Controversias sobre el Papel de los Contrastes de Hipótesis en la Investigación Experimental

Carmen Batanero Bernabeu,
Departamento de Didáctica de la Matemática, Universidad de Granada
batanero@ugr.es, http://www.ugr.es/local/batanero

Nota: Traducción del artículo Controversies around the role of statistical tests in experimental research, publicado en *Mathematical Thinking and Learning*, 2(1-2), 75-98, con autorización de Lawrence Erlbaum, Editores para reproducir la traducción en las Actas de las Jornadas Europeas de Difusión y Enseñanza de la Estadística..

A pesar del uso tan extendido de los contrastes estadísticos en la investigación experimental, su interpretación y la excesiva confianza de los investigadores en sus resultados han sido criticados en los últimos años. En este trabajo describimos la lógica de los contrastes de hipótesis en la filosofía de Fisher y Neyman-Pearson, analizamos algunas interpretaciones erróneas frecuentes de los conceptos básicos que subyacen en el contraste de hipótesis así como los mecanismos filosóficos y psicológicos que contribuyen a las mismas. Seguidamente revisamos algunas críticas usuales a los tests de hipótesis, concluyendo que la mayor parte de ella no se refieren a los mismos tests, sino a su uso por parte de los investigadores. Coincidimos con Levin (1997 a) en que debemos convertir el contraste estadístico en un proceso más inteligente que ayude a los investigadores en su trabajo. Finalmente, sugerimos algunas formas en que la educación estadística podría contribuir a la mejor comprensión y uso de la inferencia estadística.

1. Introducción

Las ciencias empíricas, en general, y en particular la psicología y la educación, dependen en gran medida de la demostración de la existencia de efectos a partir del análisis estadístico de datos. La inferencia estadística se inició hace unos 300 años, aunque fue popularizada a partir de los trabajos de Fisher, Neyman y Pearson sobre los contrastes estadísticos. En la actualidad, la mayor parte de los investigadores que emplean la inferencia, utilizan una mezcla de la lógica sugerida por estos tres autores. Sin embargo, debido a que la lógica de la inferencia estadística es difícil, su uso e interpretación no es siempre adecuado y han sido criticados en los últimos 50 años. Yates (1951), por ejemplo, sugirió que los científicos dedicaban demasiada atención al resultado de sus contrastes, olvidando la estimación de la magnitud de los efectos que investigaban. Una amplia revisión de estas críticas puede encontrarse en Morrison y Henkel (1979).

Esta controversia se ha intensificado en los últimos años en algunas instituciones profesionales (Menon, 1993; Thompson, 1996; Ellerton, 1996, Robinson y Levin, 1997, 1999 Levin, 1998 a y b, Wilkinson et al., 1999) que sugieren importantes cambios en sus políticas editoriales respecto al uso del contraste de hipótesis. Por ejemplo la American Psychological Association resalta en su manual de publicación del año 1994 que los contrastes estadísticos no reflejan la importancia o la magnitud de los efectos y animan a los investigadores a proporcionar información sobre el tamaño de estos efectos (APA, 1994, pg. 18). Más recientemente, la Task Force on Statistical Inference organizada por la APA ha publicado un artículo para iniciar la discusión en el campo, antes de revisar el manual de publicación de la APA (Wilkinson, 1999). Una decisión de este comité ha sido que la revisión cubra cuestiones metodológicas más generales y no sólo el contraste de hipótesis. Entre otras cuestiones, se recomienda publicar los *valores-p* exactos, las estimaciones de los efectos y los intervalos de confianza.

En la American Education Research Association, Thompson (1996) recomienda un uso más adecuado del lenguaje estadístico en los informes de investigación, enfatizando la interpretación del

tamaño de los efectos y evaluando la replicabilidad de los resultados. Estas instituciones, así como la American Psychological Society han constituido comités específicos para estudiar el problema, los cuales recomiendan no abandonar el contraste de hipótesis, sino complementarlo con otros análisis estadísticos (Levin, 1998 b, Wilkinson et al., 1999). Un resumen comprensivo de estos debates, así como de las alternativas sugeridas, se presenta en Harlow, Mulaik y Steiger (1997).

A pesar de estas recomendaciones, los investigadores experimentales persisten en apoyarse en la significación estadística, sin tener en cuenta los argumentos de que los tests estadísticos por si 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 (Falk y Greenbaum, 1995). En este trabajo analizamos estos problemas y sugerimos posibles medios en los que la educación estadística podría contribuir a la mejor comprensión y aplicación de la inferencia estadística.

2. La Lógica de los Tests Estadísticos: Un Ejemplo

En esta sección presentaremos una situación típica en la que un investigador recurre a la estadística para dar apoyo a una hipótesis referida a su campo de estudio y resumimos los pasos y lógica del contraste de hipótesis. Usaremos el ejemplo a lo largo del trabajo para contextualizar la discusión.

Ejemplo 1: De acuerdo con algunas teorías de aprendizaje, las representaciones contribuyen a la construcción del significado de los objetos matemáticos, de modo que un contexto rico en representaciones, que facilite el cambio de una representación a otra, favorecerá el aprendizaje. Un investigador que acepta esta teoría tiene buenas razones para esperar que el uso de los ordenadores refuerce el aprendizaje de la estadística, porque los ordenadores proporcionan potentes herramientas y sistemas de representación de los conceptos estadísticos. Para probar su conjetura, supongamos que el investigador selecciona una muestra aleatoria de 80 estudiantes entre todos los alumnos que ingresan a la universidad un curso dado. Aleatoriamente divide los 80 estudiantes en dos grupos de igual tamaño y encuentra que los dos grupos tienen un conocimiento inicial equivalente de la estadística. Organiza un experimento, donde el mismo profesor, con los mismos materiales, imparte un curso introductorio de estadística a los dos grupos durante un semestre. El grupo C (grupo de control) no tiene acceso a los ordenadores, mientras que la enseñanza en el grupo E (grupo experimental) se basa en un uso intensivo de los ordenadores.

Al final del periodo, el mismo cuestionario se pasa a los dos grupos. Si el aprendizaje con los dos métodos de enseñanza fuese igual de efectivo, no debiera haber diferencia entre la puntuación media en el cuestionario μ_e, de la población teórica de todos los estudiantes a los que se enseña con ayuda del ordenador y la puntuación media en el cuestionario μ_c, de la población teórica de estudiantes que no tienen acceso a un ordenador. Si consideramos que los grupos C y E son muestras representativas de estas poblaciones teóricas, y no hay diferencia de efectividad entre los dos métodos de enseñanza, la diferencia en las puntuaciones medias del cuestionario en los dos grupos debería ser próxima a cero. Si el investigador encuentra una diferencia positiva entre las puntuaciones medias en los dos grupos $x_e - \bar{x}_c$, ¿podría deducir que su conjetura era cierta? Es importante resaltar que el interés del investigador va más allá de las muestras particulares de alumnos (grupos C y E). Está interesado en comprobar el efecto del aprendizaje con ordenadores en la población, en general, y a través de ello, encontrar un apoyo empírico para su hipótesis sobre el efecto del contexto en el aprendizaje.

El ejemplo anterior ilustra una situación típica de uso de la inferencia estadística para determinar si los datos experimentales (las puntuaciones de los alumnos de los grupos E y C) apoyan o no una *hipótesis substantiva* (el aprendizaje está influenciado por los sistemas de representación disponibles). Como no podemos probar directamente la hipótesis substantiva, organizamos un experimento para obtener datos y contrastar una *hipótesis de investigación* deducida de la anterior (los ordenadores refuerzan el aprendizaje de la estadística). Tampoco podemos confirmar directamente la hipótesis de investigación, porque el aprendizaje es un constructo inobservable que no podemos evaluar directamente.

En consecuencia, elegimos un instrumento (el cuestionario) directamente relacionado con el aprendizaje, y que produce resultados observables (las respuestas de los alumnos). Si la hipótesis de investigación es cierta, esperamos que las puntuaciones del grupo de alumnos enseñados con ayuda del ordenador sean más altas que las de los estudiantes que no tienen acceso al mismo (*hipótesis experimental).* De cualquier modo, ya que hay múltiples factores que afectan el aprendizaje, además del ordenador, algunos estudiantes del grupo control tendrán mejores resultados que los de otros estudiantes del grupo experimental. Por tanto, necesitamos un procedimiento para comparar la distribución global en las dos poblaciones de estudiantes en vez de comparar los casos aislados.

Esta comparación se hace usualmente considerando las puntuaciones medias en las dos poblaciones teóricas y especulando sobre su posible diferencia. Si la hipótesis experimental es cierta, esperaremos que esta diferencia sea positiva (*hipótesis estadística alternativa)* aunque quizás no podamos precisar el valor de la diferencia. En este caso no podemos trabajar directamente con la hipótesis alternativa y razonamos en su lugar como si las dos poblaciones tuvieran el mismo rendimiento, es decir asumimos que la *hipótesis nula* es cierta (la diferencia de puntuaciones medias en las dos poblaciones es cero). En el ejemplo, usaremos un test unilateral, porque hemos especificado la dirección de la diferencia con la hipótesis nula. En este caso la hipótesis nula es, en realidad $\mu_e \leq \mu_c$, (el complemento de la hipótesis alternativa). Desde el punto de vista matemático, sin embargo, solo necesitamos calcular el valor crítico para el test bilateral, ya que siempre que un resultado sea significativo para la hipótesis $\mu_e=\mu_c$, también lo será para la hipótesis $\mu_e<\mu_c$. Por tanto, y para simplificar la exposición, supondremos en lo que sigue que la hipótesis nula es $\mu_e=\mu_c$.

Para probar este supuesto calculamos un estadístico de contraste relacionado con el parámetro de interés, a partir de los datos de nuestras muestras. Al aceptar que la hipótesis nula es cierta, determinamos la distribución de este estadístico (una distribución *T* con 78 grados de libertad) que servirá para calcular el valor crítico y tomar una decisión acerca de si debemos o no aceptar nuestra hipótesis de investigación inicial.

Hay dos concepciones sobre los contrastes estadísticos: a) las pruebas de significación, que fueron introducidas por Fisher y b) los contrastes como reglas de decisión entre dos hipótesis, que fue la concepción de Neyman y Pearson. La diferencia no se debe a los cálculos, sino al razonamiento subyacente. Siguiendo a Moore (1995) describiremos, en primer lugar, el razonamiento típico de una prueba de significación, que sería adecuada para el ejemplo 1, y en la que habría que seguir los pasos siguientes:

1. Describir el efecto que estamos buscando en función de los parámetros de una o varias poblaciones (la puntuación medias en el cuestionario de los estudiantes enseñados con ordenador μ_e es mayor que la de los alumnos que no tienen acceso al ordenador μ_c). El efecto que sospechamos es el verdadero describe la hipótesis alternativa: $H_1 \equiv \mu_e > \mu_c$.
2. Establecer la hipótesis nula de que el efecto no se presenta: $H_0 \equiv \mu_e = \mu_c$, (que no hay diferencias entre la puntuación media μ_e, de los estudiantes enseñandos con ordenador y la puntuación media μ_c de los estudiantes sin acceso al ordenador). La prueba de significación se diseña para evaluar la fuerza de la evidencia en contra de la hipótesis nula.
3. Calcular un estadístico a partir de los resultados en la muestra (estadístico calculado). La distribución del estadístico queda especificada cuando asumimos que la hipótesis nula es cierta

(en el caso del ejemplo sería una distribución *T* con 78 grados de libertad). Supongamos que en el Ejemplo 1 obtenemos los siguientes valores para las medias de las dos muestras: $\bar{x}_e = 115.1$; $\bar{x}_c = 101.78$ y que $s^2_e = 197.66$; $s^2_c = 215.19$ son los estimadores insesgados de las varianzas en las poblaciones; la diferencia media de la puntuación en los dos grupos para estos datos es $\bar{x}_e - \bar{x}_c = 13.32$, la estimación conjunta de la varianza $s^2 = 202.48$, y usando las formulas standard obtendríamos un valor $t = 4.16$. La cuestión que el contraste de significación trata de contestar es la siguiente: Supongamos que la hipótesis nula es cierta y que, en promedio, no hay diferencia entre la puntuación media de las muestras tomadas de las dos poblaciones, ¿es entonces el resultado muestral $t = 4.16$ demasiado grande? O ¿podríamos obtener fácilmente este valor simplemente por causa de las fluctuaciones aleatorias del muestreo?

4. La probabilidad de obtener un valor *t* tan extremo o más que el valor *t* calculado cuando la hipótesis nula es cierta se llama *valor-p*. En nuestro ejemplo, el valor-p es extremadamente pequeño (menor que .001). Si la hipótesis nula es cierta y el valor-p muy pequeño, los resultados son altamente improbables y se llaman estadísticamente significativos. En este caso, y si los datos coinciden con la dirección especificada por la hipótesis alternativa, asumimos que nuestros datos proporcionan evidencia en contra de la hipótesis nula (ello no implica que creamos que la hipótesis nula es imposible; la hipótesis se aceptará en la comunidad científica sólo a partir de un programa de experimentos repetidos en los que repliquemos nuestros resultados; la ciencia se construye a partir de hallazgos acumulativos).
5. Incluso cuando la hipótesis nula es cierta, esperaremos algunas discrepancias entre las puntuaciones medias en los grupos experimental y control en el Ejemplo 1, debido a las fluctuaciones aleatorias del muestreo. No hay una regla fija acerca de cuan pequeño debe ser el valor-p para que un resultado se considere estadísticamente significativo, aunque convencionalmente adoptamos un valor fijo con el que comparamos el valor-p para decidir sobre su significación estadística. Es el nivel de significación α, o máximo valor-p admisible para considerar los datos como significativos, que se usa para calcular los valores críticos. Supongamos que tomamos $\alpha = .05$ en el ejemplo 1. El valor crítico es la máxima diferencia que esperaríamos entre las dos muestras (grupos E y C) con probabilidad .05 (α), en caso de que las dos poblaciones tengan el mismo rendimiento. Este valor crítico se obtiene de la distribución teórica del estadístico en caso de que la hipótesis nula sea cierta (distribución *T* con 78 g.l.).

El contraste de hipótesis como proceso de decisión

En el ejemplo 1 hemos usado un contraste de significación para evaluar la fuerza de la evidencia en contra de una hipótesis nula. Hay, sin embargo, otras situaciones donde la inferencia se usa para tomar una decisión entre dos acciones posibles.

> Ejemplo 2: Supongamos que una escuela secundaria quiere evaluar la efectividad de un nuevo método de enseñanza sobre el aprendizaje de la estadística de sus estudiantes. La escuela selecciona aleatoriamente 80 alumnos entre todos los estudiantes del último curso y divide aleatoriamente los 80 estudiantes en dos grupos de igual tamaño. El mismo profesor enseña estadística a ambos grupos durante un semestre. En el grupo C (grupo de control) se usa el método usual en la escuela, mientras que la enseñanza el grupo E (grupo experimental) se basa en los nuevos materiales. Al final del periodo el mismo cuestionario se aplica a los dos grupos (que fueron juzgados como equivalentes en su conocimiento inicial mediante un pretest). La escuela quiere cambiar al nuevo sistema si la diferencia media de puntuaciones en las dos poblaciones de estudiantes es positiva. Aquí el interés también va más allá de las dos muestras particulares (los grupos E y C), ya que el método sería aplicado a otros estudiantes.

En el ejemplo 2 se usaría un contraste estadístico con un razonamiento diferente, como procedimiento para tomar una decisión. Se seguirían los pasos siguientes:

1. Establecer las hipótesis nula $H_0 \equiv \mu_e = \mu_c$ y alternativa $H_1 \equiv \mu_e > \mu_c$, como en el Ejemplo 1.
2. Calcular el estadístico a partir de los datos de las muestras, como en el Ejemplo 1.
3. Se tomaría una decisión: O bien rechazamos la hipótesis nula (y aceptaríamos H_1) o bien no rechazaríamos H_0.
4. La decisión se hace comparando el p-valor con el nivel de significación α, es decir, comparando el valor *t* calculado con el valor crítico. En el Ejemplo 2 (suponiendo los mismos datos numéricos) el valor calculado $t = 4.16$ es mayor que el valor crítico $t_c = 1.665$, y por tanto, la hipótesis nula sería rechazado.

Es importante resaltar que rechazar la hipótesis nula no implica necesariamente que sea falsa, ya que es posible cometer dos tipos de error al tomar una decisión a partir de los resultados del contraste. En primer lugar, sería posible que no hubiese diferencia real en las puntuaciones medias de los estudiantes enseñados con los dos métodos y que, debido a la variabilidad aleatoria del muestreo, hayamos obtenido en nuestros grupos particulares E y C un valor *t* que ocurre con baja probabilidad. Como sabemos, que la probabilidad de un suceso sea muy baja no implica que el suceso sea imposible. Cometeríamos un *Error Tipo I* si rechazásemos una hipótesis nula que de hecho sea verdadera y la *probabilidad de Error Tipo I* es numéricamente igual al nivel de significación α.

Por otro lado, si el resultado no es significativo, ello no implica que las dos poblaciones tengan resultados igualmente buenos en el cuestionario. Incluso cuando los estudiantes enseñados con el nuevo método tengan mejores resultados, podríamos no obtener un resultado significativo en nuestras muestras particulares si el efecto de la enseñanza es pequeño o si hay demasiada variabilidad en los datos. Ocurre un *Error Tipo II* cuando el investigador acepta la hipótesis nula cuando, de hecho, es falsa. Puesto que hay muchas posibilidades diferentes para la diferencia de media en la hipótesis alternativa, la probabilidad de error Tipo II, β es variable. Normalmente estaremos interesados en algunos valores particulares de esta probabilidad y la calcularemos sólo para los casos más desfavorables.

El complemento de β□ se llama potencia del contraste y es la probabilidad de rechazar la hipótesis nula en caso de que sea falsa. Es también variable, ya que depende del valor verdadero del parámetro (la diferencia de medias en las poblaciones en nuestro caso). Conviene enfatizar la naturaleza condicional de las probabilidades de los dos tipos de error, ya que es en la interpretación de estas probabilidades condicionales donde encontramos más errores y concepciones erróneas respecto al contraste de hipótesis.

3. Errores Comunes en la Interpretación del Nivel de Significación y el valor p

La investigación sobre la comprensión de los métodos de inferencia muestra la existencia de concepciones erróneas ampliamente extendidas, tanto entre los estudiantes universitarios, como entre los científicos que usan la inferencia estadística en su trabajo diario. Estas concepciones erróneas se refieren principalmente al nivel de significación α, que se define como la probabilidad de rechazar la hipótesis nula en caso de que sea cierta. La interpretación errónea más extendida de este concepto consiste en intercambiar los dos términos de la probabilidad condicional, es decir, en interpretar el nivel de significación como la probabilidad de que la hipótesis nula sea cierta si hemos tomado la decisión de rechazarla. Birnbaum (1982), por ejemplo, informó que sus estudiantes encontraban razonable la siguiente definición: "*Un nivel de significación del 5% indica que, en promedio, 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados*". Falk (1986) comprobó que la mayoría de sus estudiantes creían que α era la probabilidad de equivocarse al rechazar la hipótesis nula. Resultados semejantes se describen en el estudio de Pollard y Richardson (1987) realizado con investigadores.

Vallecillos (1994) planteó los siguientes ítems a una muestra de 436 estudiantes universitarios de diferentes especialidades (estadística, medicina, psicología, ingeniería y empresariales) que habían estudiado el tema:

Item 1: Un nivel de significación del 5% significa que, en promedio 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados (verdadero /falso). Justifica tu respuesta.

Item 2: Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que la hipótesis nula es cierta la rechazaremos (verdadero / falso). Justifica tu respuesta.

En el ítem 2 se presenta una interpretación frecuencial del nivel de significación (y es correcto), mientras que en el ítem 1 se han intercambiado los dos sucesos que definen la probabilidad condicional (y es incorrecto). Sin embargo, sólo el 32% de los estudiantes de la investigación de Vallecillos (1994) dio una respuesta correcta al ítem 1 y el 54% dio una respuesta correcta al ítem 2. De 135 estudiantes que justificaron su respuesta, el 41% dio un argumento correcto en los dos ítems. Un error prevalente en todos los grupos de estudiantes fue el intercambio de los términos de la probabilidad condicional, juzgando por tanto correcto el ítem 1 y falso el ítem 2. Entrevistas a un grupo reducido de estudiantes mostró que esta creencia aparecía en algunos estudiantes que eran capaces de discriminar entre una probabilidad condicional y su inversa (Vallecillos y Batanero, 1996). Otros estudiantes no distinguían las dos probabilidades condicionales, es decir, consideraban que ambos ítems eran correctos.

Que las probabilidades condicionales con términos intercambiados no coinciden, en general, se ilustra en la Tabla 1 que se refiere a la elección de estadística como tema optativo en una escuela. La probabilidad de que una chica tomada al azar estudie estadística y la probabilidad de que un estudiante de estadística sea una chica son diferentes:

$$P(\text{estudie estadística} / \text{chica}) = 3/4;\ P(\text{chica} / \text{estudia estadística}) = 3/8$$

Tabla 1. *Número de Chicos y Chicas en un Curso de Estadística*

	Chicas	Chicos	Total
Estadística	300	500	800
No Estadistica	100	100	200
Total	400	600	1000

Es importante resaltar que, incluso cuando fijemos el nivel de significación α, es decir, la probabilidad de rechazar la hipótesis (supuesto que es cierta) y podamos calcular la probabilidad de obtener un valor del estadístico de contraste menor que un valor particular (supuesta la hipótesis nula cierta), la probabilidad de que la hipótesis nula sea cierta una vez la hemos rechazado y la probabilidad de que la hipótesis nula sea cierta una vez que hemos obtenido el valor del estadístico de contraste no pueden conocerse.

La probabilidad a posteriori de la hipótesis nula dado un resultado significativo depende de la probabilidad a priori de la hipótesis nula, así como de las probabilidades de obtener un resultado significativo, dadas las hipótesis nula y alternativa. Desafortunadamente estas probabilidades no pueden determinarse. Más aún, una hipótesis es o cierta o falsa y, por tanto, no tiene mucho sentido calcular su probabilidad en un paradigma inferencial clásico (donde damos una interpretación frecuencial a las probabilidades objetivas). Sólo en la inferencia Bayesiana pueden calcularse las probabilidades a posteriori de la hipótesis, aunque son probabilidades subjetivas, Lo más que podemos hacer, y es usando inferencia Bayesiana, es revisar nuestro grado de creencia personal en la hipótesis en vista de los resultados.

Otras interpretaciones erróneas del nivel de significación y el valor *p* son:

(a) Algunas personas piensan que el valor *p* es la probabilidad de que el resultado se deba al azar. Podemos ver claramente que esta concepción es errónea del hecho de que incluso si la

hipótesis nula es cierta (e.g. si no hubiera diferencias de rendimiento en el ejemplo 1) un resultado significativo puede ser debido a otros factores, como, por ejemplo, que los estudiantes del grupo experimental trabajasen más que sus compañeros al prepararse para la evaluación. Podemos observar aquí la importancia del control experimental para intentar asegurar que todas las condiciones (excepto el tipo de enseñanza) se mantienen constantes en los dos grupos. El valor *p* es la probabilidad de obtener el resultado particular u otro más extremo cuando la hipótesis nula es cierta *y* no hay otros factores posibles que influencien el resultado. Lo que rechazamos en un contraste de hipótesis es la hipótesis nula y, por tanto, no podemos inferir la existencia de una causa particular en un experimento a partir de un resultado significativo.

(b) Otro error común es la creencia en la conservación del valor del nivel de significación cuando se realizan contrastes consecutivos en el mismo conjunto de datos, lo que produce el problema de las comparaciones múltiples. A veces aplicamos un alto número de pruebas de significación a un mismo conjunto de datos. El significado del nivel de significación (ver el Item 2 anterior) es que, si llevamos a cabo 100 comparaciones sobre el mismo conjunto de datos y usamos en todos ellos el nivel de significación .05, habrá que esperar que 5 de las 100 pruebas sean significativas por puro azar, incluso cuando la hipótesis nula sea cierta. Esto dificulta la interpretación de los resultados (Moses, 1992).

(c) El uso frecuente de niveles de significación .05 y .01 es cuestión de convenio y no se justifica por la teoría matemática. Si consideramos el contraste de hipótesis como proceso de decisión (la visión de Neyman y Pearson), debemos especificar el nivel de significación antes de llevar a cabo el experimento y esta elección determina el tamaño de las regiones críticas y de aceptación que llevan a la decisión de rechazar o no la hipótesis nula. Neyman y Pearson dieron una interpretación frecuencial a esta probabilidad: Si la hipótesis nula es cierta y repetimos el experimento muchas veces con probabilidad de Error Tipo I igual a .05 rechazaremos la hipótesis nula el 5% de las veces que sea cierta.

En su libro "Diseño de experimentos" Fisher(1935) sugirió seleccionar un nivel de significación del 5%, como convenio para reconocer los resultados significativos en los experimentos. En sus trabajos posteriores, sin embargo, Fisher consideró que cada investigador debe seleccionar el nivel de significación de acuerdo a las circunstancias, ya que "*de hecho ningún investigador mantiene un nivel de significación fijo con el cual rechaza las hipótesis año tras año y en todas las circunstancias"* (Fisher, 1956, p. 42). Por el contrario, Fisher sugirió publicar el valor *p* exacto obtenido en cada experimento particular, lo que, de hecho, implica establecer el nivel de significación después de llevar a cabo el experimento.

A pesar de estas recomendaciones, la literatura de investigación muestra que los niveles arbitrarios de .05, .01, .001 se usan casi en forma universal para todo tipo de problemas. Skipper, Guenter y Nass (1970) sugirieron que esto trae como consecuencia la diferenciación de los resultados de investigación que se publicarán o no y llama la atención sobre las posibles implicaciones sobre los problemas investigados. A veces, si la potencia del contraste es baja y el error Tipo II es importante, sería preferible una probabilidad mayor de Error Tipo I.

(d) La interpretación incorrecta del nivel de significación se une normalmente a una interpretación incorrecta de los resultados significativos, un punto donde hubo también desacuerdos entre Fisher y Neyman -Pearson. Un resultado significativo implica para Fisher que los datos proporcionan evidencia en contra de la hipótesis nula, mientras que para Neyman y Pearson solo establece la frecuencia relativa de veces que rechazaríamos la hipótesis nula cierta a la larga (Error Tipo I). Por otro lado, debemos diferenciar entre significación estadística y significación práctica. En el Ejemplo 1 obtuvimos una diferencia media en puntuaciones entre los dos grupos de 13.32, que fue significativa. Sin embargo, podríamos haber obtenido una significación estadística mayor con un efecto experimental menor y una muestra de tamaño mayor. La significación práctica implica significación estadística más un efecto experimental suficientemente elevado.

4. Los Diferentes Niveles de Hipótesis en la Investigación

El nivel de significación no es el único concepto mal comprendido en las pruebas de hipótesis. Algunas investigaciones también muestran confusión entre los papeles de las hipótesis nulas y alternativas (Vallecillos, 1994, 1995), así como entre la hipótesis estadística alternativa y la hipótesis de investigación (Chow, 1996). Chow diferencia diversas hipótesis implicadas en los diversos niveles de abstracción de la investigación experimental orientada a la confirmación de teorías, como la descrita en el Ejemplo 1.

(a) *Hipótesis substantiva* (que, en el Ejemplo1, es que el aprendizaje se ve influenciado por las herramientas semióticas disponibles para tratar un concepto). Una hipótesis substantiva es una explicación especulativa de un fenómeno. Normalmente no podemos investigarla directamente porque se refiere a un constructo o a un mecanismo inobservable. Para poder investigar la hipótesis substantiva debemos deducir algunas implicaciones observables de la misma.

(b) *Hipótesis de investigación* (que los ordenadores mejoran el aprendizaje de la estadística). Es una implicación observable de la hipótesis alternativa. Si no obtenemos apoyo para la hipótesis de investigación, la hipótesis substantiva no se verá apoyada.

(c) Con frecuencia la hipótesis de investigación no es lo suficientemente específica para dirigir una investigación empírica. Es necesario diseñar una variable dependiente bien definida (en el Ejemplo 1 la puntuación total) obtenida a partir de una tarea experimental (el cuestionario) propuesta a algunos sujetos (los grupos control y experimental al finalizar la enseñanza). Con esta base podemos construir una *hipótesis experimental* (que el rendimiento en el cuestionario será mejor en la población experimental).

Tabla 2.*Diferentes Niveles de Hipótesis en la Investigación Experimental*

Hipótesis implicada	Ejemplo
Hipótesis substantiva	El aprendizaje depende de las representaciones disponibles
Hipótesis de investigación	Los ordenadores favorecen el aprendizaje de la estadística
Hipótesis experimental	Las puntuaciones en el cuestionario son más altas en los estudiantes que usan el ordenador
Hipótesis estadística alternativa	$H_1 \equiv \mu_e > \mu_c$
Hipótesis nula	$H_0 \equiv \mu_e = \mu_c$

(d) Una implicación de la hipótesis experimental se usará para llevar a cabo el análisis estadístico (que la puntuación media en el cuestionario será más alta en los estudiantes enseñados con ayuda del ordenador que en los estudiantes que no tienen acceso al ordenador). Esta implicación es la *hipótesis estadística alternativa*, $H_1 \equiv \mu_e > \mu_c$, que no coincide con la hipótesis experimental, sino que es una consecuencia de ella, a nivel estadístico.

Finalmente, el complemento lógico de la hipótesis estadística alternativa es que la puntuación media en los dos poblaciones de estudiantes es la misma, $H_o \equiv \mu_e = \mu_c$. El establecer la hipótesis nula sirve para especificar la distribución del estadístico de contraste en el muestreo y comenzar la cadena de razonamientos (Tabla 3) que nos llevarán a rechazar o no la serie de hipótesis que hemos descrito y que se muestran en la Tabla 2.

Siguiendo a Chow (1996) presentamos en la Tabla 3 la serie de implicaciones deductivas anidadas de las cuales sólo la más interna (Implicación 5) se relaciona con el proceso de contraste

de significación. Este es el núcleo central de la cadena completa de implicaciones que en su conjunto constituyen el proceso de inferencia científica para dar apoyo a una hipótesis substantiva. En consecuencia, la concepción del contraste de hipótesis como prueba de significación se ajusta en forma natural a este tipo de investigación, mientras que la concepción de los contrastes estadísticos como proceso de decisión serían preferibles en las situaciones prácticas donde debemos tomar una decisión, como en el Ejemplo 2 o en control de calidad.

Tabla 3. *Cadena de Razonamientos Implicados en la Obtención de Apoyo a una Hipótesis Sunstantiva*

Implicación 1	Si el aprendizaje depende de las representaciones disponibles, entonces los ordenadores favorecen el aprendizaje de la estadística
Implicación 2	Si los ordenadores favorecen el aprendizaje de la estadística, entonces las puntuaciones en el cuestionario son más altas en los estudiantes que usan el ordenador
Implicación 3	Si las puntuaciones en el cuestionario son más altas en los estudiantes que usan el ordenador, entonces $\mu_e > \mu_c$
Implicación 4	Si no ocurre que $\mu_e > \mu_c$, entonces $\mu_e = \mu_c$
Implicación 5	Si $\mu_e = \mu_c$, entonces un valor significativo de $\bar{x}_e - \bar{x}_c$ es altamente improbable
Observation	$\bar{x}_e - \bar{x}_c$ es significativo,
Comclusión 5	$\bar{x}_e - \bar{x}_c$ es significativo, por tanto rechazamos que $\mu_e = \mu_c$
Conclusión 4	Rechazamos que $\mu_e = \mu_c$, por tanto asumimos que $\mu_e > \mu_c$
Conclusión 3	$\mu_e > \mu_c$; así que, supuesto que hubo un control experimental adecuado, suponemos que las puntuaciones en el cuestionario son más altas en los estudiantes que usan el ordenador
Conclusión 2	Las puntuaciones en el cuestionario son más altas en los estudiantes que usan el ordenador; supuesto que el test es una medida válida y fiable del aprendizaje, entonces los ordenadores favorecen el aprendizaje de la estadística
Conclusión 1	Los ordenadores favorecen el aprendizaje de la estadística, supuesto que la única diferencia en los dos métodos de enseñanza es el sistema de representaciones disponible, entonces la hipótesis susbtantiva es apoyada por nuestros datos

Las implicaciones 1 a 4 en la Tabla 3 no están apoyadas por la teoría estadística, sino por consideraciones teóricas del campo bajo estudio y por un control experimental adecuado que asegure que todas las variables concomitantes relevantes se mantienen constantes y que el cuestionario dado a los estudiantes es una medida válida y fiable del constructo estudiado (aprendizaje). Según Chow (1996), muchas de las críticas en contra del contraste estadístico son injustas, pues se refieren no al procedimiento estadístico (implicación 5 en la Tabla 3), sino al resto de los componentes del proceso inferencial (implicaciones 1 a 4). Reemplazar o complementar los contrastes estadísticos con otros métodos tales como intervalos de confianza o análisis de potencia no resolverá los problemas de adecuado control experimental o falta de marco teórico pertinente en un campo particular de estudio.

5. Algunas Cuestiones Filosóficas

Hemos identificado ya algunas de las razones que dificultan la comprensión de los tests estadísticos. Por un lado, el contraste estadístico implica una serie de conceptos como hipótesis nula

y alternativa, errores Tipo I y II, probabilidad de los errores, resultados significativos y no significativos, población y muestra, parámetro y estadístico, distribución de la población y distribución muestral. Algunos de estos conceptos son mal interpretados o confundidos por los estudiantes e investigadores experimentales. Más aún, la estructura formal de los contrastes estadísticos es superficialmente parecida a la de la prueba por contradicción, pero hay diferencias fundamentales, no siempre bien comprendidas entre estos dos tipos de razonamientos.

En una prueba por contradicción razonamos en la forma siguiente:

Si A, entonces B no puede ocurrir.
B ocurre; entonces deducimos que A es falso.

En un contraste estadístico tratamos de aplicar un razonamiento similar en la siguiente forma:

Si A, entonces es muy poco verosímil que B ocurra;
Ocurre B, y rechazamos A

Sin embargo no sería una conclusión válida deducir que es muy improbable que A sea cierto y aquí es donde encontramos la confusión.

A estas dificultades se añade la controversia que rodea a la inferencia estadística en filosofía de la inferencia y la dificultad de encontrar relaciones lógicas entre teorías y hechos. La ciencia se construye a partir de observaciones empíricas y no podemos tomar datos de las poblaciones completas, sino sólo de muestras de las mismas. Esperamos del contraste de hipótesis más de lo que nos puede dar y bajo estas expectativas subyace el problema filosófico de hallar criterios científicos para justificar el razonamiento inductivo, como estableció Hume. Hasta ahora las contribuciones de la inferencia estadística en esta dirección no han dado una solución completa al problema (Black, 1979; Burks, 1977; Hacking, 1975; Seidenfeld, 1979).

Por otro lado, hay dos concepciones diferentes sobre los contrastes estadísticos que a veces se mezclan o confunden. Fisher concibió las pruebas se significación para confrontar una hipótesis nula con las observaciones y para él un valor p indicaba la fuerza de la evidencia contra la hipótesis (Fisher, 1958). Sin embargo, Fisher no creyó que los contrastes estadísticos proporcionaran inferencias inductivas de las muestras a las poblaciones, sino más bien una inferencia deductiva de la población de todas las muestras posibles a la muestra particular obtenida en cada caso.

Para Neyman (1950), el problema de contraste de una hipótesis estadística se presenta cuando las circunstancias nos fuerzan a realizar una elección entre dos formas de actuar. Aceptar una hipótesis sólo significa decidir realizar una acción en lugar de otra y no implica que uno necesariamente crea que la hipótesis es cierta. Para Neyman y Pearson, un contraste estadístico es una regla de comportamiento inductivo; un criterio de toma de decisión que nos permite aceptar o rechazar una hipótesis al asumir ciertos riesgos.

Hoy muchos investigadores emplean los métodos, herramientas y conceptos de la teoría de Neyman-Pearson con un fin diferente, medir la evidencia a favor de una hipótesis dada (Royal, 1997). El razonamiento más interior de la Tabla 3 (Implicación 5, observación y conclusión 5) puede describir el razonamiento usual en los contrastes de hipótesis hoy día, que consiste en:

(a) Una decisión binaria - decidir si el resultado es o no significativo. Esta decisión se lleva a cabo comparando el valor *p* con el nivel de significación, que se establece antes de recoger los datos.
(b) Un procedimiento inferencial que engloba un silogismo condicional (implicación 5 en la Tabla 3): Si $\mu_e = \mu_c$, entonces un valor significativo $\bar{x}_e - \bar{x}_c$ es muy improbable; $\bar{x}_e - \bar{x}_c$ es significativo, por tanto rechazamos $H_0 \equiv \mu_e = \mu_c$.
(c) Otro procedimiento inferencial que incluye un silogismo disyuntivo. O bien $\mu_e > \mu_c$, o $\mu_e = \mu_c$; si rechazamos que $\mu_e = \mu_c$, entonces $\mu_e > \mu_c$

Como consecuencia, la práctica actual de los contrastes estadístico tiene elementos de Neyman-Pearson (ya que es un procedimiento de decisión) y de Fisher (es un procedimiento inferencial, en el que usamos los datos para proporcionar evidencia a favor de la hipótesis), que se aplican en diferentes fases del proceso.

Otras características tomadas de Neyman-Pearson son que H_0 es la hipótesis de no diferencia, que el nivel de significación α debe escogerse antes de analizar los datos y debe mantenerse constante, así como los dos tipos de error. De Fisher conservamos la sugerencia de que la inferencia se basa en una probabilidad condicional; la probabilidad de obtener los datos supuesta cierta H_0, y que H_0 y H_1 son mutuamente exclusivas y complementarias. Deberíamos añadir que algunos investigadores suelen dar una interpretación Bayesiana a los resultados de los contrastes de hipótesis (clásicos), a pesar de que el enfoque de la estadística Bayesiana es muy diferente de las teorías tanto de Fisher como de Neyman y Pearson.

6. Factores Psicológicos que Contribuyen a la Prevalencia de Errores Comunes

La anterior práctica de los contrastes estadísticos ha sido llamada marco ortodoxo de Neyman-Pearson (Oakes, 1986) o lógica híbrida de Neyman-Pearson (Gingerenzer, 1993). Este último autor piensa que dicha lógica puede explicar la creencia en que la inferencia estadística proporciona una solución algorítmica al problema de la inferencia inductiva, y el consecuente comportamiento mecánico que con frecuencia se presenta en relación con los contrastes estadísticos.

Como hemos descrito, Fisher y Neyman/ Pearson tuvieron diferentes interpretaciones de los contrastes estadísticos, incluyendo la forma es que se deben determinar los niveles de significación y la interpretación de un resultado significativo. Según Gingerenzer et al. (1989), la disputa entre estos autores se ha ocultado en las aplicaciones de la inferencia estadística en psicología y otras ciencias experimentales, donde se ha asumido una única solución para la inferencia. Libros de texto como el de Guilford (1942) contribuyeron a difundir una mezcla de las lógicas de los contrastes de significación de Fisher con algunos componentes de Neyman-Pearson, dando una interpretación Bayesiana al nivel de significación y otros conceptos relacionados.

Con una analogía esclarecedora, Gingerenzer et al. (1989) comparan las características de Neyman-Pearson en la práctica actual de los contrastes estadísticos con el superego del razonamiento estadístico, porque prescriben los que debería hacerse y no da libertad a los investigadores. Requieren la especificación de hipótesis precisas, niveles de significación y potencia antes de recoger los datos y la probabilidad de error debe interpretarse en el contexto de muestreo repetido. Los componentes de Fisher se comparan al ego del razonamiento estadístico. Es conveniente para los investigadores, que quieren llevar a cabo su investigación y publicar sus trabajos, incluso a costa de determinar el nivel de significación después del experimento, establecer una hipótesis alternativa difusa o no establecerla antes de coger los datos, e interpretar la probabilidad de error como la probabilidad de error en su propio experimento.

El tercer componente en el comportamiento del investigador descrito por Gingerenzer et al. (1989) es el deseo Bayesiano de asignar probabilidades a las hipótesis en base a los datos de investigación (el id de la lógica híbrida). Cuando encontramos un resultado significativo nos preguntamos si este resultado puede ser debido al azar o si por el contrario es consecuencia de nuestra manipulación experimental. Falk (1986) encuentra natural la interpretación del nivel de significación como la probabilidad a posteriori de error, una vez que hemos rechazado la hipótesis en la que el investigador está interesado. Gingerenzer et al. (1989) sugieren que es el conflicto entre estos tres componentes psicológicos lo que explica nuestros usos incorrectos de la inferencia estadística y la institucionalización del nivel de significación como medida de la calidad de la investigación en revistas científicas y manuales de estadística.

Por otro lado, los sesgos en el razonamiento inferencial son sólo un ejemplo del pobre razonamiento de los adultos en los problemas probabilísticos, que ha sido extensamente estudiado

por los psicólogos en relación con otros conceptos, como la aleatoriedad, probabilidad y correlación (Kahneman, Slovic, y Tversky, 1982; Nisbett y Ross, 1980). En el caso específico de interpretación incorrecta de los resultados de la inferencia estadística, Falk y Greenbaum (1995) sugieren la existencia de mecanismos psicológicos profundos que llevan a las personas a creer que eliminan el azar y minimizan su incertidumbre cuando obtienen un resultado significativo. Describen *la ilusión de la prueba probabilística por contradicción* o *ilusión de alcanzar la improbabilidad*, que consiste en la creencia errónea de que la hipótesis nula se vuelve improbable cuando se obtiene un resultados significativo, basada en una generalización abusiva del razonamiento lógico a la inferencia estadística (Birnbaum, 1982; Lindley, 1993).

Mientras una contradicción definitivamente prueba la falsedad de la premisa de partida, la creencia que al obtener datos cuya probabilidad condicional bajo una hipótesis dada es baja implica que la hipótesis condicionante es improbable es una falacia. La ilusión de la prueba probabilística por contradicción es, sin embargo, aparentemente difícil de erradicar, a pesar de la clarificación que se hace en muchos libros de estadística. En otros casos, esta concepción errónea está implícita en los libros de texto, como muestran Falk y Greenbaum (1995).

Según Falk (1986), las concepciones erróneas respecto al nivel de significación se relacionan también con las dificultades en discriminar las dos direcciones de las probabilidades condicionales, lo que se conoce como *la falacia de la condicional transpuesta* (Diaconis y Friedman, 1981), que desde hace tiempo se reconoce como extendida entre los estudiantes e incluso los profesionales. Además, Falk (1986) sugiere que la ambigüedad verbal al presentar α como *P(Error Tipo I)* puede provocar confusión entre las dos direcciones opuestas de la probabilidad condicional entre los estudiantes, que pueden pensar estar tratando con la probabilidad de un suceso simple. Falk sugiere que "Error Tipo I" es una expresión desafortunada que no debería usarse en si misma. Un "suceso condicional" no es un concepto legítimo y sólo las probabilidades condicionales están inequívocamente definidas, aunque esta confusión también aparece en algunos libros de texto.

Aunque α es una probabilidad condicional bien definida, la expresión "Error Tipo I" no está redactada como una condicional, ni indica cual de las dos combinaciones posibles de los sucesos que intervienen se refiere. En consecuencia, cuando rechazamos H_0 y nos preguntamos por el tipo de error que podríamos cometer, el concepto "Error Tipo I" nos viene a la mente en forma inmediata ya que la distinción crucial entre las dos direcciones opuestas de la probabilidad condicional ha sido difuminada. Esto nos lleva a interpretar el nivel de significación como la probabilidad de la conjunción de dos sucesos "la hipótesis nula es cierta" y "la hipótesis nula es rechazada" (Menon, 1993).

Durante muchos años se han lanzado críticas en contra del contraste estadístico y ha habido muchas sugerencias de eliminar este procedimiento de la investigación académica. Sin embargo, los resultados significativos siguen siendo publicados en las revistas de investigación, y los errores referentes a los contrastes estadísticos siguen llenando los cursos y libros de estadística, así como los informes publicados de investigación. Falk (1986) sugiere que los investigadores experimentan una confidencia ilusoria en los contrastes estadísticos debido a la sofisticación de los términos y fórmulas matemáticas, que contribuyen a nuestra sensación de que la significación estadística garantiza la objetividad. Un problema adicional es que otros procedimientos estadísticos sugeridos para reemplazar o complementar los contrastes estadísticos (como los intervalos de confianza, la estimación de la magnitud de los efectos experimentales, el análisis de la potencia y la inferencia Bayesiama) no resuelven los problemas filosóficos y psicológicos que hemos descrito.

7. Revisión de Algunas Críticas Comunes en Contra de los Contrastes de Hipótesis

Hemos analizado con detalle la lógica de los contrastes estadísticos, su papel en la inferencia científica y los factores filosóficos y psicológicos que contribuyen a su comprensión y uso

inadecuados. En esta sección estudiaremos algunas críticas frecuentes en contra del contraste estadístico.

1. Lo que establecemos en la hipótesis nula del ejemplo 1 es que no hay diferencia entre las medias de dos poblaciones. Es evidente para muchos críticos que la hipótesis nula nunca es cierta y por tanto los contrastes estadísticos no son válidos, al basarse en una premisa falsa (que la hipótesis nula es cierta).

Es fácil deducir que esta crítica no es pertinente, ya que el hecho de que la hipótesis nula no sea cierta no invalida la lógica del contraste estadístico. Esta lógica no se ve afectada por el hecho de que la hipótesis nula sea cierta o falsa, porque lo que afirmamos en un contraste es que un resultado significativo es improbable, en caso de que la hipótesis nula sea cierta. Esto es una propiedad matemática de la distribución muestral que no tiene nada que ver con la certeza o falsedad de la hipótesis nula.

2. En la práctica identificamos la hipótesis de interés con la hipótesis estadística alternativa. Pero la hipótesis estadística alternativa no indica la magnitud exacta de la diferencia entre las medias de las poblaciones. La significación estadística no informa sobre la significación práctica de los datos.

Cuando aplicamos esta crítica a los contrastes de significación (Ejemplo 1) estaríamos cayendo en la confusión entre los diferentes niveles de hipótesis implicados en un procedimiento inferencial, que se muestran en las Tablas 2 y 3. El fin de la investigación experimental orientada a la confirmación de teorías es proporcionar apoyo a la hipótesis substantiva. Como vemos en el ejemplo 1, la magnitud de las diferencias entre las medias de las poblaciones no depende de la hipótesis substantiva. La diferencia se refiere a la distribución del estadístico en el muestreo, esto es, a la hipótesis estadística alternativa. No hay una única correspondencia entre la hipótesis substantiva y la hipótesis estadística alternativa, que se deduce de un experimento particular y de un instrumento particular. Las teorías deben evaluarse con un razonamiento cuidadoso y un profundo juicio (Harlow, 1977).

En el contexto de toma de decisión (Ejemplo 2), por el contrario, la magnitud del efecto podría ser relevante para la decisión. En estos casos, los contrastes estadísticos son todavía útiles para hacer la decisión, aunque deben complementarse con el análisis de la potencia y/o la estimación de la magnitud de los efectos, dependiendo de las preguntas de interés en la investigación (Levin, 1998a).

3. La elección del nivel de significación es arbitraria; por tanto algunos datos podrían ser significativos a un nivel dado y no serlos a un nivel diferente.

Es cierto que el investigador escoge el nivel de significación, pero esta arbitrariedad no implica, sin embargo, que el procedimiento sea inválido o inútil. Además es también posible, si se sigue el enfoque de Fisher, usar el p-valor exacto para rechazar las hipótesis nulas a diferentes niveles, aunque en la práctica actual del contraste estadístico se aconseja elegir el nivel de significación antes de recoger los datos para dar mayor objetividad a la decisión.

4. La significación estadística no informa de la probabilidad de que la hipótesis sea cierta ni del verdadero valor del parámetro. Por ello muchos investigadores sugieren reemplazar los contrastes por intervalos de confianza.

Es verdad que los contrastes no informan acerca de la probabilidad de que la hipótesis sea cierta, pero tampoco los intervalos de confianza informan sobre esta probabilidad. Los intervalos de confianza son intervalos en los que el verdadero valor del parámetro se encuentra en un porcentaje dado de muestras, aunque no aseguran en que intervalo estará el parámetro en nuestro experimento

particular. Por tanto no substituyen sino más bien complementan los contrastes de hipótesis y están sujetos a las mismas controversias e interpretaciones erróneas que aquellos.

5. Los errores Tipo I y Tipo II están inversamente relacionados. Para los críticos los investigadores parecen ignorar los errores Tipo II mientras prestan una atención indebida a los errores Tipo I.

Aunque las probabilidades de los dos tipos de error están inversamente relacionadas, hay una diferencia fundamental entre ellas. mientras que la probabilidad de error Tipo I α es una constante que puede elegirse antes de realizar el experimento, la probabilidad de error Tipo II es función del verdadero valor desconocido del parámetro. Para resolver este problema, el análisis de la potencia supone diferentes valores posibles del parámetro y calcula la probabilidad de error Tipo II para estos diferentes valores. Esta práctica es útil para algunas aplicaciones de la inferencia, como en la toma de decisiones (ejemplo 2) y control de calidad. Sin embargo podemos aplacar aquí las mismas objeciones que en el punto s respecto al experimento orientado a apoyar una hipótesis substantiva teórica dada (ejemplo 1) donde no hay indicación sobre el valor particular de los parámetros de las variables experimentales dependientes. Es decir, los dos tipos de error no juegan el mismo papel en la corroboración de teorías científicas, aunque pueden ser igualmente importantes en otras aplicaciones de la inferencia, como el control de calidad.

5. El significado de un resultado no significativo no es claro. Para algunos críticos esto se debe a que el contraste que se usa no tiene suficiente potencia.

Podemos aplicar aquí el mismo razonamiento seguido en el punto 4. Es claro que las hipótesis nulas y alternativa, el rechazo y aceptación de la hipótesis nula, los resultados significativos y no significativos no juegan un papel simétrico en los contrastes de significación. Mientras que un resultado significativo contradice la hipótesis nula por su baja probabilidad, un resultado no significativo es altamente probable cuando la hipótesis nula es cierta, pero también podría deberse a otros factores. Esto también puede ocurrir en la prueba por contradicción, donde hacemos el siguiente razonamiento:

Si A es cierta, B es falsa
Entonces si B es cierta, A es falsa

Si B ocurre, podemos concluir que A no es posible, pero cuando ocurre B no podemos deducir que A sea necesariamente cierta; hay aquí también una asimetría entre las consecuencias de B y no B.

8. La enseñanza y Aprendizaje de Conceptos de Inferencia

En este trabajo hemos descrito la lógica de los contrastes de significación, su papel en la investigación experimental, las dificultades psicológicas y filosóficas relacionadas con ellos y, finalmente, hemos revisado algunas críticas frecuentes contra las pruebas de significación estadística. Estas críticas no pueden aplicarse a los procedimientos matemáticos en el contraste de hipótesis, donde no hay contradicciones. Por el contrario se relacionan con los usos incorrectos del contraste de significación y son consecuencias de errores conceptuales y de problemas filosóficos y psicológicos que hemos descrito a lo largo de este trabajo.

Los educadores estadísticos no son indiferentes a estos problemas, como se muestra en la sesión de trabajos invitados sobre la educación estadística y la controversia en torno a los contrastes de significación en la 52 Sesión del Instituto Internacional de Estadística y en la Mesa Redonda de IASE sobre "La formación de los investigadores en el uso de la estadística". Como describe Ito (1999), hay tres niveles diferentes en la controversia respecto a los contrastes estadísticos:

(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. Mas aún, los editores de revistas y las sociedades profesionales están sugiriendo cambios en las políticas de publicación científica, respecto a los métodos estadísticos (Lecoutre, 1999).
(c) La controversia en la enseñanza acerca de cuándo, como y con qué profundidad deberíamos enseñar la inferencia estadística.

Coincidimos con Ito en que estos tres niveles diferentes están de hecho interrelacionados, porque nuestras concepciones acerca de la teoría estadística, también afectan a nuestra aplicación y enseñanza de la estadística. Esto es importante, ya que, con la creciente investigación sobre la enseñanza y aprendizaje de la estadística, el análisis de datos se está introduciendo cada vez más en el nivel escolar (Shaughnessy, Garfield, y Greer, 1997) incluyendo también en muchos países los rudimentos de inferencia (Dahl, 1999). Nuestra opinión es que el contraste de hipótesis no debiera abandonarse en las ciencias sociales y educación, sino que debería cambiarse su enseñanza y su práctica hasta llevar a un "proceso significativo" (Levin, 1998b), que incluya replicaciones independientes de los estudios, elección de tamaños óptimos de muestras, la combinación del contraste de hipótesis con intervalos de confianza y/o estimación del tamaño de los efectos y la especificación de criterios de "éxito" antes de realizar el experimento.

Nuestro análisis muestra con claridad la complejidad conceptual de los contrastes estadísticos, y la atención particular que debe darse a la enseñanza de la inferencia si queremos prevenir en nuestros estudiantes futuras faltas de comprensión como las descritas por Vallecillos (1999). Se ha sugerido la revisión de la metodología de enseñanza en los cursos introductorios de estadística (Moore, 1997 y la discusión relacionada) para pasar a un modelo constructivista de aprendizaje, en el que el profesor guíe a sus estudiantes hacia unas competencias y conocimientos estadísticos más específicos. Nuevos libros de texto que cambian el papel del estudiante de simple oyente a una participación más activa en actividades estructuradas (e.g., Rossman, 1996) facilitan este enfoque.

Puesto que los ordenadores hacen posible una variedad de cálculos y representaciones gráficas, Moore (1997) recomienda dar a los estudiantes la oportunidad de experimentar con datos y problemas reales. Específicamente, la simulación con ordenador puede contribuir a mejorar la comprensión del estudiante de las ideas de variabilidad muestral, estadístico y su distribución, respecto a las cuales hay muchas concepciones erróneas (Rubin, Bruce, y Tenney, 1991; Well, Pollatsek, y Boyce, 1990) y que son esenciales para comprender la lógica del contraste de significación. Por ejemplo, delMas, Garfield y Chance (1999) describen el software Sampling Distribution y actividades educativas diseñadas para guiar a sus estudiantes en la exploración de las distribuciones muestrales. En su experimento, los estudiantes podían cambiar la forma de la distribución teórica de la población (normal, sesgada, bimodal, uniforme, en forma de U) y simular la distribución muestral del diferentes estadísticos para varios tamaños de muestra. Las actividades estuvieron orientadas a enfocar la atención de los estudiantes hacia el Teorema Central del Límite.

Sin embargo, e incluso cuando los resultados demostraron un cambio significativo positivo en los estudiantes como consecuencia de la instrucción, delMas, Garfield y Chance (1999) avisan que el uso de la tecnología y actividades basadas en resultados de la investigación no siempre produce una comprensión efectiva de las distribuciones muestrales. Los autores sugieren que las nuevas actividades y el aprendizaje del software puede ser demasiado exigente para algunos, y la nueva información sobre el software puede interferir con el aprendizaje de los estudiantes sobre las distribuciones muestrales, cuya comprensión requiere la integración de las ideas de distribución, promedio, dispersión, muestra y aleatoriedad. Es claro que necesitamos más investigación para

comprender como podemos usar la tecnología para ayudar a los estudiantes en su proceso de aprendizaje (ver el trabajo de Ben-Zvi, en este número de la revista). En particular necesitamos encontrar buenas situaciones didácticas en las que los estudiantes sean confrontados con sus concepciones erróneas, como la confusión entre una probabilidad condicional y su inversa o su creencia en la posibilidad de calcular la probabilidad de una hipótesis (dentro de la concepción objetiva de la probabilidad).

Por otro lado, el contraste estadístico sólo es una parte de proceso más general de inferencia científica, como se indica en las tablas 2 y 3. Sin embargo, frecuentemente encontramos que la estadística se enseña aisladamente, sin conectarla con un marco más general de metodología de investigación y diseño experimental. Desde nuestro punto de vista, es necesario discutir el papel de la estadística en la investigación experimental con los estudiantes y hacerlos conscientes de las posibilidades y limitaciones de la estadística en el trabajo experimental. Aún más, coincidimos con la sugerencia de Wood (1998) de enfocar el curso introductorio de estadística alrededor del razonamiento estadísticos, es decir el ciclo de aprendizaje Planificación-Conjetura-Comprobación-Acción. El análisis estadístico de datos no es un proceso mecánico y, por tanto, no debería ser enseñado o aplicado de esta forma. Puesto que la estadística no es una forma de hacer sino una forma de pensar que nos puede ayudar a resolver problemas en las ciencias y la vida cotidiana, la enseñanza de la estadística debería empezar con problemas reales mediante los cuales los estudiantes puedan desarrollar sus ideas, trabajando las diferentes etapas en la resolución de un problema real (planificar la solución, recoger y analizar los datos, comprobar las hipótesis iniciales y tomar una decisión en consecuencia).

Finalmente, recomendamos a los investigadores que reconozcan la complejidad de la aplicación de la inferencia estadística para resolver problemas reales y que necesitan la colaboración de estadísticos profesionales, además del uso de su conocimiento profesional para juzgar hasta qué punto sus cuestiones de investigación pueden contestarse por medio del análisis estadístico.

Agradecimientos: La autora agradece a Joel R. Levin y Paul K. Ito sus comentarios y valiosas sugerencias a una versión previa del trabajo. Esta investigación ha sido financiada por el proyecto BSO2000-1507 (M.E.C, Madrid).

Referencias

American Psychological Association (1994). *Publication Manual of the American Psychological Association* (Cuarta edición). Washington, DC: American Psychological Association.

Batanero, C., Serrano, L. y Green, D. R. (1998). Randomness, its meanings and implications for teaching probability. *International Journal of Mathematics Education in Science and Technology*, 29 (1), 113-123.

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics, 4,* 24–27.

Black, M. (1979). *Inducción y probabilidad*. Madrid: Cátedra.

Burks, A. W. (1977). *Chance, cause, reason: An inquiry into the nature of scientific evidence*. Chicago: University of Chicago Press.

Chow, L. S. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

delMas, R. C., Garfield, J. B., y Chance, B. (1999). *Exploring the role of computer simulations in developing understanding of sampling distributions*. Comunicación presentada en la reunión anual de la American Educational Research Association, Montreal, Canada.

Dahl, H. (1999). Teaching hypothesis testing. Can it still be useful? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tomo 58, Libro 2) (pp. 197-200). Helsinki: International Statistical Institute.

Diaconis, P. y Freedman, D. (1981). The persistence of cognitive illusions. *Behavioral and Brain Sciences, 4,* 378-399.

Ellerton, N. (1996). Statistical significance testing and this journal. *Mathematics Education Research Journal, 8*(2), 97–100.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

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

Fisher, R. A. (1935). *The design of experiments.* Edimburgh: Oliver y Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver y Boyd.

Fisher, R. A. (1958). *Statistical methods for research workers* (13 edición). New York: Hafner.

Gingerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. In G. Keren y C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methdological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gingerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., y Kr¸ger, L. (1989). *The empire of chance. How probability changed science and everyday life.* Cambridge: Cambridge University Press.

Guilford, J. P. (1942). *Fundamentals of statistics in psychology and education*. New York: Basic Books.

Hacking, I. (1975). *The logic of statistical inference*. Cambridge: Cambridge University Press.

Harlow, L. L. (1997). Significance testing: Introduction and overview. En L. L. Harlow, S. A. Mulaik, y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Harlow, L. L., Mulaik, S. A., y Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Lecoutre, B. (1999). Beyond the significance test controversy: Prime time for Bayes? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 205-208). Helsinki, Finland: International Statistical Institute.

Levin, J. R. (1998 a). To test or not to test H_0*? Educational and Psychological Measurement, 58*, 313-333.

Levin, J. R. (1998 b). What if there were no more bickering about statistical significance tests? *Research in the Schools, 2,* 45-53.

Levin, J. R., y Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review, 11*, 143-155.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15*, 22-25.

Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5(*1), 4–18.

Moore, D. S. (1995). *The basic practice of statistics.* New York: Freeman.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123–155.

Moses, L. E. (1992). The reasoning of statistical inference. In D. C. Hoaglin y D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107-122). Washington, DC: Mathematical Association of America.

Morrison, D. E., y Henkel, R. E. (Eds.). (1970). *The significance tests controversy. A reader*. Chicago: Aldine.

Neyman, J. (1950). *First course in probability and statistics*. New York: Henry Holt.

Nisbett, R., y Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgments*. Englewood Cliffs, NJ: Prentice Hall.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.

Pollard, P., y Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10*, 159-163.

Robinson, D. H., y Levin, J. T. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rossman, A. J. (1996). *Workshop statistics: Discovery with data.* New York: Springer.

Royal, R. (1997). *Statistical evidence. A likelihood paradigm*. London: Chapman y Hall.

Rubin, A., Bruce, B., y Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 314–319). Voorburg, The Netherlands: International Statistical Institute.

Seidenfeld, T. (1979). *Philosophical problems of staistical inference: Learning from R. A. Fisher.* Dordrecht, The Netherlands: Reidel.

Shaughnessy, J. M., Garfield, J., y Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, y C. Laborde (Eds.), *International handbook of mathematics education* (Volume 1) (pp. 205-237). Dordrecht, The Netherlands: Kluwer.

Skipper, J. K., Guenter, A. L., y Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social sciences. In D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 155-160). Chicago: Aldine.

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

Vallecillos, A. (1994). *Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios*. Unpublished doctoral dissertation, University of Granada, Spain.

Vallecillos, A. (1995). Comprensión de la lógica del contraste de hipótesis en estudiantes universitarios. *Recherches en Didactique des Mathématiques, 15(*3), 53–81.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 201-204). Helsinki, Finland: International Statistical Institute.

Vallecillos, A., y Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. En L. Puig y A. Gutiérrez (Eds.), *Proceedings of the Twentieth Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4) (pp. 271–378). Valencia, Spain: University of Valencia.

Vallecillos, A., y Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathématiques, 17*(1), 29-48.

Well, A. D., Pollatsek, A., y Boyce, S. J. (1990). Understanding the effects of the sample size on the variability of the means. *Organizational Behavior and Human Decision Processes, 47*, 289–312.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wood, G. R. (1998). Transforming first year university statistics teaching. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, y W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (Volume 1) (pp. 167-172). Singapore: International Statistical Institute.

Yates, F. (1951). The influence of "Statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46,* 19-34.