes para una interpretación más completa de los resultados. Autores como Schmidt (1996) exhortaron centrarse en la estimación del TE para la discusión final de los hallazgos. Wilkinson *et al.* (1999) recomendaron informar este estadístico junto con el valor de probabilidad. También, en la cuarta edición del manual de publicación de la American Psychological Association (1994), se realizan ciertas recomendaciones sobre el estilo de los informes de investigación y se enfatiza en que los investigadores deben proporcionar los valores de probabilidad aportados por las pruebas de significación estadística junto al los valores del tamaño del efecto (TE) y la potencia estadística como medida de confiabilidad del resultado. En este sentido, se establece una conexión entre lo significativo, importante y válido.

A pesar de estas recomendaciones, todavía son muchas las investigaciones que se publican que aún no las tienen en cuenta. Sin embargo, para inicios del nuevo siglo se muestra una tendencia a favor de no informar solamente el contraste de hipótesis, como único elemento para encontrar o no diferencias significativas, sino que se debe acompañar de otras medidas complementarias que permitan una discusión científica certera y práctica (Marín y Paredes 2020).

Serdar *et al.* (2021) plantean que la polémica de utilizar la PSHN como instrumento válido para el progreso científico aún se mantiene, lo cual se

congress of the American Psychological Association, where work sessions to this discussion are made. In this sense, are many the studies in which the researchers deep about the goodness and deficiencies of the NHST (table 1). In some the practical use of the test is defended and in others is questioned (Díaz-Batanero *et al.* 2019).

evidencia en las reuniones y congresos de la American Psychological Association, donde se dedican sesiones de trabajo a este debate. En este sentido, son numerosos los trabajos en los que los investigadores profundizan acerca de las bondades y las deficiencias de la PSHN (tabla 1). En algunos se defiende la utilidad práctica de la prueba y en otros se cuestiona (Díaz-Batanero *et al.* 2019).

Table 1. Different views about the practical application of the null hypothesis significance testing

Adopted view	Researchers
In favor	Levin (1993)
	Fritz (1995 y 1996)
	Greenwald *et al.* (1996)
	Abelson (1997) Cortina
	Dunlap (1997)
	Hagen (1997)
Detractors	Bakan (1966)
	Craig *et al.* (1976)
	Carver (1978 y 1993)
	Chow (1988)
	Thompson (1988, 1989, 1996, 1997 and 1999)
	Cohen (1990, 1994 and 1997)
	Falk and Greenbaum (1995)
	Schimdt (1996)
	Manzano (1997)
	Nickerson (2000)
	Valera *et al.* (2000)
	Borges *et al.* (2001)
	De la Fuente and Díaz-Batanero (2004)
	Morrison and Henkel (2006)
	Verdam *et al.* (2014)
	Harlow *et al.* (2016)
	Faulkenberry (2022)

Source: Own elaboration

Table 1 show the chronological analysis of the historical performance related with the contrast and checking of statistical hypothesis. These studies show the confusion, criticism and controversy among researchers, which at the beginning considered that it was enough the report of the p value to reject or accept a hypothesis (Ioannidis 2018). Ochoa *et al.* (2020) state that in the scientific literature it is frequency observed the error of interpret a non significant p value as prove of effect absence or association. It is also common to interpret a significant value as evidence of the existence of an effect or rate. In this sense, the absence of statistical significance ($p>0.05$) is not allow to prove the H_0either the presence of signification ($p<0.05$) of the H_1. Any decision about superiority or inferiority is subject to uncertainty, which is not solve in function that p be high or low to 0.05.

Wasserstein and Lazar (2016) show that due to the interpretation errors in the results of the hypothesis contrast and too many criticisms about the statistical significance, the American Statistical Association state their point of views to respect:

La tabla 1 muestra un análisis cronológico del comportamiento histórico relacionado con el contraste y comprobación de hipótesis estadísticas. Estos estudios evidencian la confusión, crítica y polémica entre los investigadores, que en un inicio consideraron que era suficiente el informe del valor p para rechazar o aceptar una hipótesis (Ioannidis 2018). Ochoa *et al.* (2020) plantean que en la literatura científica se observa con frecuencia el error de interpretar un valor de p no signi-ficativo como una prueba de ausencia de efecto o aso¬ciación. También es común interpretar un valor signifi¬cativo como una evidencia de la existencia de un efecto o relación. En este sentido, la ausencia de significación estadística ($p>0.05$) no permite probar la H_0 ni la presencia de significación ($p<0.05$) la de H_1. Cualquier decisión so¬bre superioridad o inferioridad está sujeta a incerti¬dumbre, que no se resuelve en función de que la p sea superior o inferior a 0.05.

Wasserstein y Lazar (2016) señalan que debido a los errores de interpretación en los resultados del contrate de hipótesis y a las numerosas críticas sobre la significación estadística, la American Statistical Association expuso sus puntos de vista al respecto:

1. The p value do not measure the probability of the studied hypothesis will be true neither the probability of the results should only due at random.

2. Scientific conclusions and political or enterprises decisions should not only based on if the p value exceeded a threshold value.

3. A p value or statistical significance does not measure the important of a result.

4. The p value do not has a good evidence measure for a model or hypothesis.

5. An appropriate inference requires a whole report, where other necessary statisticians were analyzed next to the statistical significance.

Statistical power. Bono and Arnau (1995), when reviewing the development of the concept of a test power, point out that in the theory develop by Neyman and Pearson in 1928, the power of a statistical test is the probability of find significant results. Their estimation, according to these authors, is determined by three basic components: sample size, level of significance (α) and ES to be detected.

There are two ways to estimate the power: *a priori* (prefixed) and *a posteriori*. The first shows the researcher about the sample size needed for adequate power. To this end, power tables have been constructed. The a posteriori power is important in the interpretation of the results of completed studies (Guerra *et al.* 2019).

Scheffé (1959) discusses the power of the F Fisher test in analysis of variance models (ANAVA), with fixed effects. It refers to the power tables, calculated for the values of $\alpha = 0.01$ and 0.05, and reproduces power graphs for the F Fisher test.

Menchaca (1974, 1975), Venereo (1976), Caballero (1979) and Menchaca and Torres (1985) contributed tables of sample sizes and number of replications in analysis of variance models, associated with designs completely randomized, random blocks, Latin square and turnover design. They include the maximum standardized difference between two means (Δ), the number of treatments (t), the level of significance (α) and the power of the test. These tables represent valuable work tools for researchers from different branches. Currently, with the advance of computer science, there are statistical packages that include the calculation of power, such as InfoStat, G Power and SPSS, among others (Guerra *et al.* 2019).

Despite the contributions of different specialists in the topic, the articles are lacking of the report of the statistical power as a truthfulness indicator of the research, which has been one of the criticisms most highlighted through years (Cohen 1992, Clark-Carter 1997, Frías *et al.* 2000 and Bakker and Wicherts 2011).

Cohen (1988, 1992) papers state by convention a minimum power of 0.80, due that usually is more serious to show that there is an effect when there is not, than to show that there is not effect when there is. Authors like Funder and Ozen (2019) report that, when the power

1. El valor de p no mide la probabilidad de que la hipótesis estudiada sea cierta ni la probabilidad de que los resultados se deban sólo al azar.

2. Conclusiones científicas y decisiones empresariales o políticas no se deben basar solamente en si el valor de p sobrepasa un valor umbral.

3. Un valor de p o significación estadística no mide la importancia de un resultado.

4. El valor de p no provee una buena medida de evidencia para un modelo o hipótesis.

5. Una inferencia apropiada requiere un informe completo, donde se analicen otros estadísticos necesarios junto a la significación estadística.

Potencia estadística. Bono y Arnau (1995), al revisar el desarrollo del concepto de potencia de una dócima, señalan que en la teoría desarrollada por Neyman y Pearson en 1928, la potencia de una dócima estadística es la probabilidad de encontrar resultados significativos. Su estimación, según indican estos autores, queda determinada por tres componentes básicos: tamaño de muestra, nivel de significación (α) y TE a detectar.

Existen dos formas de estimar la potencia: *a priori* (prefijada) y *a posteriori*. La primera le indica al investigador sobre el tamaño necesario de muestra para una potencia adecuada. Con este fin, se han construido tablas de potencia. La potencia a posteriori es importante en la interpretación de los resultados de estudios terminados (Guerra *et al.* 2019).

Scheffé (1959) aborda la potencia de la dócima F de Fisher en modelos de análisis de varianza (ANAVA) con efectos fijos. Hace referencia a las tablas de potencia, calculadas para los valores de $\alpha = 0.01$ y 0.05, y reproduce gráficos de potencia para la dócima F de Fisher.

Menchaca (1974, 1975), Venereo (1976), Caballero (1979) y Menchaca y Torres (1985) aportaron tablas de tamaños de muestra y número de réplicas en modelos de análisis de varianza, asociados a los diseños completamente aleatorizados, bloques al azar, cuadrado latino y de cambio. En ellos incluyen la máxima diferencia estandarizada entre dos medias (Δ), la cantidad de tratamientos (t), el nivel de significación y la potencia de la dócima. Estas tablas representan valiosas herramientas de trabajo para investigadores de diferentes ramas. En la actualidad, con el avance de la informática, existen paquetes estadísticos que incluyen el cálculo de la potencia, como el InfoStat, G Power y el SPSS, entre otros (Guerra *et al.* 2019).

A pesar de los aportes de diferentes especialistas en el tema, aún los artículos carecen del informe de la potencia estadística como el indicador de veracidad de la investigación, lo que se ha convertido en una de las críticas que más se destaca a través de los años (Cohen 1992, Clark-Carter 1997, Frías *et al.* 2000 y Bakker y Wicherts 2011).

Los trabajos de Cohen (1988, 1992) plantean por convención una potencia mínima de 0.80, debido a que habitualmente es más grave señalar que existe un efecto cuando no lo hay, que señalar que no existe efecto cuando sí lo hay. Autores como Funder y Ozen (2019) informan

value is lower than 0.80, it cannot conclude that the study be totally useless, so it should made valid conclusions from the sample size.

It should highlight the importance of the statistical power when a study is designed, in a way that the sample size used guarantees a high probability of detecting differences if there are really are. To perform studies of low statistical power is not ethically acceptable, so it can lead to results of uncertain scientific validity.

Effect sixe (ES). Cohen (1988) defined ES as the degree in which a phenomenon is in the population or the degree in which the null hypothesis is false. This statistical measure evaluate in a logical way the magnitude of an aspect of interest in a quantitative study and, therefore, make easy the assessment of its practical importance (Botella and Zamora 2017). In brief, is not enough with only identify the occurrence or not of certain effect, it also require to determine its magnitude or size to know their relevance or practical signification (Ponce *et al.* 2021).

In general way, the ES indexes can be classified in three big general categories: indexes of the mean families, indexes of the relation or association family and risk indexes (relative or absolute) (Ventura 2018). Rivera (2017) showed that, in the scientific literature are different formulations for the ES calculation, according to the phenomenon under study. The final interpretation of the results seen to be based on a value scale, according to the statistical test performed in the research (table 2) (Serdar *et al.* 2021).

que, cuando el valor de potencia es menor que 0.80, no se puede concluir que el estudio sea totalmente inútil, sino que se deben hacer conclusiones válidas a partir del tamaño de muestra.

Se debe señalar la importancia que reviste la potencia estadística cuando se diseña un estudio, de manera que el tamaño de muestra que se utilice garantice una elevada probabilidad de detectar diferencias si realmente existen. Llevar a cabo estudios de baja potencia estadística no es éticamente aceptable, pues puede conducir a resultados de dudosa validez científica.

Tamaño del efecto (TE). Cohen (1988) definió al TE como el grado en que un fenómeno está presente en la población o el grado en que la hipótesis nula es falsa. Esta medida estadística evalúa de forma coherente la magnitud de un aspecto de interés en un estudio cuantitativo y, por ende, facilita la valoración de su importancia práctica (Botella y Zamora 2017). En síntesis, no es suficiente con sólo identificar la ocurrencia o no de cierto efecto, también se requiere determinar su magnitud o tamaño para conocer su relevancia o significación práctica (Ponce *et al.* 2021).

De manera general, los índices del TE se pueden clasificar en tres grandes categorías generales: índices de la familia de medias, índices de la familia de la relación o asociación e índices de riesgo (relativo o absoluto) (Ventura 2018). Según señala Rivera (2017), en la literatura científica se encuentran disponibles diferentes formulaciones para el cálculo del TE, según el fenómeno en estudio. La interpretación final de los resultados suele estar basada en una escala de valores, según la prueba estadística que se realice en la investigación (tabla 2) (Serdar *et al.* 2021)

Table 2. Value scale according to the statistical test performed in the research to interpret the calculated ES value

Test	Relevant effect size	Effect Size (ES)		
		Small	Medium	Large
t-test for means	Cohen's d	0.20	0.50	0.80
Chi-Square	Cohen's ω	0.10	0.30	0.50
r x c frequency tables	Cramer's V or Phi	0.10	0.30	0.50
Correlation studies	R	0.20	0.50	0.80
2 x 2 table case control	Odd Ratio (OR)	1.50	2.00	3.00
2 x 2 table cohort studies	Risk Ratio (RR)	2.00	3.00	4.00
One-way an(c)ova (regression)	Cohen's f	0.10	0.25	0.40
ANOVA (for large sample)	η^2	0.01	0.06	0.14
ANOVA (for small size)	Ω^2	0.10	0.30	0.50
Friedman test	Average spearman Rho	0.02	0.13	0.26
Multiple regression	η^2	0.04	0.25	0.64
Coefficient of determination	R^2	0.04	0.25	0.64

Effect size (ES), according to the acronym in English
Source: Serdar *et al.* (2021)

The ES statistical provide the information about how the independent variable or variables explain the dependent variable. Low ES values means that the independent variables did not predict in adequate way

El estadístico TE proporciona información sobre qué tan bien la variable o variables independientes explican la variable dependiente. Valores bajos del TE significan que las variables independientes no predicen de manera

because they are only slightly relate with the dependent variable. High ES values represent that the independent variables are good predictors of the dependent variable. So, the ES is an important statistical indicator to evaluate the efficacy of any treatment or intervention on a determined response (Ventura 2018). In addition, Bologna (2014) state that the ES measures, when been standardized exceeded the problem of the hypothesis tests regarding their dependence with the sample size and they are use to make comparisons among researchers about a same topic when taking the results to a metric in common.

The publication manual of the American Psychological Association (2001) conclude that is necessary to report the ES with the p value to answer three basic questions of the research: a) ¿ if there a real effect or the result should attribute at random?, b) if the effect is real, ¿ how big is it?, and c) ¿ if the effect big enough for to be considered important or useful?

By all the previous, the ES is considered as a complementary analysis of the NHST which helps to correct the limitations showed by this test. However, despite their practical use it is not frequent their use in the researcher reports.

Use of the complementary measures in the literature. From the nineties, the statistical experts have been conscious that the NHST is, in many aspects, inadequate to interpret the researcher results. Therefore, the uses of other complementary measure in the results report are not even completely achieving (Ochoa *et al.* 2019).

The sixth edition of the publication manual of the American Psychological Association (2010) showed the need of seriously take into account the statistical power providing information that show that the study has the enough power to find effects of substantive interest. However, the continuous lack of interest by the power of the statistical tests only will change when the editor of the important journals demand this analysis in their editorial policy (Frías *et al.* 2002).

In studies performed by different authors from 2010, there is an increase in the use of the ES, mainly in psychology journals, so they demand the use of this statistician as rule. Authors like Odgaard and Fowler (2010) reviewed the intervention studies publishing in 2003, 2004, 2007 and 2008 in the Journal of Consulting and Clinical Psychology, and found that in general 75 % of the studies report of any index of the ES.

Sun *et al.* (2010) analyzed the articles published between 2005 and 2007 in five journals (Journal of Educational Psychology, Journal of Experimental Psychology: Applied, Journal of Experimental: Psychology Human Perception and Performance, Journal of Experimental Psychology: Learning, Memory & Cognition, and School Psychology Quarterly) and found that only 40 % of them report any index of the ES.

adecuada porque sólo están relacionadas ligeramente con la variable dependiente. Altos valores de TE representan que las variables independientes son muy buenas predictoras de la variable dependiente. Por tanto, el TE es un indicador estadístico importante para evaluar la eficacia de cualquier tratamiento o intervención sobre una respuesta determinada (Ventura 2018). Además, Bologna (2014) plantea que las medidas del TE, al ser estandarizadas superan el inconveniente de las pruebas de hipótesis en cuanto a su dependencia con el tamaño de muestra y sirven para realizar comparaciones entre investigaciones sobre un mismo tema al llevar los resultados a una métrica en común.

El manual de publicación de la American Psychological Association (2001) concluye que es necesario informar el TE junto con el valor de p para responder tres preguntas básicas de la investigación: a) ¿existe un efecto real o los resultados deberían atribuirse al azar?, b) si el efecto es real, ¿qué tan grande es?, y c) ¿es el efecto lo suficientemente grande para considerarse importante o útil?

Por todo lo anterior, se considera al TE como un análisis complementario de las PSHN que ayuda a corregir las limitaciones expuestas por dicha prueba. Sin embargo, a pesar de su utilidad práctica no es frecuente su uso en los reportes de investigación.

Utilización de las medidas complementarias en la literatura. Desde la década de los 90, los especialistas en estadística han sido conscientes de que la PSHN es, en muchos aspectos, insuficiente para interpretar los resultados de las investigaciones. Sin embargo, aún no se logra en la totalidad el empleo de otras medidas complementarias en el reporte de los resultados (Ochoa *et al.* 2019).

La sexta edición del manual de publicación de la American Psychological Association (2010) señaló la necesidad de tomar en cuenta seriamente la potencia estadística suministrando información que evidencie que el estudio tiene la suficiente potencia para hallar efectos de interés sustantivo. Sin embargo, el continuado desinterés por la potencia de las pruebas estadísticas sólo cambiará cuando los editores de las principales revistas exijan este análisis en su política editorial (Frías *et al.* 2002).

En estudios realizados por diferentes autores desde el año 2010, sí se observa incremento en el uso del TE, principalmente en revistas de sicología, pues ya demandan la utilización de este estadístico por norma. Autores como Odgaard y Fowler (2010) revisaron los estudios de intervención publicados en 2003, 2004, 2007 y 2008 en el Journal of Consulting and Clinical Psychology, y encontraron que en general 75 % de los estudios informaron de algún índice del TE.

Sun *et al.* (2010) analizaron los artículos publicados entre 2005 y 2007 en cinco revistas (Journal of Educational Psychology, Journal of Experimental Psychology: Applied, Journal of Experimental: Psychology Human Perception and Performance, Journal of Experimental Psychology: Learning, Memory & Cognition, y School Psychology Quarterly) y encontraron que sólo 40 % de los mismos informaron algún índice del TE.

McMillan and Foley (2011) consulted a total of 417 articles published among 2008 and 2010 in four specialized journals of education and psychology (Journal of Educational Psychology, Journal of Experimental Education, Journal of Educational Research, and Contemporary Educational Psychology) and found that 74 % of the studies informed any measure of the ES. These authors concluded that if the use of ES indexes in the researcher reports has being increased; the discussions about their significance are being poor by the lack of argument or ignorance of what this value represent in the study.

Sesé and Palmer (2012) analyzed the use the statisticians in the in the articles published in 2010 in eight journals (Journal of Behavioural Medicine, Behaviour, Research and Therapy, Depression and Anxiety, Behavior Therapy, Journal of Anxiety Disorders, International Journal of Clinical and Health Psychology, British Journal of Clinical Psychology, and British Journal of Health Psychology). These authors found that the ES indexes report in 61.04 % of the articles.

Caperos and Pardo (2013) verified the articles published in four Spanish journals of many discipline (Anales de Psicología, Psicológica, Psicothema, and Spanish Journal of Psychology), indexed in the database Journal Citation Reports (JCR). Their results show that only 24.3 % of the performed NHST were enclose of the a statistician of the ES and of the statistical power.

Rendón *et al.* (2021) concluded that one of the seven most common mistakes in the articles is to omit the report of the statistical power and the ES. Currently there are some mean stream journals that do not accept the publication of articles of quantitative research in which these statisticians are not reported. From 2020 journals as Memory and Cognition, Educational and Psychological Measurement, Measurement and Evaluation in Counseling and Development, Journal of Experimental Education and Journal of Applied Psychology, decided to regulate the use of complementary measurements to the NHST in the statistical analysis for the correct interpretation and practical importance of the results (Serdar *et al.* 2021).

Conclusions

It is concluded that the NHST is not enough to perform a rigorous assessment of the obtained data in a research. It is considered necessary to include in the studies reports other statistical tests, as the test power and the effect size, to offer a more complete interpretation of the results. Despite that many authors has made reference to the subject, there is the need of calculate these measures to evaluate the quality of the scientific researchers. It is recommended to the editors of scientific journals to include these statisticians in the editorial rules.

McMillan y Foley (2011) consultaron 417 artículos, publicados entre 2008 y 2010 en cuatro revistas especializadas de educación y psicología (Journal of Educational Psychology, Journal of Experimental Education, Journal of Educational Research, y Contemporary Educational Psychology) y encontraron que 74 % de los estudios informaron alguna medida del TE. Estos autores concluyeron que, si bien se había incrementado el uso de los índices TE en los informes de investigación, los debates sobre su significación siguen siendo deficientes por falta de argumentación o desconocimiento de lo que representa este valor en el estudio.

Sesé y Palmer (2012) analizaron el uso de estadísticos en los artículos publicados en el 2010 en ocho revistas (Journal of Behavioural Medicine, Behaviour, Research and Therapy, Depression and Anxiety, Behavior Therapy, Journal of Anxiety Disorders, International Journal of Clinical and Health Psychology, British Journal of Clinical Psychology, y British Journal of Health Psychology). Estos autores encontraron que los índices del SE TE informaron en 61.04 % de los artículos.

Caperos y Pardo (2013) examinaron los artículos publicados en cuatro revistas españolas de múltiples disciplinas (Anales de Psicología, Psicológica, Psicothema, y Spanish Journal of Psychology), indexadas en la base de datos Journal Citation Reports (JCR). Sus resultados indican que sólo 24.3 % de las PSHN ejecutadas se acompañaron de un estadístico del TE y de la potencia estadística.

Rendón *et al.* (2021) concluyen que uno de los siete fallos más comunes en los artículos es omitir el reporte de la potencia estadística y el TE. En la actualidad existen algunas revistas académicas de corriente principal que no admiten la publicación de artículos de investigación cuantitativa donde no se reporten estos estadísticos. A partir del año 2020, revistas como Memory and Cognition, Educational and Psychological Measurement, Measurement and Evaluation in Counseling and Development, Journal of Experimental Education y Journal of Applied Psychology, decidieron reglamentar el uso de medidas complementarias a la PSHN en los análisis estadísticos para la correcta interpretación e importancia práctica de los resultados (Serdar *et al.* 2021).

Conclusiones

Se concluye que la PSHN no es suficiente para realizar una valoración rigurosa de los datos obtenidos en una investigación. Se considera necesario incluir en los informes de trabajo otras pruebas estadísticas, como la potencia de la dócima y el tamaño del efecto, para ofrecer una interpretación más completa de los resultados. A pesar de que muchos autores se han referido al tema, aún existe la necesidad de calcular estas medidas para evaluar la calidad de las investigaciones científicas. Se recomienda a los editores de revistas científicas que se incluyan estos estadísticos entre las normas editoriales.

Conflict of interest:

The authors declare that there is not conflict of interest among them.

Authors' contribution:

Mildrey Torres Martínez: conceptualization and paper writing.

Magaly Herrera Villafranca: conceptualization and paper review.

Yaneilys García Avila: search of information and paper review.

Conflicto de intereses:

Los autores declaran no presentar conflictos de intereses en relación con la preparación y publicación de la revisión.

Contribución de los autores:

Mildrey Torres Martínez: Conceptualización y Redacción-borrador original

Magaly Herrera Villafranca: Conceptualización y Redacción-borrador original

Yaneilys García Avila: Conceptualización y Redacción-borrador original

References

Abelson, R.P. 1997. "On the surprising longevity of flogged horses: Why there is a case for the significance test". Psychological Science, 8(1): 12-15, ISSN: 1467-9280. https://doi.org/10.1111/j.1467-9280.1997.tb00536.x.

American Psychological Association. 1994. Manual of the American Psychological Association, 4th ed., Washington D.C, United States: American Psychological Association, 368p. ISBN: 9781557982414, Available: <https://apastyle.apa.org>, [Consulted: April 10, 2022].

American Psychological Association. 2001. Manual of the American Psychological Association, 5th ed., Washington D.C, United States: American Psychological Association, 439p. ISBN: 9781557987901, Available: <https://apastyle.apa.org>, [Consulted: June 14, 2022].

American Psychological Association. 2010. Manual of the American Psychological Association, 6th ed., Washington D.C, United States: American Psychological Association, 272p. ISBN: 9781433805615, Available: <https://apastyle.apa.org>, [Consulted: June 16, 2022].

Antúnez, P., Rubio, E.A. & Kleinn, C. 2021. "Hypothesis testing in forestry, agriculture and ecology: Use and overuse of the 0.05 and 0.01". Ecosistemas y Recursos Agropecuarios, 8(1): 1-5, ISSN: 2007-901X. https://doi.org/10.19136/era.a8n1.2616.

Bakan, D. 1966. "The effect of significance testing in psychological research". Psychological Bulletin, 66(6): 423-437, ISSN: 1939-1455. https://doi.org/10.1037/h0020412.

Bakker, M. & Wicherts, J.M. 2011. "The (mis) reporting of statistical results in psychology journals". Behavior Research Methods, 43(3): 666–678, ISSN: 1554-3528. https://doi.org/10.3758/s13428-011-0089-5.

Bologna, E. 2014. "Estimación por intervalo del tamaño del efecto expresado como proporción de varianza explicada". Evaluar, 14(1): 43-46, ISSN: 1667-4545. https://doi.org/10.35670/1667-4545.v14.n1.11521.

Bono, R. & Arnau Gras, J. 1995. "Consideraciones generales en torno a los estudios de potencia". Anales de Psicología, 11(2): 193-202, ISSN: 1695-2294.

Borges, A., San Luis, C., Sánchez, J.A. & Cañadas, I. 2001. "El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa". Psicothema, 13(1): 174-178, ISSN: 0214-9915. https://doi.org/10.7334/psicothema2001.14462.025.

Botella, J. & Zamora, A. 2017. "El meta-análisis: una metodología para la investigación en educación". Educación XXI, 20(2): 17-38, ISSN: 1139-613X. https://doi.org/10.5944/educXXI.18241.

Caballero, A. 1979. "Tamaños de muestras en diseños completamente aleatorizados y bloques al azar donde la unidad experimental esté formada por grupos de animales". Cuban Journal of Agricultural Science, 13 (3): 225-235, ISSN: 2079-3480.

Caperos, J.M. & Pardo, A. 2013. "Consistency errors in p-values reported in Spanish psychology journals". Psicothema, 25(3): 408-414, ISSN: 0214-9915. https://doi.org/10.7334/psicothema2012.207.

Carver, R.P. 1978. "The case against statistical significance testing". Harvard Educational Review, 48(3): 378-399, ISSN: 0017-8055. https://doi.org/10.17763/haer.48.3t49026164281841.

Carver, R.P. 1993. "The case against statistical significance testing revisited". Journal of Experimental Education, 61(4): 287-292, ISSN: 0022-0973. https://doi.org/10.1080/00220973.1993.10806591.

Chow, S.L. 1988. "Significance test or effect size? " Psychological Bulletin, 103(1): 105-110, ISSN: 1939-1455. https://doi.org/10.1037/0033-2909.103.1.105.

Clark-Carter, D. 1997. "The account taken of statistical power in research published in the British Journal of Psychology". British Journal of Psychology, 88(1): 71-83, ISSN: 2044-8295. https://doi.org/10.1111/j.2044-8295.1997.tb02621.x.

Cochran W. y Cox, G. 1999. Diseños experimentales. 2nd ed., México: Editorial Trillas, S.A. 75p., ISBN: 968-24-3669-9. Available: <https://www.urbe.edu/UDWLibrary/InfoBook.do?id=5068>, [Consulted: August 3, 2022].

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed., New York, United States: Routledge, 590p., ISBN: 9780805802832, Available: <https://www.routledge.com/books/Statistical-power-analysis-for-the-behavioral-sciences>, [Consulted: August 8, 2022].

Cohen, J. 1990. "Things I have learned (so far) ". American Psychologist, 45(12): 1304-1312, ISSN: 1935-990X. https://doi.org/10.1037/0003-066X.45.12.1304.

Cohen, J. 1992. "A power primer". Psychological Bulletin, 112(1): 155-159, ISSN: 1939-1455. https://doi.org/10.1037/0033-2909.112.1.155.

Cohen, J. 1994. "The earth is round ($p < 0.05$) ". American Psychologist, 49(12): 997-1003, ISSN: 1935-990X. https://doi.org/10.1037/0003-0.66X.49.12.997.

Cohen, J. 1997. Much ado about nothing. Conference presented at the annual meeting of the American Psychological Association, Chicago, United States.

Cortina, J.M., & Dunlap, W.P. 1997. "Logic and purpose of significance testing". Psychological Methods, 2(2): 161-172, ISSN: 1939-1463. https://doi.org/10.1037/1082-989X.2.2.161.

Craig, J.R., Eison, C.L. & Metze, L.P. 1976. "Significance tests and their interpretation: An example utilizing published research and omega-squared". Bulletin of the Psychonomic Society, 7(3): 280-282, ISSN: 0090-5054. https://doi.org/10.375/bf03337189.

De la Fuente, E.I. & Díaz-Batanero, C. 2004. "Controversias en el uso de la inferencia en la investigación experimental". Metodología de las Ciencias del Comportamiento, 5(1): 161-167, ISSN: 1575-9105.

Díaz-Batanero, C., Lozano-Rojas, O.M. & Fernández-Calderón, F. 2019. La controversia sobre el contraste de hipótesis: Situación actual en psicología y recomendaciones didácticas. En: Contreras, J.M., Gea, M.M., López M.M. & Molina E. (eds.), Actas del Tercer Congreso Internacional Virtual de Educación Estadística. España, Available: <www.ugr.es/local/fqm126/civeest.html>, [Consulted: July 12, 2022]

Falk, R., & Greenbaum, C. W. 1995. "Significance tests die hard: the amazing persistence of a probabilistic misconception". Theory and Psychology, 5(1): 75-98, ISSN: 1461-7447. https://doi.org/10.1177/0959354395051004.

Faulkenberry, T.J. 2022. Psychological statistics, the basics. 1st ed., New York, United States: Routledge, 122p., ISBN: 97811032020952, Available: <https://www.routledge.com/books/Psychological-statistics,-the-basics>, [Consulted: October 18, 2022].

Fisher, R.A. 1925. Statistical methods for research workers. 1st ed., Escocia: Genesis Publishing, 269p., ISBN: 4444000761336. Available: <https://www.iberlibro.com/buscar-libro/titulo/statistical-methods-research-workers/autor/sir-ronald >, [Consulted: May 18, 2022].

Fisher, R.A. 1935. The design of experiments. 1st ed., London: Oliver and Boyd, 256p., ISBN: 0028446909. Available: <https://www.iberlibro.com/buscar-libro/titulo/statistical-methods-research-workers/autor/sir-ronald >, [Consulted: June 5, 2022].

Fisher, R.A. 1950. Contributions to mathematical statistics. New York, United States: John Wiley & Son, 600p., ISBN: 9780678008898. Available: Rothamsted Research, https://repository.rothamsted.ac.uk, [Consulted: September 10, 2022].

Fisher, R.A. 1955. "Statistical methods and scientific induction". Journal of the Royal Statistical Society, Series B, 17(1): 245-251, ISSN: 1369-7412.

Frías Navarro, M.D., Pascual Llobel, J. & García Pérez, J.F. 2000. "Tamaño del efecto del tratamiento y significación estadística". Psicothema, 12(Suplemento): 236-240, ISSN: 0214 - 9915.

Frías, M.D., Pascual, J. & García, J.F. 2002. "La hipótesis nula y la significación práctica". Metodología de las Ciencias del Comportamiento, 4(1): 181-185, ISSN: 1575-9105.

Fritz, R.W. 1995. "Accepting the null hypothesis". Memory & Cognition, 23(1): 132-138, ISSN: 0090-502X. https://doi.org/10.3758/BF03210562.

Fritz, R.W. 1996. "The appropriate use of null hypothesis testing". Psychological Methods, 1(4): 379-390, ISSN: 1939-1463. https://doi.org/10.1037/1082-989X.1.379.

Funder, D.C. & Ozer, D.J. 2019. "Evaluating effect size in psychological research: Sense and nonsense". Advances in Methods and Practices in Psychological Science, 2(2): 156–168, ISSN: 251-2467. https://doi.org/10.1177/2515245919847202.

Greenwald, A.G., Gonzalez, R., Harris, R.J. & Guthrie, D. 1996. "Effect size and p-values: What should be reported and what should be replicated? " Psychophysiology, 33(2): 175-183, ISSN: 1469-8986. https://doi.org/10.1111/j.1469-8986.1996.tb02121.x.

Guerra, W.C., Herrera, M., Fernández, L. & Rodríguez, N. 2019. "Modelo de regresión categórica para el análisis e interpretación de la potencia estadística". Cuban Journal of Agricultural Science, 53(1): 13-20, ISSN: 2079-3480.

Hagen, R.L. 1997. "In praise of the null hypothesis statistical test". American Psychologist, 52(1): 15-24, ISSN: 1935-990X. https://doi.org/10.1037/0003-066X.52.1.1.

Harlow, L.L., Mulaik, S.A. & Steiger, J.H. 2016. What if there were no significance tests? 2nd ed., New York, United States: Routledge, 444p., ISBN: 9781317242857, Available: https://www.routledge.com/books/ What-if-there-were-no-significance-tests?, [Consulted: August 8, 2022].

Hickey, G.L., Grant, S.W., Dunning, J. & Siepe, M. 2018. "Statistical primer: Sample size and power calculations-why, when and how? " European Journal of Cardio-Thoracic Surgery, 54(1): 4-9, ISSN: 1873-734X. https://doi.org/10.1093/ejcts/ezy169.

Ioannidis, J.P.A. 2018. "The Proposal to Lower P Value Thresholds to .005". Journal of the American Medical Association, 319(14): 1429-1430, ISSN: 0098-7484. https://doi.org/10.1001/jama.2018.1536.

Kuffner, T.A. & Walker, S.G. 2019. "Why are p-Values Controversial? " The American Statistician, 73(1): 1-3, ISSN: 1537-2731. https://doi.org/10.1080/00031305.2016.1277161.

Levin, J.R. 1993. "Statistical significance testing from three perspectives". Journal of Experimental Education, 61(4): 378-382, ISSN: 1940-0683. https://doi.org/10.1080/00220973.1993.10806597.

Manzano, V. 1997. "Usos y abusos del error de Tipo I". Psicológica: Revista de metodología y psicología experimental, 18(2): 153-169, ISSN: 1576-8597.

Marín, L. & Paredes, D. 2020. Valor p, correcta e incorrecta interpretación. Revista Clínica de la Escuela de Medicina de la Universidad de Costa Rica, 10(1): 45-52, ISSN: 2215-2741.

McMillan, J,H. & Foley, J. 2011. "Reporting and discussing effect size: Still the road less treveled". Practical Assessment Research Evaluation, 16(14): 1-12, ISSN:1531-7714. https://doi.org/10.7275/b6pz-ws55.

Menchaca, M.A. 1974. "Tablas útiles para determinar tamaños de muestras en diseño de Clasificación Simple y de Bloques al Azar". Cuban Journal of Agricultural Science, 8 (1): 111-116, ISSN: 2079-3480

Menchaca, M.A. 1975. "Determinación de tamaños de muestra en diseños Cuadrados Latinos". Cuban Journal of Agricultural Science, 9 (1): 1-3, ISSN: 2079-3480.

Menchaca, M.A. & Torres V. 1985. Tablas de uso frecuente en la Bioestadística. Instituto de Ciencia Animal. Cuba.

Morrison, D.E. & Henkel, R.E. 2006. The significance test controversy: a reader. 1st ed., Chicago, United States: Aldine, 352p., ISBN: 9780202300689, Available: https://www.abebooks.com/The-significance-test-controversy:-a-reader, [Consulted: August 6, 2022].

Neyman, J. & Pearson, E.S. 1928. "On the use and interpretation of certain test criteria for purposes of statistical inference". Biometrika, 20A: 175-240, ISSN: 0006-3444. https://doi.org/10.1093/biomet/20A.3-4.263.

Nickerson, R.S. 2000. "Null hypothesis significance testing: a review of an old and continuing controversy". Psychological methods, 5(2): 241-301. ISSN: 1939-1463. https://doi.org/10.1037/1082-989x.5.2.241.

Ochoa, C., Molina, M. & Ortega, E. 2019. "Inferencia estadística: probabilidad, variables aleato-rias y distribuciones de probabilidad". Evidencias en Pediatría, 15(2): 27-32, ISSN: 1885-7388.

Ochoa, C., Molina, M. & Ortega, E. 2020. "Inferencia estadística: contraste de hipótesis". Evidencias en Pediatría, 16(1): 11-18, ISSN: 1885-7388.

Odgaard, E.C. & Fowler, R L. 2010. "Confidence intervals for effect sizes: compliance and clinical significance in the Journal of Consulting and Clinical Psychology". Journal of Consulting and Clinical Psychology, 78(3): 287–297, ISSN: 0022-006X. https://doi.org/10.1037/a0019294.

Ponce, H.F., Cervantes, D.I. &Anguiano, B. 2021. "Análisis de calidad de artículos educativos con diseños experimentales". Revista Iberoamericana para la Investigación y el Desarrollo Educativo. 12(23): 49-79, ISSN: 2007-7467. https://doi.org/10.23913/ride.v12i23.981.

Rendón, M.E, Zarco, I.S. & Villasís, M.A. 2021. "Métodos estadísticos para el análisis del tamaño del efecto". Revista Alergia de México, 68(2): 128-136, ISSN: 2448-9190. https://doi.org/10.29262/ram.v658i2.949.

Rivera, F. 2017. Convivencia del nivel de significación y tamaño del efecto y otros retos de la práctica basada en la evidencia. Boletín Psicoevidencias, No. 48. Junta de Andalucía y Consejería de Salud, Andalucía, España, ISSN: 2254-4046.

Rothman, J. 1978. A show of confidence. New England Journal of Medicine, 299(24): 1362-1363, ISSN: 0028-4793. http://dx.doi. org/10.1056/NEJM197812142992410.

Scheffé, H. 1959. The Analysis of Varianza. New York, United States: John Wiley & Sons, Inc, 477p., ISBN: 0-471-75834-5, Available: https://www.abebooks.com/The-significance-test-controversy:-a-reader, [Consulted: January 6, 2023].

Schmidt, F.L. 1996. "Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers". Psychological Methods, 1(2): 115-129. ISSN: 1082-989X. https://doi.org/10.1037/1082-989X.1.2.115.

Serdar, C.C., Cihan, M., Yücel, D. & Serdar, M.A. 2021. "Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies". Biochemia Medica Journal, 31(1): 1-27, ISSN: 1330-0962. https://doi.org/10.11613/BM.2021.010502.

Sesé, A. & Palmer, A. 2012. "El uso de la estadística en psicología clínica y de la salud a revisión". Clínica y Salud, 23(1): 97-108, ISSN: 2174-0550.

Sun, S., Pan, W. & Wang, L.L. 2010. "A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology". Journal of Educational Psychology, 102(4): 989-1004, ISSN: 1939-2176. https://doi.org/10.1037/a0019507.

Thompson, B. 1988. "A note about significance testing". Measurement and Evaluation in Counseling and Development, 20(4): 146-148, ISSN: 1947-6302. https://doi.org/10.1080/07481756.1988.12022864.

Thompson, B. 1989. "Asking «what if» questions about significance tests". Measurement and Evaluation in Counseling and Development, 22(2): 66-68, ISSN: 1947-6302. https://doi.org/10.1080/07481756.1989.12022912.

Thompson, B. 1996. "AERA editorial policies regarding statistical significance testing: Three suggested reforms". Educational Researcher, 25(2): 26-30, ISSN: 0013-189X. https://doi.org/10.2307/1176337.

Thompson, B. 1997. If statistical significance tests are broken/misused, what practices should supplement or replace them? Conference presented at the annual meeting of the American Psychological Association, Chicago, United States.

Thompson, B. 1999. "If statistical significance tests are broken/misused, what practices should supplement or replace them? " Theory and Psychology, 9(2): 165-181, ISSN: 1461-7447. https://doi.org/10.1177/095935439992.

Valera, S., Sánchez, J. & Marín, F. 2000. "Contraste de hipótesis e investigación psicológica española: Análisis y propuestas". Psicothema, 12(2): 549-582, ISSN: 0214-9915.

Venereo, A. 1976. "Número de réplicas en diseños cuadrados latinos balanceados para la estimación de efectos residuales". Cuban Journal of Agricultural Science, 10(3): 237-246, ISSN: 2079-3480.

Ventura, J. 2018. "Otras formas de entender la d de Cohen". Revista Evaluar. 18(3):73-78, ISSN: 1667-4545. https://doi.org/10.35670/1667-4545.v18.n3.22305.

Verdam, M.G., Oort, F.J. & Sprangers, M.A. 2014. "Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing". Quality of Life Research, 23(1): 5-7, ISSN: 1573-2649. https://doi.org/10.1007/s11136-013-0437-2.

Wasserstein, R.L. & Lazar, N.A. 2016. "The ASA's Statement on p-Values: Context, Process, and Purpose". The American Statistician, 70(2): 129-133, ISSN: 1537-2731. https://doi.org/10.1080/00031305.2016.1154108.

Wilkinson, L., & TFSI - Task Force on Statistical Inference. 1999. "Statistical methods in psychology journals: Guidelines and explanations". American Psychologist, 54(8): 594-604, ISSN: 0003-066X. https://doi.org/10.1037/0003-066X.54.8.59.

Received: June 10, 2023
Accepted: September 15, 2023