Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico

Carmen **Batanero** Bernabeu
Facultad de Educacion, Universidad de Granada
España
batanero@ugr.es

Resumen

La inferencia estadística es uno de los temas más enseñados, a la vez, peor comprendido y aplicado a nivel universitario. Recientemente se incluyen contenidos de inferencia en el Bachillerato, e incluso la enseñanza secundaria en algunos países, surgiendo la necesidad de encontrar una transposición didáctica de estos temas asequible a los alumnos no universitarios. En esta conferencia se resumen algunas de las dificultades frecuentes de comprensión de la inferencia clásica, sugiriendo la importancia de educar el razonamiento estadístico en forma progresiva, antes de abordar el estudio formal de la inferencia. Se describen, asimismo, algunas aproximaciones alternativas a la enseñanza de la inferencia que pueden contribuir a la educación de este razonamiento, preparando al estudiante para una mejor comprensión y aplicación de la inferencia en la universidad y trabajo futuro.

Palabras clave: inferencia estadística, dificultades, enfoques alternativos, enseñanza no universitaria.

Según Hacking (1990), uno de los descubrimientos decisivos del siglo XX fue la constatación de que el mundo no es determinista. No sorprende, por tanto, que la inferencia estadística sea uno de los temas más enseñados en la universidad, al ser una herramienta fundamental en la política y administración y en la investigación en todas las áreas de conocimiento, pues permite interpretar la información, obtener predicciones y conclusiones y tomar decisiones adecuadas.

Por otro lado, en los últimos años observamos una tendencia creciente a incluir contenidos de inferencia estadística en el currículo de matemáticas de los últimos años de la escuela secundaria y Bachillerato. Por ejemplo, los estándares americanos NCTM (2000) sugieren que los estudiantes en los grados 6-8 deben usar las observaciones sobre las diferencias entre dos o más muestras para hacer conjeturas sobre las poblaciones correspondientes. En los grados 9-12 la

simulación se debe usar para explorar la variabilidad de los estadísticos muestrales (como la media) y comenzar a comprender lo que es la distribución muestral. En España la especialidad en Ciencias Sociales del Bachillerato (MEC, 2007) incluye en el segundo año (alumnos de 17-18 años), los siguientes contenidos:

- Implicaciones prácticas de los teoremas: Central del límite, de aproximación de la binomial a la normal y Ley de los grandes números.
- Problemas relacionados con la elección de las muestras. Condiciones de representatividad. Parámetros de una población. Distribuciones de probabilidad de las medias y proporciones muestrales.
- Intervalo de confianza para el parámetro *p* de una distribución binomial y para la media de una distribución normal de desviación típica conocida.
- Contraste de hipótesis para la proporción de una distribución binomial y para la media o diferencias de medias de distribuciones normales con desviación típica conocida.

Estas directrices plantean un desafío didáctico, pues la investigación nos alerta que muchos estudiantes, incluso a nivel universitario, tienen concepciones que les impiden hacer una adecuada interpretación de los resultados proporcionados por la inferencia estadística (Vallecillos, 1999; Batanero, 2000; Castro, Vanhoof, Noortgate, & Onghena, 2007; Harradine, Batanero, & Rossman, en prensa). Igualmente se ha denunciado el uso e interpretación incorrecta de la inferencia por parte de investigadores a lo largo de muchos años (por ejemplo, Morrison & Henkel, 1970; Abelson, 1997; Harlow, Mulaik, & Steiger, 1997; Borges, San Luis, Sánchez, & Cañadas, 2001). Una de las posibles razones de esta situación es que la enseñanza es con frecuencia rutinaria, enfatiza las fórmulas y definiciones sin prestar toda la atención que requieren las actividades de interpretación y contexto de donde se tomaron los datos. Aunque los estudiantes lleguen a dar las definiciones y usar los algoritmos con competencia aparente, pueden tener dificultades de comprensión o de conexión de los conceptos estadísticos fundamentales y no sabrán elegir el procedimiento que deben aplicar cuando se enfrenten a un problema real de análisis de datos.

En este trabajo comenzamos describiendo algunos de los errores más denunciados en el uso de la inferencia en su acepción frecuencial. Analizamos seguidamente algunas ideas que podrían servir para introducción el tema de forma progresiva y con una menor formalización. Finalizamos con algunas reflexiones sobre la enseñanza del tema.

Diferentes aproximaciones a la inferencia

La enseñanza actual de la inferencia soslaya la problemática filosófica asociada y las diversas aportaciones que la estadistica ha proporcionado para resolver dicha problemática. De este modo, se presenta la inferencia frecuencial como una metodologia unificada, ocultando las diferentes aproximaciones y las controversias que dentro de la misma estadística ha tenido esta metodologia (Batanero, 2000).

Los problemas filosóficos a que hemos aludido se relacionan con la posibilidad de obtener conocimiento general (teorías científicas) a partir de casos particulares (inducción empírica), esto es, con la posibilidad de justificar el razonamiento inductivo y sus conclusiones, problema de gran importancia en las ciencias empiricas. Este problema ha ocupado a los filósofos y estadísticos

por largo tiempo, sin que se haya obtenido una solución aceptada por consenso (Rivadulla, 1991, Cabria, 1994).

Popper (1967) propuso que una cierta teoría puede racionalmente considerarse como cierta frente a otras con las que se halla en competencia, si, a pesar de numerosos intentos, no se consigue refutarla. Este autor sugirió poner a prueba las hipótesis científicas, mediante experimentos u observaciones y comparar los patrones deducidos de la teoría con los datos obtenidos. La teoria seria provisionalmente confirmada si, los datos recogido siguiesen estos patrones, aunque 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 de la hipotesis tiene mayor fuerza que su confirmación.

Estas ideas de Popper tuvieron una gran influencia en el desarrollo de la inferencia estadística, que fue desarrollada 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*), diversos autores intentaron enunciar proposiciones probables (*verdad probable*), tratando de calcular la probabilidad de que una hipótesis fuese cierta (Ridadulla, 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 & Lectoutre, 2001). 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.

Por otro lado, dentro de la inferencia frecuencial 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. Estas aproximaciones no se diferencian en lo que concierne a los cálculos, pero si en la filosofía y objetivos. La enseñanza ignora estas diferencias y presenta los contrastes de hipótesis como si se tratase de una única metodología.

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 (Ridavulla, 1991), para los cuáles no se puede rechazar la hipótesis.

El razonamiento que apoya un test de significación 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 distribución 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. Si el valor obtenido pertenece a la región de rechazo se rechaza la hipñotesis y en caso contrario no se rechaza. Observamos que en este enfoque no se identifica una hipótesis alternativa concreta (Batanero & Díaz, 2006). Tampoco 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.

Los contrastes de hipótesis de Neyman y Pearson

Neyman y Pearson conceptualizaron el contraste de hipótesis como un proceso de decisión que permite elegir entre una hipótesis dada H_0 y otra hipótesis alternativa H_1 (Rivadulla, 1991). 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:

Error *tipo I:* Rechazar una hipótesis nula cuando es cierta. 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 del 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. 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. En este enfoque, el contraste proporciona un criterio para decidir entre una de las dos hipótesis, se reconocen los errores tipo II, y las probabilidades de error tienen una interpretación frecuencial.

Errores usuales en la interpretación de la inferencia frecuencial

El contraste de hipótesis

El mayor numero de errores e interpretaciones incorrectas en la inferencia estadística están relacionados con el contraste de hipótesis, lo que lleva a una situación paradójica, pues, por un lado, se requiere un resultado significativo para obtener un artículo publicado en muchas revistas y, por otro lado, los resultados significativos son malinterpretados en estas publicaciones (Falk & Greenbaum, 1995).

Un concepto que se suele comprender erróneamente es el nivel de significación, α, que, como se ha indicado, es la probabilidad de rechazar la hipótesis nula, supuesta cierta. La interpretación más común de este concepto consiste en cambiar los dos términos de la probabilidad condicionada, es decir, interpretar el nivel de significación como la probabilidad de que la hipótesis nula sea cierta, una vez que la decisión de rechazarla se ha tomado. A este respecto Birnbaum (1982) informó de que sus alumnos encontraron razonable la siguiente definición: Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados. Falk (1986) informo que la mayoría de sus estudiantes creían que α era la probabilidad de equivocarse al rechazar la hipótesis nula. Vallecillos (1994) confirma estos errores en una investigación utilizando una amplia muestra de estudiantes de diferentes especialidades universitarias. Resultados similares fueron encontrados por Krauss y Wassner (2002) en profesores de universidad implicados en la enseñanza de métodos de investigación.

En los contrastes de hipótesis también se confunden las funciones de las hipótesis nula y alternativa, así como las hipótesis estadística alternativa con la hipótesis de investigación (Chow, 1996). Falk y Greenbaum (1995) sugieren la existencia de mecanismos psicológicos que llevan a la creencia de que la obtención de un resultado significativo supone minimizar la incertidumbre. Vallecillos (1999) describió cuatro concepciones distintas sobre el tipo de prueba de que proporciona el contraste de hipótesis: (a) contraste como una regla en la toma de decisiones, (b) contraste procedimiento para la obtención de apoyo empírico a la hipótesis de investigación; (c) contraste como prueba probabilística de la hipótesis y (d) contraste como prueba matemática de la verdad de la hipótesis. Mientras que las dos primeras concepciones son correctas, muchos estudiantes en su investigación, entre ellos algunos profesores en formación, se inclinan por las dos últimas.

La creencia de que rechazar la hipótesis nula supone demostrar que es errónea, también se encontró en la investigación por Liu y Thompson (2009) al entrevistar a ocho profesores de secundaria, que parecían no comprender la lógica de la inferencia estadística. Liu y Thompson observan que las ideas de probabilidad y atipicidad son fundamentales para comprender la lógica de la prueba de hipótesis, donde se rechaza una hipótesis nula cuando una muestra de esta población se considera lo suficientemente inusual a la luz de la hipótesis nula: El muestreo es un sistema de ideas interrelacionadas que implica repetir la selección al azar, la variabilidad y la distribución. Mientras que una muestra aleatoria simple es una parte fundamental de la inferencia estadística, probablemente más importante es la apreciación de las que podrían haberse elegido. (Saldahna & Thompson, 2002).

Intervalos de confianza

Para paliar los errores anteriores se propone complementar o sustituir los contrastes

estadísticos por intervalos de confianza. Sin embargo, 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, & Fidler, 2004).

Fidler y Cumming (2005) preguntaron a estudiantes de licenciatura y postgrado en ciencias su interpretación de resultados estadísticamente no significativos en un estudio de baja potencia, dando los datos en dos formas diferentes (la primera vez usando valores *p,* y la segunda mediante intervalos de confianza). Los autores indican que los estudiantes interpretaron incorrectamente los valores *p* más a menudo que los intervalos de confianza. Concluyen que es preferible la enseñanza de la inferencia a través de intervalos de confianza (IC), en lugar de a través del contraste de hipótesis.

Pasos en la construcción del razonamiento inferencial

Dado los errores descritos, así como el gran número de conceptos cuya comprensión se requiere para un adecuado uso de la inferencia, es claro que el desarrollo de un razonamiento estadístico adecuado requiere un periodo de varios años. El desarrollo del razonamiento inferencial debe construirse en forma progresiva desde la educación secundaria y Bachillerato, para poder abordar en la universidad el estudio de los contrastes de hipótesis e intervalos de confianza de una forma más adecuada. En este sentido, las nuevas propuestas curriculares proporcionan una oportunidad de introducir gradualmente ideas sobre inferencia, aumentando el nivel de formalización progresivamente. En lo que sigue, describimos algunas etapas y alternativas didácticas para la construcción de este razonamiento.

Muestreo

El alumno debe comprender en primer lugar el proceso de muestreo. El concepto de muestra nos introduce a la inferencia y establece un puente entre la estadística y probabilidad. Es una idea importante, porque todo nuestro conocimiento y juicios sobre el mundo o las personas están basados en el muestreo, ya que, usualmente, sólo podemos estudiar u observar una parcela de la realidad en la que estamos interesados.

Aunque parezca una idea sencilla, muchas personas no alcanzan un razonamiento correcto sobre el muestreo y no siguen las reglas matemáticas normativas que guían la inferencia científica formal, cuando toman una decisión bajo incertidumbre (Kahneman, Slovic y Tversky, 1982). En su lugar, se utilizan heurísticas que reducen la complejidad de los problemas de probabilidad, pero que causan errores y son resistentes al cambio. Por ejemplo, en la *heurística de la representatividad* las personas estiman la verosimilitud de un suceso teniendo sólo en cuenta su representatividad respecto a la población a la cual pertenece. Un error asociado es la *creencia en la ley de los pequeños números* por la que se espera una convergencia de las frecuencias relativas en pequeñas muestras (Sedlemeier, 1999; Jones y Thornton, 2005).

Muchos currículos actuales de secundaria ofrecen la posibilidad de remediar esta situación al incluir algunos contenidos sobre los diferentes métodos de muestreo aleatorio. Será importante que el profesor tenga cuidado con no transmitir la idea de que una muestra aleatoria es una copia de la población y proporcione a los estudiantes la posibilidad de observar la variabilidad del muestreo. Para entender el propósito de extraer una muestra aleatoria simple en inferencia es necesario asimilar dos ideas aparentemente antagónicas: la representatividad y la variabilidad

(Batanero, Godino, Vallecillos, Green, & Holmes 1994). El fin de tomar una sola muestra sería cuantificar el grado de atipicidad en relación con todas las otras muestras que podrían haber sido extraídas (Saldahna & Thompson, 2002).

Introducción intuitiva de ideas de inferencia

Algunos autores sugieren una alternativa informal a la enseñanza de la inferencia estadística (por ejemplo, Rubin, Hammerman, & Konold, 2006; Rossman, 2008) que consistiría en la introducción menos formalizada de un conjunto de ideas fundamentales que sustentan la comprensión de la inferencia estadística formal. Rossman (2008) sugiere una introducción informal siguiendo los pasos siguientes: (a) Partir de una hipótesis dada acerca de los datos (b) Uso de la simulación o de cálculos de probabilidad elemental para concluir que los datos observados son muy poco probables si el modelo fuera cierto (cálculo intuitivo de un p-valor), y (c) Rechazar la hipótesis inicial (modelo) basado en el p-valor muy pequeño, en lugar de creer que un suceso muy raro ha ocurrido por casualidad. Este proceso de razonamiento, es muy natural para los estudiantes, y de hecho sigue la concepción de Fisher de pruebas de significación. Podemos analizarlo a partir del siguiente ítem, usado en las investigaciones de Green (1991) con chicos de 11 a 16 años, que es similar a los utilizados con adultos en las pruebas sobre percepción de la aleatoriedad:

> **Ítem 1**. Se pidió a dos niñas lanzar una moneda equilibrada 150 veces y anotaran los resultados. Estos son los datos que presentaron al profesor ¿Hicieron trampas Clara o Luisa? ¿Por qué?
>
> Clara: c+c++cc++cc+c+c++c++c+ccc+++ ccc++c++c+c+c++cc+ccc+c+c+cc+++c c++c+c++cc+c++cc+c++cc+cc+c+++c ++cc++c++c+c+cc+c++cc+c+c++cc c+cc++c+c++cc+++c+++c+c++ccc++
>
> Luisa: +cc+++c++++c+cc+++cc+cc+++cc+ccc+++c++++++c+c+c+c++++ccccccc+ ccc+c+cc+ccccc+ccc++ccc+c+cccc ccccc++c+cccccccc+++++cccc++c+ c+cc+cc+cc++++++c+cc++ccc++ccc

Una de las estrategias que pueden seguir los estudiantes para resolver el problema anterior, es contar el número de caras de cada una de las secuencias y comparar con el número esperado (75 caras si consideramos que la moneda está bien equilibrada). Este sería el modelo o hipótesis nula y habría que comprobar si los datos observados (las secuencias de Clara y Luisa) se acercan a los patrones esperados bajo esta hipótesis. Nosotros hemos realizado este recuento, presentando los resultados en la Tabla 1. Observamos que ningunas de las dos secuencias tiene una frecuencia de caras y cruces exactamente igual a la esperada (teórica), pero, de todos modos, si se hubiese obtenido exactamente la frecuencia teórica, nos hubiese resultado sospechoso. Esperamos que la frecuencia observada (de Clara y Luisa) se asemeje a la teórica, pero no demasiado. Intuitivamente parece que Luisa se aparta demasiado y, por tanto, hizo trampas.

Tabla 1.

Resultados teóricos y observados en el ítem 1

	Cara	Cruz
Clara	72	78
Luisa	67	83
Teórica	75	75

Continuando el análisis, analizamos la frecuencia de resultados posibles cuando contamos los resultados de dos en dos o tres en tres (Tablas 2 y 3). Ahora se observa más claramente la mayor diferencia de Clara respecto a la distribución teórica (por ejemplo, nunca obtiene tres caras seguidas, cuando teóricamente se esperarían 6 rachas de este tipo). También podemos ver que Clara se aparta más que Luisa de lo esperado al representar gráficamente el número de caras obtenidas en cada dos o tres lanzamientos en las tres secuencias (Figura 1). Por tanto, ya que los datos de Clara son muy improbables, en caso de ser cierta nuestra hipótesis nula, los rechazamos y decidimos que es ella la que ha hecho trampas.

Tabla 2.

Frecuencias de posibles sucesos al contar los resultados 2 a 2

	CC	C+	+C	++
Clara	12	30	18	15
Luisa	25	21	12	17
Teórica	19	19	19	19

Tabla 3.

Frecuencias de posibles sucesos al contar los resultados 3 a 3

	CCC	CC+	C+C	C+C	+CC	+C+	++C	+++
Clara	0	2	13	9	6	7	10	1
Luisa	8	11	6	3	6	4	5	6
Teórica	6	6	6	6	6	6	6	6

Este estudio, que hemos hecho en forma elemental, estaría al alcance de los alumnos de secundaria y también podría hacerse a nivel más formal, en la universidad utilizando el contraste chi-cuadrado. En cualquiera de los dos casos serviría para introducir las ideas básicas que subyacen en el modelo de test de significación de Fisher.

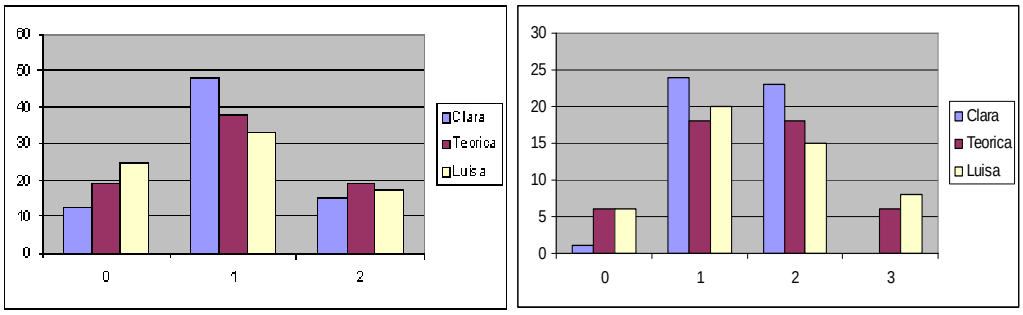

Figura 1. Número de caras en dos y en tres lanzamientos en las secuencias

La distribución muestral y el Teorema Central del Limite

Otro punto en el aprendizaje inferencia estadística es comprender la variación de un

estadístico dado (por ejemplo, la media) en diferentes muestras de la misma población (distribución de muestreo). Al pensar en la inferencia estadística se debe ser capaz de diferenciar claramente entre tres distribuciones:

- La distribución de probabilidad que describe los valores de una variable aleatoria de la población (por ejemplo, el peso al nacer de un recién nacido). Esta distribución por lo general depende de algunos parámetros. Por ejemplo, la distribución normal se especifica mediante μ y σ, la media y la desviación estándar poblacional.
- La distribución de los datos de los valores de una variable estadística de una sola muestra tomada al azar de la población (el peso en una muestra de 100 recién nacidos). De esta muestra se pueden utilizar las estadísticas $\bar{x}$ y s, la media muestral y la desviación estándar para estimar los valores de los parámetros de la población.
- La distribución de probabilidad que describe los valores de un estadístico en todas las muestras tomadas al azar de la población (el peso medio en todas las posibles muestras de 100 recién nacidos tomadas al azar de la población). El teorema central del límite asegura que los parámetros de esta distribución se relacionan con la de la muestra. Por ejemplo, el valor esperado de la distribución muestral de los medios es $E(\bar{x}) = \mu$, y la desviación estándar $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. Este hecho se utiliza para la prueba de hipótesis o construir intervalos de confianza sobre el valor del parámetro.

La comprensión de las distribuciones muestrales requiere que los estudiantes integren los diferentes niveles descritos, pues cada una de estas distribuciones requiere conocimientos y técnicas específicas. Cuando se piensa en la distribución de la población y la distribución de la muestra aleatoria simple, la unidad de análisis (caso) es un objeto individual; sin embargo, al estudiar la distribución de muestreo el caso es una muestra aleatoria simple (Batanero et al, 1994). Sin embargo la mayoría de los estudiantes tienen dificultades, confundiendo los diversos niveles (Saldahna & Thompson, 2002).

Por otro lado, es importante que los estudiantes comprendan intuitivamente el teorema del central límite (TCL), que es fundamental en la construcción de distribuciones de muestreo aproximadas para la media y otros parámetros. La simulación con ordenador proporciona una herramienta interesante para observar esta convergencia. Chance, Delmas y Garfield (2004) informaron sobre una serie de estudios que investigan el impacto de la interacción de los estudiantes con el software, específicamente diseñado para enseñar a las distribuciones de muestreo, en su comprensión del tema. En los primeros estudios, y a pesar de las capacidades del software, los estudiantes tendieron a buscar las reglas, pero no comprendieron las relaciones subyacentes (el teorema del límite central), que causaron los patrones visibles en el muestreo. En estudios posteriores, los autores pidieron a los estudiantes a hacer predicciones acerca de las distribuciones muestrales de las medias y luego enfrentarse a sus conjeturas, observando los resultados empíricos de la simulación. Esta estrategia demostró ser útil para mejorar el razonamiento de los alumnos acerca de las distribuciones muestrales.

Inferencia Bayesiana elemental

Como alternativa a la inferencia clásica se podría realizar una introducción intuitiva a los métodos bayesianos que, según algunos autores (por ejemplo, Lecoutre & Lecoutre, 2001) son

más intuitivos que la inferencia frecuencial para los estudiantes. El teorema de Bayes, permite transformar las probabilidades a priori (antes de realizar un experimento), una vez se observan sus consecuencias, en probabilidades a posteriori, que incorporan la información de los datos observados. Consideremos el siguiente ejemplo de diagnóstico médico:

Item 2. La probabilidad de que una mujer americana de entre 40 y 50 años sin síntomas, tenga cáncer de pecho es 0,8 %. Si una mujer americana tiene cáncer de pecho tendrá una mamografía positiva con probabilidad 90%. También el 7% de mujeres sanas dan positivo en la mamografía. Supongamos que una mamografía da positiva, ¿Cuál es la probabilidad de que la mujer en realidad tenga cáncer de pecho? (Eddy, 1982).

Si en el ejemplo llamamos *"C"* al suceso "tener cáncer" y *"+"* al suceso "obtener un diagnóstico positivo", hemos de diferenciar la *probabilidad a priori* de tener cáncer antes de hacerse la prueba *P(C)*=0.008 y la *probabilidad a posteriori P(C/+)* o probabilidad condicional de tener cáncer una vez que la prueba fue positiva. *P(+/C)*=0,8 es la *verosimilitud* de tener una prueba positiva si se tiene cáncer. Calculemos la probabilidad pedida, ayudándonos de un diagrama en árbol y pensando en términos de frecuencias absolutas, para lo cual consideraremos un grupo de 100000 mujeres de las características dada (Figura 2). Con la proporción supuesta de cáncer en la población, aproximadamente 800 de estas mujeres estarían enfermas y de ellas 720 serían detectadas en la mamografía (90%). El 7 % de ellas recibirían un resultado positivo, aunque estén sanas (falso positivo), lo que supone un total de 6944 mujeres. En total tenemos 7664 mamografías positivas en las 100000 mujeres, aproximadamente, la mayor parte de las cuales son, en realidad de personas sanas. Aplicando la regla de Laplace, obtenemos que la probabilidad de que la mujer que reciba el resultado positivo realmente tenga un cáncer es el cociente 720/7664 que da un valor de 0,0939; es decir, solo el 9% de las mujeres que obtienen una mamografía positiva en este grupo de edad realmente están enfermas.

Para el caso general donde A_i representa un conjunto de posibles sucesos que pueden dar lugar a unos datos *D*, y queremos calcular las probabilidades finales de los sucesos $P(A_i/D)$, conocidas sus probabilidades iniciales $P(A_i)$ y verosimilitudes de obtener estos datos según vengan causados por los diferentes sucesos $P(D/A_i)$, el teorema se puede expresar con la formula (1):

$$(1) \qquad P(Ai/D)=\frac{P(Ai)\,x\,P(D/Ai)}{P(A_1)\times P(D/A_1)+P(A_2)\times P(D/A_2)+...+P(A_n)\times P(D/A_n)}$$

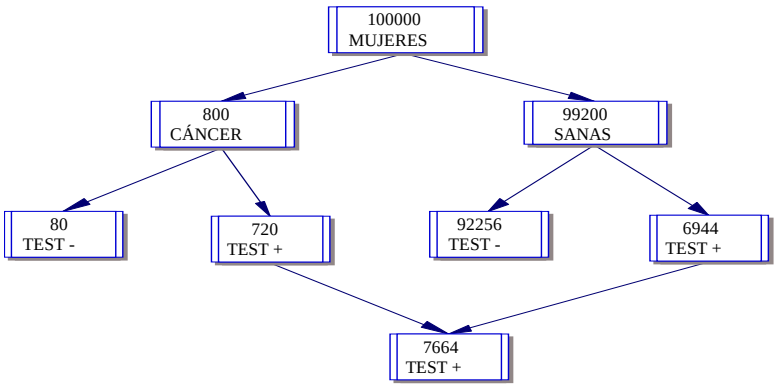

Figura 2. Partición de la población supuesta de 100000 mujeres

Como se muestra en el ejemplo, las probabilidades de los sucesos de interés (estar o no enfermo) pueden revisarse (pasar de probabilidades a priori a probabilidades a posteriori) y pierden de este modo el carácter objetivo que tenían en las concepciones clásica y frecuencial.

En inferencia clásica un parámetro θ (por ejemplo el peso medio de un recién nacido en una cierta población) se considera constante, pero desconocido y el objetivo de la inferencia es encontrar una estimación o aproximación de su verdadero valor, a partir de los datos. En inferencia bayesiana, por el contrario un parámetro θ es una variable aleatoria con una distribución a priori de probabilidades *p(θ)*, que indica el conocimiento sobre θ antes de tomar los datos. En el ejemplo, se supondría que el peso medio del recién nacido varía (en el tiempo, o geográficamente, etc.), pero se tiene una distribución inicial de probabilidades para dicho valor medio. El objetivo de la inferencia bayesiana sería usar los datos de una muestra para actualizar esta distribución a priori y mejorar nuestro conocimiento del peso medio del recién nacido. Esta actualización se realiza mediante el teorema de Bayes. Si representamos por $y = (y_1,..., y_n)$ el conjunto de datos, cuya *función de verosimilitud p(y/θ)* depende del parámetro, entonces la distribución a posteriori de θ dados los datos observados *y* viene dada por (2) (Lee, 2004)

$$(2) \quad p(\theta / y) = \frac{p(y / \theta)\, p(\theta)}{\int p(y / \theta)\, p(\theta)}$$

La aplicación sistemática del teorema de Bayes constituye el método principal de la inferencia bayesiana, cuyo objetivo básico es obtener la distribución a posteriori de los parámetros (O'Hagan y Forster, 2004). La mejor estimación del parámetro será su valor esperado (promedio) en la distribución a posteriori, que minimiza el error cuadrático esperado. Dicha distribución también permitirá calcular probabilidades de que los parámetros se encuentren en intervalos de valores dados (intervalos de credibilidad), así como calcular probabilidades de que ciertas hipótesis sean verdaderas o falsas. El teorema de Bayes podría aplicarse sucesivamente en nuevos experimentos, tomando como probabilidades iniciales del segundo experimento las probabilidades finales obtenidas en el primero y así sucesivamente.

Hay, hoy día, un creciente número de publicaciones acerca de cómo introducir los conceptos bayesiano a los estudiantes que provienen de contextos no científicos (por ejemplo, Albert y Rossman, 2001; Díaz, 2005). Sin embargo, los resultados reportados de los experimentos o investigaciones que se centran en la enseñanza de la estadística bayesiana son aún muy limitados. Por otra parte, algunas de las experiencias comunicadas indican que los alumnos pueden cometer errores en la interpretación de sus resultados (Albert, 2000; Díaz, De la Fuente y Batanero; 2008) por lo que sería necesario realizar una mayor investigación sobre este tema,

Implicaciones para la enseñanza y la investigación

El análisis realizado indica que no es suficiente enseñar a los estudiantes sobre las reglas y conceptos con el fin de llegar a la comprensión integral de la inferencia. A pesar de nuestros esfuerzos, las concepciones erróneas permanecen después de la instrucción formal en estadística. Debiéramos preguntarnos por qué la enseñanza actual de la estadística no mejora las intuiciones y qué tendríamos que cambiar para remediar la situación. Quizás "la estadística debiera enseñarse a la vez que se muestran materiales sobre estrategias intuitivas y errores de inferencia"..." esto tendría la ventaja de aclarar los principios subyacentes de la estadística y probabilidad y facilitar que se aprecie su aplicación a situaciones concretas" (Nisbett & Ross,

1980, p.281).

Numerosos applets interactivos proporcionan hoy un entorno dinámico y visual en el que los estudiantes pueden participar en el muestreo al azar y la construcción de las distribuciones muestrales. La disponibilidad actual de software y tecnología hace que sea posible dedicar el tiempo que previamente se invertía en cálculos laboriosos para propiciar una aproximación menos formal y más intuitiva a la estadística. "La capacidad estadística que se requiere no es la tradicional", "Debemos preguntarnos si la enseñanza tradicional de los estudiantes es demasiado restringida" Moore (1997, p. 124).

Dada la dificultad de integrar los conceptos involucrados en la inferencia estadística, tiene sentido sugerir que estas ideas deben ser desarrollados progresivamente en la mente de los alumnos, siguiendo los pasos sugeridos en este trabajo. Las nuevas directrices curriculares donde la educación estadística se introduce desde la escuela primaria proporcionan una oportunidad y un desafío para ayudar a los estudiantes a desarrollar su conocimiento y razonamiento estadístico. Debemos también reflexionar sobre la dosis exacta de formalización que se requieren para enseñar los conceptos estadísticos. En este sentido, la estadística puede ser paradigmática respecto a encontrar nuevas maneras de enseñar conceptos matemáticos avanzados a gran número de estudiantes e incluso para repensar el significado del pensamiento matemático avanzado (Artigue, Batanero, & Kent, 2007).

Agradecimientos: Se agradece el apoyo económico al proyecto EDU2010-14947 (MCIN) y grupo FQM126 (Junta de Andalucía).

Referencias

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.

Albert, J. (2000). Using a sample survey project to assess the teaching of statistical inference, *Journal of Statistical Education, 8*. On line: www.amstat.org/publications/jse/.

Albert, J. H., & Rossman, A. (2001). *Workshop statistics. Discovery with data. A bayesian approach.* Bowling Green. OH: Key College Publishing.

Artigue, M., Batanero, C., & Kent, P. (2007). Mathematics thinking and learning at post-secondary level. En F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1011-1049). Greenwich, CT: Information Age Publishing, Inc., & National Council of Teachers of Mathematics.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning,* 2(1-2), 75-98.

Batanero, C. y Díaz, C. (2006). Methodological and didactical controversies around statistical inference. *Actes du 36iémes Journées de la Societé Française de Statistique.* CD ROM. Paris: Societé Française de Statistique.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology*, *25*(4), 527–547.

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

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.

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

Castro-Sotos, A. E., Vanhoof, S., Noortgate, W. & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2* 98–113

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning About Sampling Distributions. In D.Ben-Zvi and J. Garfield (eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). The Netherlands: Kluwer.

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

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

Díaz, C. (2007). Introducción a la Inferencia Bayesiana. Granada: La autora.

Díaz, C., de la Fuente, I., & Batanero, C. (2008). Implications between learning outcomes in elementary Bayesian inference. In R. Gras (Ed.), *Statistical implicative analysis: theory and applications* (pp. 163-183). Springer. Studies in Computational Intelligence 127.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. En D. Kahneman, P. Slovic y Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

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

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

Fidler, F. & Cumming, G. (2005, August). Teaching confidence intervals: problems and potential solutions. Trabajo presentado en la *International Statistical Institute, 55th Session*. Lisbon.

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

Green, D. R. (1991). A longitudinal study of children's probability concepts. En D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 320-328). Voorburg: International Statistical Institute.

Hacking, I. (1990). *The taming of chance*. Cambridge, MA: Cambridge University Press.

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

Harradine, A., Batanero, C., & Rossman, A. (En prensa). Students and teachers' knowledge of sampling and inference. En C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education.* Springer.

Jones, G. A. y Thornton, C. A. (2005). An overview of research into the teaching and learning of probability. En G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 65-92).. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Krauss, S., & Wassner, K. (2002). How significance tests should be presented to avoid the typical misinterpretations. En B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics.* Cape Town, South Africa: International Association for Statistics Education. Online: www.stat.auckland.ac.nz/~iase/publications.

Lecoutre, B., & Lecoutre, M. P. (2001). Uses, abuses and misuses of significance tests in the scientific commumity: Won't the Bayesian choice be unavoidable? *International Statistical Review*, 69(3)-399-417.

Lee, P. M. (2004). *Bayesian statistics. An introduction.* York, UK: Arnold.

Liu, Y., & Thompson, P. W. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies, 4* (2), 126-138.

MEC (2007). Real Decreto 1467/2007, de 2 de noviembre, por el que se establece la estructura del bachillerato y se fijan sus enseñanzas mínimas (Royal Decreet establihing the estructure and content of the high school curriculum).

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 635,* 123-165.

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

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: NCTM. Online: standards.nctm.org/.

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

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19. Online: http://www.stat.auckland.ac.nz/serj.

O'Hagan, A. y Forster, J. (2004). *Bayesian inference.* Vol. 2B en Kendall's Advanced Theory of Statistics. London: Arnold.

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

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

Rubin, A., Hammerman, J. K. L & Konold, C. (2006). Exploring informal inference with interactive visualization software. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics.* Cape Town, South Africa: International Association for Statistics Education. Online: www.stat.auckland.ac.nz/~iase/publications.

Saldanha. L., & Thompson, P. (2002) Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Sedlmeier, P. (1999). *Improving statistical reasoning. Theoretical models and practical implications.* Mahwah, NJ: Erlbaum.

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

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