DESARROLLO DEL RAZONAMIENTO SOBRE PRUEBAS DE SIGNIFICACIÓN DE ESTUDIANTES DE BACHILLERATO EN UN AMBIENTE TECNOLÓGICO[1]

High School Students' reasoning development about tests of significance in a technological environments

Sánchez, E.[a], García-Ríos, V.N.[a] y Mercado, M.[b]

[a]Departamento de Matemática Educativa, Cinvestav, México
[b]Colegio de Ciencias y Humanidades, UNAM

Resumen

Se describe el desarrollo del razonamiento de 36 estudiantes de bachillerato, organizados en parejas, acerca de la técnica de pruebas de significación estadística con el apoyo de un software educativo de estadística (Fathom); dicho desarrollo es visto a través de sus respuestas a 4 problemas de pruebas de significación; estos se resolvieron en sendas sesiones en las que además se realizaron cortas intervenciones del profesor, se enseñó a utilizar el software y se discutieron dudas de los problemas vistos en la sesión previa respectiva. Las respuestas a cada problema se clasificaron en niveles SOLO. Los resultados muestran avances en la calidad de las respuestas de los participantes, superándose en cada actividad algunos errores cometidos en la previa, esto lleva conjeturar que los estudiantes se van apropiando del esquema de pruebas de significación. No obstante, se presentan algunas dificultades similares a las ya reportadas en la literatura.

Palabras clave: *Razonamiento, Pruebas de significación, niveles SOLO.*

Abstract

The development of 36 high school students' reasoning, organized in pairs, about the technique of statistical significance tests with the support of educational statistical software (Fathom) is described. Such development is seen through the responses to 4 problems of test of significance which were solved respectively in four sessions. In these session also there were short interventions of the teacher, who taught to use the software and discussed doubts of the problems seen in the respective session before. The answers to each problem were classified in SOLO levels. The results show progress in the quality of the responses of the participants, overcoming in each activity some errors committed in the previous one, this leads us to conjecture that students are appropriating the schema of tests of significance. However, some difficulties similar to those already reported in the literature are present in the students' responses.

Keywords: *Reasoning, Test of significance, SOLO levels.*

INTRODUCCIÓN

La inferencia estadística se suele introducir desde el nivel bachillerato y, después, se estudia en la mayoría de carreras universitarias. Su importancia se justifica por ser una herramienta fundamental para la investigación que permite, en todas las áreas de conocimiento, procesar e interpretar la información, hacer predicciones y tomar decisiones racionales (Batanero, 2011). Por esta razón es importante atender los problemas de su enseñanza y aprendizaje. Los dos grandes temas de la inferencia estadística clásica son el contraste de hipótesis y los intervalos de confianza; estos están constituidos por una red compleja

Sánchez, E., García-Ríos, V.N. y Mercado, M. (2017). Desarrollo del razonamiento sobre pruebas de significación de estudiantes de bachillerato en un ambiente tecnológico. En J.M. Muñoz-Escolano, A. Arnal-Bailera, P. Beltrán-Pellicer, M.L. Callejo y J. Carrillo (Eds.), *Investigación en Educación Matemática XXI* (pp. 447-456). Zaragoza: SEIEM.

formada por varios conceptos como probabilidad, muestreo aleatorio, estadístico, parámetro, distribución muestral del estadístico, confianza, hipótesis nula e hipótesis alternativa, error tipo I, error tipo II, *p*-valor y nivel de significación (Liu y Thompson, 2009). En esta investigación exploramos el razonamiento de los estudiantes acerca del tema específico de las pruebas de significación (con el apoyo del software Fathom) ya que requiere un subconjunto limitado de esos conceptos y las ideas subyacentes a dicho tema pueden formar un antecedente sólido para el desarrollo del razonamiento acerca del contraste de hipótesis y, en general, acerca de la inferencia. Nos hemos restringido a las pruebas de significación de Fisher (1956), pues en éstas la idea subyacente es nítida: Si la evidencia ofrecida por una muestra es rara o inusual bajo determinada hipótesis entonces se pone en duda dicha hipótesis. Nos proponemos responder ¿cómo desarrollan los estudiantes de bachillerato esta idea en la resolución de problemas y con apoyo de herramienta tecnológica?

ANTECEDENTES

La reseña de investigaciones sobre concepciones erróneas de la Inferencia Estadística de Castro-Sotos et al. (2007) tiene un apartado dedicado al contraste de hipótesis. Los autores organizan su exposición sobre las dificultades y concepciones erróneas acerca del contraste de hipótesis en 5 componentes, lo que ilustra la dificultad del tema. No es posible ni conveniente reproducir aquí tales componentes y las referencias citadas de cada uno, pero sí que merece la pena mencionar algunos trabajos pioneros. Vallecillos y Batanero (1997) encontraron que las concepciones de los estudiantes acerca de los conceptos del contraste de hipótesis están lejos de la lógica normativa que subyace a esta técnica, presentándose dificultades en varios de ellos. Una dificultad es la elección o definición apropiada de las hipótesis a partir de la situación, Vallecillos (1999) informa sobre el problema que representa para los estudiantes la determinación correcta de las hipótesis nula y alternativa. También, los conceptos de nivel de significación y *p*-valor, que son clave en las pruebas de significación, resultan muy difíciles para los estudiantes (Haller y Krauss, 2002). La concepción errónea más común del nivel de significación y el *p*-valor consiste en intercambiar los dos términos de la probabilidad condicional que los define (Falk, 1986). Liu y Thompson (2009) ponen énfasis en la importancia de enmarcar el proceso en una concepción estocástica de la probabilidad, en la que el dato o evidencia presentada en el problema sea concebido como un valor particular de una distribución del estadístico (distribución muestral). Tales dificultades van a encontrarse también en el presente estudio pero en el contexto de actividades con apoyo del software, que les proporciona ciertos matices.

En estudios previos al presente, los autores hemos encontrado que cuando los estudiantes intentan resolver problemas de pruebas de significación sin instrucción previa, suelen ignorar los datos, se basan en creencias personales para obtener una conclusión y no son capaces de utilizar un lenguaje probabilístico en la formulación de ésta (García-Ríos y Sánchez, 2014). Un procedimiento ingenuo que llevan a cabo cuando utilizan los datos en pruebas de significación de proporciones consiste en rechazar la hipótesis si ésta no coincide con la proporción de la muestra. En otro estudio, en el que se realizaron intervenciones de enseñanza, los estudiantes se mostraron receptivos al enfoque de pruebas de significación de Fisher, superando algunos de los errores detectados en el estudio previo, no obstante, se mantuvo la dificultad en la determinación del papel de *p*-valor y la zona de rechazo, es decir, para responder la pregunta ¿Exactamente cuándo el valor del estadístico es inusual o raro? (García-Ríos y Sánchez, 2015)

ESQUEMA DE PRUEBAS DE SIGNIFICACION

El concepto de prueba de significación se relaciona con las nociones de probabilidad, población y muestra, parámetro y estadístico, hipótesis nula, distribución muestral del estadístico, *p*-valor y región de rechazo. Suponemos que los estudiantes pueden familiarizarse poco a poco con algunas de estas nociones en el proceso de resolución de problemas, mediante intervenciones oportunas del profesor para aclarar la lógica del proceso de prueba de significación, el papel del software y sus aplicaciones en el proceso de prueba y con la aclaración y discusión de dudas acerca de los problemas.

Siguiendo la exposición de Batanero y Díaz (2015) la técnica de *prueba de significación* se expuso por primera vez en Fisher (1935) y tiene por objetivo apoyar una hipótesis acerca de la población (por ejemplo, que un efecto o cambio se ha producido en una población) con base en la información proporcionada por una muestra. El proceso a grandes rasgos consiste en lo siguiente: Se formula la *hipótesis nula*, la cual generalmente es conservadora, en el sentido de que afirma que no se produjo el cambio que se desea apoyar; dicha hipótesis se refiere a un *parámetro* de la población. Se determina una *zona crítica* o *región de rechazo* consistente en un evento, en principio posible, relacionado con un estadístico (un estimador del parámetro), pero muy improbable de ocurrir bajo la hipótesis nula (con probabilidad no mayor a un número α dado, generalmente 0.05). Se obtiene una muestra aleatoria de la población y se calcula el *estadístico* y el *p-valor*; es decir, la probabilidad de obtener el valor del estadístico encontrado u otro más extremo suponiendo H_0 verdadera. Si el *p*-valor es menor que α o, equivalentemente, el valor del estadístico cae en la zona crítica, entonces se rechaza la hipótesis nula; en caso contrario no se rechaza. El concepto de *distribución muestral* del estadístico no aparece explícitamente en esta descripción general, sin embargo, es un concepto crucial de la técnica para entender el significado del *p*-valor y la región de rechazo. En su lugar, se realiza una transformación del estadístico (tipificación) y el teorema central de límite para calcular el *p*-valor y determinar la zona crítica con ayuda de la distribución normal. Este es quizá uno de los aspectos más oscuros de toda la técnica para los estudiantes y la posibilidad de poner la noción de distribución muestral en el centro de la discusión de una prueba de significación es probablemente la principal aportación de la utilización de un software educativo (en este caso, Fathom).

En el diseño de actividades con ayuda del software para un acercamiento informal a las pruebas de significación por parte de los estudiantes, conviene considerar las siguientes cuatro componentes:

- *Realización y uso de la Simulación*. Esta componente está formada por la formulación de la hipótesis nula y por los siguientes dos pasos: a) La obtención mediante un proceso de simulación de una distribución del estadístico bajo el supuesto de que la hipótesis nula es verdadera. b) El uso de dicha distribución como si fuera la distribución muestral del estadístico, para estimar probabilidades.
- *Estimación del p-valor o determinación de una zona crítica*. Esta componente consiste en alguna de dos acciones: a) Utilizar la distribución muestral para definir una región de rechazo o, b) A partir de la distribución muestral simulada estimar el *p*-valor y compararlo con el nivel de significación.
- *Formulación de la Conclusión*. A partir de la región de rechazo o del *p*-valor obtener la consecuencia correcta, consistente en rechazar o no rechazar la hipótesis.
- *Conciencia de la Incertidumbre*. Darse cuenta de que el proceso de contraste no determina si la hipótesis nula es verdadera o falsa, sino que permite tomar una decisión racional en función de la evidencia exhibida u obtenida.

APROXIMACIÓN INFORMAL AL RAZONAMIENTO INFERENCIAL

La descripción del esquema de las pruebas de significación y los conceptos que la forman revelan la gran complejidad del tema, confirmada ésta con estudios empíricos que dan cuenta de las grandes dificultades de los estudiantes para apropiarse de dicho esquema y utilizarlo de manera flexible en sus investigaciones posteriores. Por esta razón, en el ámbito de la investigación en educación estadística ha emergido la preocupación de proponer situaciones y problemas de enseñanza para propiciar un acercamiento informal a los temas de inferencia antes de estudiar sus aspectos más formales. Por ejemplo, se ha propuesto el concepto de Razonamiento Inferencial Informal (RII), en un intento por elaborar dicha preocupación. Zieffler, Garfield y Reading (2008) definieron el RII como la forma en que los estudiantes usan su conocimiento informal de estadística para crear argumentos basados en muestras observadas que apoyen inferencias sobre una población desconocida. Para Batanero y Díaz (2015)

"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". En el presente estudio se comparte esta idea y, por esto, lo inscribimos en la búsqueda de encontrar acercamientos informales al razonamiento inferencial. Esta búsqueda tiene aún más sentido cuando los sujetos de estudio son estudiantes de nivel bachillerato. En consecuencia, para nosotros, una aproximación informal al razonamiento de las pruebas de significación es usar la simulación para crear una distribución muestral empírica, en lugar de la distribución teórica. Con dicha simulación se calcula el *p*-valor con el enfoque empírico de la probabilidad al determinar la frecuencia de resultados de la muestra o más extremos. Con este razonamiento no es necesario realizar cálculo formal de probabilidad, ni estandarizar (tipificar) la distribución por lo que no se requiere calcular puntuaciones *z* o usar la tabla de la distribución normal estándar, es decir, se emplea un razonamiento informal del contraste de hipótesis.

MÉTODO

Participantes. A 36 estudiantes, agrupados en 18 parejas (referidas como R1 a R18), de 11º grado (16-17 años de edad) de un bachillerato de la Ciudad de México se les administraron cuatro problemas de pruebas de significación. Los participantes no habían cursado la asignatura de estadística pero realizaron actividades para aprender a operar el software y a generar distribuciones muestrales a partir de una aplicación en Fathom previamente diseñada por los autores.

Instrumento. Se presentaron las cuatro situaciones-problemas que se detallan en seguida y en cada una de las cuales se formula una pregunta que conduce a una prueba de significación.

- Situación 1. La propaganda de Coca Cola presume que la mayoría (más del 50%) de la población que bebe refresco de cola prefiere Coca Cola en lugar de Pepsi Cola. Para investigar la veracidad de dicha afirmación, se hizo un experimento donde a 60 personas de dicha población escogidas al azar, se les dio dos vasos de refresco (uno con Coca y otro con Pepsi) y debían decidir cuál les había gustado más. De los 60 participantes 35 personas prefirieron Coca Cola. ¿Es razonable la hipótesis "más del 50% de la población que beben refresco de cola en México prefiere Coca y no Pepsi"?
- Situación 2. En la situación anterior, si se hace el experimento a 180 personas elegidas al azar y de ellas 104 prefieren Coca Cola. ¿Es razonable la hipótesis "más del 50% de la población que beben refresco de cola en México prefiere Coca que Pepsi"?
- Situación 3. Si la producción diaria de la máquina de una fábrica tiene más del 10% de artículos defectuosos es necesario mandarla a reparar. Para revisar la calidad de la maquina el supervisor toma una muestra aleatoria de 120 piezas del día y observa que contiene 16 piezas defectuosas. ¿Debe decidir enviar a reparar la máquina?
- Situación 4. El fabricante de una medicina afirmó que su producto es 80% eficaz para aliviar una alergia. Para verificar esta hipótesis se realizó un experimento con 100 personas que padecían de la alergia, y la medicina alivio a 74. ¿Es correcta la hipótesis "el tratamiento es 80% eficaz para aliviar la alergia?"

Procedimiento de ejecución. Se llevaron a cabo 4 sesiones de 2 horas cada una; cada sesión dividida en dos partes. La sesión 1, consistió en la introducción, presentación y exploración del software Fathom. En la primera parte de las restantes sesiones se realizaba lo siguiente: introducción de nuevos conceptos y discusión del problema abordado en la sesión anterior. En el grupo se analizaba los errores más comunes detectados en las respuestas del problema abordado en la sesión anterior y se pedía que reflexionaran en ellos y sugirieran la forma de evitarlos. La segunda parte de cada sesión consistía en resolver uno de los problemas con apoyo del software; se dejaba a los equipos trabajar por ellos mismos y llenar las hojas de trabajo; el profesor sólo intervenía para resolver pequeñas dudas. A continuación se describe con más detalle lo que se hizo en cada sesión:

- Sesión 1. Introducción al Fathom. Presentación del programa para simular un muestreo aleatorio de una población teniendo en cuenta la proporción de un rasgo de la población. Explicación de la lógica de una prueba de significación. Uso del programa para resolver el primer problema
- Sesión 2. Se discutieron dificultades surgidas en la sesión anterior. La más frecuente tenía que ver con la creencia de que la simulación genera muestras de la población real y no de una hipotética; es decir, asumen implícitamente que la hipótesis nula es cierta.
- Sesión 3. Se discutieron los problemas de la zona crítica y el *p*-valor; en las sesiones anteriores determinaron de manera intuitiva regiones para evaluar cuándo el dato muestral es extremo o atípico. En ésta se les propuso considerar niveles de significación del 5%, así como evaluar un *p*-valor como inusual o raro cuando es menor o igual al 5%.
- Sesión 4. Al discutir las dificultades de la actividad anterior, se puso énfasis en la coordinación de los diferentes porcentajes relevantes en el proceso de la prueba de significación, a saber, la proporción de la muestra (valor del estadístico), el *p*-valor expresado en porcentaje, el nivel de significación (5%).

Procedimiento de análisis. Los datos son el conjunto de respuestas que las parejas de estudiantes ofrecieron a cada una de las tareas. Para analizarlos se utilizó la metodología SOLO (Structure of Observed Learning Outcomes) de Biggs y Collis (1982, 1991), la cual propone identificar para cada tarea los aspectos relevantes que adecuadamente combinados llevan a la solución correcta del problema. En nuestro caso, los aspectos relevantes son las cuatro componentes definidas en la sección de Esquemas de Pruebas de Significación. Con base en la presencia o ausencia de estos aspectos en las respuestas de las parejas de estudiantes, éstas se clasifican en diferentes niveles. Nosotros los llamamos niveles de razonamiento acerca de las pruebas de significación. Mediante un análisis a priori de los problemas y considerando las componentes definidas en el esquema de pruebas de significación, se definieron de la siguiente manera los niveles SOLO:

- *Nivel Pre-estructural*. Las respuestas no muestran ninguno de las componentes que constituyen el esquema de prueba de significación.
- *Nivel Uniestructural*. Las respuestas sólo presentan una de las componentes del esquema de prueba de significación
- *Nivel Multiestructural*. Las respuestas se obtienen teniendo en cuenta al menos dos componentes del esquema de prueba de significación sin una integración entre ellas. Generalmente la respuesta contiene inconsistencias que no permiten llegar a la conclusión correcta. En las respuestas de este nivel no se refleja conciencia de la incertidumbre.
- *Relacional*. Las respuestas se basan en las tres primeras componentes del esquema de pruebas de significación, integradas de manera adecuada, pero no reflejan conciencia de la incertidumbre.
- *Abstracto extendido*. Las respuestas en este nivel integran adecuadamente las cuatro componentes del esquema de prueba de significación.

En la Tabla 1 se resumen los elementos que es importante tener en cuenta en la solución de cada situación. La segunda columna es la hipótesis que se va a probar, la tercera el tamaño de la muestra, la cuarta es la región crítica, la quinta columna se presenta el *p*-valor teórico, mientras que en la sexta un *p*-valor estimado mediante un proceso de simulación; finalmente en la séptima columna la conclusión a la que se debe de llegar.

Tabla 1. Elementos a considerar en las pruebas de significación de las cuatro situaciones

	Hipótesis nula	N	Región	*p*-valor	$\hat{P}$-valor	Conclusión de la prueba
Situación 1	P = 0.50	60	X >= 35	0.098	0.129	No se rechaza H_0
Situación 2	P = 0.50	180	X >= 104	0.019	0.032	Se rechaza H_0
Situación 3	P = 0.10	120	X >= 16	0.111	0.088	No se rechaza H_0
Situación 4	P = 0.8	100	X <= 74	0.067	0.080	No se rechaza H_0

RESULTADOSEN ESTA SECCIÓN SE EXPONEN LA DISTRIBUCIÓN DE LAS FRECUENCIAS DE RESPUESTA DE LOS ESTUDIANTES POR PROBLEMA Y NIVEL Y EN SEGUIDA EJEMPLOS DE RESPUESTA EN CADA UNO DE LOS NIVELES.

Frecuencias de respuesta por problema y nivel

Las frecuencias de la clasificación de las respuestas a todos los cuatro problemas (A1, A2, A3, y A4) en niveles SOLO se presentan en la Figura 1. Como los problemas fueron aplicándose de manera sucesiva se nota un efecto de aprendizaje, ya que en el problema A1, una respuesta se clasifica en el nivel Pre-estructural y13 se agrupan en el nivel Uniestructural, quedando 3 respuestas en el nivel Multiestructural y sólo 1 en el Relacional, mientras que para el problema A2, disminuyen a 5 las respuestas en el nivel Uniestructural. Esto significa que para el segundo problema más estudiantes que en el problema previo, además de realizar la simulación, buscan determinar el p-valor y/o una zona crítica; esta tendencia a disminuir las respuestas clasificadas en Uniestructural continúa con el problema A3, aunque aumenta en una unidad en el problema A4. Es interesante observar que en los problemas A2, A3 y A4, las frecuencias más altas de respuesta corresponden al nivel Relacional y, si se consideran las frecuencias de los niveles relacional y abstracto de manera conjunta, la tendencia es creciente. En resumen, hay un sensible avance en la calidad de las respuestas de los estudiantes conforme van atendiendo los problemas.

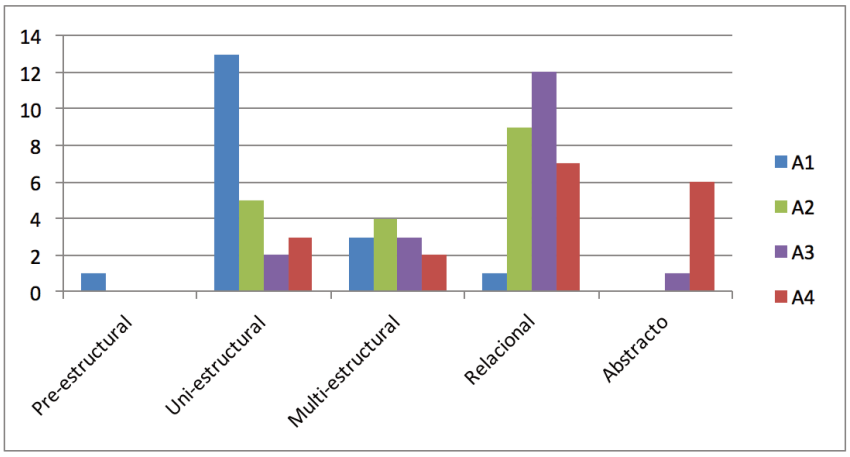

Figura 1. Niveles de razonamiento

Nivel Pre-estructural. La pareja de estudiantes que ofreció la única respuesta Pre-estructural lo hicieron en respuesta al problema 1; su razonamiento sólo tiene en cuenta que en la muestra la mayoría prefiere Coca Cola (35 de 60). Esta respuesta se puede interpretar de dos maneras; la primera es que la pareja de estudiantes no ve la diferencia entre muestra y población, pensando que el problema consiste sólo en traducir 35 de 60 a porcentajes (58%); la segunda, que la muestra replica las proporciones de la población. La respuesta fue la siguiente:

> … La propaganda está en lo correcto al poder presumir que la mayoría de la población prefiere su refresco en lugar de Pepsi ya que mencionan que "LA MAYORÍA de la población prefiere su refresco", recalcando esto podemos entender que como dicen, hacen referencia a más de la mitad de ésta, es decir

> a partir de un 51% de la población ya podemos comprender que es la MAYORÍA. Y debido a que los resultados del experimento arrojaron que 35 personas de 60 prefieren el refresco Coca Cola, lo que es igual a un 59% del total (0.59 × 60 = 35.4), podemos concluir que no erran en lo que presumen ya que están en todo lo correcto…

En otras exploraciones hemos encontrado que este es un razonamiento que se realiza con frecuencia en estudiantes que no han estudiado inferencia.

Nivel Uniestructural. Los estudiantes que ofrecen respuestas en este nivel han entendido que pueden llegar a la respuesta obteniendo una distribución mediante la simulación de la situación. Un punto que la mayoría hace correctamente es la elección del tamaño de la muestra, no obstante, en algunas ocasiones no eligen el valor conveniente de la proporción (es decir, la hipótesis nula), en otras no interpretan razonablemente bien la distribución obtenida, por ejemplo, entienden a ésta como si fuera la población. En el caso del estudiante, cuya respuesta reproducimos abajo, realiza simulaciones que, si fuera consciente de lo que hace, lo llevaría a constatar que al asignar una proporción mayor al evento "preferir la Coca" obtiene una distribución que indica que más muestras presentan una proporción favorable a dicho evento y si asigna una proporción menor se obtiene más muestras con proporciones que no son favorables al evento; no obstante, confunde la distribución muestral con la población, es decir, en lugar de decir que "más muestras tienen una proporción favorable a la Coca" opina que "más gente prefiere la Coca".

> Al momento de utilizar el programa para realizar las encuestas utilizando una proporción de 0.54 podemos notar que efectivamente la Coca es más consumida que la Pepsi. Se realizaron varios ejemplos cambiando la proporción de la población, para llegar a la conclusión de que si se utiliza una proporción menor o igual de p = 0.50 podemos notar que la gente prefiere la Pepsi, en cambio si se utiliza una proporción mayor o igual de p = 0.51 podemos notar que la Coca cola es consumida más.

Un aspecto central de las dificultades para superar exitosamente la primera componente del esquema de prueba de significación es determinar la hipótesis nula y notar que la distribución muestral se obtiene suponiendo que se cumple esta hipótesis. Al hacerlo, se despliega un razonamiento hipotético deductivo inquiriendo "¿Qué pasaría si la hipótesis nula fuera verdadera?" En varias respuestas que se han clasificado en Uniestructural, los estudiantes hacen simulaciones en las que se nota la ausencia de este razonamiento hipotético deductivo, por ejemplo, cuando eligen el valor de la proporción dada, como la proporción para hacer la simulación. Este problema se presenta incluso en algunos casos que hemos clasificado como Multiestructural.

Nivel Multiestructural. Los estudiantes que ofrecen respuestas que se clasifican en este nivel obtienen una distribución (adecuada o no al problema) y estiman el *p*-valor o una zona de la distribución para rechazar o apoyar la hipótesis. Hacerlo implica un razonamiento con frecuencias o probabilidades: ¿Con qué frecuencia o probabilidad se obtendría en una muestra la proporción dada o más? ¿De dónde a dónde se tiene el 95% de resultados (resultados normales)? ¿De dónde a dónde se puede decir que los resultados son 'raros'? Las respuestas en este nivel no obtienen la conclusión correcta ya sea porque no eligen adecuadamente la hipótesis nula, no estiman adecuadamente el *p*-valor (unos estiman $P\left(X = \hat{p}_0 \middle| H_0\right)$ en lugar de $P\left(X - \hat{p}_0 \middle| H_0\right)$ donde X es el estadístico "la proporción de éxitos en la muestra" y $\hat{p}_0$ es la proporción de la muestra dada) o no es adecuada la zona de rechazo que proponen; no obstante, ya prefiguran el razonamiento correcto. En el ejemplo siguiente (respuesta del equipo R7) los estudiantes determinaron una distribución muestral para muestras de tamaño 120 bajo la hipótesis de que 10% son artículos defectuosos. Consideraron el estadístico "el número de artículos defectuosos en la muestra" y obtuvieron la distribución que se muestra en la Figura 2. Con base en ésta estimaron un *p*-valor inapropiado ($P\left(X = \hat{p}_0 \middle| H_0\right)$) que les arrojo una probabilidad de 4.4%. Como esta probabilidad es menor al 5% consideraron que el resultado es raro, por lo que concluyeron que la máquina tiene que ser reparada; si hubieran calculado $P\left(X - \hat{p}_0 \middle| H_0\right)$ no habrían encontrado razones para rechazar la hipótesis nula.

> Es más del 10% las piezas que se están obteniendo defectuosas. Debido a que en las pruebas realizadas en cuestión de calidad se llegó a que el valor de 16 piezas del dato base es un valor atípico con un 4.4% de las pruebas, por lo cual a partir de este valor se considera el rango de más de 10%. En conclusión el supervisor debe reparar la máquina. Estamos seguros [de la conclusión], ya que en nuestro rango de "más del 10%" se comprobó que hay más pruebas con un valor mayor al establecido para el funcionamiento correcto de las máquinas, por lo cual estos valores son más atípicos que 16.

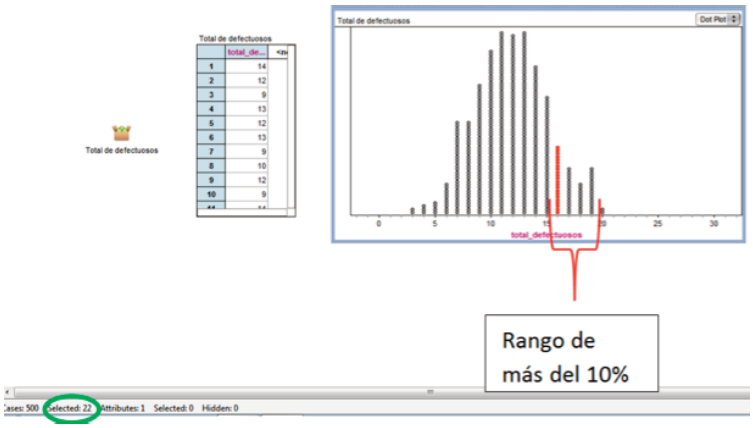

Figura 2. Razonamiento de R7 en actividad 3

El ejemplo muestra que una dificultad de los estudiantes es la de ver el resultado de la muestra experimental como parte de un conjunto y no de manera aislada. Para calcular el *p*-valor la pregunta es ¿pertenece el resultado a un conjunto raro? Y no ¿el resultado individual es raro? Dicho conjunto se determina pensando en la probabilidad de que pase lo que ocurrió o algo peor o, en su defecto, determinando el evento raro (zona crítica) y viendo si el valor particular del estadístico cae ahí.

Nivel Relacional. Los estudiantes cuyas respuestas se clasifican en este nivel consideran las tres primeras componentes del esquema de pruebas de significación y llegan a una respuesta correcta. La siguiente respuesta se refiere a la Situación 4 y fue proporcionada por el equipo R9. Aunque su explicación no es muy clara, se deduce que estimaron, mediante una simulación, el *p*-valor bajo la hipótesis H_0: $p = 0.80$, y obtuvieron 0.09 (el valor teórico es 0.067); como este valor es mayor a 0.05 clasificaron correctamente el resultado como típico, por lo que no se rechaza la hipótesis nula.

> El experimento dice que al 74% de 100 personas les fue eficaz la medicina, pero para comprobarlo decidimos ver si el resultado era típico o atípico. El resultado tenía que ser típico pues si era atípico significaba que era más o menos eficaz del 80% cosa que no queríamos. Utilizamos la regla del 5% que dicen los científicos y obtuvimos que los resultados alrededor de 74 son típicos para el 80%, pues se obtuvo un resultado de un 9%, el cual es mayor a 5%, por lo que –según la regla de los científicos– es un resultado típico. Nuestra conclusión no puede estar incorrecta en este caso no, porque creemos que aplicamos la formula correctamente y tomamos los rangos correctos

Sin embargo, con relación a la pregunta: ¿Qué tan seguro estás de tu conclusión? Responden que creen que están en lo correcto. Esto nos muestra que entienden la pregunta de manera diferente a la intención de los autores al formularla. La intención era que reflexionaran acerca de la imposibilidad de estar seguros de que la hipótesis nula sea la hipótesis verdadera, ya que la certeza sólo se puede alcanzar estudiando a toda la población o mediante una muestra suficientemente grande. Pero la pregunta se interpreta como si dijera ¿qué tan seguro estás de que seguiste el procedimiento correctamente? La mayoría de los estudiantes cuyas respuestas fueron clasificadas en este nivel responden a esta pregunta.

Nivel abstracto extendido. En las respuestas de este nivel los estudiantes, además de cumplir con los requisitos de una respuesta relacional, dejan claro que el resultado del proceso de una prueba de significación no asegura la falsedad o la veracidad de la hipótesis. Por ejemplo, en el caso de la pareja R16, después de explicar los pasos que siguieron hasta llegar a la conclusión, agregan al final:

> Estamos completamente seguros de nuestra conclusión ya que se realizaron las gráficas y los cálculos suficientes para comprobarlo. Sin embargo, puede ser equivocada y se podrían hacer más encuestas para aumentar la seguridad del resultado

A MODO DE CONCLUSIÓN

Con relación a la primera componente del esquema de pruebas de significación: *Realización y uso de la simulación*, se puede decir que los estudiantes aprenden que la solución al problema se asocia con una distribución (muestral) que se puede construir o generar mediante simulación; este es un aspecto, como señalan Liu y Thompson (2009), crucial para entender la lógica de las pruebas de significación. Las respuestas de los estudiantes que llegan a este nivel reflejan que estos han superado la creencia de que la proporción de la muestra es la misma que la proporción de la población. Sin embargo, emerge la dificultad de definir la hipótesis de manera adecuada; dificultad que coincide con lo señalado en la literatura (Vallecillos y Batanero, 1997). Por ejemplo, algunos estudiantes utilizan el valor del estadístico como parámetro para generar la distribución simulada, es decir, consideran como hipótesis dicho valor.

Con relación a la segunda componente, *Estimación del p-valor* o *determinación de una zona crítica*. para estimar el *p*-valor o determinar una zona crítica se requiere interpretar la distribución y estimar a partir de ésta, probabilidades de eventos formados por conjuntos de valores de la variable. Las dificultades que se presentan son diferentes a las de invertir los términos de la expresión condicional con la que se define el *p*-valor, reportados por Falk (1986), pues en el ambiente tecnológico no se formulan expresiones como la de probabilidad de rechazar la hipótesis suponiendo que ésta es verdadera. Por otro lado, uno de los errores encontrados en nuestra investigación consiste en creer que el *p*-valor es la probabilidad de obtener el valor del estadístico y no la probabilidad del intervalo que va de este valor a más extremos. Otro tipo de error consiste en comenzar a contabilizar los resultados que se consideran inusuales o raros a partir del valor del estadístico. Finalmente, se dan casos en que utilizan como extremo de la zona crítica la moda de la distribución.

Con relación a la tercera componente, *formulación de la conclusión*, las actividades anteriores culminan en la formulación de la conclusión en la que el razonamiento se relaciona con evaluar si el valor dado del estadístico es típico o atípico, en el primer caso la hipótesis no se rechaza y en el segundo se rechaza. Generalmente se llega a esta etapa cuando se ha sido coherente en las dos anteriores.

Finalmente, con relación a la cuarta componente, *consciencia de la incertidumbre*, el aspecto más difícil de la inferencia que emergió en el enfoque del presente estudio es el de adquirir la conciencia de que seguir el procedimiento correcto no necesariamente lleva a tomar la decisión correcta. Después de un proceso de inferencia no es posible tener certeza de no cometer un error; tener dicha conciencia lleva a entender los dos tipos de errores: no rechazar la hipótesis nula cuando es falsa o rechazarla cuando es verdadera.

Referencias

Batanero. C. (2011). Del análisis de datos a la inferencia: Reflexiones sobre la formación del razonamiento estadístico. *XIII CIAEM-IACME*, Recife, Brasil.

Batanero, C. y Díaz, C. (2015). Aproximación informal al contraste de hipótesis. 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, España.

Biggs, J. B., y Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO Taxonomy*. New York: Academic Press.

Biggs, J.B. y Collis, K.F. (1991). Multimodal Learning and the Quality of Intelligent Behavior. En H. Rowe (Ed.), *Intelligence. Re-conceptualization and Measurement* (pp.57–76). Hilsdale, NJ, USA: Lawrence Erlbaum Associates Publishers and Australian Council for Educational Research.

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

Falk, R. (1986). Conditional Probabilities: insights and difficulties. En R. Davidson y J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics*. (pp. 292 – 297). Victoria, Canada: International Statistical Institute.

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

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

García-Rios, V. N., y Sánchez, E. (2014). Razonamiento inferencial informal: el caso de la prueba de significación con estudiantes de bachillerato. En M. T. González, M. Codes, D. Arnau y T. Ortega (Eds.). *Investigación en Educación Matemática XVIII* (pp. 345-357). Salamanca: SEIEM.

García-Rios, V. N., y Sánchez E. (2015). Dificultades en el razonamiento inferencial intuitivo. 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. 207-214). Granada: 2015.

Haller, H., y Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–20.

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

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.

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

Zieffler, A., Garfield, J., delMas, R., y Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistical Education Research Journal, 7*(2), 40– 58.

[1] Agradecimiento: Proyecto EDU2016-74848-P (FEDER, AEI). Proyecto Conacyt No. 254301.