Análisis de los errores de aplicación de la inferencia estadística

Analysis of application errors of statistical inference

María del Mar López-Martín[1], Elena Molina-Portillo[2], José Miguel Contreras[2] y Felipe Ruz[2]

[1]Universidad de Almería, [2]Universidad de Granada, España

Resumen

La estadística inferencial es ampliamente utilizada en las investigaciones científicas con el fin de sustentar las hipótesis de estudio. Sin embargo, su interpretación en estos contextos denota cierta carencia del "sentido estadístico" necesario a nivel profesional. Este hecho ha sido abordado en diversas investigaciones con cierta controversia, ya que su lógica involucra gran cantidad de conceptos, provocando que su uso e interpretación a veces no sea el adecuado. El presente trabajo tiene como objeto analizar la interpretación de dos herramientas estadísticas utilizadas con gran frecuencia: los test estadísticos y los intervalos de confianza. Posteriormente concluimos el estudio con la revisión de las principales dificultades que conlleva la interpretación de dichos conceptos inferenciales.

Palabras clave: test estadísticos, intervalos de confianza, errores de interpretación

Abstract

Statistical inference is widely applied in scientific research to support the study hypothesis. However, its interpretation in these contexts denotes a lack of the "statistical sense" necessary at the professional level. This fact has been addressed in previous investigations with some controversy, since its logic involves a large number of concepts, which makes its use and interpretation sometimes not appropriate. The main of this paper is to analyze the interpretation of two statistical tools frequently used: statistical tests and confidence intervals. Subsequently, we conclude our study with the review of the main difficulties involved in the interpretation of these inferential concepts.

Keywords: statistic tests, confidence interval, misunderstanding

1. Introducción

En el ámbito científico, y concretamente en los estudios de investigación, toma especial relevancia la estadística inferencial con la pretensión de fundamentar las hipótesis iniciales de investigación y las conclusiones obtenidas. En este sentido es fundamental que el investigador posea una adecuada cultura estadística, un dominio en el análisis de datos y un amplio conocimiento del razonamiento o pensamiento estadístico involucrado en su procedimiento, es decir, que posea un amplio *sentido estadístico* (Batanero, 2013).

Tradicionalmente se ha fomentado producir un ciudadano estadísticamente culto al finalizar la secundaria (Franklin y cols., 2007). Dado el interés por plantear preguntas de investigación, recopilar y analizar datos, los últimos compendios estadísticos (Carver y cols., 2016) promueven el pensamiento y razonamiento estadístico restando interés sobre la cultura estadística (Schield, 2017). En esta línea, cobra especial relevancia el análisis del sentido estadístico, principalmente en contextos científicos o profesionales.

Actualmente, la implementación de las técnicas estadísticas se complementa con la utilización de software estadístico especializado que, si bien tienen la ventaja de agilizar los procesos, no requiere los fundamentos estadísticos necesarios.

López-Martín, M.M., Molina-Portillo, E., Contreras, J. y Ruz, F. (2019). Análisis de los errores inferenciales en el ámbito científico. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística*. Disponible en www.ugr.es/local/fqm126/civeest.html

Por tal motivo, se demanda que el usuario de dichas herramientas tenga conocimientos avanzados de interpretación de los objetos matemáticos, relacionados con la inferencia estadística, utilizados habitualmente en las investigaciones.

Los test estadísticos y los intervalos de confianza son métodos frecuentemente implementados en inferencia estadística. El uso extendido de los test estadísticos podría deberse, en cierta medida, al pensamiento ampliamente extendido de que "la ciencia avanza planteando hipótesis y haciendo observaciones sobre ella" y de manera informal se emplea en la vida cotidiana para tomar decisiones (Prieto y Herranz, 2005). En este contexto, el test estadístico es una técnica que debe ser utilizada por profesionales cualificados, ya que su amplia difusión ha conllevado un uso ambiguo por parte de la comunidad científica. En particular, se ha usado y abusado del concepto de significación estadística y del *p*-valor para decidir si se publican resultados en revistas, sólo porque se obtienen "resultados estadísticamente significativos" (Díaz, Batanero y Wilhelmi, 2008). Este hecho, entre otros, ha dado pie a que asociaciones tales como American psychological association (APA, 2010) recomienden encarecidamente el uso de los intervalos de confianza para fundamentar los estudios de investigación (Cumming, 2013). En consecuencia, en las últimas décadas el uso de los intervalos de confianza en los trabajos de investigación ha tenido gran auge ya que, además de proporcionar la misma la información respecto a la significación estadística, conduce a conclusiones correctamente razonadas y decisiones debidamente justificadas (Fidler y Cumming, 2005).

El presente trabajo consiste en una reflexión sobre la controversia existente a la hora de interpretar los procedimientos asociados tanto a los test estadísticos como a los intervalos de confianza. Para tal fin, se analiza en la Sección 2 el sentido estadístico (Batanero, 2013) necesario en el ámbito científico. Posteriormente, la sección 3 introduce la inferencia estadística centrándonos en los conceptos asociados a los test estadísticos e intervalos de confianza. En la sección 4, se analizan algunos trabajos de investigación que reflejan la complejidad que entrama una correcta interpretación de los conceptos estadísticos involucrados. Para concluir, en la sección 5, se realiza una reflexión sobre el contenido objeto de este trabajo.

2. El sentido estadístico

Actualmente, la estadística está implementada en la mayoría de los currículos oficiales durante toda la etapa educativa. En parte, esto es debido a que instituciones internacionales relevantes tales como la Organización de Naciones Unidas (ONU), la Organización para la Cooperación y el Desarrollo Económico (OCDE) o la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) hacen hincapié en la importancia de la estadística en la sociedad. Sin embargo, en la práctica, la enseñanza de la estadística, se transmite de forma que los estudiantes no llegan a entender totalmente su sentido ni alcance (Shaughnessy, 2007).

Para establecer las bases en las que se debe sustentar la didáctica de esta materia, Batanero (2013) define el "sentido estadístico" como la unión entre cultura y razonamiento estadístico. Dado que ambas consideran fundamental unas actitudes adecuadas, en las últimas investigaciones se presentan las actitudes hacia la estadística como una de las componentes que describen el sentido estadístico (Batanero, Begué y Gea, 2018). A continuación se detallan cada una de las componentes:

La primera componente se corresponde con la terminología bibliográfica introducida por Batanero (2005), "cultura estadística". Ésta engloba la estadística básica que cualquier ciudadano debe poseer para desenvolverse plenamente en la sociedad y requiere de un dominio avanzado de las denominadas "ideas estadísticas fundamentales".

La segunda componente involucra al pensamiento o razonamiento estadístico. Existen varios modelos para definir dicho razonamiento estadístico tales como el descrito por Wild y Pfannkuch (1999), quienes distinguen cuatro dimensiones: (i) el ciclo de investigación; (ii) los modos fundamentales de razonamiento; (iii) el ciclo de interrogación y (iv) una serie de actitudes.

La tercera y última componente es la relacionada con las actitudes hacia la estadística. Batanero, Begué y Gea (2018) señalan que dichas actitudes pueden ser positivas o negativas, con distinto grado de intensidad y referidas tanto a la utilidad del tema como a la propia percepción de la competencia para aprenderlo, el interés hacia el mismo y una conveniente disposición para su aplicación.

Estas componentes se desarrollan progresivamente, según el grado requerido, a lo largo de la etapa educativa obligatoria, así como durante el bachillerato, siendo consolidadas tanto en los grados universitarios que lo requieran como en los ámbitos profesionales especializados. En el modelo del sentido estadístico nos centramos, más que en una introducción informal de estos conceptos, en el dominio de las principales técnicas formales de inferencia estadística aplicadas al campo de especialización específico. Por tanto, el presente trabajo tiene como objetivo poner el foco de atención en el ámbito científico que requiere un mayor grado de especialización de las tres componentes.

3. La inferencia estadística

La inferencia estadística es una rama de la ciencia estadística cuyo objeto es obtener conclusiones sobre las características de una población, partiendo del estudio de una o varias muestras representativas de la misma. Entre los distintos procedimientos que proporciona, en el marco de la inferencia paramétrica, destacamos los test estadísticos y los métodos de estimación por intervalos de confianza. Estas técnicas, desde una perspectiva clásica frecuencial, permiten obtener información sobre parámetros poblacionales desconocidos (como la media, proporciones, varianza, etc.) a partir de la evidencia muestral disponible.

Estos procedimientos se asocian a trabajos desarrollados a partir de 1920 por Fisher, Neyman y Pearson (Batanero y Borovcnik, 2016) en los que se fundamentan dichas técnicas, que a día de hoy son ampliamente utilizadas. Sin embargo, tanto los test estadísticos como los intervalos de confianza son procedimientos que entraman cierta complejidad, pues ponen en juego un gran número de conceptos propios de la inferencia estadística (Liu y Thompson, 2009). Entre ellos, destacan las nociones de aleatorización, muestra, variabilidad en el muestreo, parámetro, estadístico, distribución poblacional, distribución muestral, *p*-valor, nivel de significación, hipótesis nula y alternativa, intervalos de confianza, margen de error y nivel de confianza, entre otros. A continuación, se exponen las propiedades y características más destacadas de ambos procedimientos estadísticos.

3.1. Test de significación de Fisher y contraste de hipótesis de Neyman y Pearson

Los test estadísticos permiten conocer si una propuesta sobre los posibles valores que puede tomar uno o varios parámetros puede considerarse cierta. Entre las distintas formas de proceder se destacan los test de significación desarrollados por Fisher (1925) y los contrastes de hipótesis de Neyman y Pearson (1928a, 1928b). Aunque el procedimiento de cálculo en ambas metodologías sea muy similar, el razonamiento y los fines subyacentes en sus aproximaciones a la inferencia son diferentes (Díaz, Batanero y Wilhelmi, 2008).

El test de significación de Fisher es un procedimiento cuyo objetivo se centra en rechazar una hipótesis. Para tal fin, se calcula el grado de credibilidad de la hipótesis nula, conocido habitualmente como *p*-valor, y definido como la probabilidad de encontrar un valor del estadístico muestral igual o más extremo que el observado, suponiendo que la hipótesis nula es cierta. Concretamente, Fisher consideró que cuanto más pequeño fuera el *p*-valor, más fuerte sería la razón para dudar de la hipótesis. Sin precisar los límites, estableció que un *p*-valor elevado, tal vez superior a 0.20, indicaba una evidencia débil, y enfatizó que tal falta de significación estadística, definitivamente, no debe tomarse en el sentido de que la hipótesis sea cierta (Cumming, 2013).

Por otro lado, en la metodología del contraste de hipótesis desarrollada por Neyman y Pearson se introduce, junto a la hipótesis nula (hipótesis de Fisher), una segunda hipótesis de forma mutuamente excluyente, denominada hipótesis alternativa, con el objeto de tomar una decisión sobre la pregunta de investigación planteada. Para tal fin consideran el Error Tipo I (definido como la probabilidad de rechazar la hipótesis nula siendo ésta verdadera) y Error Tipo II (definido como la probabilidad de no rechazar la hipótesis nula sabiendo que ésta es falsa), seleccionando el criterio de decisión que minimice ambos errores. A diferencia de los test de significación de Fisher, el investigador fija de antemano el nivel de significación (o Error Tipo I), permitiendo obtener la región de rechazo o realizar la comparativa con el *p*-valor, con objeto de rechazar o no la hipótesis nula.

Sin embargo, aunque ambas metodologías teóricamente son diferentes, el avance de la tecnología y la difusión de los programas estadísticos han conseguido, en apariencia, desdibujar esas diferencias (Romero, 2012). Además, ha provocado que, dado el uso excesivo y muchas veces incorrecto de los test estadísticos, diferentes asociaciones profesionales como la American psychological association (APA), American educational research association (AERA) y national council on measurement in education (NCME) recomienden complementar este procedimiento con los intervalos de confianza respectivos, para así obtener mayor información sobre la hipótesis analizada (Olivo, 2008).

3.2. Los intervalos de confianza

El propósito de esta herramienta consiste en proporcionar un intervalo de valores que contenga, con alta probabilidad, el verdadero valor de un parámetro poblacional de interés desconocido, a partir de la información obtenida mediante una o varias muestras representativas de la población. De igual forma que toda estimación puntual debe ir acompañada de su precisión, todo estimador debería ir acompañado de un intervalo de confianza que contenga el valor del parámetro de interés con alta probabilidad (Peña, 2014). Concretamente, la construcción de un intervalo de confianza se centra en la búsqueda de dos números reales, $E_{inferior}$ y $E_{superior}$, que definan los extremos de un intervalo, con un nivel de confianza, denotado por $(1-\alpha)\times 100\%$, de manera que se

verifique que $P(E_{inferior} < \theta < E_{superior}) = 1 - \alpha$, donde θ denota el parámetro poblacional que se desea estimar.

En la construcción del intervalo se ponen en juego diversos objetos matemáticos tales como: población, variable aleatoria, parámetro, muestra, estadístico, distribución muestral y dispersión, entre otros. Sin embargo, aunque todos tienen un papel fundamental en la construcción del mismo destacamos, tanto por su significado como por su interpretación, el nivel de confianza $(1 - \alpha) \times 100\%$. El significado asociado al mismo es: si se consideran distintas muestras aleatorias de la misma población, todas con igual tamaño muestral, entonces se verifica que de todos los intervalos de confianza con un nivel de $(1 - \alpha) \times 100\%$ obtenidos a partir de cada una de ellas, en promedio, el $(1 - \alpha) \times 100\%$ de los intervalos contendrán el verdadero valor del parámetro desconocido.

Una de las propiedades más destacadas de los intervalos de confianza es el carácter aleatorio de sus extremos ya que, al depender de los elementos muestrales, varían a medida que alteramos la muestra. Este hecho adquiere gran importancia en su interpretación pues, para el caso particular de una muestra dada, los extremos del intervalo quedan determinados por dos valores de forma unívoca, perdiendo como consecuencia su carácter aleatorio.

Varias son las investigaciones que han centrado su interés en la interpretación de los extremos del intervalo de confianza (Cumming y Fidler, 2005; Fidler y Cumming, 2005; Olivo, 2008). Sin embargo, algunas de ellas han sido criticadas por otros autores. Por ejemplo, Hoekstra y cols. (2014) interpretan el intervalo de confianza al 95% para la media poblacional como: "si se repite un experimento una y otra vez, entonces el 95% de las veces los intervalos de confianza contienen la verdadera media" (p. 1160). Dicha definición fue criticada por Miller y Ulrich (2016) quienes ponen de manifiesto que "aunque es formalmente correcta, no parece del todo satisfactoria porque no menciona los límites inferior y superior que se obtuvieron realmente del análisis de los datos. Cualquier interpretación de los datos muestrales debe resumir de alguna manera la información proporcionada por la muestra" (p. 126).

Desde nuestro punto de vista, la definición más completa sería la dada por Batanero y Borovcnik (2016) donde los autores establecen que

> "La interpretación correcta de un intervalo de confianza al 95% es que, cuando se toman muestras aleatorias repetidas del mismo tamaño de la misma población y se calcula el intervalo de confianza para cada una de estas muestras, en promedio, el 95% de estos intervalos incluirá el parámetro poblacional. Después de seleccionar una muestra en particular y calcular el intervalo para ese parámetro, este intervalo puede o no cubrir el valor del parámetro; es una cuestión de incertidumbre" (p. 189).

A modo ejemplo, la Figura 1 muestra la representación de los intervalos de confianza, a un nivel del 95%, para la media poblacional obtenidos a partir de la simulación de 100 muestras aleatorias simples e independientes con igual tamaño, extraídas de una población con distribución normal de varianza desconocida. Obsérvese que aproximadamente el 95% de los intervalos contiene el verdadero valor del parámetro a estimar y el resto (representados en rojo) no lo cubre. En la práctica confiamos, aunque no con toda seguridad, que la selección de nuestra muestra sea tal que obtengamos como resultado uno de los intervalos que contienen al verdadero valor del parámetro, asumiendo el riesgo de equivocarnos en un 5%.

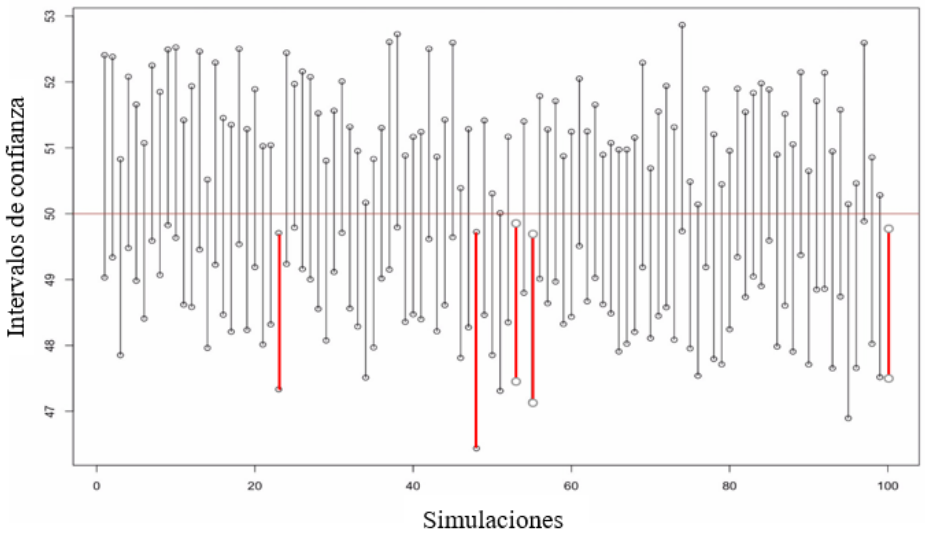

Figura1. Simulación de intervalos de confianza para 100 muestras aleatorias simples

4. Análisis de los errores de interpretación en estadística inferencial

La inferencia estadística probablemente recoge el mayor número de investigaciones sobre errores de interpretación, pues pone en juego nociones abstractas. Dado que los programas estadísticos especializados mecanizan los procedimientos de cálculo, es imprescindible que el usuario de estas herramientas realice una interpretación adecuada de los resultados, en el marco de su investigación. Sin embargo, en el ámbito científico, esto no siempre sucede permitiendo así identificar errores al argumentar los resultados de aplicar las técnicas de inferencia estadística. Dichos errores son debidos principalmente tanto a la falta de destreza en el análisis de datos como a la carencia en el desarrollo del razonamiento estadístico.

En la literatura se encuentran diversos errores relacionados con una incorrecta comprensión de las distintas distribuciones implicadas en la inferencia estadística (Harradine, Batanero y Rossman, 2011) y los asociados a los contrastes de hipótesis y/o intervalos de confianza. A continuación, se realiza un análisis detallado de los errores relacionados con los dos procedimientos inferenciales sobre los que se centra el presente trabajo.

4.1. Errores asociados a la comprensión los test estadísticos

Las investigaciones centradas en la comprensión de los test estadísticos muestran la existencia de concepciones erróneas centradas, principalmente, en la interpretación del *p*-valor y del nivel de significación, α. Estas dificultades tienen su origen, en parte, en una deficiente comprensión del concepto de probabilidad condicional (Falk, 1986).

El principal error de comprensión del *p*-valor, radica en considerarlo como la probabilidad de que la hipótesis nula sea cierta. Recordemos que éste se define como la probabilidad de obtener un valor del estadístico muestral al menos tan extremo como el observado, sabiendo que la hipótesis nula es cierta. Éste, junto con otros tipos de errores asociados a los test de hipótesis, ha sido ampliamente analizado en múltiples trabajos, destacando los desarrollados por Vallecillos (1999) y Batanero (2000). Posteriormente, coincidiendo con dichas investigaciones, Kline (2004) pone de manifiesto las falacias que giran en torno a la interpretación de dicho término. Entre otras, los autores destacan

los errores relacionados con: i) considerar que el *p*-valor es la probabilidad de que el resultado se deba a un error del muestreo, es decir, a un resultado debido al azar; ii) el *p*-valor es la probabilidad de que la hipótesis nula sea cierta dada la información muestral; iii) si la hipótesis nula es rechazada, entonces el *p*-valor es la probabilidad de que esta decisión sea incorrecta; iv) el complementario del *p*-valor (1-*p*) es la probabilidad de que la hipótesis alternativa sea cierta dados los datos; v) el complementario del *p*-valor (1-*p*) es la probabilidad de que un resultado se replicará bajo condiciones constantes (p. 63-65).

Diversos estudios también indagan en los errores de comprensión del nivel de significación. Recordemos que este se define como la probabilidad de rechazar la hipótesis nula sabiendo que es cierta. La principal confusión detectada, relacionada con su interpretación incorrecta, se debe mayoritariamente a la confusión en la probabilidad condicionada (Falk, 1986, Vallecillos, 1999). Concretamente, es debida al intercambio de ambos términos en la probabilidad condicionada, $P(H_0\ cierta \mid H_0\ rechazada)$, describiendo una probabilidad que no puede ser calculada. Por su parte, Falk (1986) señala que otro error interpretativo, relacionado con la probabilidad condicionada, se produce al pensar que el nivel de significación corresponde a la probabilidad de equivocarse al rechazar la hipótesis nula.

El uso de un nivel de significación del 0,05 es cuestión de convenio y no tiene justificación científica. Esto unido al uso excesivo de "*p*-valor < 0,05", sin dar el valor numérico del *p*-valor, junto a una incorrecta interpretación del mismo, ha provocado un abuso excesivo del término *significación* (Prieto y Herranz, 2005). En consecuencia, obtener un nivel de significación menor al 5% deriva en la creencia de concluir con resultados más significativos. Por tal motivo, y siguiendo las recomendaciones de las asociaciones profesionales, se debe facilitar junto al *p*-valor el intervalo de confianza asociado a la estimación del parámetro bajo estudio con el fin de obtener informes de investigación más sólidos.

4.2. Errores de comprensión de los intervalos de confianza

Los intervalos de confianza son ampliamente utilizados como un simple indicador de la significación estadística de los resultados. Este hecho, junto con el equívoco de obviar la probabilidad, aunque pequeña, de que el intervalo hallado no contenga al verdadero valor de dicho parámetro, constituyen dos grandes imprecisiones en el uso de esta técnica inferencial.

Dado que los extremos del intervalo se definen en función de los estadísticos muestrales, dichos límites adquieren un carácter aleatorio. Es importante señalar que una vez determinado el intervalo de confianza para una muestra en particular, desaparece esa aleatoriedad, pues la probabilidad será de uno o cero si el parámetro está o no comprendido entre los límites del intervalo (Cumming y Fidler, 2005, Dracup, 2005; Greenland y cols., 2016).

Otro error relevante es la incorrecta interpretación del nivel de confianza, pues se tiende a pensar que el nivel de confianza indica la probabilidad de que el intervalo hallado contenga el verdadero valor del parámetro. Recordemos que el nivel de confianza al 95% significa que, si tomamos muchas muestras aleatorias, todas de igual tamaño, en promedio, el 95% de los intervalos calculados contendrán al verdadero parámetro poblacional (Batanero y Borovcnik, 2016).

5. Reflexiones finales

A pesar de la relevancia de la inferencia en el ámbito científico, algunas investigaciones han evidenciado que muchos de los estudios estadísticos, enfocados en avalar la hipótesis de investigación, reflejan errores en la interpretación de los conceptos estadísticos subyacentes (Cumming, Williams y Fidler, 2004; Hoekstra y cols., 2014), mostrando un bajo "sentido estadístico" en los investigadores. Principalmente en dicho ámbito, se ha de controlar adecuadamente los objetos matemáticos implicados, asegurar una correcta interpretación de los resultados obtenidos y garantizar una adecuada contextualización de los datos.

Sin embargo, autores como Wasserstein y Lazar (2016) ponen de manifiesto que la comunidad estadística ha estado más preocupada por la replicabilidad de las conclusiones científicas más que en entrar en las definiciones y distinciones de estos términos, generando confusiones y dudas en relación a la validez de la ciencia. Este hecho puede verse influenciado también por el uso de programas estadísticos pues no promueven la interpretación adecuada de los resultados estadísticos obtenidos, concluyendo inadecuadamente sobre la hipótesis que sustenta la investigación.

En la línea de lo promovido en los últimos compendios (GAISE, 2016), consideramos necesario fomentar el razonamiento o pensamiento estadístico en las etapas de formación de los investigadores enfocando el análisis de datos y la asimilación de los objetos matemáticos implicados en el marco de la estadística inferencial en investigación experimental (Batanero, 2000).

6. Referencias

American Psychological Association. (2010). *Publication manual of the American psychological association* (6th ed.). Washington, DC: Author.

Batanero, C. (2000). Controversies around the role of statistical test in experimental research. *Mathematical Thinking and Learning, 2*(1-2), 75-98.

Batanero, C. (2005). Los retos de la educación estadística. *Yupana, 1,* 27-40.

Batanero, C. (2013). Sentido estadístico: componentes y desarrollo. En J. M. Contreras, G. R. Cañadas, M. M. Gea y P. Arteaga (Eds.). *Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria* (pp. 55-61). Granada: Departamento de Didáctica de la Matemática de la Universidad de Granada.

Batanero, C., Begué, N. y Gea, M. M. (2018, Octubre). *¿Cómo desarrollar el sentido del muestreo en los estudiantes?* Ponencia presentada en el tercer encuentro colombiano de educación estocástica. Bogotá, Colombia.

Batanero, C. y Borovcnik, M. (2016). *Statistics and probability in high school*. Springer.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Holmes, G., Velleman, P., Witmer, J. y Wood, B. (2016). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: American Statistical Association.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G., Williams, J y Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*(4), 299-311.

Díaz, C., Batanero, C. y Wilhelmi, M. R. (2008). Errores frecuentes en el análisis de datos en educación y psicología. *Publicaciones, 38*, 9-23.

Dracup, C. (2005). Confidence intervals. En B. S. Everitt y D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*, (Vol. 1 A–D pp. 366–375): Wiley.

Falk, R. (1986). Conditional probabilities: insights and difficulties. En R. Davidson y J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics.* (pp. 292 – 297). Victoria, Canada: International Statistical Institute.

Fidler, F. y Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th International Statistics Institute Session*. Sydney, Australia: International Statistical Institute. Online: www.stat.auckland.ac.nz/~iase/publications.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver y Boyd.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. y Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association. Online: ww.amstat.org/Education/gaise/.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. y Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337-350.

Harradine, A., Batanero, C. y Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. En C. Batanero, G. Burrill y C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 235-246). Netherlands: Springer.

Hoekstra, R., Morey, R. D., Rouder, J. N. y Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157-1164.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington D.C.: APA.

Liu, Y. y Thompson, P. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies: An International Journal*, *4*(2), 126-138.

Miller, J. y Ulrich, R. (2016). Interpreting confidence intervals: a comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review*, *23*(1), 124-130.

Neyman, J. y Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20A*, 175-263.

Neyman, J. y Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, *20A*, 264-294.

Olivo, E. (2008). *Significados del intervalo de confianza en la enseñanza de la ingeniería en México*. Tesis Doctoral. Universidad de Granada.

Peña, D. (2014). *Fundamentos de estadística*. España. Alianza editorial.

Prieto, L. y Herranz, I. (2005). *¿Qué significa "estadísticamente significativo"?: la falacia del criterio del 5% en la investigación*. Madrid: Díaz de Santos.

Romero, N. (2012). La revolución en la toma de decisiones estadísticas: el p-valor. *Telos*, *14*(3), 439 - 446.

Schield, M. (2017). *GAISE 2016 promotes statistical literacy*. *Statistics Education Research Journal*, *16*(1), 50-54.

Shaughnessy, J. M. (2007). *Research on statistics learning and reasoning.* En F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1010). Greenwich, CT: Information Age y NCTM.

Shaughnessy, J. M., Chance, B. y Kranendonk, H. (2009). *Focus in high school mathematics: Reasoning and sense making in statistics and probability*. Reston, VA: NCTM.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Proceedings of the 52 session of the International Statistical Institute* (Vol.2, pp. 201–204). Helsinki: International Statistical Institute.

Wasserstein, R. L. y Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Wild, C., y Pfannkuch, M. (1999). Statistical thinking in empirical. *International Statistical Review*, *67*(3), 223-265.