PRUEBAS DE HIPÓTESIS Y EL VALOR P: USOS E INTERPRETACIONES

Rodríguez, María Inés (*); Albert Huerta Armando(**); Agnelli, Héctor (*)

(*) Universidad Nacional de Río Cuarto, Argentina; (**) Tecnológico de Monterrey, México.

mrodriguez@exa.unrc.edu.ar; albert@itesm.mx; hagnelli@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 en las prácticas de enseñanza al adoptarse aspectos de las distintas corrientes contrapuestas. Este trabajo se ubica en la perspectiva socioepistemológica, presentando un análisis de las prácticas de investigación de una comunidad científica referida a tesis doctorales de ciencias biológicas vinculadas a las pruebas de hipótesis y al valor p. Se evidenció la presencia de creencias y convenciones en la aplicación e interpretación, no siempre adecuadas con respecto al saber de referencia, así como, cierta confusión conceptual de los enfoques de Fisher y de Neyman-Pearson y problemas con el uso del valor p.

Palabras Clave

test, valor p, nivel α.

Introducción

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 o de la filosofía de la ciencia como los de Rivadulla (1991). También desde la psicología ha sido abordada esta problemática por diversos autores como 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 relativo a problemas probabilísticos, que ha sido extensamente estudiado

por los psicólogos en relación con otros conceptos, como la aleatoriedad, probabilidad y correlación.

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".

La vigencia de la preocupación sobre esta problemática queda reflejada por ser tema principal de eventos importantes tales como: el *5º Foro sobre el Razonamiento, Pensamiento y Alfabetización Estadística (2007)*, (SRTL-5) realizado en 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 siguiente foro, (SRTL-6) se realizará en el corriente año en Australia, siendo el tema de su convocatoria: "El rol del contexto y la evidencia informal en el Razonamiento Inferencial".

Además de las dificultades conceptuales vinculadas a la filosofía y a la psicología, las pruebas de hipótesis presentan dificultades epistemológicas puestas de manifiesto en las controversias entre las ideas acerca de la inferencia estadística de Ronald Fisher (1925) por un lado y por otro, las de Jerzy Neyman y Egor Pearson (1928). Por lo general, estas ideas no son tenidas en cuenta en el proceso de enseñanza de los tests ya que sólo se muestra una mezcla poco armonizada de estos enfoques. Al respecto, Ito (1999), señala los tres aspectos siguientes, que interactúan:
(a) La disputa dentro de la misma estadística, ya que existen tres enfoques diferentes para abordar la problemática de las pruebas de hipótesis.

(b) La controversia en la aplicación de la estadística, donde, en la práctica las pruebas de hipótesis son una mezcla informal de los contrastes de significación originales de Fisher y la teoría de Neyman-Pearson (N-P).

(c) La controversia en la enseñanza acerca de cuándo, cómo y con qué profundidad deberíamos enseñar la inferencia estadística.

Por otra parte, Bishop (2001), en su servicio de consultoría 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 universitarias, 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 de los usuarios, acerca de las pruebas de hipótesis y que son reportadas en Rodríguez (2006) y Rodríguez y Albert (2007).

La presente investigación se ubica en la perspectiva socioepistemológica, la cual ofrece una visión incluyente de las variables del tipo social y cultural que participan en la construcción del conocimiento, Crespo y Farfán (2005). Siguiendo este enfoque y tratando de ahondar en los señalamientos de Bishop, nos proponemos como objetivo estudiar las prácticas de investigación de una comunidad científica de ciencias biológicas referidas a la inferencia estadística, en particular los tests de hipótesis y el valor *p,* a través de sus trabajos de tesis doctorales.

Antecedentes

Problemática asociada a los tests de hipótesis

La significación estadística derivada de la aplicación de pruebas de hipótesis origina serios problemas de interpretación entre los investigadores usuarios de la estadística. Algunos autores (Hubbard y Bayari, 2003; Gigerenzer, Krauss, y Vitouch, 2004), argumentan que la razón principal

que sustenta esta problemática radica en que muchos textos de metodología estadística presentan confusiones sobre el real alcance de este concepto. En los textos y en consecuencia en muchos de los cursos de estadística, se mezclan de manera híbrida dos medidas incompatibles de significación estadística. Una de estas medidas es el *valor p* de Fisher, otra la *tasa de error α*, de la teoría de N-P. Estas medidas reflejan las diferencias que presentan los test de significación de Fisher y los test de hipótesis de N-P. Mientras Fisher sostenía la validez de la inferencia inductiva, N-P sostenían la necesidad de aplicar lo que ellos denominaban el comportamiento inductivo.

La idea fisheriana básica para conducir un test de significación está basada en la elección de un estadístico apropiado que resuma la información muestral y la determinación de su distribución bajo la suposición de que la hipótesis nula sea verdadera. Realizado el experimento se calcula el valor p. Éste, mide la probabilidad de encontrar un valor del estadístico del test, tan o más extremo que el hallado, condicionado a que la hipótesis nula (en general no existencia de efecto) sea verdadera. La disyuntiva esencial que se plantea al realizar un test de significación es: o la hipótesis nula no es verdadera o ha ocurrido un resultado excepcionalmente raro. Una medida de evidencia en contra de Ho está dada por el *valor p*. Se rechaza Ho si el *valor p* es suficientemente pequeño. En caso contrario no hay conclusión. Así, el *valor p* es una medida de evidencia inductiva. Fisher estaba convencido de que la estadística podía desempeñar un importante papel en el contexto de la inferencia inductiva al generar conclusiones relativas a las poblaciones a partir del conocimiento obtenido mediante muestras. De esta manera para él el *valor p* asumía un rol epistémico.

La teoría de test de hipótesis elaborada por N-P, si bien en principio perseguía el objetivo de mejorar las ideas de Fisher, difiere marcadamente del paradigma de la inferencia inductiva. Ellos vieron a los tests de hipótesis, como un procedimiento para tomar decisiones y guiar la conducta de los decisores. Mientras Fisher sólo establecía la hipótesis nula, Ho, N-P especificaban dos hipótesis, la nula Ho y la alternativa, Ha y establecían la manera de decidir entre dos posibles acciones: aceptar Ho o rechazar Ho a favor de Ha. Al tomar estas decisiones se puede incurrir en dos tipos de errores. Se comete el error tipo I cuando se incurre en un falso

rechazo de Ho y se comete el error tipo II cuando se acepta Ho siendo ésta falsa. La construcción de los tests de N-P apuntan a minimizar la probabilidad de cometer estos errores. A la probabilidad de cometer el error de tipo I se la designa con la letra α. La probabilidad del evento complementario a cometer el error tipo II, es llamada la potencia del test. Es decir, la potencia del test es la probabilidad de rechazar Ho cuando ella es falsa. En la teoría de N-P se buscan los test más potentes dentro de todos aquellos que tienen la misma probabilidad α de cometer el error tipo I. Fijar un nivel α, significa que estos test aplicados repetidamente no rechazarán equivocadamente la Ho, el α100 % de las veces. Esta probabilidad es una tasa de error y se establece antes de realizar el test, a diferencia de los valores *p* que son dato-dependientes. El valor de significación α es una característica del test que señala el comportamiento a la larga del test elegido.

En la teoría de N-P al *valor p* no le cabe ningún papel. Sin embargo, a través de los años se produjo una asimilación de las ideas involucradas en las pruebas de significación de Fisher y los test de hipótesis de N-P y este sincretismo es el que hoy mayoritariamente se enseña sin reparar en las diferencias que los nutren. Esquemáticamente podríamos decir que el procedimiento actual de los test de hipótesis es el de N-P, pero con el cálculo adicional fácilmente entregado por casi todos los paquetes estadísticos, del *valor p* y el proceso de decisión se lleva adelante comparando este *valor p* con el nivel de significación α, elegido. Este planteo es correcto en la medida que se utilice sólo para tomar una decisión, pero es incorrecto cuando además se especifica el *valor p* como una medida de evidencia en contra de Ho en el contexto de un test de N-P. En general es recomendable usar los test de N-P, cuando se está tratando un problema de decisión y en este caso es correcto reportar la tasa de error α. Si en cambio se plantea un test para encontrar evidencias en contra de la hipótesis nula lo adecuado es reportar el valor p.

Por otra parte, cuando se reportan los valores p es importante no incurrir en errores de interpretación. A continuación se expresan algunas de las creencias equivocadas acerca del significado del valor p que se presentan más frecuentemente entre los usuarios de test de hipótesis.

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". Como ya se ha señalado, el valor p es la probabilidad de que el estadístico del test sea tan o más extremo que el valor obtenido, condicionado a que la hipótesis nula H_0 sea verdadera. Lo anterior se pude simbolizar como $p = \Pr(D / Ho)$, utilizando la letra D para enfatizar que el valor observado del estadístico depende de los datos aportados por el experimento.

Pero, salvo casos excepcionales, $p = \Pr(D / Ho) \neq \Pr(Ho / D)$ siendo esta última la probabilidad de que H_0 sea verdadera condicionada al resultado observado. Luego el valor p no es la probabilidad (en distintos grados) de la validez hipótesis nula, sino que es la probabilidad de los datos en caso de ser cierta esa hipótesis nula. Existe una tendencia a ver estas dos probabilidades como equivalentes, a esta situación Dawes (1988) la llamó "la confusión de la probabilidad inversa". Una de las consecuencias que trae aparejada esta confusión es creer que $p = \Pr(Ho)$

Si realmente se está interesado en conocer $P(Ho / D)$ se debe aplicar la regla de Bayes:

$$P(Ho / D) = \frac{P(D / Ho)P(Ho)}{P(D / Ho)P(Ho) + P(D / Ha)P(Ha)}$$

Ahora se hace necesario contemplar la existencia de la hipótesis alternativa (aquí como complemento de Ho), y disponer de información acerca de Ho antes de realizar el

experimento. Cuando el experimentador realiza un test no conoce si Ho es verdadera o falsa, pero podría entonces asumir a priori, por ejemplo, que $P(Ho) = P(Ha) = 1/2$, en este caso

$$P(Ho/D) = \frac{p}{p + P(D/Ha)}$$

y entonces puede ocurrir que $P(Ho/D) > p$. Por lo tanto igualar equivocadamente el valor p con las probabilidades a posteriori de Ho tiene la consecuencia práctica de sobrestimar las evidencias en contra de Ho.

2. *Creencia acerca de que el rechazo de Ho establece la verdad de la teoría que predice que Ho 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. Los lectores de trabajos de investigación desean tener confianza en que los resultados publicados constituyen lo que Fisher

llamó un fenómeno “demostrable”- aquél que puede ser reproducido por los investigadores-. Un resultado aislado pude ser juzgado “demostrable” si su soporte empírico sugiere que puede ser replicado. Es importante poner en claro que significa replicación. Para ello hay que distinguir entre dos contextos de la inferencia estadística: uno el de estimación y el otro el de test de hipótesis. En el contexto de test de hipótesis se considera que un test replica a otro cuando conducen a igual conclusión respecto a la misma hipótesis nula. En el contexto de estimación, puntual o de intervalo, la replicación de estimaciones significa que ambas verifican algún criterio de proximidad.

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. Por otra parte y con referencia a la replicabilidad del rechazo de la hipótesis nula, si bien un pequeño valor p del orden de 0.005 puede ser evidencia indirecta de replicabilidad (Greenwald, 1996), en general es recomendable no utilizar este criterio.

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.

5. *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.

Metodología

Este trabajo se centró en estudiar algunas de las prácticas de investigación que realiza una comunidad de investigadores en ciencias biológicas de una universidad argentina a través de su producción de tesis de doctorado. Según el enfoque socioepistemológico:"la práctica social no es lo que hace en sí el individuo o el grupo, sino aquello que les hace hacer lo que hacen", Cantoral (2006). De este modo, se pretende identificar las normativas de estas prácticas para considerarlas e introducirlas en el desarrollo de actividades didácticas.

En particular observamos las prácticas caracterizadas por una regularidad epistémica: la inferencia estadística. Esto resulta útil, entre otras cosas, porque muchos de los autores de las tesis, llevaron cursos de probabilidad y estadística en la misma universidad en el nivel licenciatura y puede verse, en cierta manera, su impacto en la investigación.

Respecto a la característica epistémica de estas prácticas nos interesa observar: tipos de pruebas que utilizan valor p; verificación de supuestos, cálculo de la potencia de la prueba y valor p. Respecto a las prácticas en sí mismas, nos interesa observar, si es posible: tradición predominante, convenciones y argumentaciones. Se analizaron doce tesis doctorales, las cuales mostraron evidencia de prácticas sociales de esta comunidad científica, que prevalecen sobre los conocimientos impartidos en el aula y en los textos.

Resultados

Respecto a la característica epistémica de estas prácticas, la inferencia estadística, la siguiente tabla muestra los resultados encontrados.

Cantidad de Tesis	Tipo de prueba	Verificación de supuestos	Cálculo de potencia	Explicitación del valor p
2	Comparación de pares de medias	No	No	$P<0,01$ $P<0,05$

1	ANOVA y Comparación de pares de medias	Sí	No	*P<0,05*
2	Comparación de pares de medias	No	No	*P<0,05*
2	ANOVA y Comparación de pares de medias	Menciona transformación de los datos, sin justificarla.	No	*P<0,01* *P<0,05*
2	ANOVA y Comparación de pares de medias	No	No	*P<0,05*
3	Ninguna	------	------	------

Los tres últimos de los trabajos sólo usaron elementos de estadística descriptiva para el manejo de su información, sin embargo de la lectura de los mismos se desprende que hacen inferencia pues los resultados obtenidos son generalizados a una población.

En los trabajos que se utiliza inferencia estadística, las metodologías utilizadas fueron el test t de Student y análisis de la varianza (ANOVA). En general se puede advertir la no verificación de los supuestos. En algunos casos se podría deducir que esta etapa fue cumplimentada por que se realizaron transformaciones de los datos, aunque no se explicitan los motivos por los cuales se efectuaron las mismas. En ningún caso se exhibe explícitamente la hipótesis nula del test, aunque de la lectura se comprende que ésta es asumida como la de no diferencia de efectos. No se precisa ninguna hipótesis alternativa y está ausente el cálculo de potencia. No se reportan los valores p precisamente sino que se hace su comparación con los valores 0.01 ó 0.05.

Se pudo identificar la predominancia de la tradición Fisher por la forma en que son conducidos los test en el sentido de que se busca rechazar la hipótesis de igualdad de efectos sin explicitar ninguna hipótesis alternativa. Aunque también está presente la tradición Neyman-Pearson por el uso que se da al valor p al compararlo con α sin indicar su valor preciso, hace perder a p su

valor evidencial y la decisión de rechazar Ho es tomada en el sentido de N-P, sin importar cuán pequeño es p. Al no considerarse el test en el sentido de N-P, es decir al no hacer explicitas hipótesis alternativa alguna no se hace evidente la necesidad de calcular de antemano la potencia del test y en consecuencia el tamaño de muestra adecuado para alcanzar esta potencia.

Se identificó una norma: "se debe dar cuenta del valor p" y dos convenciones: "el nivel de significancia debe ser por lo menos de 0.05" y "los resultados son válidos científicamente si el valor p es menor a 0.05". Fue notable la ausencia de argumentaciones sobre el porqué del uso de las técnicas estadísticas empleadas y sus supuestos.

Es notable la presencia de la creencia acerca de que el rechazo de H_0 establece la verdad de la teoría que predice que H_0 es falsa. En casi todas las tesis a partir del resultado estadístico se concluye, sin tomar en cuenta otros aspectos biológicos, ni la magnitud del efecto, que ha quedado establecido con valor de verdad determinística un nuevo enunciado biológico. Por otra parte, al no usar cálculo de potencia y no determinar tamaños de muestra adecuados, en general se trabaja con muestras pequeñas y se tiene la creencia de que, dado que el valor p obtenido es pequeño, se espera la replicabilidad del resultado aislado hallado.

Discusión

La no verificación de los supuestos requeridos para la validez de la aplicación de las pruebas estadísticas es un error grave, ya que en caso de no verificarse los mismos la credibilidad de las conclusiones se torna dudosa. Uno de los supuestos principales es el que asume que los datos provienen de una muestra aleatoria o de un diseño experimental correspondientemente aleatoreizado. También son supuestos básicos los vinculados con los aspectos distribucionales de los datos.

La mezcla de las dos convenciones citadas en resultados, conduce a que no se reporte el valor real del valor p, esto limita mucho su capacidad de análisis y argumentación en su

interpretación, además de mezclar dos tradiciones que no siempre convergen en resultados e interpretación.

Conclusiones

En las prácticas de investigación de la comunidad de investigadores de ciencias biológicas estudiada resultó muy frecuente el uso de la metodología inferencial. En las regularidades epistémicas sobre estadística inferencial en esta comunidad científica existe un sincretismo de tradiciones Fisher y Neyman-Pearson que les ha conducido a conflictos epistémicos y de interpretación.

Por otra parte, la creencia de que el rechazo de Ho establece la verdad de la teoría muestra poca comprensión sobre el recurso de la inferencia estadística. Esto se debe, en parte, a que hay una limitada reflexión sobre cuándo se aplican las técnicas de estadística inferencial y sus limitaciones.

Centrándonos especialmente en las prácticas y no sólo en el objeto de análisis, hemos comprobado que los hábitos en el uso de la estadística, propios de los investigadores de la comunidad de pares, prevalecen sobre los conceptos que han sido enseñados y evaluados en los cursos. Por lo tanto, se recomienda que durante el desarrollo de los cursos del nivel licenciatura como de postgrado se haga énfasis en la argumentación, casi ausente en la comunidad estudiada, pues el desarrollo de ella les permitirá un mayor desarrollo del razonamiento conceptual y de interpretación. En particular consideramos que es importante hacerles 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 y hacer explicitas las diferencias entre el valor p y el nivel de significación α.

Reconocimientos

Este trabajo se realizó durante la ejecución del proyecto de investigación conjunto en el marco de la cooperación SECYT (ARG)- CONACYT (MEX): *Estudio teórico y experimental sobre*

dificultades conceptuales, procedimentales y metodológicas de la estadística inferencial en el nivel universitario y del Proyecto de tesis del Doctorado en Ciencias que se está realizando en *CICATA.*

Bibliografía

Bishop, G. & Talbot, M. (2001). Statistical thinking for novice researchers in the biological sciences. En Batanero, C. (Ed.). *Training Researchers in the Use of Statistics,* 215-226.

Crespo, C.C. & Farfán, R.M. (2005). Una visión socioepistemológica de las argumentaciones en el aula. El caso de las demostraciones por reducción al absurdo. *Revista Latinoamericana de Investigación en Matemática Educativa,* 8(3), 287-317.

Cantoral, R.; Farfán, R.M.; Lezama, J y Martínez-Sierra G. (2006). Socioepistemología y representación: algunos ejemplos. *Revista Latinoamericana de Investigación en Matemática Educativa,* Número Especial, 83-102.

Erwin, E. (1998). The logic of null hypothesis testing. *Behavioral and Brain Sciences*. 21 (2), 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.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Gigerenzer, G., Krauss, S., Vitouch, O. (2004). The null ritual: What you always to know about significance testing but were afrais ask. In Kaplan, D. (Ed.). *The Sage handbook of quantitative methodology for the social sciences*. Sage Publications, 391-408.

Goodman, S.(1993). P values, hypothesis tests, and likelihood. Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology,* 137. 485-496.

Greer, B. (2000). *Mathematical Thinking and Learning*, 1532-7833, Vol. 2, 1, 1-9.

Greenwald, A.; Gonzalez, R.; Harris, R.; Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology.* 33, 175-183.

Hayes, A. (1998). Reconnecting data analysis and research design: Who needs a confidence interval? *Behavioral and Brain Sciences,* 21 (2), 203-204.

Hubbard, R.; Bayarri, M. (2003*).* Confusion over measures of evidence (p´s) versus errors (α´s) in classical statistical testing. *The American Statistician,* 57, 171-182.

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.

Kahneman, D., Slovic, P., Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

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

Neyman, J. & Pearson, J.(1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika,* 20A,175-240.

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

Thompson, B. (2002). What future quantitative social science research could look like:Confidence intervals for effect sizes. *Educational Researcher.* 31. 25–32.

Rodríguez, M. I. (2006). Estudio de sesgos en el razonamiento de conceptos de prueba de hipótesis estadística en alumnos universitarios. *Actas Jornadas Internacionales de Estadística y VII Congreso Latinoamericano de Sociedades de Estadística,* Argentina.

Rodríguez, M. I. & Albert, J. A. (2007). Prueba de hipótesis estadística. Estudio de dificultades conceptuales en estudiantes de grado y de postgrado. *XI Memoria de la Escuela de Invierno de Educación Matemática.* Mérida, Yucatán: Red de Centros de Investigación en Matemática educativa, 328-343.