Aproximación informal al contraste de hipótesis

Carmen Batanero[1] y Carmen Díaz[2]

[1]batanero@ugr.es, Universidad de Granada
[2]carmen.diaz@dpsi.uhu.es, Universidad de Huelva

Resumen

La inferencia estadística es una herramienta esencial en muchas ramas de la actividad humana, por lo que su enseñanza es generalizada en la universidad y formación profesional, en incluso en alguna modalidad de Bachillerato. A la vez encontramos una extensa bibliografía que critica su uso inadecuado. Ello ha originado una línea de investigación y desarrollo de lo que se conoce como *inferencia informal* que está cobrando un gran auge. En este trabajo se resumen las diferentes aproximaciones actuales al contraste de hipótesis y algunas dificultades frecuentes de comprensión de la inferencia. Finalmente se muestra un ejemplo de aproximación informal a la enseñanza del contraste, siguiendo la metodología de Fisher y se analizan las condiciones para que una aproximación informal pueda considerarse como inferencia. Estas aproximaciones informales, bien planteadas pueden contribuir a la educación del razonamiento del estudiante, antes de iniciar el estudio formal.

Palabras clave: Inferencia estadística, errores de comprensión, aproximaciones informales, inferencia informal.

1. Introducción

La inferencia estadística se incluye en la actualidad la mayoría de las carreras universitarias, y estudios de postgrado. También se introduce en Bachillerato en la especialidad en Ciencias Sociales del Bachillerato (MEC, 2007), donde encontramos la asignatura Matemática II Aplicadas a las Ciencias Sociales, con 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 (p. 45476).

La importancia que se da a este tema se refleja en el hecho de que en las pruebas de acceso a la Universidad de Matemáticas Aplicadas a las Ciencias Sociales se ha venido proponiendo con relativa frecuencia un problema sobre contraste de hipótesis. Un ejemplo típico es el siguiente, propuesto en las pruebas de acceso en Andalucía en Junio de 2013 (prueba de reserva a), calificado con 2.5 puntos:

> Problema 1: Un director sanitario sostiene que el Índice de Masa Corporal (IMC) medio de los adolescentes de su distrito no supera el nivel 25 (sobrepeso). Para contrastar su afirmación toma una muestra aleatoria de 225 adolescentes que da como resultado un IMC medio de 26. Sabiendo que el IMC sigue una distribución Normal con desviación típica 5 discuta, mediante un contraste de

En J. M. Contreras, C. Batanero, J. D. Godino, G.R. Cañadas, P. Arteaga, E. Molina, M.M. Gea y M.M. López (Eds.), Didáctica de la Estadística, Probabilidad y Combinatoria, 2 (pp. 135-144). Granada, 2015.

hipótesis con $H_0 \equiv \mu \leq 25$, si la afirmación del director sanitario es correcta, con un nivel de significación del 5%.

Para resolver el problema, el alumno debe recordar que, para realizar un contraste de hipótesis sobre la media de la población μ, ha de utilizar la media de la muestra; en este caso $\bar{x} = 26$. También que, cuando la variable sigue una distribución normal $N(\mu,\sigma)$ la media muestral tiene una distribución normal $N(\mu,\frac{\sigma}{\sqrt{n}})$, donde *n*=225 es el tamaño de la muestra. Por tanto, la distribución de la media muestral tiene como desviación típica $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{225}} = \frac{5}{15} = 1.33$. Tipificando la variable media muestral obtenemos una distribución normal tipificada N (0,1): $\frac{\bar{x} - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{\bar{x} - 25}{1/3} = (26 - 25) \times 3 = 3$

En dicha distribución normal tipificada la probabilidad de obtener un valor 3 o superior sería $P(Z \geq 3) = .0013$. Lo razonable sería rechazar la afirmación del director sanitario, pues si el índice medio de masa corporal en su distrito fuese igual o menor a 25; sólo 13 de cada 10000 muestras de 225 adolescentes darían un índice medio igual o mayor que 26. Como vemos, este problema requiere el uso de muchos conceptos y procedimientos: muestra y población, media muestral y poblacional, distribución de la media muestral, probabilidad condicional, hipótesis nula y alternativa, región de aceptación y rechazo, distribución normal, tipificación, uso de las tablas de la distribución normal tipificada.

En realidad, la resolución correcta de este problema no implica que el estudiante comprenda y discrimine todos estos conceptos ni que haya adquirido suficiente razonamiento estadístico, sino que recuerda y sabe aplicar una serie de fórmulas, que quizás no comprenda. Esta enseñanza soslaya también la problemática filosófica asociada y los errores de aplicación de la inferencia. Además, presenta la inferencia frecuencial como una metodología única, ocultando las diferentes aproximaciones y las controversias que dentro de la misma estadística ha tenido la inferencia (Batanero, 2000). En lo que sigue analizamos las aproximaciones más importantes a la inferencia, los errores más comunes en su interpretación y un ejemplo en que aplicamos el enfoque didáctico denominado *inferencia informal* (ver, por ejemplo, Rubin, Hammenran y Konold, 2006; Zieffler, Garfield, delMas y Reading, 2008; Rossman, 2008) analizando las condiciones para que dicho enfoque constituya en si mismo una verdadera inferencia estadística.

2. Inferencia estadística: diferentes aproximaciones

La inferencia estadística se creó como respuesta al problema consistente en obtener un conocimiento general (una nueva teoría científica) a partir del análisis de casos particulares (inducción empírica). Esto es, nace con la búsqueda de métodos que permitan justificar el razonamiento inductivo y la extensión de sus conclusiones, problema de gran importancia en las ciencias empíricas. Este problema ha ocupado a los filósofos y estadísticos por largo tiempo, sin que se haya obtenido una solución aceptada por consenso (Cabria, 1994; Rivadulla, 1991).

En el problema descrito el director sanitario quiere probar que el índice de masa corporal de los adolescentes en su distrito es adecuado, es decir, no supera el valor fijado de 25. Para comprobar si esta hipótesis es cierta, obtiene una muestra aleatoria de adolescentes en su distrito, los pesa y calcula el IMC medio. Como estos adolescentes son una muestra de todos los posibles en el distrito, estamos en un caso de razonamiento inductivo, porque queremos generalizar lo observado en casos particulares (la muestra) a algo más general (la población).

¿Cómo podríamos validar una generalización inductiva a partir de datos empíricos?

Popper (1967) propuso aceptar una teoría como provisionalmente cierta frente 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 teoría seria provisionalmente confirmada si los datos recogidos siguiesen estos patrones. En el ejemplo dado, si al medir el IMC en muestras sucesivas de adolescentes, siempre obtenemos un IMC medio dentro del rango esperado (menor que 25), se acepta la afirmación del director sanitario. Pero es importante ver que esta aceptación es sólo provisional. La confirmación de una teoría a partir de datos empíricos nunca es definitiva, porque los datos futuros podrían contradecirla. Así, basta que en una muestra el IMC sea muy alto para que no pueda sostenerse la afirmación del director sanitario. En cambio, si los datos del experimento se apartasen del patrón esperado, la teoría sería refutada, por lo que el rechazo de la hipótesis tiene mayor fuerza que su confirmación (basta que en una única muestra, como ha pasado al resolver el problema, los datos contradigan al director sanitario).

Estas ideas de Popper tuvieron una gran influencia en el desarrollo de la inferencia estadística. Ya que mediante un razonamiento inductivo no es posible llegar a la certidumbre de una proposición (verdad cierta), diversos autores intentaron calcular la probabilidad de que una hipótesis sea cierta (verdad probable) (Ridadulla, 1991; Batanero, 2000).

2.1. Probabilidad de una hipótesis e inferencia bayesiana

Es importante resaltar que la probabilidad de una hipótesis no tiene sentido en inferencia clásica frecuencial, donde la probabilidad se interpreta como límite de la frecuencia relativa en un gran número de repeticiones independientes de un experimento. La razón es que una hipótesis será cierta o falsa siempre; no puede ser cierta un porcentaje de veces en una serie de pruebas. En el ejemplo dado o es cierto o es falso que el IMC medio en la población no supera el valor 25.

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 que oscila entre 0 (falsedad absoluta) y 1 (certeza absoluta) (Gingerenzer, 1993; Lecoutre & Lecoutre, 2001). Es decir, si pedimos al director sanitario que nos exprese en una escala 0-1 su grado de creencia en que el IMC de los adolescentes no supera el valor 25, el director podría decirnos que su grado de creencia es .9. Podemos interpretar esta probabilidad subjetiva como una apuesta: el director sanitario está dispuesto a apostar 9 frente a 1 su afirmación. Para él hay 9 posibilidades de que el índice sea menor o igual a 25 a que supere este valor. De hecho, en la inferencia bayesiana podemos considerar dos probabilidades diferentes para una hipótesis:

- *Probabilidad inicial,* grado de creencia inicial en la hipótesis antes de recoger datos de experimentos donde se trate de poner la hipótesis a prueba (grado de creencia en la hipótesis antes de hacer el experimento; supongamos sigue siendo .9).
- *Probabilidad final,* grado de creencia en la hipótesis una vez se han recogido los datos. El teorema de Bayes servirá para combinar la probabilidad inicial con los datos y llegar a la probabilidad final de la hipótesis. Si el experimento tiene éxito, puede que cambie mi grado de creencia a .95. Si, como en el caso del problema, fracasa, el grado de creencia disminuye. No explicaremos como se llega a la probabilidad final, pero el lector interesado puede consultar el material de Díaz (2005).

Como la probabilidad subjetiva expresa un grado de creencia personal, otras personas, según su conocimiento podrían asignar inicialmente diferente probabilidad subjetiva a una hipótesis. En el ejemplo dado, se podría dar una probabilidad inicial .5 si no se tienen motivos para

aceptar o rechazar la afirmación del director sanitario. El teorema de Bayes corrige poco a poco las diferentes asignaciones iniciales cuando aumentamos el tamaño de la muestra, de forma que se tiende a una probabilidad final no muy diferente, si la muestra es grande, aunque las probabilidades iniciales varíen.

2.2. El test de significación: refutación empírica de una hipótesis

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 sus objetivos.

El *test de significación* fue propuesto por Fisher en su libro "The design of experiments", publicado en 1935, es un procedimiento que permite rechazar una hipótesis, con un cierto *nivel de significación.* 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_0 (hipótesis nula) es cierta. Se suele tomar como hipótesis nula o de no efecto la contraria de la que se pretende probar. (En nuestro ejemplo, la hipótesis nula sería que el IMC medio no supera el valor 25). Generalmente la hipótesis se refiere a al valor supuesto de un parámetro, pero no se tiene acceso a toda la población, sino sólo a una muestra de la misma. En el ejemplo, estamos interesados en el IMC medio μ en toda la población de adolescentes, pero solo podemos tomar una muestra; calcularemos la media $\bar{x}$ en la muestra y lo compararemos con el supuesto en la población.

Para poner la hipótesis a prueba se organiza un experimento aleatorio asociado a H_0 y se considera un cierto suceso *S* que puede darse o no en este experimento (obtener un valor de la media en la muestra muy improbable de ser cierta la hipótesis). El experimento en el ejemplo sería recoger la muestra aleatoria de 225 adolescentes y calcular la media de su IMC.

Se sabe que si H_0 fuese cierta (si el IMC medio en la población no supera el valor 25), hay muy poca probabilidad de que ocurra *S* (obtener un valor medio en la muestra muy alejado de 25). Realizado el experimento ocurre precisamente *S,* pues al tipificar el valor $\bar{x} = 26$ se ve que es muy improbable. Hay dos posibles conclusiones:

- Bien la hipótesis H_0 era cierta y ha ocurrido *S*, a pesar de su baja probabilidad (pues un suceso improbable no es imposible).
- Bien la hipótesis *Ho* era falsa.

Al igual que en el ejemplo, 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. Al comparar la media en la muestra en nuestro ejemplo con la media supuesta de medias en la población podemos medir la discrepancia entre datos e hipótesis.

El razonamiento que apoya un test de significación parte de la suposición de que la hipótesis nula es cierta. En dicho caso, el estadístico define una distribución muestral, al variar los datos aleatoriamente (Cabriá, 1994; Batanero, 2000). En nuestro caso, al tipificar la media muestral podemos usar la distribución normal N (0,1) para calcular la probabilidad de obtener un valor igual o mayor que 3 (valor obtenido en la tipificación). Trabajamos con $P(Z \geq 3)$ en vez de con $P(Z = 3)$ porque este último valor es siempre igual a cero y para tener en cuenta también los casos más extremos que el observado.

En resumen, un test de significación efectúa una división entre los posibles valores del

estadístico en dos clases: resultados estadísticamente significativos (para los cuales se rechaza la hipótesis) y no estadísticamente significativos para los cuáles no se puede rechazar la hipótesis (Ridavulla, 1991). En el enfoque de Fisher el interés es rechazar la hipótesis nula; pero no se identifica una hipótesis alternativa concreta (Batanero y 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 (nivel de significación) lo fija el investigador según su juicio subjetivo y su experiencia. Suele ser común tomar un nivel de significación α=.05.

2.3. El contrastes de hipótesis de Neyman y Pearson: regla de decisión entre dos hipótesis

Neyman y Pearson por su parte estaban interesados en 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). Este enfoque tiene más sentido cuando se trata de una prueba que se repite muchas veces en las mismas condiciones. Supongamos, por ejemplo, un contexto de control de calidad: Estamos en un proceso de llenado de paquetes de azúcar de 1 kg. Suponemos una distribución normal N (0,1). Estamos interesados en diferenciar entre dos hipótesis:

H_0: El proceso está controlado. Los paquetes tienen un peso medio de 1kg

H_1; Se ha descontrolado el proceso. Los paquetes tienen un peso medio mayor o menor de 1 kg.

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. Las dos situaciones implican un coste: Si el proceso está descontrolado y supongo que es correcto, puedo estar vendiendo más o menos peso del que cobro; aunque vender un menor peso podría parecer no grave, puedo perder mi calidad o mi imagen. Si el proceso funciona bien, entonces si supongo que se ha descontrolado y paro la producción para arreglar la maquinaria, tengo un coste innecesario. Todos los días, para control de calidad tomo una muestra de 30 paquetes y los peso; cada día repito el contraste de hipótesis para ver si paro o no la producción. Solo peso los 30 paquetes (uso una muestra), pues sería muy caro pesarlos todos.

Al tomar una de estas decisiones sobre las hipótesis (decidir si el proceso está controlado o no) a partir de los resultados del contraste (del peso medio de los paquetes en la muestra) se consideran dos tipos de error:

- *Error tipo I:* Rechazar una hipótesis nula cuando es cierta (Parar el proceso de producción, cuando el proceso está controlado). Se suele establecer un nivel de significación α que asegura que la probabilidad de cometer este tipo de error sea menor que un número preestablecido. Generalmente se trabaja con α=.05.
- *Error tipo II*: Aceptar una hipótesis nula que de hecho es falsa (considerar que el proceso está controlado, cuando no lo está). Beta (β) es la probabilidad de cometer este tipo de error y su complemento ($1 - \beta$) la *potencia* del contraste. Mientras que α es un número preestablecido, β es variable, porque su valor depende del valor del parámetro (generalmente desconocido). En el ejemplo, dependerá del verdadero valor del peso medio del paquete. Por ejemplo, si en vez de envasar 1 kg. el proceso fabrica paquetes de 999 gramos, la probabilidad de error tipo II es alta, porque hay poca diferencia entre el peso medio real del paquete y el supuesto. La media de la muestra será muy cercana a 1kg. Pero si estamos fabricando paquetes de 1200 gramos, el peso medio de 30 paquetes será muy cercano a 1200 gramos y será difícil suponer el proceso controlado (hacer errores tipo II).

Una vez definidas las hipótesis nula y alternativa y fijada la probabilidad de cometer error

tipo I, se toma una muestra (30 paquetes). Calculado el estadístico (peso medio en la muestra), se toma la decisión de rechazar o no rechazar la hipótesis nula, comparando la probabilidad de obtener un valor igual o más extremo que el valor del estadístico (valor p) con el nivel de significación.

3. Errores frecuentes en la interpretación del contraste de hipótesis

Son muchos los errores e interpretaciones incorrectas del contraste de hipótesis, incluso en los trabajos de investigación, situación que ha sido observada desde hace tiempo (Falk & Greenbaum, 1995; Batanero, 2000; Harradine, Batanero y Rossman, 2011).

Un concepto que se suele comprender erróneamente es el nivel de significación α. Como se ha indicado, es la probabilidad de rechazar la hipótesis nula, supuesta cierta. En forma simbólica: $\alpha = P(\text{Rechazar } H_0 \mid H_0 \text{ es cierta})$. Supongamos que trabajamos con un valor α=.05 o del 5 %. Esto quiere decir que si H_0 es cierta, la rechazamos 5 de cada 100 veces, o lo que es lo mismo en 100 días que hagamos el control de calidad y en los cuáles el proceso esté controlado, pararemos el proceso innecesariamente 5 días.

La interpretación errónea más común del nivel de significación consiste en cambiar los dos términos de la probabilidad condicionada en la expresión anterior, es decir, interpretar α como la probabilidad de que la hipótesis nula sea cierta, una vez que la decisión de rechazarla se ha tomado, esto es, suponer que $\alpha = P(H_0 \text{es cierta} \mid \text{se ha rehazado } H_0)$

Por ejemplo, Birnbaum (1982) pidió a sus estudiantes que definiesen el nivel de significación. La respuesta más frecuente fue: "*Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados*". En la investigación de Falk (1986) la mayoría de sus estudiantes creían que α era la probabilidad de equivocarse al rechazar la hipótesis nula. Vallecillos (1994) también encontró este error en una investigación con más de 400 estudiantes en la Universidad de Granada; y el error se repetía en diferentes tipos de estudiantes (por ejemplo, de medicina, ingeniería o psicología). Resultados similares fueron encontrados por Krauss y Wassner (2002) en profesores de universidad responsables de la enseñanza de métodos de investigación.

En los contrastes de hipótesis también se confunden las funciones las hipótesis nula y alternativa. Es decir, algunos estudiantes piensan que la hipótesis nula es la que queremos demostrar (no la que queremos rechazar). Posiblemente se deba esta creencia a que en matemáticas casi siempre se trata de probar una hipótesis (aunque en el método de reducción al absurdo se trata de rechazarla).

Vallecillos (1999) describió cuatro creencias distintas en sus estudiantes sobre el tipo de prueba de que proporciona el contraste de hipótesis:

a. El contraste de hipótesis es una regla que te ayuda en la toma de decisiones; esta creencia sería correcta y refleja el método de Neyman y Pearson, donde queremos decidir entre dos hipótesis.

b. El contraste procedimiento para la obtención de apoyo empírico a la hipótesis de investigación. Esta visión también es correcta y refleja la propuesta de Fisher, donde sólo se repite el contraste una vez (se hace un experimento).

c. El contraste de hipótesis es una prueba probabilística de la hipótesis. Permite calcular la probabilidad de que una hipótesis sea cierta. Esta creencia sería cierta sólo cuando aplicamos inferencia bayesiana y además se trataría de una probabilidad subjetiva

(personal) o grado de creencia. Cuando trabajamos el método de Fisher o de Neyman y Pearson esta creencia es falsa pues, como hemos dicho, la probabilidad de una hipótesis no tiene sentido en inferencia frecuencial.

d. El contraste estadístico es un método matemático; como tal, y al ser la matemática una ciencia exacta, al finalizar hemos probado la verdad o falsedad de una hipótesis. Esta creencia es siempre errónea; la tienen algunos alumnos que tienen poca base matemática y una fe ciega en las matemáticas. Suelen tener dificultades de comprensión, aprenden el cálculo de memoria y piensan que el resultado debe ser cierto o falso. Falk y Greenbaum (1995) sugieren que esta creencia se debe a la existencia de mecanismos psicológicos; algunas personas desean minimizar la incertidumbre en la decisión y suponen que la obtención de un resultado estadísticamente significativo les ayuda a ello.

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 atípica si la hipótesis nula es cierta. También hay que comprender el muestreo como un sistema de ideas interrelacionadas; selección al azar, replicación, variabilidad y distribución. Mientras que comprender la idea de muestra aleatoria simple es fundamental, probablemente es más importante entender que cada muestra es sólo una de las posibles entre las que podrían haberse elegido (Saldahna & Thompson, 2002).

4. Una aproximación informal al contraste de hipótesis

La complejidad del razonamiento sobre el contraste de hipótesis es evidente, dada la diversidad de conceptos implicados que el alumno ha de comprender y discriminar: Población y muestra, estadístico y parámetro, hipótesis (nula y alternativa para el contraste de Neymann y Pearson), nivel de significación, valor-p, potencia, región crítica y de aceptación.

No es extraño que actualmente observemos algunas propuestas de acercarse de manera informal al contraste de hipótesis. Estas propuestas tratan de introducir algunas de las ideas principales y el razonamiento del contraste y, a la vez, liberar al alumno de los cálculos asociados, recurriendo a la simulación. De acuerdo a Rossman (2008) una inferencia estadística es una conclusión que extiende un resultado observado en unos datos a un contexto más amplio (población) y debe estar justificado por un modelo de probabilidad que ligue los datos a este contexto más amplio. A continuación resolveremos el problema 1 utilizando la simulación, en vez de las tablas de la distribución normal. Seguiremos el método de Fisher, es decir, no nos preocupamos de la hipótesis alternativa o el error tipo II y además suponemos que únicamente nos interesa hacer una vez el contraste. Utilizaremos el simulador de la distribución muestral de la media disponible en: www.rossmanchance.com/applets/SampleMeans/SampleMeans.html (colección de Rossman y Chance). Otro simulador semejante puede descargarse de http://www.tc.umn.edu/~delma001/stat_tools/software.htm y es también sencillo construir un simulador de medias muestras con Excel.

Para resolver el problema con ayuda del simulador, el alumno también debe recordar que, para realizar un contraste de hipótesis sobre la media de la población μ, ha de utilizar la media de la muestra; en este caso $\bar{x} = 26$. Pero no ha de realizar cálculo formal de probabilidad; por tanto no necesita recordad la fórmula de la desviación típica de la media, la fórmula de tipificación o la lectura de la tabla de la distribución normal y el problema se simplifica

bastante.

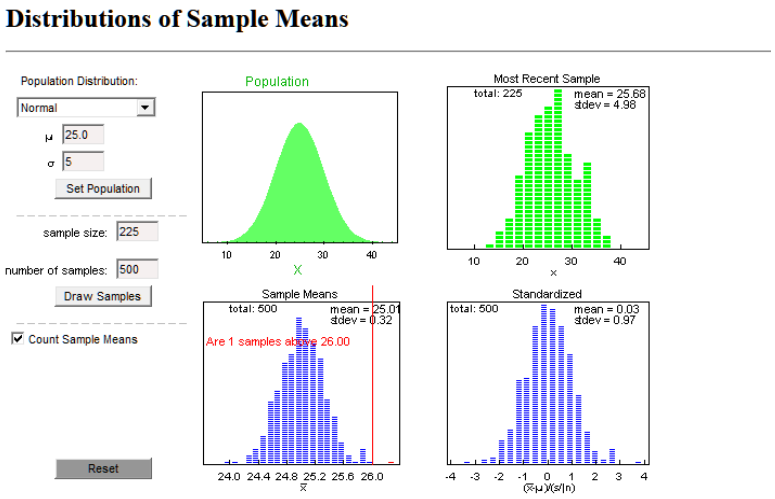

Figura 1. Simulación de la distribución muestral en el problema 1 (500 muestras)

En el simulador únicamente tiene que introducir el dato supuesto de la media de la población, μ=26 y de la desviación típica σ=5. El programa representa la gráfica de la distribución supuesta en la población (ventana izquierda superior). A continuación se pide extraer muestras; hay que dar el tamaño de la muestra n=225. En la ventana superior derecha aparece la última muestra extraída y la distribución tipificada debajo de ella. En la ventana inferior izquierda aparece la distribución empírica de la media muestral que se va completando según se repite la simulación de extracción de muestras (nosotros hemos pedido simular 500 muestras, cada una de 226 elementos). Podemos también pedir al simulador que cuente el número de muestras con valor igual o superior al dado para nuestra muestra en el supuesto dado por enunciado del problema (media superior a 26). De hecho, en nuestra simulación sólo hemos obtenido 1 entre 500 con media igual o superior a 26. La simulación nos proporciona una estimación de la probabilidad de obtener este suceso, si la hipótesis nula fuese cierta. Dicha probabilidad (valor p) será:

$$\text{Valor p}= P(\bar{x} \geq 26) \mid H_0 \text{ cierta}) \cong \frac{1}{500} = .002$$

Observamos una ligera diferencia de esta estimación de la probabilidad con la probabilidad exacta, calculada en el apartado 1. Ello es debido a que, en vez de utilizar la distribución exacta de la media muestral, estamos simulándola; por tanto, introducimos un pequeño error en el cálculo. Sin embargo, si el número de simulaciones, como en el ejemplo, es alto, los valores exactos y estimados de la probabilidad son muy parecidos.

En consecuencia, la simulación (en vez del cálculo formal) puede utilizarse al comenzar la enseñanza del contraste de hipótesis para poder concentrar al alumno en el aprendizaje de la lógica del proceso y en los conceptos que, todavía necesita: muestra y población, media muestral y poblacional, distribución de la media muestral, hipótesis nula, valor *p*. Es importante hacer ver al alumno que el valor *p* viene dado por una probabilidad condicional, porque para estimarlo estamos suponiendo que la hipótesis nula es cierta. Esto es sencillo, pues utilizando el mismo simulador podemos cambiar la hipótesis de partida; por ejemplo, suponer que la media de la población μ=25.5. En este caso, repitiendo la simulación obtenemos 28 muestras con IMC

medio igual o superior a 26.

$$\text{Valor p}= P(\bar{x} \geq 26) \mid \mu = 25{,}5) \cong \frac{28}{500} = .056$$

Esta probabilidad sigue siendo pequeña, pero no tanto como antes. Si se hubiera fijado como límte α=.05 para rechazar la hipótesis nula, en este caso no se podría rechazar la hipótesis; si la media de la población fuese 25.5, no sería ya tan raro obtener un IMC en la muestra igual a 26.

5. Algunas reflexiones

La gran cantidad de errores identificados en la práctica del contraste de hipótesis sugiere que no es suficiente enseñar a los estudiantes sobre las reglas y conceptos con el fin de llegar a una comprensión suficiente de este tema. Es claro que la enseñanza actual de la estadística no mejora las intuiciones y que contamos con recursos para mejorar la situación. Numerosos applets interactivos, como el utilizado en este trabajo proporcionan hoy un entorno dinámico y visual en el que los estudiantes pueden concretizar y visualizar el muestreo aleatorio, el concepto de distribución muestral y su estimación empírica. 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, comenzando ya desde la educación secundaria, con actividades de simulación de muestras aleatorias. Los conceptos de estimación puntual y de intervalos de confianza también pueden ser introducidos con una metodología similar.

Es importante también que el profesor comprenda los límites y no sólo las posibilidades de la simulación. La simulación introduce un error de estimación añadido a los errores tradicionales en el estudio del contraste de hipótesis (errores tipo I y II) o en los intervalos de confianza (complementario del coeficiente de confianza). Simultáneamente, es importante que el profesor enfatice los pasos requeridos en el procedimiento de contraste, discuta con los estudiantes la naturaleza de las probabilidades que calcula (condicionales), así como la interpretación correcta de las probabilidades de error. Desafortunadamente, algunas propuestas didácticas sobre inferencia informal, no enfatizan suficientemente estos puntos y se limitan a enseñar la inferencia (informal) nuevamente como un conjunto de recetas: a) simular la extracción de muestras; b) formar una distribución muestral empírica; c) calcular la probabilidad de obtener el valor observado del estadístico usando esta distribución muestral empírica; d) tomar una decisión.

Al no resaltar la naturaleza condicional del valor *p*, no se relaciona la probabilidad adecuadamente con la hipótesis supuesta y se inducen nuevamente los errores consistentes en el cambio de términos en la probabilidad condicional que define el valor *p* y el nivel de significación. Al no explicitar los conceptos subyacentes, estos quedan sin fijarse en la mente de los alumnos. En otras propuestas de "inferencia informal", simplemente se pide tomar una decisión subjetiva sobre el rechazo de una hipótesis utilizando sólo procedimientos descriptivos; lo cual no sería una verdadera inferencia estadística. Finalmente en muchos casos, se trata de sustituir todo el razonamiento probabilístico mediante la tecnología, sin enfatizar los procesos aleatorios involucrados en la simulación. Tampoco esto sería una inferencia estadística, pues esta debe estar justificada por un modelo de probabilidad que ligue los datos a la población (Rossman, 2008)

Agradecimiento: Proyecto EDU2013-41141-P y grupo FQM126 (Junta de Andalucía).

Referencias

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

Batanero, C. (2013). Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico. *Cuadernos de Investigación y Formación en Educación Matemática, 8*(11), 277-291.

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.

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

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

Díaz, C. (2005). *Apuntes sobre inferencia bayesiana*. Granada: La autora.

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.

Harradine, A., Batanero, C. y Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (pp. 235-246). Springer Netherlands.

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*. Madrid: Autor.

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

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

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19.

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.

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.

Zieffler, A., Garfield, J. B., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 5-19. Online: www.stat.auckland.ac.nz/serj/.