PATRONES EN EL DESARROLLO DEL RAZONAMIENTO INFERENCIAL INFORMAL: INTRODUCCIÓN A LAS PRUEBAS DE SIGNIFICANCIA EN EL BACHILLERATO

Patterns in the development of informal inferential reasoning: introduction to significance tests in high school

Silvestre, E. y Sánchez, E.

CINVESTAV-IPN

Resumen

En este trabajo se presentan los resultados iniciales de la aplicación de una actividad correspondiente a la introducción a las pruebas de significancia a estudiantes de bachillerato (17-18 años) con la intención de analizar el tipo de acciones y razonamiento que exhiben partiendo de que han sido instruidos en temas básicos de inferencia estadística mediante un enfoque informal (uso de simulaciones aleatorias y distribuciones muestrales). Se observó un uso y comprensión de conceptos estadísticos como variabilidad, distribución y valores críticos para generar un criterio de toma de decisiones, así como ciertos errores y limitaciones que podrían obstaculizar el desarrollo de un razonamiento inferencial informal apropiado.

Palabras clave: *Razonamiento inferencial informal, pruebas de significancia, distribución muestral.*

Abstract

In this paper we present the initial results of the application of an introductory activity correspondent to an introduction to significance tests in order to analyze the type of actions and reasoning that high school students (17-18 years) exhibit considering that they have been previously instructed on basic topics of inferential statistics with an informal approach (usage of random simulations and empirical sampling distributions). We observed a usage and understanding of statistical concepts such as variability, distribution and critic values aimed to generate a criteria for decision making and some errors and limitations that could inhibit an appropriate development of an informal inferential reasoning.

Keywords: *Informal inferential reasoning, significance tests, sampling distribution.*

INTRODUCCIÓN

Algunos investigadores han sugerido que diversos conflictos y limitaciones que se presentan, no sólo por parte de estudiantes en general sino también por profesores y profesionistas, a la hora de realizar inferencias estadísticas están relacionados con una formación y comprensión deficiente (o ausente) sobre conceptos estadísticos como variabilidad, variabilidad muestral y distribuciones muestrales. Por ejemplo, la revisión proporcionada por Herradine, Batanero & Rossman (2011) señala que las personas en general tienden a presentar creencias e intuiciones erróneas sobre muestreo e inferencia (por ejemplo, la heurística de la representatividad, ley de los números pequeños, incapacidad para inferir a partir de un número grande de muestras, entre otros) que la enseñanza tradicional no suele considerar e incidir dado que tiende a enfocarse en cálculos y procedimientos. En adición a esto, en el caso de México se incorpora un contenido muy limitado (o inexistente) sobre estadística inferencial al currículum matemático para niveles pre-universitarios,

Silvestre, E. y Sánchez, E. (2016). Patrones en el desarrollo del razonamiento inferencial informal: introducción a las pruebas de significancia en el bachillerato. En J. A. Macías, A. Jiménez, J. L. González, M. T. Sánchez, P. Hernández, C. Fernández, F. J. Ruiz, T. Fernández y A. Berciano (Eds.), *Investigación en Educación Matemática XX* (pp. 509-518). Málaga: SEIEM.

lo cual provee muy pocas o nulas oportunidades al estudiante de enfrentarse en un ambiente escolarizado a situaciones problema donde se exponga la necesidad de estimar con base en resultados muestrales. Esto se considera puede generar un "vacío" en su formación al no permitirle generar conocimientos que promuevan el desarrollo gradual de un de razonamiento tipo "puente" entre la lógica subyacente (operatividad y relaciones entre conceptos) y la utilización e interpretación de métodos básicos de inferencia formal (como un intervalo de confianza o una prueba de significancia). En reconocimiento de una problemática similar, ciertos investigadores han propuesto aproximaciones o métodos didácticos alternativos fuertemente relacionados al uso y comprensión de conceptos estadísticos, como variabilidad y distribuciones muestrales empíricas, así como una utilización recurrente de simulaciones aleatorias computarizadas para desarrollar métodos (informales) como solución en situaciones problema de inferencia estadística. Estas aproximaciones didácticas son denominadas a menudo como inferencia estadística informal (IEI).

El incidir en esta problemática nos ha motivado a plantear la pregunta de investigación: ¿qué tipo de razonamiento, basado en la observación y análisis de acciones y estrategias de solución, presentaría un estudiante de nivel bachillerato (momento en nuestro país donde suele aparecer el primer curso introductorio de Probabilidad y Estadística), que ha llevado a cabo un proceso de aprendizaje sobre distribuciones muestrales que incluye el uso de simulaciones aleatorias bajo un enfoque informal, frente a situaciones problema que requieren la aplicación (informal) de una prueba de significancia? Para abonar a la respuesta de esta pregunta, en esta comunicación se presentan los resultados y análisis preliminares de una actividad, que forma parte de una Trayectoria Hipotética de Aprendizaje general (THA) (Simon, 1995; Simon y Tzur, 2004), que corresponde a una introducción a las pruebas de significancia. Es importante tener en mente que el objetivo de esta aplicación no es de formación o aprendizaje, sino el de explorar e identificar patrones sobre el razonamiento de los estudiantes (errores, limitaciones, sesgos, etc.) cuando éstos se enfrentan por primera ocasión a una prueba de significancia y partiendo de que sus conocimientos previos incluyen a la distribución muestral, probabilidad y valores críticos.

A continuación se describen las consideraciones metodológicas del estudio; el diseño, desarrollo y resultados de la actividad; así como una serie de reflexiones generales derivadas del análisis del desempeño de los estudiantes.

ALGUNA LITERATURA RELACIONADA

Sobre distribuciones muestrales y el uso de la tecnología en la enseñanza de la estadística

Diversos elementos interrelacionados intervienen en el proceso de construcción, aplicación e interpretación de herramientas de inferencia estadística, tales como muestra y población, parámetro y estimador, variabilidad y distribución, probabilidad y confianza, entre otros. En particular, estudios realizados como el de Chance, delMas & Garfield (2004) y el de Liu y Thompson (2007) destacan que una comprensión deficiente o limitada sobre el concepto de distribución muestral puede opacar el desarrollo de un razonamiento apropiado sobre la inferencia estadística. Actualmente se ha desarrollado paquetería especializada en apoyar el proceso de formación de un estudiante en conceptos de inferencia estadística y probabilidad como Fathom, que permite, entre otras aplicaciones, la generación de simulaciones aleatorias que dan lugar a la emergencia de conceptos abstractos como la distribución muestral; mediante la modificación de parámetros de la distribución (tamaño de muestra, parámetro poblacional, número de muestras) se brinda al estudiante la posibilidad de estudiar el comportamiento de este concepto evitando un tratamiento que tradicionalmente es teórico. Este recurso es clave para potencialmente promover el desarrollo del razonamiento estadístico de los estudiantes (Chance, Ben-Zvi, Garfield & Medina; 2007).

Sobre razonamiento inferencial informal

El trabajo realizado por distintos grupos de investigadores ha dado como resultado una variedad de definiciones sobre la IEI y el razonamiento en que se apoya (RII). Zieffler, delMas & Reading (2008) parten de distintas caracterizaciones de estos conceptos para brindar una sobre lo que significa el RII: "la forma en que los estudiantes usan sus conocimientos informales de estadística para crear argumentos basados en muestras observadas para sustentar las inferencias sobre la población desconocida" (p.44). Rossman (2008) describe una inferencia estadística como una generalización que va más allá de los datos disponibles hacia una población más grande, o la realización de una conclusión más profunda sobre la relación entre las variables (esto mediante un modelo probabilístico que relacione a los datos con dicha generalización o conclusión). También sugiere que las simulaciones aleatorias son una forma efectiva e informal para introducir a los estudiantes a la lógica de la inferencia estadística. Makar, Bakker & Ben-Zvi (2011) caracterizan la IEI como una generalización probabilística sobre los patrones revelados por la información (datos) disponibles; dicha generalización es el resultado de la aplicación de un razonamiento inferencial informal (RII). En este trabajo se adopta esta caracterización para el RII.

Sobre pruebas de significancia

Los errores sobre la comprensión e interpretación de las pruebas de significancia han sido uno de los más documentados en el último par de décadas en lo que se refiere a la didáctica de la estadística. Destacamos el estudio realizado por Vallecillos (1999) donde describe algunas concepciones erróneas sobre el tipo de prueba que ofrecen las pruebas de significancia (como lo es el asumirla como una prueba matemática de la verdad de cierta hipótesis). Batanero y Díaz (2015) describen errores comunes tales como la comprensión inadecuada del nivel de significancia (α) al considerarla como $\alpha = P(H_0$ es cierta | se ha rechazado $H_0)$ cuando en realidad es $\alpha = P(\text{rechazar } H_0 \mid H_0$ es cierta). En este último trabajo también se distinguen las diferencias entre los métodos de Fisher, Neyman y Pearson para una prueba de significancia y se brinda una descripción de lo que sería realizar una prueba de significancia bajo el método de Fisher (interés en rechazar una hipótesis nula con un cierto nivel de significancia) mediante un enfoque informal; finalmente se incluyen algunas consideraciones de cuidado sobre posibles limitaciones conceptuales que este procedimiento informal puede generar (como la falta de énfasis en la probabilidad condicional correspondiente al p-valor).

MÉTODO

Participantes/sujetos. Los antecedentes de los 42 estudiantes que participaron en el estudio son un curso de Estadística y Probabilidad I (completado en el semestre previo) donde se atendieron los contenidos de: representaciones gráficas, medidas de tendencia central y de variabilidad, enfoque clásico y frecuencial de la probabilidad, cálculo de probabilidad para eventos simples y compuestos; y lo correspondiente a los tópicos incluidos en la THA: distribución muestral empírica de proporciones, efectos del tamaño de muestra en la distribución, valores críticos (introducidos desde el inicio de la THA como valores muestrales a partir de los cuales los siguientes, menores o mayores, se consideran valores poco frecuentes/probables de ocurrir dado que representan el 10% o 5% de todos los obtenidos en la simulación) y cálculo de probabilidad de obtener cierto valor muestral a través de frecuencias relativas.

Recolección de datos. Los datos recabados constaron de hojas de trabajo (físicas) en donde los estudiantes respondieron los cuestionamientos; observaciones de dos profesores/investigadores que acudieron a la sesión (además de las de los presentes autores); y algunas entrevistas breves que fueron aplicadas durante y después de la aplicación de la actividad. La actividad fue realizada en una sesión de aproximadamente 1.5 horas.

Método de análisis. A partir de los datos recabados se realizó una codificación y categorización inicial que está basada en patrones recurrentes de acciones específicas que realizaron los estudiantes para resolver la actividad. Dichas categorías hacen referencia a la identificación y uso de los conceptos de variabilidad (muestral) y valores críticos (empíricos) de una distribución muestral. Estos códigos y categorías son utilizados para brindar mayoritariamente una descripción sobre el tipo de respuestas obtenidas a partir de lo cual se plantean patrones acerca del razonamiento exhibido por los estudiantes.

Descripción y estrategia de aplicación de la actividad. Los componentes (en términos de una THA) de la actividad son: a) objetivo de aprendizaje: que el estudiante realice una prueba de significancia mediante el enfoque de Neyman y Pearson de manera informal, es decir, que elija una de dos hipótesis H_0 o H_1 (nula o alternativa) como la más factible o plausible con base en la frecuencia relativa de obtención de un cierto valor muestral.

b) la actividad/situación problema se divide en tres momentos. En el primero el profesor organiza a los estudiantes en parejas y plantea la situación problema cuyo texto corresponde a: "Una fábrica dedicada la generación de componentes para computadora posee cuatro máquinas especializadas en producir tarjetas madre para laptop. De manera irremediable, cada máquina produce de forma aleatoria una cierta cantidad de tarjetas defectuosas en cierto período de tiempo. El departamento de control y calidad indica que las máquinas pueden presentar hasta 10% de tarjetas defectuosas en su producción, de lo contrario éstas deben ser enviadas a revisión para una posible reparación. Ahora, un equipo de técnicos recolectó en cierto día una muestra aleatoria de 120 tarjetas de cada máquina para analizar el número de tarjetas defectuosas que se presentan, obteniendo así los resultados: máquina A - 42 tarjetas defectuosas, máquina B - 21, máquina C - 27 y máquina D - 15. Con base en estos resultados muestrales, ¿cuáles máquinas consideras son necesarias mandar a revisión?".

A partir del planteamiento de esta pregunta se abre un espacio de aproximadamente 15 minutos para que los estudiantes respondan el mismo de manera libre, produciendo y entregando así la primera hoja de trabajo al profesor. Una vez hecho esto, se inicia el segundo momento de la actividad en donde el profesor utiliza Fathom para generar un simulador de porcentajes muestrales dado un cierto parámetro poblacional conocido (figura 1, izquierda). Este simulador de porcentajes muestrales utiliza el contexto de una actividad de la THA que había sido previamente trabajada donde se simula la toma de muestras de una población finita constituida de frijoles negros y pintos (50% cada uno). Tras haber realizado y exhibido algunas simulaciones del porcentaje muestral, se retoma la pregunta inicial: "A partir de lo observado en el simulador, ¿cambiarías algo en tu respuesta anterior? (cuáles máquinas enviar a revisión y porqué)". Se abre nuevamente un espacio para que los estudiantes respondan el cuestionamiento y entreguen la segunda hoja de trabajo al profesor.

Después de la entrega de la segunda hoja de trabajo, el tercer momento de la actividad se inicia donde el profesor utiliza un archivo previamente diseñado en Fathom donde se exhibe la simulación de una distribución muestral empírica en particular de 300 muestras aleatorias, cada una de 120 tarjetas provenientes de una máquina Z que se sabe cumple con tener un 10% de tarjetas defectuosas en su producción (figura 1, derecha). En esta gráfica se incluyen los valores críticos que corresponde al 95% de confianza y el promedio de la distribución muestral (líneas verticales).

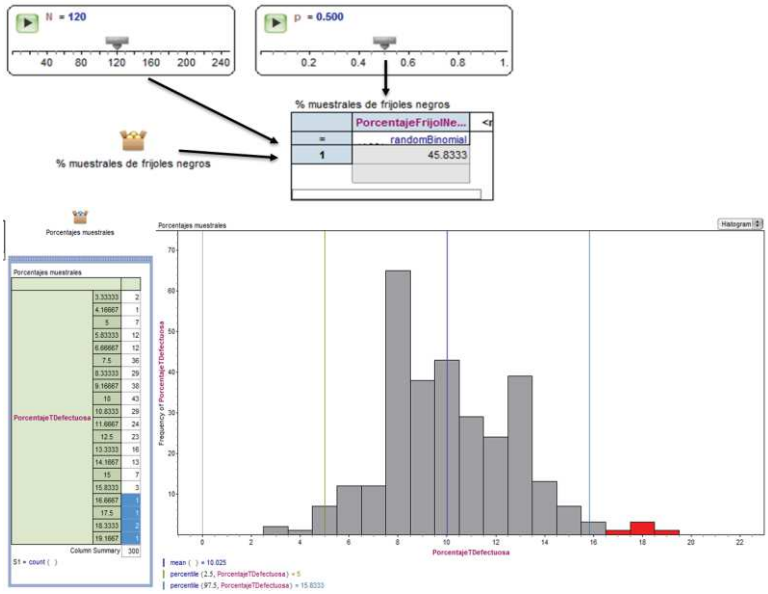

Figura 14. Simulación de un porcentaje muestral proveniente de una población con parámetro de 50% (izquierda) y simulación de una distribución muestral empírica de una población (máquina Z) con parámetro del 10% (derecha).

Tras haber realizado una explicación breve sobre los componentes del archivo (distribución muestral y valores agregados) que pertenecen a la máquina Z, se brinda a los estudiantes un espacio para que respondan la pregunta "Con base en lo observado en las simulaciones de la distribución muestral, ¿cuáles máquinas consideras son necesarias mandar a revisión?" y entreguen su tercer y última hoja de trabajo al profesor, concluyendo así con la aplicación de la actividad.

Como puede observarse, en el último momento de la actividad se plantean de forma implícita H_0 (el número de tarjetas defectuosas está controlado dado que es menor o igual al 10%) y H_1 (el número de tarjetas defectuosas excede el 10% permitido) a través de la comparación de los valores muestrales con la máquina Z. También es importante tener en consideración que los estudiantes no diseñan ni manipulan ninguno los archivos de Fathom mencionados durante la aplicación de la actividad, y que el nivel de significación está previamente establecido ya que se trabajaron valores críticos con niveles de 5% y 10% en actividades previas de la THA (en contextos y situaciones diferentes a las de una prueba de significancia).

Finalmente, el componente c) de la actividad corresponde a las hipótesis de aprendizaje para los cuestionamientos planteados. De acuerdo al orden de aparición de éstos, las hipótesis son: 1) los estudiantes no considerarán la variabilidad muestral que puede presentarse respecto al número de tarjetas defectuosas y enviarán todas las máquinas a revisión dado que sobrepasan el 10% permitido; 2) una vez que se observa que puede presentarse variabilidad muestral, los estudiantes enviarán a revisión las máquinas A y C dado que es muy poco probable (basados en la observación de la frecuencia relativa de aparición) que la variabilidad muestral genere porcentajes muestrales tan lejanos al 10%, pero generará dudas sobre esta decisión para B y D (¿producto de la variabilidad muestral o un porcentaje poblacional mayor al 10%?); 3) basado en su conocimiento previo de la distribución muestral, probabilidad y los valores críticos determinará que todas las máquinas, excepto D, deben ser enviadas a revisión dado que sus porcentajes muestrales están fuera del rango de los valores críticos, siendo entonces poco probable que esas máquinas (A,B y C) posean un porcentaje menor o igual al 10% de tarjetas defectuosas.

RESULTADOS

A continuación se describen las acciones realizadas por los estudiantes para cada momento de la actividad. En el primer momento de la actividad se obtuvo:

Tabla 1. Resultados del primer momento de la actividad.

Identifica y acepta variabilidad muestral (7 parejas)	*No acepta variabilidad muestral (14 parejas)*
Manda A, B y C a revisión pero duda, posterga o no envía D, 5 parejas. Manda a revisión A y C pero duda en B y D, 2 parejas.	Manda todas a revisión ya que exceden el 10%, 12 parejas. Manda todas a revisión pero señala la cercanía de D al 10%, 2 parejas.

Para este cuestionamiento se registraron 14 parejas (66.7%) cuya decisión fue enviar todas las máquinas a revisión dado que su valor muestral excede el 10% permitido; de éstas, dos señalan que el "error" de D hace no urgente la revisión de esta máquina pero igualmente mantienen la decisión de revisarla al igual que el resto. Como ejemplo de este tipo de respuestas, la pareja 8 (P8) respondió "mandaríamos a reparar todas las máquinas ya que el límite permitido de tarjetas defectuosas es el 10% y todas las máquinas rebasaron ese porcentaje, pues el máximo de tarjetas defectuosas que deberían salir serían 12 de una muestra de 120". En cambio, siete parejas (33.3%) no envían a revisión todas las máquinas de una forma tan determinista y centran su atención en D o en D y B; de éstas cinco señalan explícitamente que el error en D es aceptable ya que la diferencia con el 10% es considerada muy pequeña debido a posibles variaciones y dos más brindan argumentos similares para D y B. Como ejemplo, P2 mencionó:

> Se mandarían a revisión las máquinas A, B y C de manera inmediata debido a que exceden el 10% que se permitía como margen de error. De la misma manera se mandaría a revisar la máquina D, pero esa, se haría dicha operación en un tiempo posterior, debido a que no hay mucha variación en el margen de error.

A pesar de haber generado una discusión y reflexión al momento del planteamiento del problema donde se hace énfasis en la naturaleza aleatoria de la cantidad de tarjetas defectuosas producidas en un cierto período de tiempo, estas respuestas sugieren que más de la mitad de los estudiantes no consideró la idea de que puede presentarse variabilidad muestral independientemente del porcentaje de tarjetas defectuosas que posea esa máquina en particular.

En el segundo momento de la actividad se obtuvo lo siguiente seccionado a partir de los resultados del momento previo:

Tabla 2. Resultados del segundo momento de la actividad según la clasificación inicial.

	Identifica y acepta variabilidad muestral (7 parejas)	*No acepta variabilidad muestral (14 parejas)*
Cambia decisión inicial	Incluye B en la no revisión, 2 parejas. Duda sobre revisar cualquier máquina, 1 pareja.	Considera ahora la variabilidad muestral y por lo tanto las máquinas B y D no se envían/dudan sobre mandarlas a revisión, 8 parejas. Considera lo anterior sólo para D, 3 parejas.
No cambia decisión inicial	Considera B y D/solo D se encuentra dentro de un rango aceptable de variación alrededor del 10%. 4 parejas.	Determina que la información que brinda el simulador no es relevante/útil, 3 parejas.

Para el grupo que parece identificó y aceptó inicialmente variabilidad muestral, dos parejas buscan refinar o precisar su decisión previa al observar las variaciones que se producen en las muestras a través del simulador y argumentan que los errores presentados son admisibles dentro de algún rango aceptable de variación alrededor del 10%; mientras que otra pareja parece entrar en conflicto y toma la postura "extrema" opuesta y sugiere reducir la variabilidad muestral aumentando el tamaño de muestra para obtener resultados muestrales más confiables. Asimismo cuatro parejas de este grupo

refuerzan su decisión previa incorporando a su argumentación esta idea de un rango aceptable de variación alrededor del 10%. Como ejemplo de estas respuestas, P1 contestó (no cambia su decisión inicial):

> Porque se consideró que puede haber mucha variabilidad en el margen de error en las muestras, de manera que el % no va a ser igual ni exacto, puede aumentar o disminuir, pero la máquina D sigue teniendo el porcentaje menor y no tendría caso mandar a arreglarla, porque su % puede seguir siendo aceptable dentro del margen de error.

Por otro lado, del grupo que no consideró variabilidad muestral inicialmente, ocho parejas cambian su decisión inicial para D y B al incorporar esta idea de un rango aceptable de variación y tres parejas más lo hacen de la misma manera sólo para D. Como ejemplo de estas respuestas, P20 respondió (cambia decisión previa):

> Sí, en la máquina B no la mandaríamos a arreglar porque el simulador nos hizo recordar la variabilidad que puede haber en la muestra y no sobrepasa por mucho el límite establecido, ya que puede bajar o subir el % de tarjetas defectuosas siempre y cuando no sobresalga mucho. En la D considerando lo anterior no la mandaríamos a arreglar.

A pesar de que en general se observa que la mayoría de los estudiantes identifica e incorpora de alguna manera la variabilidad muestral a su decisión hasta este momento, tres parejas mantienen su postura inicial al no identificar esta idea dado que no relacionan lo que muestra el simulador de porcentajes muestrales (porcentaje muestral de frijoles de cierto tipo) con esta situación en particular. Finalmente, en el tercer momento de la actividad se exhibe una simulación de la distribución muestral empírica de la máquina Z (10% de artículos defectuosos) y se repite por última ocasión la pregunta inicial (cuáles máquinas enviar a revisión y porqué); obteniéndose los siguientes resultados generales:

Tabla 3. Resultados generales del tercer momento de la actividad

Utiliza valores críticos en su decisión (17 parejas)	Manda a revisar A, B y C pero no D. Utiliza apropiadamente el rango entre valores críticos (empíricos), 8 parejas.
	Manda a revisar A, B y C pero no D. Aplica criterio de cercanía del valor muestral al 10% y utiliza apropiadamente el rango entre valores críticos empíricos, 4 parejas.
	Manda a revisar A y C, no D y duda en B. Utiliza criterio de rango entre valores críticos pero duda en B, reconoce queda fuera de los límites, 2 parejas.
	Manda a revisar D o B y D pero no A y C. Aplica criterio de rango de valores críticos a la inversa, 3 parejas.
No utiliza valores críticos en su decisión (4 parejas)	Manda a revisar A, B y C pero no D. Utiliza criterio de cercanía respecto al 10%, 3 parejas.
	Mandar revisar A y C, no D y B. Utiliza criterio de cercanía al 10% y asume B y D como dentro de un rango aceptable, 1 pareja.

Encontramos que el 81% de las parejas utiliza los valores críticos en su regla de decisión sobre cuáles máquinas mandar a revisión. Ocho parejas se basan completamente en este criterio y mencionan que dado que los valores muestrales de A, B y C se encuentra fuera del rango de los valores críticos entonces son considerados valores muestrales muy poco probables (representan menos del 5% del total de las muestras) para una máquina con 10% de tarjetas defectuosas, por lo tanto estas máquinas deben ser enviadas a revisión; caso contrario en D que al pertenecer a este intervalo se considera un resultado muestral probable de obtener y por lo tanto no debe ser enviada a revisión. Como ejemplo de estas respuestas, P10 respondió:

> Mandaríamos a reparar las máquinas A, B y C porque los % de tarjetas defectuosas de estas máquinas quedan fuera de los valores críticos (representan el 2.5% de todos los valores obtenidos). La máquina D sería la única que no mandaríamos a reparar debido a que su % de tarjetas defectuosas entran en los valores críticos que representan el 95% de todos los valores obtenidos.

Asimismo, P3 respondió:

> Es seguro que se mandarían a reparar las máquinas A, B y C ya que son las que presentan mayor cantidad de tarjetas defectuosas y se encuentran fuera del rango de valor crítico superior (el rango 10-15%), por lo que la máquina D tiene menos probabilidad de producir tarjetas defectuosas.

Cuatro parejas utilizan este criterio pero no se deslindan del criterio de cercanía; es decir, consideran que si la distancia entre el valor muestral y el poblacional es muy elevada, bajo un criterio no especificado, entonces dicha máquina debe ser enviada a revisión y viceversa. Como primer ejemplo de esto, P17 contestó:

> Concluimos en que es necesario mandar a revisión las máquinas A, B y C ya que sus %'s sobrepasan por mucho el margen de error (10%), por lo que la máquina D sigue teniendo un % de error considerable dentro del rango, considerando la gráfica está dentro de los valores más probables de que las tarjetas no salgan defectuosas. Podemos decir que en las otras máquinas es menor la probabilidad de que se obtenga un 10% de error, por lo tanto un mayor número de tarjetas saldrían defectuosas.

Dos parejas aplican nuevamente el criterio del rango entre valores críticos pero no tienen claridad sobre una conclusión al observar que el valor muestral de B es el inmediato fuera del rango, siendo este un valor que sí aparece dentro de los porcentajes muestrales simulados y no los de A y C. Tres parejas aplican el criterio del rango entre valores críticos a la inversa, esto es, concluyen que si el valor muestral se encuentra dentro del rango entonces se espera produzca una mayor cantidad de tarjetas defectuosas y viceversa si se encuentra fuera de este, por lo tanto se determina se debe mandar a revisión B y D pero no A y C. Finalmente, el 19% de las parejas no incluyó en su decisión el criterio de los valores críticos de ninguna forma; tres parejas parecen ignorar estos valores y presentan dificultades para articular con mayor precisión argumentos basados en las simulaciones de la distribución muestral, apoyándose en mayor medida en el criterio de la cercanía, por ejemplo, P13 respondió:

> Mandaríamos a revisar las máquinas A, B y C ya que el % que se observa en las muestras es muy elevado lejano al 10% permitido. La máquina D no la mandaríamos a revisar ya que el 12.5% que se observa en la muestra es muy cercano al 10%. Después de realizar más muestras el valor se mantendría muy cercano a este 10%, caso contrario de las A, B y C que está muy alejado.

Una pareja aplica este criterio y menciona la existencia de cierto rango aceptable de variación pero de nuevo no parece hacer referencia alguna al rango de valores críticos que se menciona y exhibe en la gráfica.

CONCLUSIONES Y DISCUSIÓN

Como primer punto de reflexión, se enfatiza la estrategia mayoritariamente empleada por los estudiantes que corresponde a una utilización aceptable/apropiada de los valores críticos como valor de referencia para aceptar o rechazar una hipótesis; siendo esto contrario a la hipótesis de aprendizaje sobre la utilización de las frecuencias relativas como criterio de decisión (en la cual no se especifica de manera clara cuándo un valor muestral es considerado como probable o no probable).

Nótese que el empleo de los valores críticos no fue propuesto en ningún momento por el profesor durante la aplicación de la actividad y se considera más bien, que esto es producto del razonamiento inducido por las actividades previas de la THA en donde emergieron y se utilizaron conceptos fundamentales de la estadística (Burril & Biehler, 2011), tales como variabilidad, distribución,

inferencia y muestreo. Esto sugiere que es posible hacer que sea el mismo estudiante quien proponga de manera razonada parte del método/procedimiento (formal) de una prueba de significancia, a pesar de que sea esta la primera ocasión en la que se enfrenta a este tipo de situaciones problema. En cierta medida esto contribuye a evitar la introducción de procedimientos de estadística inferencial en forma de receta (identificar y plantear H_0 y H_1, calcular los valores críticos correspondientes al 95% de confianza para una distribución muestral teórica de tipo binomial con n=120 y p=.1, calcular el estadístico de prueba para observar en qué región de la distribución se encuentra para así aceptar o rechazar H_0 y H_1), lo cual suele ocurrir en la enseñanza tradicional de estos tópicos.

Como segundo punto de reflexión, se observaron ciertos conflictos y limitaciones en el desempeño de los estudiantes. Por ejemplo, algunas parejas utilizaron expresiones verbales y textuales que no coinciden con lo que en realidad hacen referencia (algunos estudiantes utilizan indistintamente las expresiones "tamaño de muestra" por "número de muestras" cuando hacen referencia a la primer idea; otros mencionan "más porcentajes muestrales" para hacer referencia a la reducción de variabilidad cuando se aumenta el tamaño de muestra). Otros presentaron confusión sobre la cuantificación de las áreas de la distribución muestral con respecto a los valores críticos (se mencionan áreas de 85%, 90% como la que contiene el intervalo y 5% o 10% cuando se hace referencia al área que se encuentra en la cola derecha).

De este tipo de dificultades sobresale la falta de consideración de variabilidad muestral y ciertas confusiones y/o ausencias en el uso de un lenguaje probabilístico adecuado para expresar una conclusión sobre las máquinas a enviar a revisión (por ejemplo las respuestas de P10, P3 y P17); también se observa el uso de la expresión "mandar/enviar a reparar" en sustitución de "mandar/enviar a revisar" (por ejemplo en las respuestas de P1, P10 y P20), a pesar de que los profesores/investigadores utilizaron cuidadosamente la primera expresión durante todo el desarrollo de la actividad.

Si bien los estudiantes demuestran ser capaces de comprender apropiadamente cierta parte del procedimiento de una prueba de significancia, las dificultades señaladas en el párrafo anterior nos sugieren que éstos pueden estar concibiendo que dicho método posee un carácter determinista respecto a la naturaleza de verdad o falsedad de la hipótesis nula y alternativa. Por ejemplo, es posible que los estudiantes asuman que esta solución implica determinar con certeza que la máquina D sí posee un porcentaje de tarjetas defectuosas del 10%, exhibiendo quizás un sesgo que excluye la identificación de la naturaleza probabilística interviniente en el proceso de razonamiento a lo largo del proceso de solución. Se considera que en esta situación problema se enfatiza que el estudiante debe tomar una decisión ante la presencia de variabilidad e incertidumbre y esto puede hacerle ignorar la posibilidad de cometer un error una vez se ha aceptado o rechazado la hipótesis nula o alternativa (error tipo I en el caso de las máquinas B, C y D y del tipo II para la máquina A). Después de todo, los profesores/investigadores no señalaron explícitamente en ningún momento de la realización de la actividad que este problema enmascara trabajar con una probabilidad condicional clave que es el p-valor (en el caso de la máquina A la simulación arrojó un p-valor = $P(p \geq .175 \mid P = .1) \cong 2/300$); esto con el objetivo de poder observar las acciones que realizan los estudiantes de forma natural ante la información, recursos didácticos y conocimientos previos de los que disponen.

Como reflexiones finales, concluimos que las actividades trabajadas en la THA y la presentada en esta comunicación, pueden promover el desarrollo de un razonamiento inferencial informal que a su vez complementaría o se convertiría en la base de apoyo para comprender y utilizar apropiadamente tanto ideas estadísticas fundamentales, así como métodos formales de inferencia. Sin embargo, el uso constante de distribuciones muestrales empíricas y simulaciones aleatorias no garantiza que los estudiantes estén exentos de cometer errores conceptuales como los aquí expuestos.

Desde una perspectiva de formación o aprendizaje, la aplicación de este tipo de actividades en el aula permite hacer explícitas las creencias, intuiciones y conocimientos de los que disponen los estudiantes respecto a la inferencia estadística. Esto constituye una oportunidad para el profesor de incidir directamente en los posibles errores y limitaciones conceptuales que pueden presentarse (como la no identificación de la probabilidad condicional implicada en el p-valor), abonando así a la eficiencia y efectividad de un proceso de enseñanza y aprendizaje en condiciones reales en las que éste es llevado a cabo. Encontramos también que la incorporación de este tipo de actividades en edades tempranas de formación académica permite robustecer las experiencias y conocimientos de los estudiantes sobre la naturaleza y relaciones entre conceptos clave como variabilidad, distribución e inferencia; acciones que actualmente en nuestro país son considerablemente limitadas.

Desde una perspectiva de la investigación en educación estadística, sugerimos es necesario continuar con una exploración detallada de cómo es que los estudiantes razonan al momento de enfrentarse a escenarios donde es necesario realizar inferencias estadísticas; información que puede 1) ayudar a comprender las razones y motivos que originan diferentes errores conceptuales que comúnmente se presentan y 2) apoyar la creación de nuevas propuestas didácticas que buscan desarrollar y promover un razonamiento inferencial informal.

Referencias

Batanero, C. & 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. 207-214). Granada, 2015.

Burril, G. & Biehler, R. (2011). Fundamental Statistical Ideas in the School Curriculum and in Training Teachers. In C. Batanero, G. Burrill y C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education* (pp. 57-69), 10. New York: Springer.

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions. En D. Ben-Zvi & J. Garfield (Eds.). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Amsterdam: Kluwer Academic Publishers.

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning. *Technology Innovations in Statistics Education*, 1(1).

Harradine, A., Batanero, C. & Rossman, A. (2011). Students and teachers' knowledge of sampling and inference. In C. Batanero, G. Burrill y C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education* (pp. 235-246). New York: Springer.

Liu, Y., & Thompson, P. (2007) Teachers' Understandings of Probability, *Cognition and Instruction*, 25:2-3, 113-160.

Makar, K., Bakker, A. & Ben-Zvi, D. (2011). The Reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, 13(1-2), 152-173.

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

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26, 114–145.

Simon, M., & Tzur R. (2004). Explicating the Role of Mathematical Tasks in Conceptual Learning: An Elaboration of the Hypothetical Learning Trajectory, *Mathematical Thinking and Learning*, 6(2), 91-104.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Proceedings of the International Statistical Institute 52nd Session*. Helsinki: International Statistical Institute. Online: www.stat.auckland.ac.nz/~iase/publications.

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