PUBLICACIONES, 38, 2008
pp. 9-23

ERRORES FRECUENTES EN EL ANÁLISIS DE DATOS EN EDUCACIÓN Y PSICOLOGÍA

Carmen Díaz
Universidad de Huelva
Carmen Batanero
Universidad de Granada
Miguel R. Wilhelmi
Universidad Pública de Navarra

Recibido: 8 febrero 2008 / Aceptado: 12 marzo 2008

RESUMEN

En las investigaciones educativas basadas en una metodología cuantitativa, el análisis de datos se lleva a cabo, generalmente, dentro del marco de la inferencia clásica frecuencial, principalmente a partir del contraste estadístico de hipótesis. En los últimos años se ha señalado el uso e interpretación incorrecta de los contrastes estadísticos. En este trabajo analizamos algunos de los errores más frecuentes en la inferencia estadística y damos algunas sugerencias para superarlos.

Palabras clave: errores, análisis de datos, uso y transmisión de conocimientos estadísticos.

ABSTRACT

In educational research based on quantitative methodology, data analysis is carried out generally within the frame of frequentist inference, mainly by using hypothesis tests whose incorrect use has been criticized in the past years. In this paper we analyse some of the most frequent errors in statistical inference and finish with some suggestions to overcome them.

Key words: Biases, data analysis, use and transmission of statistical knowledge.

1. INTRODUCCIÓN

La inferencia estadística ha jugado un papel destacado en diversas ciencias humanas, como la Educación, la Psicología o la Sociología, que basan sus investigaciones en datos recogidos en muestras de poblaciones mayores a las que quieren extender sus conclusiones. El uso e interpretación de la estadística en estas investigaciones no son

siempre adecuados, como se muestra en una amplia literatura (Morrison y Henkel, 1970; Abelson, 1997; Harlow, Mulaik y Steiger, 1997; Ares, 1999; Borges, San Luis, Sánchez, y Cañadas, 2001; etc.).

En este trabajo analizamos la posibilidad que ofrece la estadística para validar el razonamiento inductivo a partir de datos empíricos y los principales errores conceptuales en la interpretación de diversos objetos matemáticos que intervienen en la inferencia estadística. Finalizamos con algunas sugerencias para mejorar la práctica de la estadística en la investigación psicológica y educativa.

2. EL PROBLEMA DE LA JUSTIFICACIÓN DEL RAZONAMIENTO INDUCTIVO EMPÍRICO Y LAS SOLUCIONES APORTADAS POR LA ESTADÍSTICA

La problemática filosófica de la inferencia estadística se relaciona con la posibilidad de obtener conocimiento general (teorías científicas) a partir de casos particulares (inducción empírica), esto es, con la dificultad de justificar el razonamiento inductivo y sus conclusiones. Este problema ha ocupado a los filósofos y estadísticos por largo tiempo, sin que hasta la fecha se haya obtenido una solución aceptada por consenso (Rivadulla, 1991, Cabria, 1994).

Un autor que tuvo una fuerte influencia en el debate sobre el método inductivo fue Popper, quien propuso que una cierta teoría puede racionalmente considerarse como cierta frente a otras con las que se halla en competencia, si, a pesar de nuestros intentos, no conseguimos refutarla. Popper (1967) sugirió poner a prueba las hipótesis científicas, mediante experimentos u observaciones y comparar los patrones deducidos de la teoría con los datos obtenidos. En caso de que, al hacer estas pruebas, los datos apoyasen la teoría, ésta recibiría una confirmación, que sólo sería provisional, pues los datos futuros podrían contradecirla. En cambio si los datos del experimento se apartasen del patrón esperado, la teoría sería refutada, por lo que el rechazo y la aceptación de las teorías no tendrían el mismo estatuto lógico.

Estas ideas de Popper tuvieron una gran influencia en el desarrollo de la inferencia estadística. Algunos matemáticos las utilizaron para tratar de apoyar el razonamiento inductivo, recurriendo a la probabilidad. Ya que mediante un razonamiento inductivo no es posible llegar a la certidumbre de una proposición (*verdad cierta*), estos autores intentaron enunciar proposiciones probables (*verdad probable*), tratando de calcular la probabilidad de que una hipótesis fuese cierta (Rivadulla, 1991; Batanero, 2000).

Es importante resaltar que la probabilidad de una hipótesis no tiene sentido en inferencia clásica frecuencial, donde la probabilidad se interpreta como el límite de la frecuencia relativa. Ello es debido a que una hipótesis será cierta o falsa siempre, no un porcentaje de veces en una serie de pruebas. Sin embargo, es posible asignar una probabilidad a las hipótesis dentro del marco de la inferencia bayesiana, donde la probabilidad se concibe como un grado de creencia personal (Gingerenzer, 1993; Lecoutre, 1999). En este último caso podremos diferenciar dos usos del concepto de probabilidad de una hipótesis:

- *Probabilidad inicial*, creencia inicial en la hipótesis antes de recoger datos de experimentos donde se trate de poner la hipótesis a prueba.
- *Probabilidad final*, es decir, creencia en la hipótesis una vez se han recogido los datos.

A continuación analizamos los soluciones aportadas al problema de la inducción por Fisher, Neyman y Pearson y la escuela Bayesiana. Aunque el procedimiento de cálculo en las metodologías de Fisher y Neyman-Pearson sea muy similar, el razonamiento y los fines subyacentes en sus aproximaciones a la inferencia son diferentes. A su vez estas dos metodologías son incompatibles con la bayesiana, donde los conceptos de probabilidad, parámetro e inferencia tienen una interpretación distinta de la que reciben en inferencia frecuencial.

2.1. El test de significación de Fisher

Un *test de significación* es para Fisher un procedimiento que permite rechazar una hipótesis, con un cierto *nivel de significación*. En su libro *The design of experiments,* publicado en 1935, Fisher introduce su teoría de las pruebas de significación, que resumimos en lo que sigue.

Supongamos que se quiere comprobar si una cierta hipótesis H_o (hipótesis nula) es cierta. Generalmente la hipótesis se refiere a una propiedad de la población (por ejemplo, el valor supuesto de un parámetro) pero no se tiene acceso a toda la población, sino sólo a una muestra de la misma. Para poner la hipótesis a prueba se organiza un experimento aleatorio asociado a H_o y se considera un cierto suceso *S* que puede darse o no en este experimento, y del cual se sabe que tiene muy poca probabilidad, si H_o es cierta. Realizado el experimento ocurre precisamente *S*. Hay dos posibles conclusiones:

- Bien la hipótesis H_o era cierta y ha ocurrido *S*, a pesar de su baja probabilidad.
- Bien la hipótesis H_o era falsa.

Generalmente el experimento consiste en tomar una muestra de la población sobre la que se realiza el estudio y calcular un estadístico, que establece una medida de discrepancia entre los datos y la hipótesis. En caso de que se cumpla la hipótesis, el estadístico define una distribución, al variar los datos aleatoriamente (Cabriá, 1994; Batanero, 2000). Un test de significación efectúa una división entre los posibles valores de este estadístico en dos clases: resultados estadísticamente significativos (para los cuales se rechaza la hipótesis) y no estadísticamente significativos (Rivadulla, 1991), para los cuáles no se puede rechazar la hipótesis.

Aunque, aparentemente este procedimiento parece de tipo inductivo, en realidad es deductivo. El razonamiento parte de la suposición de que la hipótesis nula es cierta. Bajo este supuesto, se calcula la distribución del estadístico en todas las posibles muestras de la población. A partir de esta se calcula la probabilidad del valor particular del estadístico

obtenido en la muestra y se determina a cual de las dos clases (resultado estadísticamente significativos y no estadísticamente significativos) pertenece, lo cual es un razonamiento deductivo. Es importante destacar tres aspectos en la lógica del test de significación:

- El objetivo del test de significación es falsar la hipótesis nula.
- No se identifica por una hipótesis alternativa concreta ni, por supuesto, el error asociado a la misma (De la Fuente y Díaz, 2003).
- No hay un criterio estándar sobre qué es un "suceso improbable". El valor de la probabilidad por debajo de la cuál rechazamos la hipótesis lo fija el investigador según su juicio subjetivo y su experiencia.

2.2. Los contrastes de hipótesis de Neyman y Pearson

Neyman y Pearson conceptualizan el contraste de hipótesis como un proceso de decisión que permite elegir entre una hipótesis dada H_0 y otra hipótesis alternativa H_1 (Valera, Sánchez y Marín, 2000). Por ello contemplan dos posibles decisiones respecto a H_0: rechazar esta hipótesis, asumiendo que es falsa y aceptando la alternativa, o abstenerse de esa acción. Al tomar una de estas decisiones sobre las hipótesis a partir de los resultados del contraste se consideran dos tipos de error (Nortes Checa, 1993; Peña y Romo, 1997):

- *Error tipo I:* Rechazar una hipótesis nula que de hecho sea verdadera. Este es el error que, desde el punto de vista estadístico, se ha considerado más grave. Para evitarlo, se suele establecer un criterio de prueba que asegura que la probabilidad de cometer este tipo de error sea menor que un número α preestablecido o *nivel de significación*.
- *Error tipo II:* aceptar la hipótesis nula que de hecho es falsa. Beta (β) es la probabilidad de cometer este tipo de error y el complemento de beta ($1 - \beta$) sería la *potencia* del contraste. Mientras que α es un número preestablecido, β es variable, porque su valor depende de cual es el valor del parámetro (generalmente desconocido).

Una vez definidas las hipótesis nula y alternativa y fijada la probabilidad de cometer error tipo I, se elige el contraste de mayor potencia (Nortes Checa, 1993). Calculado el estadístico, se toma la decisión de rechazar o no rechazar la hipótesis nula, comparando el *p*-valor con el nivel de significación o, equivalentemente, comparando el valor del estadístico calculado con el valor crítico. Es importante resaltar las siguientes características de la metodología de Neyman y Pearson:

- El contraste proporciona un criterio para decidir entre una de las dos hipótesis.
- Se reconocen los errores tipo II.
- Las probabilidades de error tienen una interpretación frecuencial. Se trata de hallar un procedimiento que a la larga controle el porcentaje de los dos tipos de error

cuando se repite muchas veces el procedimiento en la misma población. Este tipo de situación ocurre, por ejemplo, en el diagnóstico médico o en el control de calidad.

- Las probabilidades de error son *probabilidades iniciales* y no *finales*. Por tanto, no es posible calcular inductivamente la probabilidad de la hipótesis a partir de los datos, ya que el procedimiento de Neyman y Pearson es un procedimiento deductivo.
- La relación entre las probabilidades de error que hace que, al aumentar una disminuya la otra, si se mantiene constante el tamaño de la muestra.

2.3. Inferencia bayesiana

El posible cálculo inductivo de la probabilidad de una hipótesis lo encontramos en *el teorema de Bayes*, que permite calcular las *probabilidades finales* de una o varias hipótesis, a partir del conocimiento de sus *probabilidades iniciales* y de los datos obtenidos experimentalmente (Rivadulla, 1991: Bolstad, 2004). En su forma más simple, este teorema se expresa en la forma siguiente:

Tenemos un suceso B (los datos) y queremos saber si pueden explicarse por una de las causas $A_1, A_2,\ldots, A_n$ (una serie de hipótesis científicas rivales). Se conocen las probabilidades iniciales $P(A_1), P(A_2),\ldots, P(A_n)$ de cada una de las hipótesis rivales, así como las probabilidades $P(B \mid A_1), P(B \mid A_2),\ldots, P(B \mid A_n)$ o verosimilitud de obtener los datos B dependiendo de cual de las hipótesis es cierta. Entonces, la probabilidad $P(A_j|B)$ (probabilidad final de que la hipótesis A_j sea la verdadera, una vez que hemos obtenido los datos B) viene dada por la siguiente expresión:

$$P(A_j \mid B) = \frac{P(B \mid A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)}$$

La escuela bayesiana postula que el teorema de Bayes (y toda la inferencia bayesiana que se obtiene generalizándolo y desarrollándolo) es un instrumento adecuado para obtener un conocimiento inductivo, pues las probabilidades iniciales pueden ser transformadas en probabilidades finales a la luz de los sucesos observados (Box y Tiao, 1992). No hay que olvidar, sin embargo, que estas probabilidades tienen un carácter subjetivo, pues son probabilidades que cambian con la nueva información (condicionales). Se trata de grados de creencia personales de un investigador en las hipótesis en juego.

3. ERRORES EN EL USO E INTERPRETACIÓN DE LA INFERENCIA

Un problema frecuente en la investigación empírica es la mezcla de las metodologías descritas al emplear la inferencia estadística para poner una hipótesis a

prueba (Gingerenzer, 1993; Falk y Greenbaum, 1995). Se procede de la manera siguiente:

a) Elementos tomados de los contrastes de hipótesis de Neyman-Pearson:
- Se recogen y analizan los datos para "decidir entre dos hipótesis".
- Se consideran dos tipos de error y se fija el nivel de significación α, manteniéndolo constante.

b) Elementos tomados del test de significación de Fisher:
- La finalidad es "proporcionar evidencia en contra de la hipótesis nula".
- Sólo se toma una muestra, por lo que el *p*-valor se interpreta en forma epistémica, basándose solamente en la muestra particular y no en una serie de ensayos repetidos de la población.

c) Elementos tomados de la inferencia bayesiana. Se da una interpretación bayesiana a los resultados, interpretando el *p*-valor como probabilidad de que la hipótesis alternativa sea cierta, una vez rechazada la nula; a pesar de que los principios de la inferencia bayesiana contradicen a los utilizados tanto por Fisher como por Neyman-Pearson.

En definitiva, ni los investigadores ni los profesores de estadística son siempre conscientes de la existencia de diferentes conceptualizaciones dentro de la estadística y las mezclan en la práctica y en la enseñanza del análisis de datos. Esta costumbre ha sido ampliamente criticada (Menon, 1993; Wilkinson, 1999; Altman, 2002), por las posibles repercusiones que, sobre los resultados de la investigación, puede tener. A continuación clasificamos y resumimos los errores de interpretación de la inferencia estadística más frecuentemente descritos, remitiendo al lector a trabajos complementarios en que se puede encontrar un análisis más completo del tema.

3.1. Interpretación del nivel de significación

La interpretación incorrecta más extendida es cambiar los términos de la probabilidad condicional en la definición del nivel de significación α (probabilidad de rechazar la hipótesis nula siendo cierta), interpretándolo como la probabilidad de que la hipótesis nula sea cierta, habiendo tomado la decisión de rechazarla. Por ejemplo, Birnbaum (1982) informó que los 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*" aunque la definición correcta sería "un nivel de significación del 5% indica que, en promedio, 5 de cada 100 veces que la hipótesis nula sea cierta la rechazaremos". El mismo tipo de error se encontró en un estudio con 436 estudiantes universitarios de diferentes especialidades (estadística, medicina, psicología, ingeniería y empresariales) (Vallecillos, 1994), incluso en aquellos que eran capaces de discriminar entre una probabilidad condicional y su inversa (Vallecillos y Batanero, 1996). También se han descrito los mismos errores en profesores de metodología e investigadores en Psicología (Lecoutre, 1999, 2006; Lecoutre, Lecoutre y Poite-

vineau, 2001; Haller y Kraus, 2002). Como indica Cohen (1994), (reproducido también en Harlow, Mulaik y Steiger, 1997), un test de significación:

> *No nos dice lo que queremos saber, y queremos saber tanto lo que queremos saber, que, en nuestra desesperación, creemos que lo dice. Lo que queremos saber es "Dado estos datos, ¿cuál es la probabilidad de que H_0 sea cierta?". Pero, como la mayoría sabemos, lo que nos dice es "Dado que H_0 es cierta, ¿cual es la probabilidad de tener estos datos (u otros más extremos)?* (p. 997).

3.2. Interpretación de resultados significativos

La interpretación incorrecta del nivel de significación se une normalmente a la confusión entre significación estadística y significación práctica. Un resultado significativo implica para Fisher que los datos proporcionan evidencia en contra de la hipótesis nula, mientras que para Neyman y Pearson sólo establece la frecuencia relativa de veces que rechazaríamos la hipótesis nula cierta a la larga (*error tipo I*). Pero la significación práctica implica significación estadística más un efecto experimental (diferencia del valor del parámetro en función de una cierta variable experimental) suficientemente elevado.

La significación estadística por sí sola no implica una significación práctica, pues podemos encontrar datos estadísticamente significativos con un pequeño efecto experimental, siempre que tomemos una muestra grande (Frías, Pascual y García, 2000). Esto no es bien comprendido por los investigadores, como muestra un experimento con 20 investigadores en Psicología y 25 profesionales estadísticos (Lecoutre, 1999), donde se concluye que los psicólogos tenían una confianza excesiva en los contrastes y olvidaban la información previa (por ejemplo, el tamaño de la muestra o el valor del efecto) en la interpretación de los resultados significativos.

Respecto al *p*-valor (probabilidad de obtener el valor dado del estadístico, en caso de ser cierta la hipótesis nula) se piensa que este valor indica la probabilidad de que el valor obtenido del estadístico se deba al azar. Pero, cuando rechazamos la hipótesis nula, no podemos inferir la existencia de una causa particular que llevó al resultado (Batanero, 2000). Por ejemplo, una diferencia significativa de las medias de un grupo experimental y otro de control puede ser debida a un tratamiento particular, pero también puede ser que la razón de las diferencias sea intrínseca a los grupos, por ejemplo, un grupo puede estar formado por individuos más inteligentes o con mejores medios y por ello el primero puede tener mejores resultados en un cuestionario. El diseño experimental trata por eso de controlar las variables extrañas.

3.3. Comparaciones múltiples

Otra confusión consiste en creer en la conservación del valor del nivel de significación cuando se realizan contrastes consecutivos en el mismo conjunto de datos (Moses, 1992). La definición del nivel de significación nos indica que, si llevamos a cabo 100 comparaciones sobre el mismo conjunto de datos y usamos en todos ellos el nivel de sig-

nificación 0,05, habrá que esperar que 5 de las 100 pruebas sean significativas por azar, incluso cuando la hipótesis nula en cada una sea cierta. Por ello, si sobre un mismo conjunto de datos llevamos a cabo muchos contrastes (por ejemplo, comparamos las medias de diferentes variables por género, edad, clase social, centro escolar, etc.) y nos aparece un resultado significativo, se dificulta la interpretación, pues algunas de estas diferencias aparecerán simplemente por azar aunque no haya diferencias reales (Moses, 1992).

3.4. Elección de valores particulares para el nivel de significación

Asimismo se piensa que hay una justificación matemática para dar y tomar un valor del nivel de significación de 0,05 o de 0,01. Esta creencia es infundada, pues no hay ninguna razón matemática para tomar estos valores o cualquier otro pero, aunque se recomienda no fijar el nivel de significación y, en lugar de ello, informar sobre el valor exacto del *p*-valor, los niveles de significación anteriores se usan casi de forma universal. Una consecuencia absurda es la diferenciación de los resultados de investigación entre aquellos que son significativos al 0,05 y los que no lo son, llegando al caso de no publicar un trabajo si la significación es mayor (por ejemplo 0,06) (Skipper, Guenter y Nass, 1970). Se olvida el hecho de que, si la potencia del contraste es baja y el error tipo II es importante, sería preferible una probabilidad mayor de error tipo I.

3.5. Error tipo II y potencia

Una consecuencia de la interpretación incorrecta del nivel de significación como probabilidad *a posteriori* de la hipótesis (interpretación bayesiana) es la creencia en la replicación de los resultados. Kahneman, Slovic, y Tversky (1982) describieron la "creencia en la ley de los pequeños números", consistente en pensar que las características de una población se han de reproducir incluso en una muestra pequeña. Los investigadores olvidan con frecuencia informar de la potencia, incluso cuando el contraste no llega a rechazar la hipótesis nula (Valera, Sánchez, y Marín, 2000). Sin embargo, sería necesario hacer un análisis de potencia antes del estudio, y explicar el criterio elegido para determinar el tamaño del efecto mínimo que se considera relevante (por ejemplo, si se usó información de trabajos anteriores) (Wilkinson, 1999). Asimismo se debe informar sobre el tamaño de muestra, las hipótesis y procedimientos analíticos que llevan al análisis de la potencia (Cohen, 1994; Valera, Sánchez, Marín y Velandrino, 1998).

3.6. Confusión respecto a los diferentes niveles de hipótesis

Diversos autores alertan de que también se confunden los papeles de la hipótesis nula y alternativa, e incluso la hipótesis estadística alternativa y la hipótesis de investigación (Chow, 1996; Vallecillos, 1994). Hay varias hipótesis implicadas en los distintos niveles de abstracción de la investigación experimental orientada a la confirmación de teorías (Chow, 1996). Es necesario aclarar a los estudiantes los distintos niveles de hipótesis dentro de una investigación:

- *Hipótesis substantiva*: es la explicación teórica que nos planteamos acerca del fenómeno de estudio. Suele hacer referencia a un constructo teórico inobservable (por ejemplo la inteligencia, actitud, etc. depende de una cierta variable), por lo que su estudio directo no es posible. Para poder investigar la hipótesis substantiva debemos deducir algunas implicaciones observables de la misma.
- *Hipótesis de investigación*: es una deducción observable de la hipótesis substantiva (por ejemplo, puesto que el rendimiento de los estudiantes se puede deducir de su inteligencia, podemos estudiar si el rendimiento de los estudiantes depende de la variable en cuestión). Trataremos de encontrar apoyo para la hipótesis sustantiva a través de la hipótesis de investigación.
- *Hipótesis experimental*: en muchas situaciones la hipótesis de investigación es todavía demasiado ambigua y necesitamos especificarla mejor para poder estudiarla. Así llegamos a la hipótesis experimental, que se define en términos de variables independientes y dependientes bien definidas (puesto que "rendimiento" es todavía una variable muy ambigua, aunque al menos es observable, podemos comparar las puntuaciones en dos pruebas de rendimiento en matemáticas en los grupos de sujetos que varían respecto a la variable de interés).
- *Hipótesis estadística alternativa*: de la hipótesis experimental pasamos a la hipótesis estadística, que es una consecuencia de la hipótesis experimental a nivel estadístico. Hace referencia a una población de sujetos, descrita mediante un modelo matemático que se especifica por uno o varios parámetros (en el ejemplo, suponemos que la distribución de puntuaciones en la prueba de rendimiento en matemáticas de dos grupos de alumnos sigue una distribución normal, y que la media de uno de los grupos será mayor que la del otro).
- *Hipótesis estadística nula*: el complemento lógico de la hipótesis alternativa. En estadística es usual trabajar con la hipótesis nula de no diferencia. Si la hipótesis alternativa era que la puntuación media en el cuestionario era diferente en dos grupos de sujetos, la hipótesis nula especificaría que no hay diferencias de medias entre los grupos control y experimental.

La teoría estadística se ocupa del último nivel (hipótesis estadísticas nula y alternativa), pero alumnos e investigadores confunden todas estas hipótesis y cuando encuentran un resultado significativo, piensan que se refiere a la hipótesis de investigación o incluso a la sustantiva (Chow, 1996). Es decir, si se rechaza que la media de los dos grupos de estudiantes en la prueba de rendimiento en matemáticas es igual, se deduce (injustificadamente) que el rendimiento (en general) o bien la inteligencia de uno de los grupos es mayor que la del otro.

3.7. Intervalos de confianza

En resumen, la lógica de los contrastes estadísticos es difícil e involucra la comprensión de muchos objetos matemáticos que se confunden entre sí. Por ello, para paliar los errores anteriores se propone complementar los resultados de los contrastes estadís-

ticos o bien sustituirlos por alguna medida de la variabilidad en el muestreo, tal como el error estándar (Kish, 1970; Hunter, 1997), o los intervalos de confianza. Es posible que su escaso uso se deba al hecho de que usualmente son demasiado amplios y los investigadores desconocen que se puede mejorar la precisión aumentando la fiabilidad de los procedimientos y el tamaño de las muestras (Cohen, 1994). Por otro lado, los intervalos de confianza tienen la misma interpretación frecuencial que los contrastes, ya que el coeficiente de confianza sólo nos indica la proporción de intervalos calculados de la misma población con tamaño de muestra dado que cubrirían el valor del parámetro, pero no si el intervalo calculado lo cubre o no (Cumming, Williams y Fidler, 2004).

3.8. Errores en el muestreo

También se producen errores que inciden en los resultados e interpretación de la estadística por una incorrecta identificación de la población en estudio, tomar una muestra de tamaño insuficiente o interpretar resultados de muestras no aleatorias como si el muestreo hubiese sido aleatorio (White, 1980).

Una omisión típica es no especificar con claridad la población objeto de estudio. En los estudios descriptivos no se precisa el uso de la inferencia y en los casos prácticos de muestreo aleatorio repetido (como en control de calidad) la pertinencia de la inferencia es clara. Sin embargo, cuando se toma una única muestra y se usa la inferencia, se generaliza a una población (universo hipotético) que no está bien especificada; se trataría de la población que se obtendría al repetir ilimitadamente la investigación dada en las mismas condiciones temporales, culturales, sociales, y cognitivas (Hagod, 1970).

3.9. Otros problemas

Un contraste de hipótesis sin ninguna información adicional es poco menos que inútil. Por ejemplo, es de poca utilidad dar únicamente la lista de coeficientes de correlación significativos marcados con asteriscos, sin indicar el valor del coeficiente, aunque esta es una práctica frecuente (Abelson, 1997).

Por otro lado, en muchas áreas de investigación, bien debido a la dificultad de tomar muestras aleatorias o al tipo de variables analizadas, es muy difícil alcanzar las condiciones exigidas para aplicar los contrastes en forma correcta (Selvin, 1970). Por ejemplo, al aplicar un test de ji-cuadrado para estudiar la posible asociación entre dos variables en una tabla de contingencia, los supuestos podrían ser violados, bien por tener una frecuencia pequeña en cada casilla, bien por haber reagrupado "convenientemente" filas o columnas en la tabla para que aparezca una cierta asociación, o bien por realizar la prueba en tablas binarias cuando se tiene un conjunto de datos multivariante, en que una relación podría implicar más de una de dichas variables (Lipset, Trow y Coleman, 1970).

Otra dificultad al aplicar los contrastes de hipótesis en las ciencias humanas es el control experimental. Puesto que no es posible controlar todas las variables relevantes, la estadística sugiere la necesidad de aleatorización, que garantiza la existencia de técnicas para medir la probabilidad de que un efecto haya sido producido aleatoriamente. En la

práctica esta situación no se da. Por un lado, el número de observaciones dificulta que se pueda controlar simultáneamente todas las variables de interés. Por otro lado, quizás no todas las variables relevantes hayan sido medidas y algunas variables "confundidas" podrían no ser controladas (Selvin, 1970).

4. ALGUNAS RECOMENDACIONES

Los problemas descritos han sido advertidos desde hace años en Psicología y Educación y nos obligan a replantearnos nuestras prácticas en el análisis de datos empíricos. Por ejemplo, la interpretación errónea del *p*-valor como probabilidad de que la hipótesis nula sea falsa es casi universal y viene acompañada del error de suponer que su complementario es la probabilidad de que la siguiente replicación del experimento tendrá éxito (Cohen, 1994). Una forma de paliar estos problemas sería basarnos más en los nuevos métodos de análisis exploratorio de datos, que utilizan gráficos dinámicos para visualizar y comprender mejor el conjunto particular de datos que tenemos. Ello nos permitiría pensar mejor qué tratamiento estadístico se ha de implementar, en función de la forma y otras características de los datos, y no usar la estadística de forma mecánica. Debido al carácter inductivo de nuestras ciencias, debiéramos apoyarnos más en la replicación de las investigaciones, en vez de considerar que lo importante es la novedad de la investigación.

El interés despertado por la correcta práctica de la estadística se muestra en el número monográfico en *Psychological Science* (Vol. 8, n.1, Enero 1997), en el que se acepta que la obtención de resultados significativos fue un criterio para seleccionar trabajos en esta revista cuando se recibía un número excesivo, aunque no el único (Estes, 1997). Este autor sugiere no imponer ni prohibir ningún método estadístico, sino crear una atmósfera que favorezca la adaptación de la metodología al avance de la ciencia. Respecto a la sugerencia de abandonar el contraste de hipótesis, para evitar los errores de interpretación, la compara a un médico que renunciase a tratar una enfermedad si hubiese muchos enfermos. Los síntomas, en este caso, sugieren al autor que muchos investigadores carecen del conocimiento matemático y lógico mínimo que les prepare para comprender incluso los métodos estadísticos más sencillos y que los esfuerzos para mejorar la práctica de la estadística podrían empezar con la mejora de su enseñanza (p. 19).

Mientras el uso de los contrastes como regla de decisión dicotómica puede llevar a resultados erróneos, hay otros casos en que un test puede ser muy valioso, por ejemplo, para comprobar la bondad de ajuste de los datos a un modelo (Abelson, 1997).

Otra muestra del interés por mejorar la práctica de la estadística es la creación de la *Task Force on Statistical Inference* en la *American Psychological Association* (Wilkinson, 1999; Fidler, 2002). Este comité ha hecho varias recomendaciones, las más importantes son:

- *Contrastes estadísticos*. Es difícil imaginar una situación en que sea preferible una decisión dicotómica (aceptar- rechazar) a informar sobre el *p-valor* o el inter-

valo de confianza. Se recomienda complementar los contrastes con intervalos de confianza y no usar nunca la expresión "aceptar la hipótesis nula".

- *Comparaciones múltiples*. Se sugiere seguir procedimientos especiales (por ejemplo, el test de Bonferroni) para tratar las situaciones en que tenemos que hacer varias comparaciones en la misma muestra.
- *Potencia y tamaño de la muestra*. Se debe proporcionar información sobre el tamaño de muestra y por qué se eligió dicho tamaño. Informar sobre el tamaño de los efectos que se considera mínimo admisible para una significación práctica, establecer las hipótesis con claridad y describir los procedimientos analíticos que llevan al análisis de la potencia.
- *Población y muestra*. Puesto que todo estudio depende de las características de la población y muestra, estas deben ser definidas con claridad, así como el procedimiento de muestreo o la forma de asignar sujetos a los tratamientos. También recomienda controlar las variables relevantes cuando no ha habido diseño aleatorio y en caso de que esto no sea posible usar el término "grupo control" con precaución.
- *Recomendaciones generales*. Se deben describir mejor los datos, incluyendo los datos faltantes, asegurarse de que los resultados publicados no se producen debido a anomalías o valores atípicos. Se enfatiza también la importancia del marco teórico como guía de la investigación y la necesidad de suficientes estudios de tipo exploratorio, antes de avanzar en estudios orientados a la confirmación de teorías. Alertan también del mal uso de los potentes programas de cálculo que pueden llevar a la elaboración de artículos e informes de investigación sin comprender cómo se hacen los cálculos o qué significan.

Finalmente, en los últimos años son muchos los trabajos que sugieren que la inferencia bayesiana proporciona una respuesta más ajustada a los problemas de inducción en ciencias experimentales y que la interpretación de los conceptos bayesianos es más intuitiva para el investigador (Lindley, 1993; Rouanet, 1998; Lecoutre, 1999; Díaz y Batanero, 2006). El análisis bayesiano proporciona al investigador lo que pide, la probabilidad (subjetiva) final de que la hipótesis sea cierta, o la probabilidad de que un parámetro en la población tome un valor dado (por ejemplo, que una diferencia de medias en dos grupos sea mayor que un cierto valor). Por el contrario, el nivel de significación da la probabilidad (objetiva e inicial) de un suceso, bajo la condición de que la hipótesis nula sea cierta, aunque como hemos dicho puede que esta hipótesis no sea cierta.

Por otro lado, el enfoque bayesiano tiene en cuenta la perspectiva del investigador, su conocimiento del problema, ya que el investigador tiene que especificar la distribución inicial. Una crítica a este enfoque es que diferentes investigadores pueden obtener distintos resultados de los mismos datos, lo que podría ser problemático con pequeñas muestras, pero cuando contamos con grandes muestras, el método bayesiano corregirá en sucesivos experimentos los posibles sesgos iniciales. Esta necesidad de sucesivos experimentos (muchos de ellos de carácter exploratorio) antes de avanzar en estudios orientados a la confirmación de teorías, no es una limitación del método bayesiano, sino

la constatación de una necesidad general señalada por el comité *Task Force on Statistical Inference* de la APA, como hemos señalado anteriormente.

Una característica bien conocida de la inferencia bayesiana es que permite reinterpretar muchos procedimientos frecuenciales y esta reinterpretación permite concienciar a los estudiantes sobre posibles interpretaciones erróneas de los tests de significación. Más aún, la posibilidad de investigar interactivamente las distribuciones *finales* mediante un *software* gráfico facilita la comprensión de los conceptos bayesianos. Buscar las formas de facilitar su enseñanza e interpretación es un gran desafío actual.

Agradecimiento: Trabajo financiado por el proyecto SEJ2007–60110, MEC-FEDER.

REFERENCIAS BIBLIOGRÁFICAS

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test? *Psychological Science,* 8(1), 12-14.

Altman, D.G. (2002). Poor-quality medical research. *Journal of the American Medical Association*, 287(21), 2765-2767.

Ares, V. M. (1999). La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta. *Metodología de Encuestas*, 1, 47-68.

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

Batanero, C. y Díaz, C. (2006). Methodological and Didactical Controversies around Statistical Inference. *Actes des 36ièmes Journées de la Société Française de Statistique*. CD ROM. Paris: Société Française de Statistique.

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

Bolstad, W. (2004). *Introduction to Bayesian statistics*. New York: Wiley.

Borges, A., San Luis, C., Sánchez, J.A. y Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema,* 13 (1), 174-178.

Box, G. P. y Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. Nueva York: Wiley.

Cabriá, S. (1994). *Filosofía de la estadística*. Valencia: Servicio de Publicaciones de la Universidad.

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

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist,* 49(12), 997-1003.

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

De la Fuente, E. I. y Díaz, C. (2003). Reflexiones sobre los métodos inferenciales en psicología. En *Libro de resúmenes del VIII Congreso de Metodología de las Ciencias Sociales y de la Salud* (pp. 326 – 327). Valencia: Asociación Española de Metodología de las Ciencias del Comportamiento.

Díaz, C. y Batanero, C. (2006). ¿Cómo puede el método bayesiano contribuir a la investigación en psicología y educación? *Paradigma, 27*(2), 35-53.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science*, 8(1), 18-20.

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.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational And Psychological Measurement*, 62 (5), 749-770.

Fisher, R. A. (1956). Mathematics of a lady testing tea. En J. Newman (Ed.), *The world of mathematics* Vol., III. Simon and Schuster. Traducido como "Las matemáticas de la catadora de té". En J. R. Newman (Ed.), *El mundo de las matemáticas* (Vol. 3, pp. 194-203). Barcelona: Grijalbo, 1979.

Fisher, R. A. (1935). *The design of experiments*. New York: Hafner Press.

Frías, M.D., Pascual, J. y García, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema,* 12, supl. 2, 236-240.

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

Hagod, M. J. (1970). The notion of hypothetical universe. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 65 – 79). Chicago: Aldine.

Haller, H. y Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research,* 7(1). On line: http://www.mpronline.de/issue16/art1/haller.pdf.

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

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8 (1), 3 – 7.

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

Kish, L. (1970). Some statistical problems in research design. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 127 – 141). Chicago: Aldine.

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.

Lecoutre B. (2006). Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science*, *4*, 207-232.

Lecoutre B., Lecoutre M. P. y Poitevineau J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, 69, 399-418.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15(1), 22-25.

Lipset, S. M., Trow, M. A. y Coleman, J. S. (1970). Statistical problems. En D. E. Morrison y

R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 81 – 86). Chicago: Aldine.

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

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

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

Nortes Checa, A. (1993). *Estadística teórica y aplicada*. Barcelona: PPU.

Peña, D. y Romo, J. (1997). *Introducción a la estadística para las ciencias sociales*. Madrid: McGraw-Hill.

Popper, K. R. (1967). *La lógica de la investigación científica*. Madrid: Tecnos.

Rivadulla, A. (1991). *Probabilidad e inferencia científica*. Barcelona: Anthropos.

Rouanet, H. (1998). Statistical practice revisited. En H. Rouanet et al. (Eds.), *New ways in statistical methodology* (pp. 29-64). Berna: Peter Lang.

Selvin, H. C. (1970). A critique of tests of significance in survey research. En. D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 94 – 106). Chicago: Aldine.

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. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 155-160). Chicago: Aldine.

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

Valera, A., Sánchez, J., Marín, F. y Velandrino, A.P. (1998). Potencia Estadística de la Revista de Psicología General y Aplicada (1990-1992). *Revista de Psicología General y Aplicada*, 51 (2).

Vallecillos, A (1994). *Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios*. Tesis doctoral. Universidad de Granada.

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.

White, A. L. (1980). Avoiding errors in educational research. En R. J. Shumway (Ed.), *Research in mathematics education* (pp. 47 – 65). Reston, Va: National Council of Teachers of Mathematics.

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