PRUEBAS DE HIPÓTESIS ESTADÍSTICAS: ALGUNAS CONSIDERACIONES PARA LA PRÁCTICA DOCENTE.

Rodríguez, María Inés (*); Agnelli, Héctor (*); Albert Huerta Armando(**)
(*) Universidad Nacional de Río Cuarto; (**) Tecnológico de Monterrey, México.
mrodriguez@exa.unrc.edu.ar

RESUMEN

La lógica de la inferencia estadística, en particular las pruebas o tests de hipótesis, presentan dificultades conceptuales vinculadas a la filosofía y a la psicología que la hacen susceptible de interpretaciones incorrectas. Además, desde los inicios del desarrollo de esta metodología surgió una fuerte controversia conceptual entre prominentes impulsores de la misma. Con el transcurso del tiempo, estas diferencias han quedado ocultas al adoptarse una metodología que reúne aspectos de las distintas corrientes contrapuestas. Este proceso de síntesis, construido sobre la base de conceptos no necesariamente conciliables, es el que actualmente se enseña y, en consecuencia, se aplica. Hemos comprobado, a partir del análisis realizado sobre el uso de la inferencia estadística en trabajos de investigación y tesis doctorales en ciencias biológicas, la presencia de sesgos en el uso e interpretaciones de las pruebas de hipótesis, en particular las relacionadas con el nivel de significación y el valor *p* de una prueba.

Consideramos recomendable que durante el desarrollo de los cursos de estadística, se ponga más énfasis en los aspectos conceptuales en los que se basa la inferencia estadística, seleccionando estrategias de enseñanza activa que eviten la simple transferencia de habilidades técnicas y conduzcan en cambio, a priorizar el desarrollo del razonamiento conceptual y la interpretación. Asumimos que el profesor será un buen diagnosticador de los problemas de enseñanza en la medida en que disponga de elementos de juicio fundados para ello. Intentamos con este trabajo contribuir a elucidar y divulgar algunas de las dificultades subyacentes en la metodología de las pruebas de hipótesis.

Palabras clave: inferencia, test de hipótesis, valor p, nivel de significación, efectos, pruebas de significación estadística.

INTRODUCCIÓN

Uno de los problemas principales en un curso introductorio de estadística a nivel universitario es hacer la transición del análisis de datos a la inferencia, Moore (1998). La importancia y preocupación vigente de esta problemática, dentro de la comunidad de investigadores en educación estadística, lo demuestra la realización en Agosto de 2007, del *5º Foro sobre el Razonamiento, Pensamiento y Alfabetización Estadística*, (SRTL-5) [1], que tuvo lugar en la Universidad de Warwick, Coventry, (R.U), cuyo tema central de debate y análisis fue "Razonamiento acerca de la Inferencia Estadística: Maneras innovadoras de conectar la probabilidad y los datos". El 6º Foro tendrá lugar en Australia, en el corriente año, siendo el tema de su convocatoria: "El rol del contexto y la evidencia informal en el Razonamiento Inferencial". Estos foros se realizan cada dos años y su temática se determina en función de las propuestas acerca de las cuestiones que merecen ser indagadas y resueltas con mayor urgencia, formuladas por docentes e investigadores en enseñanza de la estadística.

El rol que la inferencia estadística ha jugado en el desarrollo científico y los problemas filosóficos que ha planteado, han dado origen a multitud de trabajos en el campo de la filosofía y de la filosofía de la ciencia como los de Black (1979); Rivadulla (1991); Lindsay (1988). También desde la psicología ha sido abordada esta problemática por diversos autores como Scholz (1987); Nisbett y Ross (1980); Kahneman, Slovic y Tversky (1982); quienes nos han alertado de la existencia de estrategias incorrectas de razonamiento estadístico que implican sesgos en las conclusiones obtenidas. Al respecto, sostiene Greer (2000), los sesgos en el razonamiento inferencial son sólo un ejemplo del escaso razonamiento de los adultos en lo atinente a problemas probabilísticos, que ha sido extensamente estudiado por los psicólogos en relación con otros conceptos, como la aleatoriedad, probabilidad y correlación.

[1] http://srtl.stat.auckland.ac.nz

Respecto a la evidencia del uso muchas veces incorrecto de la inferencia estadística, Falk y Greenbaum (1995), sostienen: "a pesar de las recomendaciones, los investigadores experimentales persisten en apoyarse demasiado en la significación estadística, sin tener en cuenta los argumentos de que los tests estadísticos por sí solos no justifican suficientemente el conocimiento científico. Algunas explicaciones de esta persistencia incluyen la inercia, confusión conceptual, falta de mejores instrumentos alternativos o mecanismos psicológicos, como la generalización inadecuada del razonamiento en lógica deductiva al razonamiento en la inferencia bajo incertidumbre" (citado en Batanero, 2001).

El propósito de este trabajo es conocer más de cerca la naturaleza de las dificultades que se presentan en estudiantes de postgrado al usar pruebas de Hipótesis en sus investigaciones, para identificar mejores maneras de enseñarlas en los niveles de pregrado. Para ello se presentan algunos de los resultados de la *aproximación cognitiva* sobre las dificultades más sobresalientes reportadas por otros investigadores y las halladas también por nuestra propia investigación, particularmente sobre valor p. Por otra parte, se hace una *aproximación epistemológica* a los Test de hipótesis. Para esto, se presentan elementos sobresalientes sobre la problemática asociada a los test o pruebas de hipótesis; diferencias y similitudes entre los distintos enfoques de los test de hipótesis; consideraciones sobre algunas interpretaciones erróneas del valor p; así como algunas medidas que las comunidades científicas están tomando en consideración para superar, al menos en parte, estas dificultades a través del reporte de tamaños de efectos.

PROBLEMÁTICA ASOCIADA A TEST DE HIPÓTESIS

Frecuentemente se encuentran publicaciones de carácter científico y trabajos de tesis de postgrado en distintas disciplinas, en las que se utilizan habitualmente pruebas de hipótesis estadísticas de forma mecánica y sin comprobación de los supuestos necesarios para la validez de las conclusiones derivadas de su aplicación. Esto muestra un conocimiento incompleto de ellas, así como también desconocimiento de las implicaciones del uso inadecuado de las mismas.

En su servicio de consultoría Bishop (2001), observó que muchos estudiantes de postgrado que concurrían a realizar consultas, carecían de conocimientos estadísticos suficientes para mantener una discusión acerca del rol de la estadística en sus proyectos. Nuestra trayectoria como docentes de cursos de estadística de distintas carreras de postgrado, maestrías y doctorados en ciencias biológicas, ciencias agrarias y ciencias de la salud, en los que se aborda la enseñanza de la inferencia y también la práctica de consultoría, nos ha permitido detectar la presencia de la misma problemática. De aquí entonces surgió la necesidad de investigar las concepciones y dificultades de los alumnos y los usuarios acerca de las pruebas de hipótesis.

Para ese propósito se aplicó un cuestionario a una muestra de estudiantes de las carreras de Agronomía, Ciencias Biológicas, Microbiología y Matemáticas. El cuestionario fue una adaptación del utilizado por Vallecillos (1996), cuya validez y fiabilidad fueron por ella comprobadas. Un cuestionario similar fue utilizado para recolectar información de los alumnos de doctorado en Ciencias Biológicas y además se consideraron algunas tesis de grado y doctorales, para analizar el uso e interpretación de los métodos inferenciales en dichos trabajos.

Los resultados obtenidos en detalle fueron difundidos en trabajos anteriores Rodríguez, M.I. y Albert, A., (2007). A los fines del presente trabajo cabe señalar que las dificultades más frecuentes reveladas por los dos tipos de estudiantes están relacionadas con la interpretación y definición del nivel de significación y del valor ***p.*** Los resultados hallados fueron concordantes con los reportados por investigadores que analizaron el uso de la estadística en otras disciplinas científicas.

Consideramos que estos errores tienen su origen en dificultades conceptuales y epistemológicas del tema ya que, como lo señala Ito(1999), existen distintos aspectos que interactúan en esta problemática, que son:

(a) La disputa dentro de la misma estadística, donde diferentes métodos e interpretaciones varias de los mismos métodos fueron recomendadas por los enfoques de Fisher, Neyman-Pearson y en el enfoque Bayesiano.
(b) La controversia en la aplicación de la estadística, donde, en la práctica el contraste de significación es una mezcla informal de los contrastes de significación originales de Fisher, la teoría de Neyman-Pearson y conceptos e interpretaciones que no son partes de esta última.
(c) La controversia en la enseñanza acerca de cuándo, cómo y con qué profundidad deberíamos enseñar la inferencia estadística.

Basados en estos antecedentes, planteamos a continuación algunas consideraciones epistemológicas relativas al surgimiento y desarrollo de las pruebas de hipótesis, como así también a la interpretación del valor p, en la creencia que su conocimiento (o reflexión) puede contribuir a superar algunos de los errores y dificultades observadas en los estudiantes y en las aplicaciones de los usuarios de la estadística.

DIFERENCIAS Y SIMILITUDES ENTRE LOS DISTINTOS ENFOQUES DE LOS TESTS DE HIPÓTESIS

En el desarrollo teórico de las pruebas de hipótesis estadísticas encontramos tres enfoques, que son: el de R.A. Fisher, para quien los tests sirven para confrontar una hipótesis postulada con los datos observados, donde el *valor p* de la prueba es el que indica la fuerza de la evidencia en contra de la hipótesis postulada. Simultáneamente, Jerzy Neyman y Egon Pearson, desarrollaron un aporte a la teoría introduciendo la idea de que es necesario considerar dos hipótesis, surgiendo así las denominaciones de hipótesis nula y alternativa. Posteriormente, en la segunda mitad del siglo XX cobró un renovado impulso la inferencia Bayesiana basada en la utilización sistemática del Teorema de Bayes, incorporando la información *a priori* para intentar resolver el problema de la asignación de probabilidades de la hipótesis.

Los dos primeros enfoques se ubican dentro de la denominada *Estadística clásica o frecuencial*, mientras que el último está dentro de la considerada *Estadística bayesiana*, poco mencionada en los cursos introductorios de estadística, pero de gran desarrollo metodológico y creciente uso en la investigación aplicada en los últimos tiempos.

Consideramos importante aclarar y difundir los diferentes enfoques, debido a que en la actualidad una mezcla de la lógica sugerida por estos distintos enfoques es utilizada por la mayor parte de los investigadores en sus prácticas y por los educadores en su actividad de enseñanza. Generalmente esto ocurre por desconocimiento de sus diferencias, ya que las mismas no radican en los cálculos algorítmicos, sino más bien en el razonamiento subyacente. Por otra parte, la mayoría de los libros de texto optan por un solo enfoque y son pocos los que hacen explícitas las ideas sostenidas por los diferentes autores. Entre éstos, podemos citar el de D. Moore (1998) que si bien desarrolla el tema desde la teoría de Fischer aduciendo que es la utilizada por la mayoría de los usuarios de la estadística también alude a las ideas de Neyman y Pearson y cita la referencia para conocer el tema desde ese enfoque. Otro ejemplo es el texto de Gómez Villegas (2005) que presenta todos los temas desde dos perspectivas: la de los métodos clásicos y la Bayesiana, sin ningún menoscabo de uno frente a otro, construyendo los contrastes de hipótesis bajo el enfoque de Neyman y Pearson.

Cabe señalar también que las notas de clase de la mayoría de los cursos básicos de estadística, utilizados en distintas universidades, nacionales como también en las extranjeras, no hacen referencia alguna al desarrollo histórico del tema. Tampoco hacen mención a las motivaciones que provocaron su surgimiento, que en gran medida están vinculadas al desarrollo de otras disciplinas. Esto origina las concepciones y terminologías que fueron causa de discusiones entre sus autores, que no se encuentran superadas hasta el momento.

- **Las pruebas de significación de Fisher**

La clave de los Tests fisherianos es que no hacen referencia a la hipótesis alternativa. Realmente pueden ser considerados como un procedimiento de validación de un modelo. Se tiene la distribución del modelo, planteada en la hipótesis nula, y se examina si los datos parecen raros o adecuados a ese modelo planteado. Para ello se calcula la probabilidad de obtener un resultado tan o más "raro" que el obtenido, a este valor de probabilidad Fisher lo

denominó el valor p del test. En este contexto, el nivel de significación α es simplemente una regla de decisión por la cual se rechaza o no la hipótesis nula. En otras palabras, es un valor de probabilidad que indica cuán raros deben ser los datos para rechazar la hipótesis nula. Implícitamente el nivel α determina qué datos llevarían al rechazo de la hipótesis nula y cuáles no, Christensen (2005). En este enfoque el valor *p* es realmente un concepto más importante que el valor α. Cuando $p < \alpha$ se dice que los resultados son estadísticamente significativos.
Encontramos en Rivadulla (1991, p. 150) la siguiente definición acerca de los test de Fisher: "Un *test de significación estadístico* es, pues, un procedimiento que permite dividir los resultados experimentales en dos clases: la de aquellos que muestran una discrepancia significativa respecto de una cierta hipótesis, la *hipótesis testeada o hipótesis nula* y la de los que concuerdan con ella. Los resultados experimentales tienen por consiguiente como objeto testar la significación de las desviaciones que muestran las observaciones respecto de una hipótesis nula determinada, que especifica la frecuencia con que ocurrirán los diferentes resultados experimentales. Si la hipótesis nula es verdadera y la clase de resultados que se oponen a ella acontecen con una frecuencia de, por ejemplo 5%, dispondremos entonces de un test por medio del cual juzgar, a un nivel de significación conocido, si los datos contradicen o no la hipótesis testeada".

Para redondear este enfoque, podemos decir que para Fisher los tests de significación sirven para confrontar la hipótesis nula con los datos observados, siendo el valor *p* de la prueba el que indica la fuerza de la evidencia en contra de la hipótesis nula. Según su punto de vista, las pruebas estadísticas no deben ser utilizadas para realizar inferencias inductivas de la muestra a la población, sino más bien, una inferencia deductiva de la población de las posibles muestras a la muestra particular obtenida en cada caso. Para él habría que tener en cuenta que la hipótesis nula nunca resulta probada en el curso de un experimento, pero sí posiblemente refutada. "Las pruebas de significación se diseñan con el objetivo de valorar la fuerza de la evidencia en contra de la hipótesis nula. En general, la hipótesis nula es una afirmación de *ausencia de efecto* o de *no diferencia."* Moore (1998a, p 364).

Podemos señalar como características de las pruebas de significación estadística, según Fisher, los siguientes puntos:

- Se plantea sólo una hipótesis, la nula.
- Un test sólo sirve para rechazar una hipótesis, nunca para confirmarla.
- El *valor p* de la prueba, indica la fuerza de la evidencia en contra de la hipótesis postulada.
- No existen errores de tipo II, esto es, no se considera el error que se comete al aceptar una hipótesis nula, bajo el supuesto que sea falsa, ya que nunca se acepta esta hipótesis.

- **El test de hipótesis como proceso de decisión**

Mientras que en el enfoque anterior la prueba se utiliza para evaluar la fuerza de la evidencia en contra de una hipótesis, para Neyman y Pearson (N-P) el problema de una prueba de hipótesis estadística aparece en circunstancias en que estamos forzados a realizar la elección entre dos cursos de acción. Esta dupla defendió una versión denominada contraste de hipótesis que mezcla la idea de prueba de significación y la idea de regla para tomar decisiones. Con N-P nace la idea de que es necesario considerar otra hipótesis, surgiendo así la hipótesis alternativa, dando origen al concepto de test uniformemente más potente.
En este enfoque se plantea lo siguiente: considerando una muestra aleatoria $(x_1,\ldots, x_n)$ de una población X con función de densidad f_Θ, donde $\theta \in \Theta$, se trata de decidir con base en la información que brinda esta muestra si es posible que la misma provenga del espacio paramétrico Θ_0 o del espacio Θ_1, siendo Θ_0 y Θ_1 una partición previamente elegida del espacio paramétrico Θ. Dicho en lenguaje estadístico, queremos contrastar la hipótesis nula $H_0 : \Theta \in \Theta_0$ frente a la hipótesis alternativa $H_1: \Theta \in \Theta_1$, para decidir con la evidencia que brinda la muestra, cuál de ellas es válida.

> *Dos son, pues, las ideas que hemos de tener en cuenta a la hora de testar una hipótesis. Primera, que debemos admitir la existencia de hipótesis alternativas a la hipótesis testada, ya que no hay ninguna razón por la que tales hipótesis no deban ser tenidas en cuenta a la hora de elegir un test apropiado; y segunda, que debemos procurar evitar los errores que se pueden cometer en la aplicación del test, dicho de otro modo: hay que tratar de minimizar su frecuencia*. Neyman (1952, p.43 y 54-55, citado en Rivadulla (1991, p. 161))

Dado que ahora en cualquier test hay dos hipótesis en competencia N-P pueden definir a priori α , que indica la probabilidad rechazar equivocadamente H_0 y β la probabilidad de no rechazar H_0 cuando es falsa. Por convención se denomina error de Tipo I al rechazo incorrecto de H_0 y error de Tipo II a su "aceptación" cuando no corresponde. A la probabilidad 1-β se la conoce como la potencia del test, e indica la capacidad que éste tiene para rechazar una hipótesis nula falsa.

Resumiendo, los rasgos característicos del enfoque de Neyman-Pearson son los siguientes:

- Se da una interpretación a la hipótesis alternativa.
- Se reconocen dos tipos de errores: el de tipo I y el de tipo II.
- Se trata de buscar el test con alto valor de potencia (1- β), para un determinado valor de significación, α.

- **Enfoque Bayesiano de los tests**

Los fundamentos del enfoque Bayesiano a la inferencia están basados en los trabajos de Thomas Bayes (1701- 1761), cuya publicación póstuma fue realizada por Price en 1763, bajo el título *Un ensayo hacia la resolución de un problema en la doctrina del azar.* La aproximación Bayesiana al contraste de hipótesis continuó desarrollándose hasta la aparición de los trabajos de E. Pearson, J. Neyman y R. Fisher, que interrumpieron la utilización de los métodos Bayesianos, básicamente por depender éstos de la distribución inicial, lo que hace que distintas personas con diferente información inicial y la misma información muestral lleguen a conclusiones distintas. Es decir, ninguno de los autores citados utilizó el teorema de Bayes de la probabilidad condicional como base para la inferencia estadística, y particularmente Fisher fue reacio a ello. Continuaron con el enfoque Bayesiano, entre otros, Ramsey con su publicación *Fundamentos de Matemáticas*, en 1931; Jeffreys, con su *Teoría de la Probabilidad* en 1939; De Finetti, 1974; Lindley en 1975, quienes han promovido que actualmente los métodos Bayesianos, en particular los contrastes de hipótesis, presenten un elevado desarrollo.

El enfoque Bayesiano considera el parámetro μ como una cantidad aleatoria con una distribución de probabilidad conocida, distribución a priori, que expresa la información imprecisa que se tiene acerca de su valor. Por ejemplo, la altura media μ de todos los estudiantes de una escuela si bien puede ser incierta, no es del todo desconocida. En la perspectiva Bayesiana, el concepto de probabilidad se amplía para incluir *probabilidades subjetivas* o personales. Lo que es nuevo aquí no son las matemáticas, que permanecen iguales, sino la interpretación de la probabilidad como representación de una estimación subjetiva de la incertidumbre en vez de considerarla una frecuencia relativa en el largo plazo, como sostiene Moore (1998b) para quien la conclusión Bayesiana es sin lugar a dudas más sencilla de entender que el enunciado clásico (cap. IV, p. 24).

Las características de la escuela Bayesiana se resumen en dos:

- Por un lado utilizan el teorema de Bayes como base para la inferencia estadística.
- Por otro lado, y donde la controversia es mayor, asigna probabilidades a una hipótesis.

Es importante resaltar que la probabilidad de que la hipótesis nula sea cierta una vez que la hemos rechazado y la probabilidad de que la hipótesis nula sea cierta una vez que hemos obtenido los datos, no pueden conocerse a partir del contraste de hipótesis, en las acepciones de Fisher o de Neyman-Pearson. El único enfoque que permite calcular estas probabilidades a posteriori de la hipótesis es el Bayesiano, por el cual el experimentador puede revisar su grado de aprobación de la hipótesis, en vista de los resultados obtenidos.

CONSIDERACIONES SOBRE ALGUNAS INTERPRETACIONES ERRÓNEAS DEL VALOR *P*

A continuación señalamos algunas de las interpretaciones incorrectas más habituales relativas al significado del valor *p*.

1. Creencia acerca de que *p* es la probabilidad de que la hipótesis nula sea cierta y que 1-*p* es la probabilidad de que H_a sea verdadera.

Falk y Greenbaum (1995) han llamado a esta situación "la ilusión de la demostración probabilística por contradicción". El valor p es la probabilidad de que el estadístico del test sea tan o más extremo que el valor obtenido, digamos D, condicional sobre que la hipótesis nula H_0 sea verdadera : $p = \Pr (D / H_0)$ que es distinto a la probabilidad de que H_0 sea

verdadera condicional sobre el resultado observado Pr (H_0 /D). Existe una tendencia a ver estas dos probabilidades como equivalentes, a esta situación Dawes (1988) la llamó "la confusión de la probabilidad inversa".

2. **Creencia acerca de que el rechazo de H_0 establece la verdad de la teoría que predice que H_0 es falsa.**

Esta interpretación errónea surge al plantear: Si la teoría es verdadera, la que me interesa, entonces la hipótesis nula será rechazada. Como H_0 fue rechazada entonces la teoría debe ser verdadera. Esto es incurrir en la falacia lógica de afirmar el consecuente (si P entonces Q, como se cumple Q, entonces P)

Aún si uno interpreta la significación estadística como evidencia en contra de la hipótesis que un efecto observado fue debido al azar, la significación estadística no garantiza por si misma la conclusión de que una explicación específica no aleatoria sea verdadera. Para que esto acontezca deben ser eliminadas otras posibles explicaciones no aleatorias en competencia Erwin (1998). A esto último contribuye un diseño adecuado de la experiencia mediante el control de potenciales factores de confusión Hayes (1998).

3. **Creencia acerca de que pequeños valores de *p* constituyen evidencia a favor de la replicación de los resultados**.

Esta creencia considera que un valor pequeño de *p* se toma como una evidencia a favor de que los resultados obtenidos se repitan en otro experimento (Shaughnessy, 1994). Sin embargo, un valor pequeño de p, no garantiza que en otro experimento llevado a cabo bajo las mismas condiciones se obtenga el mismo efecto, en dirección y tamaño, o considerando la hipótesis nula, que ésta vuelva a rechazarse. Para que exista reproducibilidad se requiere alta potencia en el nuevo estudio Schmit y Hunter (1997).

4. **Creencia acerca de que pequeños valores de *p* significan un efecto de tratamiento de gran magnitud**.

El valor p no es indicativo de la magnitud del efecto: si se tiene una muestra de gran tamaño o con poca variabilidad, un efecto no importante puede conducir a un p pequeño y por otra parte una muestra chica, o con gran variabilidad en la misma, pueden llevar a un p grande, aún cuando la magnitud del efecto sea importante. Creencia acerca de que la significación estadística implica significación práctica.

El hecho de que un valor p sea pequeño y se rechace en consecuencia H_0, no implica asumir la importancia práctica que pueda tener el efecto declarado significativo por el test. Sobre este particular Thompson (1996) sugiere utilizar la frase "estadísticamente significativo" en lugar de la palabra significativo al describir el rechazo de la hipótesis nula, para evitar una posible confusión con la significación o importancia desde el punto de vista práctico.

REPORTE DE TAMAÑOS DE EFECTOS

Algunos de los problemas antes planteados pueden quedar superados al informar el tamaño de los efectos encontrados. Así, la American Psychological Association (2001) establece en sus recomendaciones para la publicación de trabajos de investigación: "El principio general a seguir es proporcionar al lector no únicamente la información referida a la significación estadística sino también dar la información suficiente para evaluar la relaciones o efectos observados" (p.26). Acerca de si es más conveniente utilizar tamaños de efecto o intervalos de confianza Kirk (1996) sugiere que cuando las mediciones están expresadas en unidades habituales se debería usar un intervalo de confianza, y en caso contrario, informar sólo el tamaño del efecto. De cualquier modo, el uso del tamaño del efecto y/o su intervalo de confianza permite evaluar de manera más efectiva la información que brindan lo datos que sólo reportando el valor *p*.

CONCLUSIONES

La presencia de tres enfoques no siempre reconciliables en la enseñanza y su uso en investigaciones de los Test de Hipótesis: el enfoque de Fisher, el enfoque de Neyman y Pearson y el Bayesiano, traen dificultades didácticas singulares que hay que tomar en cuenta para una mejor didáctica de la estadística.

Las dificultades sobresalientes del valor p en los Test de hipótesis tales como "la ilusión de la demostración probabilística por contradicción", atribuciones erróneas sobre su demostración de teorías verdaderas, creencias como la de que pequeños valores de *p* constituyen evidencia a favor de la replicación de los resultados o que se trata de un efecto de tratamiento de gran magnitud, ponen en alerta a profesores e investigadores para mejorar la enseñanza actual de la estadística.

Es recomendable que durante el desarrollo de los cursos, se ponga más énfasis en los aspectos conceptuales en los que se basa la inferencia estadística, seleccionando estrategias de enseñanza activa que eviten la simple transferencia de habilidades técnicas y conduzcan en cambio, a priorizar el desarrollo del razonamiento conceptual y la interpretación. Consideramos que es importante hacer conocer las controversias que aún persisten acerca de los distintos enfoques relativos a las pruebas de hipótesis, mostrar sus diferencias y explotar las capacidades de cada uno. En particular, parece oportuno introducir la metodología bayesiana como una alternativa, no necesariamente antagónica, a la inferencia frecuencial clásica.

REFERENCIAS BIBLIOGRÁFICAS

Batanero, C. (2001). Traducción del artículo: Controversies around the role of statistical tests in experimental researcher. *Mathematical Thinking and Learning.* 2000, 2(1-2), 75-98.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *American Statistical Association.* Vol.59, Nº2.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego: Harcourt Brace.

Erwin E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences.*21,197-198.

Falk, R., Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, 5 (1), 75-98.

Hayes, A. F. (1998). Reconnecting data analysis an research design: Who needs a confidence interval? *Behavioral and Brain Sciences 21*, 203-204.

Gómez Villegas, M.A. (2005). *Inferencia Estadística*. Ediciones Díaz de Santos.

Ito, P. K. (1999). *Reaction to invited papers on statistical education and the significance tests controversy.* Ponencia invitada en la Fifty-Second International Statistical Institute Session, Helsinki, Finland.

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

Moore, D.S. (1998a). *Estadística Aplicada Básica*. Barcelona. Antoni Bosch editor.

Moore, D.S. (1998b). Incertidumbre. En L. A. Steen (ed.). *La enseñanza agradable de las matemáticas*. México: Limusa.

Rivadulla, A. (1991). *Probabilidad e Inferencia Científica*. Barcelona: Anthropos.

Rodríguez M.I., Albert A. (2007). Prueba de hipótesis estadística:estudio de dificultades conceptuales en estudiantes de grado y de postgrado. *Memoria de la XI Escuela de Invierno de Educación Matemática. Red de Centros de Investigación en Matemática Educativa. 328-343.* ISBN: 978-970-9971-14-9.

Schmit,F. ; Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In Harlow, Mulaik, Steigler (Eds), *What if there were no significance tests?* (pp. 37-64) Hillsdale, NJ: Erlbaum.

Shaughnessy, J.J.; Zechmeister, E. B. (1994). *Research methods in psychology.* New York: McGraw-Hill.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher,* 25(2), 26-30.

Vallecillos, A. (1996). *Inferencia estadística y enseñanza: un análisis didáctico del contraste de hipótesis estadísticas.* Granada: Comares.