RAZONAMIENTO DE ESTUDIANTES UNIVERSITARIOS SOBRE VARIABILIDAD E INTERVALOS DE CONFIANZA EN UN CONTEXTO INFERENCIAL INFORMAL

UNIVERSITY STUDENTS´ REASONING ON VARIABILITY AND CONFIDENCE INTERVALS IN INFERENTIAL INFORMAL CONTEXT

Santiago Inzunsa Cazares
Universidad Autónoma de Sinaloa
sinzunza@uas.edu.mx

En este artículo se presentan resultados de una investigación con una metodología de tipo cualitativo con un grupo de 15 estudiantes universitarios de ciencias sociales, sobre el razonamiento inferencial informal que desarrollaron en un ambiente computacional sobre conceptos que intervienen en los intervalos de confianza. Los resultados señalan que los estudiantes desarrollaron un razonamiento correcto sobre la variabilidad del muestreo y lograron visualizar intervalos razonables de variabilidad en un muestreo repetido, a su vez identificaron relaciones correctas entre el tamaño de muestra y la confiabilidad en la amplitud del intervalo y el margen de error, e identificaron la aleatoriedad de un intervalo de confianza. Sin embargo, tuvieron dificultades para conceptualizar la confiabilidad como el porcentaje de intervalos que capturan al parámetro en un muestreo repetido en condiciones idénticas.

Palabras clave: Análisis de Datos y Estadística, Tecnología, Modelación

Introducción

Hacer inferencias sobre poblaciones e interpretar resultados de estudios estadísticos se ha vuelto parte de la vida profesional y cotidiana de las personas. Un ejemplo concreto muy recurrente son las encuestas de opinión que aparecen casi a diario en los medios de comunicación, en las cuales se reportan estimaciones sobre parámetros de una población, margen de error y confiabilidad, entre otros conceptos. La investigación reporta que los conceptos y el razonamiento que caracteriza a la inferencia estadística son complejos para la mayoría de los estudiantes (Castro Sotos, Vanhoof, Noortgate, & Onghena, 2007), incluso para profesores e investigadores que la aplican en su profesión (Liu & Thompson, 2004).

En el caso particular de la estimación de parámetros mediante intervalos de confianza, la literatura reporta diversas dificultades de comprensión y errores en la interpretación de resultados. Por ejemplo, un error muy persistente consiste en considerar que un intervalo de 95% de confianza indica que existe un 95% de probabilidad de que el parámetro poblacional se encuentre entre los límites del intervalo. Otros errores consisten en no reconocer la aleatoriedad y naturaleza inferencial del intervalo e ignorar el efecto del tamaño de muestra y la confiabilidad en la amplitud del intervalo y el margen de error; creer que en distintas muestras se obtendrá el mismo intervalo, entre otros (Olivo & Batanero, 2007).

Entre las principales causas que se ofrecen como explicación de la complejidad de la inferencia estadística y el razonamiento a partir de muestras, destacan la multiplicidad de conceptos abstractos que se entrelazan en una inferencia (Chance, delMas & Garfield, 2004; Pfannkuch, Wild & Parsonage, 2012); el enfoque formal deductivo a través del cual se ha abordado la enseñanza de la inferencia (Lipson, 2002); y la dificultad para ver las muestras y cálculo de estadísticos como eventos estocásticos, que en un muestreo repetido presentan una distribución que revela información importante para hacer la estimación de un parámetro (Saldanha & Thompson, 2014). Un acuerdo generalizado entre investigadores, orienta a reemplazar o complementar el enfoque formal por un enfoque más conceptual y más accesible que brinde oportunidad a los estudiantes de comprender las grandes ideas que subyacen a la inferencia estadística (Cobb & Moore, 1997; Wild, Pfannkuch &

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

Reagan, 2011); este enfoque es conocido como *inferencia estadística informal,* y el razonamiento que lo caracteriza como *razonamiento inferencial informal.* Entre sus objetivos está generar comprensión de los conceptos de la inferencia sin depender de los métodos formales basados en la teoría estadística y la probabilidad.

El avance de las tecnologías digitales proporciona grandes posibilidades para generar este cambio de enfoque en el estudio de la inferencia estadística, dado el carácter dinámico, interactividad, múltiples representaciones y capacidad de simulación que caracterizan a algunas tecnologías educativas, lo cual les confiere un potencial cognitivo que permite visualizar e interactuar con las representaciones de los datos, el proceso de muestreo, el cálculo de estadísticos y su distribución muestral; objetos matemáticos complejos a partir de los cuales se construyen los intervalos de confianza, el margen de error y la confiabilidad. En este contexto, nos hemos propuesto analizar el razonamiento inferencial informal que desarrollan estudiantes universitarios de ciencias sociales sobre la variabilidad y los intervalos de confianza en un ambiente computacional como que el provee el software TinkerPlots (Konold & Miller, 2011). En específico, nos interesa investigar si los estudiantes identifican relaciones correctas sobre el muestreo, el efecto de tamaño de muestra y la confiabilidad en los intervalos de confianza, y si logran interpretar correctamente el margen de error y la confiabilidad en una estimación.

Marco conceptual

Una inferencia estadística es una aseveración sobre una población, la cual es generada a partir de una sola muestra y con un nivel explícito de confianza. El razonamiento inferencial informal involucra ideas y relaciones como centralidad, variabilidad, tamaño de muestra y control de sesgo (Rubin, Hammerman & Konold, 2006), y se define como la habilidad para interconectar ideas de distribución, muestreo y centralidad, dentro de un ciclo de razonamiento empírico (Pfannkuch, 2006). Zieffler, Garfield, delMas y Reading (2008) lo definen como la forma en la que los estudiantes usan su conocimiento estadístico informal para hacer argumentos para apoyar inferencias acerca de poblaciones basándose en muestras. Makar, Bakker y Ben-Zvi (2011) identifican una serie de elementos clave interrelacionados que son necesarios para apoyar el razonamiento inferencial informal, como son: el conocimiento estadístico, el conocimiento del contexto del problema, normas y hábitos desarrollados con el tiempo y ambientes de aprendizaje basados en cuestionamientos e investigación.

En el contexto de los intervalos de confianza, Pfannkuch, Wild y Parsonage (2012) proponen una ruta conceptual para desarrollar la idea intervalo de confianza desde una perspectiva informal utilizando técnicas de simulación, y definen una comprensión estocástica de los intervalos de confianza como un proceso que contempla las siguientes etapas:

- Concebir un proceso de muestreo aleatorio como la selección de una cantidad de elementos de una población y el registro de cada dato de los elementos seleccionados, para después calcular un estadístico de la muestra (por ejemplo la media o mediana) y estimar el parámetro de la población.
- Imaginar repetidamente la selección de muestras de un tamaño dado y determinar si el intervalo de confianza calculado de la muestra, "captura" el valor del parámetro.
- Comprender que este proceso producirá una colección de resultados de la forma "captura" o "no captura" el verdadero valor del parámetro.
- Comprender que en el muestreo aleatorio existe variabilidad en los resultados, pero conforme se incrementa el tamaño de la muestra, la distribución de resultados adquiere una forma más estable y centra en el verdadero valor del parámetro.
- La proporción del resultado "captura" en una larga corrida es el nivel de confianza asociado al método.

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

Metodología

La investigación se llevó a cabo con 15 estudiantes de ciencias sociales que tomaban un curso básico de probabilidad y estadística. Los estudiantes tenían pocos antecedentes matemáticos en la materia, por lo que decidimos enfocar el curso hacia la modelación y simulación de eventos aleatorios y muestreo de poblaciones utilizando el ambiente computacional que proporciona el software TinkerPlots, tomando como referencia contextos reales de estudios de opinión en el área de ciencias sociales publicados por empresas encuestadoras mexicanas. Por cuestiones de espacio en el presente trabajo se analizan y discuten resultados de una de las últimas actividades del curso. Como instrumentos de recolección de información se utilizaron hojas de trabajo para cada actividad, archivos del software y entrevistas con algunos estudiantes.

El software TinkerPlots permite el análisis y visualización de datos en forma dinámica e interactiva, con un gran potencial para la modelación y simulación de eventos aleatorios, como es el caso del muestreo. Para el caso específico de los intervalos de confianza, el software dispone de una herramienta de modelación conocida como "Sampler", en la cual, a través de mecanismos aleatorios (ruletas, urnas, diagramas de barras) los usuarios generan el modelo de una población y sus parámetros; posteriormente extraen una gran cantidad de muestras y visualizan el proceso de muestreo y cálculo de estadísticos conforme éste se desarrolla, para generar la distribución muestral del estadístico en cuestión, en forma gráfica o tabular. Las actividades de enseñanza se diseñaron con el propósito de desarrollar en los estudiantes una concepción estocástica de los intervalos de confianza y desarrollar un razonamiento informal adecuado sobre los conceptos como el muestreo, tamaño de muestra, distribución muestral, confiabilidad y margen de error.

Resultados y discusión

Para el diseño de la actividad y el análisis de los resultados hemos tenido en cuenta las etapas y procesos definidos por Pfannkuch, Wild & Parsonage (2012) para desarrollar una comprensión estocástica de los intervalos de confianza.

Actividad:

El tema de la legalización del consumo de mariguana en México ha generado opiniones contrarias en sectores de la sociedad. La empresa Parametría realizó una encuesta para estimar la opinión de los mexicanos (http://www.parametria.com.mx/carta_parametrica.php?cp=4816). Utilizó una muestra aleatoria de 800 personas mayores de edad y reporta una confiabilidad de 95% y una margen de error de $\pm 3.5\%$ en los resultados.

Tabla 1: Resultados de la encuesta

Opinión sobre la legalización de la marihuana	Porcentaje
A favor de	20%
En contra	**77%**
No sabe aún	3%

Considera los resultados anteriores como los parámetros de la población objetivo, en particular considera la proporción de mexicanos que están en contra de la legalización de la marihuana, esto es, P=0.77.

Población, muestreo y variabilidad muestral

El punto de partida consiste en formular el modelo de la población, para después extraer muestras aleatorias y explorar la relación entre los resultados muestrales con los parámetros de la población. Los estudiantes formularon el modelo con facilidad, como resultado de su experiencia en los temas

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

de probabilidad previamente vistos (ver figura 1).

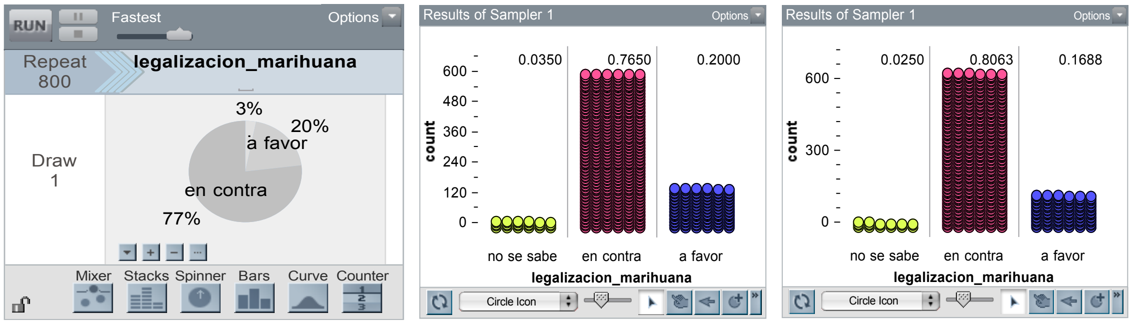


Figura 1. Modelo poblacional y dos muestras aleatorias (n=800) extraídas de la población.

La comparación de los resultados de varias muestras con los valores poblacionales permitió a los estudiantes identificar la variabilidad como una característica intrínseca del muestreo y formular un intervalo intuitivo razonable de resultados esperados. Como evidencia se presentan las respuestas que María José (MJ) y Andrea (A) proporcionaron al investigador (R) en una entrevista:

R: Una vez que construiste el modelo de la población, en la primera muestra de 800 personas [tal como lo hizo la encuestadora] obtuviste un proporción en contra de la legalización de 0.76, ¿te pareció razonable el resultado?

MJ: Sí, me parece que no varía mucho, si tomamos en cuenta que el valor verdadero de la proporción es 0.77.

R: Si repites el muestreo, ¿esperas tener resultados iguales o diferentes?

MJ: Espero resultados diferentes, pero no muy alejados de 0.77.

R: ¿Un intervalo razonable en el cual se esperas los resultados de las muestras?

MJ: Como mínimo 0.74 y como máximo 0.80.

R: ¿Por qué lo consideras así?

MJ: Porque el margen de error que proporciona la encuestadora es del 3.5%, entonces podemos tomar el 0.77 como punto medio y sumar y restar el margen de error.

R: Si en lugar de 800 personas en la encuesta se hubieran utilizado 1500 personas, ¿crees que hubiera resultado el mismo intervalo?

MJ: Los porcentajes se elevarían, el 0.77 quizá sería más grande porque la muestra es más amplia, pero también podría bajar porque se está preguntando a más personas.

R: ¿El intervalo entre 0.74 y 0.80, sería el mismo?

MJ: El margen de error disminuiría.

R: ¿Los resultados que obtuviste en las tres muestras te parecen razonables?

A: Sí, porque el valor verdadero es 0.77, no varían mucho del parámetro.

R: ¿Podrías establecer un intervalo razonable de variación para los resultados muestrales?

A: Del 0.75 al 0.79.

R: ¿En qué te basaste para establecer el intervalo?

A: Consideré que no puede ser un margen de error tan grande, si el parámetro es 0.77.

R: Por ejemplo, sin en lugar de 800 personas se hubieran encuestado 1500, ¿que pasaría con el intervalo?

A: Sería más estrecho.

Las respuestas de Andrea (A) y María José (MJ) muestran que tienen una idea correcta de la variabilidad muestral alrededor del parámetro, y que ésta disminuye conforme se incrementa el tamaño de la muestra. Construyen un razonable intervalo intuitivo de variación de los resultados muestrales esperados. En el caso de María José relaciona el intervalo esperado con el margen de error

Articles published in the Proceedings are copyrighted by the authors.

de la encuesta de manera correcta, lo cual significa que tienen un idea de intervalo formado por el estimador y la suma y resta del margen de error. Ante la pregunta sobre incrementar el tamaño de muestra, no tiene claro el efecto que tendría en la estimación, pues le atribuye mayor variabilidad, cuando en realidad muestras más grandes deben parecerse más a la población.

Distribución muestral, confiabilidad y margen de error

TinkerPlots permite visualizar el muestreo como un proceso repetible, calcular el estadístico en cada muestra y acumular los resultados en una tabla que posteriormente puede ser graficada; es decir, genera la distribución muestral para una cierta cantidad de muestras (ver figura 2).

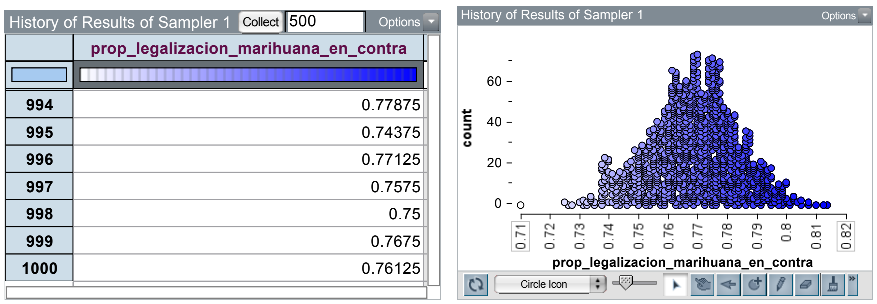
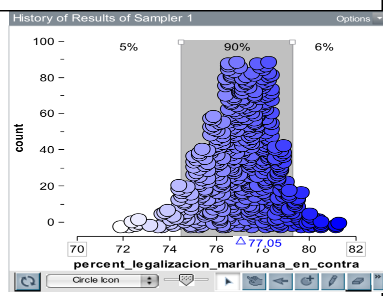
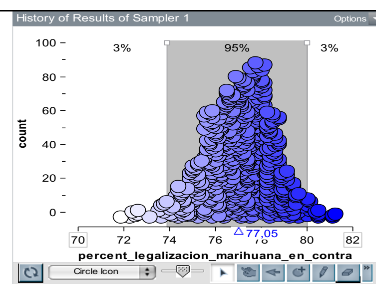
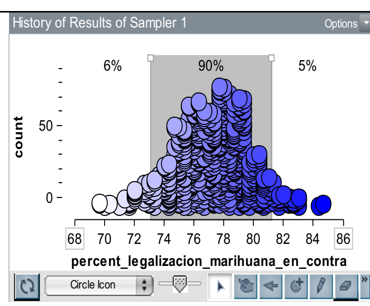
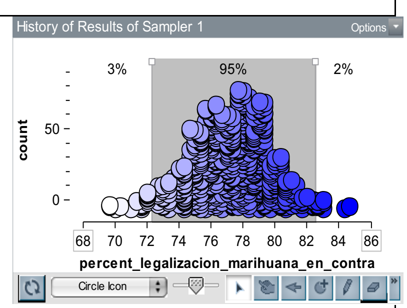


Figura 2. Distribución muestral para 500 muestras de tamaño 800 (P=0.77).

Los estudiantes seleccionaron muestras de tamaño 800 (como lo hizo la empresa encuestadora), y muestras de tamaño 300 (otra posible empresa), con el fin de comparar las distribuciones muestrales y ver el efecto del tamaño de muestra. Se agregaron bandas que sombrean una parte de la distribución muestral y que hacen el papel de intervalos gráficos capturando 90% y 95% de las muestras respectivamente (ver tabla 2).

Tabla 2: Comparación de distribuciones muestrales (n=800 y n=300) y niveles de confiabilidad (90% y 95%)

n	Confiabilidad 90%	Confiabilidad 95%
800		
300		

La comparación de distribuciones para cada confiabilidad y tamaño de muestra, permitió a los

Articles published in the Proceedings are copyrighted by the authors.

estudiantes identificar algunas relaciones importantes como se muestra en las respuestas que dieron Perla y Katya en la hoja de trabajo, y Andrea (A) y Anaid (AN) en la entrevista:

> “Entre más grande es la confiabilidad los intervalos serán más grandes”. Perla

> “El comportamiento del tamaño de la muestra en relación con la amplitud de los intervalos es a la inversa que con la confiabilidad. La distribución de las muestras de 800 es más angosta que la distribución de muestras tamaño 300”. Katya

R: ¿Qué efecto tiene el incrementar la confianza en el ancho del intervalo?
A: Entre menor es la confiabilidad se hace más estrecho el intervalo.
R: ¿Qué ventajas crees que tendría un estudio con una confiabilidad alta?
A: Que cierto porcentaje de la población muy probablemente cae dentro de ese intervalo. Pero si el intervalo es muy amplio puede que no sea muy útil, ya que intervalos grandes son más confiables pero menos precisos.
R: ¿Que se puede hacer para aumentar la precisión?
A: El tamaño de muestra se debe aumentar.
R: ¿Qué nivel de confianza preferirías en un estudio: 90% o 95%?
AN: Es mejor el de 90% porque el margen de error es mas pequeño.
R: ¿Acaso no es mas confiable uno de 95%?
AN: En el de 90 tienes menos posibilidades que caiga dentro y el 95 es más grande y tiene mas posibilidades.
R: ¿Qué pasa con el aumento del tamaño de la muestra en el intervalo?
AN: El ancho del intervalo aumenta al bajar el tamaño de muestra.

Las respuestas de las estudiantes muestran que han identificado correctamente el efecto de la confiabilidad y el tamaño de muestra en el ancho de un intervalo. Sin embargo, el significado de confiabilidad para Andrea es erróneo, al considerar que representa un porcentaje de la población que caerá dentro del intervalo, una concepción muy persistente ya documentada en otros estudios (Olivo & Batanero, 2007).

Otra idea importante que nos propusimos explorar es la aleatoriedad de un intervalo, esto significa de una muestra a otra los limites y el ancho del intervalo pueden cambiar. Para esta parte de la actividad nos propusimos desarrollar en la hoja de cálculo de TinkerPlots los cálculos que se involucran en un intervalo de confianza para una confiabilidad de 90% y 95% respectivamente y repetir la simulación para 500 o 1000 muestras (ver figura 3).

History of Results of Sampler 1 — Collect 500 — Options

	prop_legaliza...	Error_estandar	margen_de_error	limite_inferior	limite_superior	resultado
478	0.7375	0.0155561	0.0311122	0.706388	0.768612	Cae fuera
479	0.7725	0.0148216	0.0296432	0.742857	0.802143	Cae dentro
480	0.765	0.0149906	0.0299812	0.735019	0.794981	Cae dentro
481	0.77	0.0148787	0.0297574	0.740243	0.799757	Cae dentro
482	0.75375	0.015232	0.030464	0.723286	0.784214	Cae dentro

Figura 3. Hoja de cálculo con los elementos de un intervalo de confianza (500 intervalos generados).

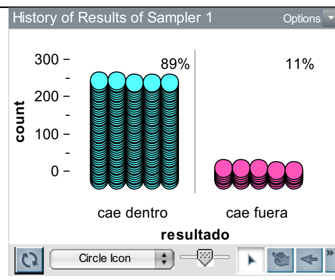
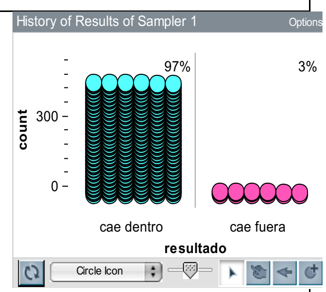
A continuación se muestran las respuestas de Andrea (A) y Anaid (AN):

R: ¿Al repetir el muestreo esperas que salga el mismo intervalo?
AN: No, porque las muestras son aleatorias.
R: ¿Varían todos los elementos de un intervalo?
AN: Si porque cada vez que lo corres da un P diferente, aunque cercano.

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.
Articles published in the Proceedings are copyrighted by the authors.

R: ¿Que relación crees que tiene el 97% de resultados capturados con la confiabilidad? (ver tabla 3).
AN: La confiabilidad es del 95% y ese es muy cercana a 97%, puede incluso ser igual.
R: El 96% de los intervalos que simulaste cayeron dentro y un 4% cayó fuera, ¿tiene alguna relación con la confiabilidad de 95%?
A: Si recuerdo que los resultados que caen fuera el intervalo son lo no contienen al parámetro
R: ¿Eso podría suceder en una encuesta real?
A: Si
R: ¿Lo tomarías como un error de la encuestadora?
A: No, como algo que pasa por el azar, y que sucede con poca frecuencia.
R: ¿En una distribución muestral esos valores donde los ubicarías?
A: En los extremos de la distribución.
R: Viendo estos resultados, ¿qué significa la confiabilidad?
A: Es el porcentaje que un encuestador puede decir que sus muestras contienen al parámetro, que son verdaderas.

Tabla 3: Gráficas con porcentajes de intervalos que capturan y no capturan al parámetro para los niveles de confiabilidad 90% y 95%

<table>
<tr><th></th><th>Confiabilidad 90%</th><th>Confiabilidad 95%</th></tr>
<tr><td>Gráficas de la columna resultados</td><td>
</td><td>
</td></tr>
</table>

Las respuestas de Anaid y Andrea señalan que tienen claro la aleatoriedad de un intervalo, porque depende de los resultados variables de una muestra. No logran establecer un significado correcto sobre la confiabilidad pese a que visualizan en la gráfica los porcentajes de muestras que capturan y no capturan al parámetro, respectivamente. Sin embargo, cabe resaltar que Andrea está consciente que en una encuesta real se pueden presentar intervalos que no capturan al parámetro, los considera poco frecuentes y los ubica correctamente en las colas de una distribución muestral.

Conclusiones

Los resultados señalan que los estudiantes razonaron correctamente sobre algunos conceptos que integran una comprensión estocástica de los intervalos de confianza definidos por Pfannkuch, Wild & Parsonage (2012), tales como la relación entre el tamaño de muestra y la variabilidad muestral, el efecto del nivel de confiabilidad y el tamaño de muestra en el ancho de un intervalo de confianza. Lograron identificaron intervalos razonables de los resultados esperados en la muestra e identificaron además, el carácter aleatorio de un intervalo, conceptos que se reportan como complejos por investigaciones previas. Sin embargo la confiabilidad resultó ser un concepto muy difícil para todos los estudiantes, y no lograron conceptualizarlo correctamente, aún cuando las actividades enfatizaron en la repetición de muestras para visualizar el porcentaje de intervalos que capturan al parámetro y relacionarlo con la confiabilidad previamente establecida. El ambiente computacional como el que proporciona TinkerPlots en complemento con actividades que promueven la relación explícita entre los conceptos que intervienen en un intervalo de confianza parecen ser adecuados para el diseño de

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

trayectorias de aprendizaje que promueven un razonamiento inferencial informal correcto en los estudiantes.

This article presents the results of a qualitative research with a group of 15 university students of social sciences on informal inferential reasoning developed in a computer environment on concepts involved in the confidence intervals. The results indicate that students developed a correct reasoning about sampling variability and visualized reasonable intervals of variability in a repeated sampling, at the same time students identified correct relationships between sample size and confidence level in the width of an interval and margin of error, and identified the randomness of a confidence interval. However, they had difficulties conceptualizing the confidence level as the percentage of intervals that capture the parameter in a sampling repeated under identical conditions.

Keywords: Data Analysis and Statistics, Technology, Modeling

Introduction

Making inferences about populations and interpreting results of statistical studies has become part of people´s professional and daily lives. A concrete and very recurrent example is the opinion polls that appear almost daily in the media, in which estimates of population parameters, margin of error and confidence level, among other concepts are reported. The research reports that the concepts and reasoning that characterizes the statistical inference are complex for most students (Castro Sotos, Vanhoof, Noortgate, & Onghena, 2007), even for teachers and researchers who apply it in their profession (Liu & Thompson, 2004).

In the particular case of parameter estimation by confidence intervals, the literature reports diverse difficulties of understanding and errors in the interpretation of results. For example, a persistent error is to consider that a 95% confidence level indicates a 95% probability that the population parameter is between the limits of the interval. Other errors are not recognizing the randomness and inferential nature of the interval, and ignoring the effect of sample size and the confidence level in the width of the interval and the margin of error; believing that in different samples, the same interval will be obtained, among others (Olivo & Batanero, 2007).

Among the main causes offered as an explanation of the complexity of statistical inference and reasoning from samples, those that stand out are the multiplicity of abstract concepts that are intertwined in an inference (Chance, delMas, & Garfield, 2004; Pfannkuch, Wild, & Parsonage, 2012); deductive formal approach used to teaching inference (Lipson, 2002); and the difficulty to recognize samples and computation of statistics as stochastic events, which in a repeated sampling show a distribution that reveals important information for estimating a parameter (Saldanha & Thompson, 2014). A generalized agreement among researchers, aims to replace or supplement the formal approach with a more conceptual and more accessible approach that provides opportunity for students to understand the great ideas behind statistical inference (Cobb & Moore, 1997; Wild, Pfannkuch, & Reagan, 2011); this approach is known as an informal statistical inference, and reasoning that characterizes, informal inferential reasoning. One of its objectives it is to generate understanding of the inference concepts without relying on formal methods based on statistical theory and probability.

Advances in digital technologies provide great potential to generate this change of approach to the study of statistical inference, given the dynamic properties, interactivity, multiple representations and simulation capabilities that characterize some educational technologies. These technological advances allow for a cognitive potential to visualize and interact with representations of the data, the sampling process, the calculation of statistics measures and their sampling distribution; complex mathematical objects from which confidence intervals, the margin of error and confidence level are

Articles published in the Proceedings are copyrighted by the authors.

built. In this context, we intend to analyze the informal inferential reasoning that university students of social sciences develop on the variability and confidence intervals in a computer environment such as that provided by the TinkerPlots software (Konold & Miller, 2011). Specifically, we want to investigate whether students identify correct relations on sampling, the effect of sample size and confidence level in the confidence intervals, and whether or not students correctly interpret the error and confidence level in the estimation of a parameter.

Conceptual framework

A statistical inference, is a statement about a population, which is generated from a single sample and with an explicit confidence level. The informal inferential reasoning involves ideas and relationships as center, variability, sample size, and control of bias (Rubin, Hammerman & Konold, 2006), and is defined as the ability to interconnect ideas of distribution, sampling and centrality, within a cycle of empirical reasoning (Pfannkuch, 2006). Zieffler, Garfield, delMas, and Reading (2008) define it as the way that students use their informal statistical knowledge to make arguments to support inferences about populations based on samples. Makar, Bakker, and Ben-Zvi (2011) identify a number of interrelated key elements that are needed to support the informal inferential reasoning, such as: statistical knowledge, knowledge of the context, rules and habits developed over time and learning environments based on questions and research.

In the context of the confidence intervals, Pfannkuch, Wild, and Parsonage (2012) propose a conceptual pathway to develop the idea of confidence interval from an informal approach using simulation techniques, and define a stochastic understanding of confidence intervals as a process that includes the following steps:

- Conceiving a random sampling process as selecting a number of elements of a population, and recording each data of the selected elements, then calculating a statistics measure (e.g. the mean or median) to estimate the population parameter.
- Imagining repeatedly taking samples of a given size and determinate whether or not the confidence interval calculated for each sample, "covers" the parameter value.
- Understanding that this process will produce a collection of outcomes that would either "cover" or "not cover" the true parameter value.
- Understanding that because of the random selection process there is variability in the outcomes, but as the sample size increases, the distribution of outcomes becomes stable and centered at the true parameter value.
- The long run proportion of ‘‘covers’’ is the confidence level associated with the method.

Methodology

The research was conducted with 15 social science students taking an introductory course in probability and statistics. The students had little mathematical background in the subject, therefore we decided to focus the course on the modeling and simulation of random events and population sampling using the computer environment that provides TinkerPlots software, using real contexts of opinion polls in the area of social sciences published by Mexican pollsters. Due to space limitation, in this paper we analyze and discuss results from the last course activities. The data collection instruments used were worksheets for each activity, computer activity files, and interviews with some students.

TinkerPlots software enables analysis and visualization of data in dynamic and interactive way, with great potential for modeling and simulation of random events, such as sampling. For the specific case of confidence intervals, the software has a modeling tool known as "Sampler," which, through random mechanisms (spinners, urns, bar charts), generates the model of a population and its parameters. Following, users select a large amount of samples and visualize the sampling process and

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

statistics measures calculation (e.g. mean, median) as they develop, to generate the sampling distribution of the statistic in question, in graphical or tabular form. The teaching activities were designed with the purpose of developing in students a stochastic conception of confidence intervals and developing an appropriate informal reasoning on concepts such as sampling, sample size, sampling distribution, confidence level, and margin of error.

Results and Discussion

For the design of the activity and the analysis of the results, we have taken into account the stages and processes defined by Pfannkuch, Wild, and Parsonage (2012) to develop a stochastic understanding of confidence intervals.

Activity

The topic of legalization of marijuana in Mexico has generated conflicting opinions in diverse sectors of society. Parametría Company conducted a survey to estimate the opinion of Mexicans (http://www.parametria.com.mx/carta_parametrica.php?cp=4816). It used a random sample of 800 people and reported a confidence level of 95% and a margin of error of ± 3.5% in the results.

Table 1: Survey results

Opinion on the legalization of marijuana	Percent
Agree	20%
Against	**77%**
Do not know yet	3%

Consider the above results as the parameters of the target population, particularly considering the proportion of Mexicans who are against the legalization of marijuana, that is, P = 0.77.

Population, sampling and sampling variability

The starting point consists in formulating the population model with the purpose of taking random samples and exploring the relationship between the sample results with the population parameters. Students formulated the model easily due to their experience on the issues of probability previously studied (see Figure 1).

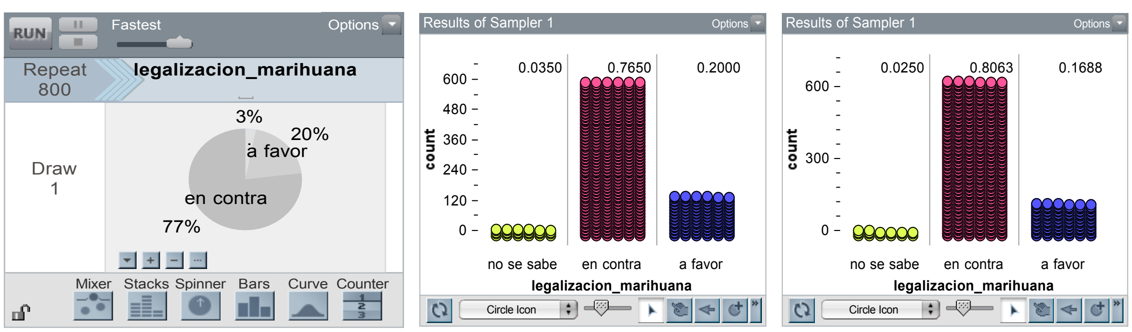


Figure 1. Population model and two random samples (n = 800) drawn from the population.

Comparing the results of several samples with population values, allowed students to identify the variability as an intrinsic feature of the sampling and formulate an intuitive and reasonable interval of expected results. As evidence we present the answers that Maria Jose (MJ) and Andrea (A) provided to the researcher (R) in an interview:

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

R: Once you built the population model, in the first sample of 800 people [as the pollster did] you obtained a proportion against the legalization of 0.76. Did the result seem reasonable to you?
MJ: Yes, I think that it is does not vary much, if we consider that the true value of the proportion against is 0.77.
R: If the sampling is repeated, do you expect to have the same or different results?
MJ: I expect different results, but not far from 0.77.
R: A reasonable interval in which you expect the results of the samples?
MJ: 0.74 minimum and maximum 0.80.
R: Why do you think so?
MJ: Because the margin of error provided by the pollster is 3.5%, then we can take 0.77 as the midpoint and add and subtract the margin of error.
R: If instead of 800 people surveyed, 1500 people had been used in the survey, do you think that would have been the same interval?
MJ: The percentages would rise, perhaps 0.77 would be bigger because the sample is wider, but it could also lower because more people are been surveyed.
R: The interval between 0.74 and 0.80, would be the same?
MJ: Decrease the margin of error
R: Do the results you obtained in the three samples seem reasonable to you?
A: Yes, because the true value is 0.77, it does not vary much from the parameter.
R: Could you establish a reasonable range of variation for sample results?
A: From 0.75 to 0.79.
R: On what base did you set the interval?
A: I considered that it may not be so great a margin of error, if the parameter is 0.77.
R: For example, if instead of 800 people they had surveyed 1500, what would happen to the interval?
A: It would be narrower.

The responses of Andrea (A) and Maria Jose (MJ) show that they have a correct idea of the sampling variability around the parameter, and it decreases as the sample size increases. They built an intuitive and reasonable interval of variation of the expected sample results. In the case of Maria Jose, she relates the expected interval with the margin of error in the survey correctly, which means they have an idea of interval formed by the estimator, adding and subtracting the margin of error. When asked about increasing the sample size, she is not clear on the effect of the estimate, because she attributed a greater variability, when in fact, larger samples should be closer to the population.

Sampling distribution, reliability and margin of error

TinkerPlots displays the sampling as a repeatable process, calculate a statistics measure in each sample and accumulate the results in a table that one can then graph; that is, it generates the sampling distribution for a certain amount of samples (see Figure 2).

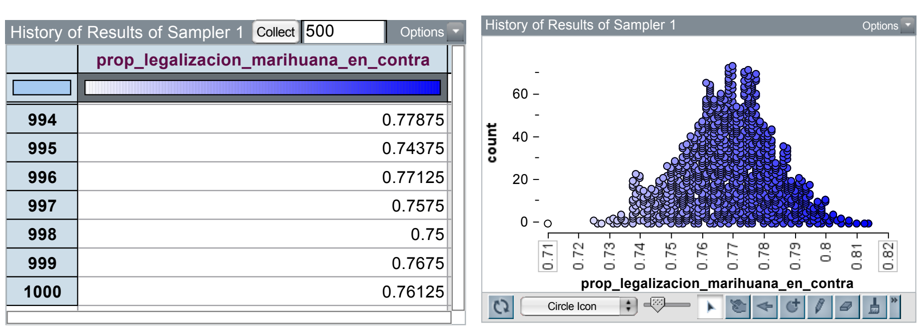


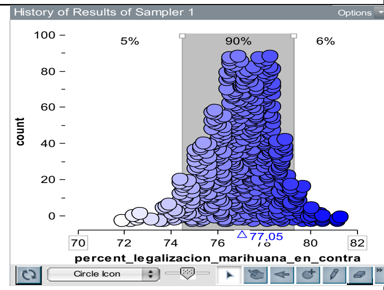
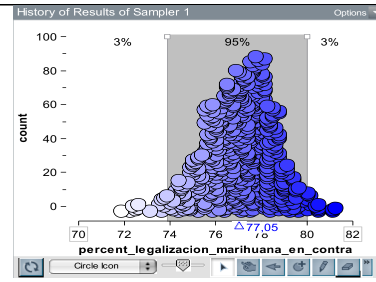
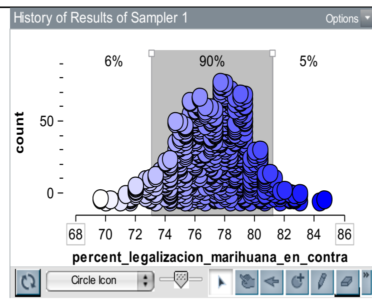
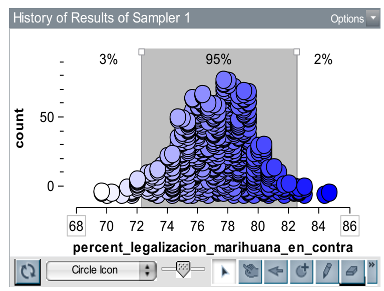
Figure 2. Sampling distribution for 500 samples of size 800 (P = 0.77).

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

Students selected samples size of 800 (as the pollster did), and samples size of 300 (another possible company), in order to compare the sampling distributions and see the effect of sample size. Shading bands were added to a portion of the sampling distribution that make the role of graphics intervals capturing 90% and 95% of the samples respectively (see Table 2)

Table 2: Comparison of sampling distributions (n = 800 and n = 300) and confidence levels (90% and 95%)

n	Confidence level 90%	Confidence level 95%
800		
300		

The comparison of distributions for each confidence level and sample size, allowed students to identify some important relationships as shown in the answers given by Perla and Katya in the worksheet, and Andrea (A) and Anaid (AN) in the interview:

"The larger confidence level, the larger intervals will be." Perla

"The behavior of the sample size in relation to the width of the intervals is reversed with respect to the confidence level. The sampling distribution of size 800 is narrower than the sampling distribution of size 300." Katya

R: What is the effect of increasing the confidence level on the width of the interval?
A: The lower is the confidence is, the narrower the interval becomes.
R: What advantages do you think a high confidence level study would have?
A: That certain percentage of the population likely falls within that interval. But if the interval is very wide, it may not be very useful, since large intervals are more reliable but less precise.
R: What can be done to increase the accuracy?
A: We should increase the sample size.
R: What confidence level would you prefer in a study: 90% or 95%?
AN: 90% is better, because the margin of error is smaller.
R: Is it not an interval of 95% more reliable?
AN: At interval of 90% you are less likely to fall within, and 95% is larger and has more possibilities.

Wood, M. B., Turner, E. E., Civil, M., & Eli, J. A. (Eds.). (2016). *Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Tucson, AZ: The University of Arizona.

Articles published in the Proceedings are copyrighted by the authors.

R: What happens to the interval when you increase the sample size?
AN: The width of the interval increases by lowering the sample size.

The responses by the students show that they have correctly identified the effect of confidence level and sample size in the width of an interval. However, the meaning of confidence level for Andrea is wrong, considering that it represents a percentage of the population that falls within the interval, a very persistent misconception documented in other studies (e.g. Olivo & Batanero, 2007).

Another important idea that we proposed to explore is the randomness of an interval, this means that the limits and width of the interval may change from one sample to another. For this part of the activity, we decided to develop, in the TinkerPlots' spreadsheet, the calculations involved in a confidence interval for a confidence level of 90% and 95% respectively, and repeat the simulation for 500 or 1000 samples (see Figure 3).

History of Results of Sampler 1 — Collect 500 — Options

	prop_legaliza...	Error_estandar	margen_de_error	limite_inferior	limite_superior	resultado
478	0.7375	0.0155561	0.0311122	0.706388	0.768612	Cae fuera
479	0.7725	0.0148216	0.0296432	0.742857	0.802143	Cae dentro
480	0.765	0.0149906	0.0299812	0.735019	0.794981	Cae dentro
481	0.77	0.0148787	0.0297574	0.740243	0.799757	Cae dentro
482	0.75375	0.015232	0.030464	0.723286	0.784214	Cae dentro

Figure 3. Spreadsheet with elements of a confidence interval (500 intervals generated).

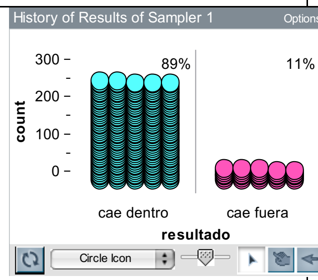
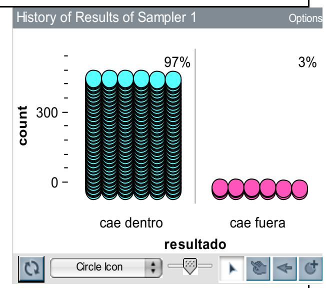
The answers of Andrea (A) and Anaid (AN) are shown below:

R: When you repeat the sample, do expect the same interval?
AN: No, because the samples are random.
R: Do all the elements of an interval vary?
AN: Yes, because every time you run it, it gives a different P, although it is close.
R: What relationship do you think 97% of captured results has with the confidence level? (See Figure 3).
AN: The confidence level is 95% and that is very close to 97%, it may even be the same.

R: In your simulation, 4% of the intervals did not capture the parameter; Does that have any relationship with the confidence level of 95%?
A: Yes, I remember that the results that fall outside the interval do not contain the parameter.
R: Could that happen in a real survey?
A: Yes
R: Would you take it as a pollster error?
A: No, I would take it as something that happens by chance, and that it happens infrequently.
R: Where would you locate those values in a sampling distribution?
A: At the ends of the distribution.
R: Seeing these results, what does reliability means?
A: It is the percentage that a pollster can say that their samples contain the parameter, which are true.

Articles published in the Proceedings are copyrighted by the authors.

Table 3: Charts with percentages of intervals that "covers" and "not covers" the parameter for confidence levels 90% and 95%

	Confidence level 90%	Confidence level 95%
Charts of the results column		

The responses of Anaid and Andrea indicate that they are clear about the randomness of the interval, because it depends on the variable results of a sample. They are unable to establish the correct meaning of the confidence level despite the fact that the graph displays the percentage of samples that cover and do not cover the parameter, respectively. However, it should be noted that Andrea is aware that in a real survey, there may be intervals that do not cover the parameter, she considered them rare and located them correctly in the tails of a sampling distribution.

Conclusions

The results indicate that the students reasoned correctly on some concepts that integrate a stochastic understanding of confidence intervals defined by Pfannkuch, Wild, and Parsonage (2012), such as the relationship between sample size and sample variability, the effect of confidence level and sample size on the width of a confidence interval. The students identified reasonable intervals of expected results in a sample and the random nature of an interval, concepts reported as complex by previous studies. However, confidence level proved to be a very difficult concept for all students, and they failed to conceptualize it correctly, even when activities emphasized repetition of samples to display the percentage of intervals that covers the parameter, and relating it to the previously established confidence level. The computer environment such as that provided by TinkerPlots joint with activities that promote the explicit relationship between the concepts involved in a confidence interval seem to be suitable for the design of learning pathways that promote a correct informal inferential reasoning in the students.

References

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

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

Cobb, G. & Moore, D. (1997). Mathematics, statistics and teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

Konold, C. y Miller, C. (2011). *TinkerPlots Dynamic Data Exploration Software*. Emeryville, CA: Key Curriculum Press Technologies.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics.* [CD-ROM]. Cape Town South Africa.

Liu, Y., & Thompson, P. (2004). Teachers' personal and pedagogical understanding of probability and statistical inference. In D. E. McDougall & J. A. Ross (Eds.), *Proceedings of the Twenty-Sixth Annual Meeting of the*

Articles published in the Proceedings are copyrighted by the authors.

North American Chapter of the International Group for the Psychology of Mathematics Education (Vol 1., pp. 408-414). Toronto, Canada: Ontario Institute for Studies in Education of the University of Toronto.

Makar, K., Bakker, A., & Ben-Zvi, D. (2013). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1-2), 152-173.

Olivo, E. & Batanero, C. (2007). Un estudio exploratorio de dificultades de comprensión del intervalo de confianza. *Revista Unión*, núm. 12, www.sinewton.org/numeros/

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* [CD-ROM]. Salvador Bahía, Brazil.

Pfannkuch, M., Wild, Ch. & Parsonage, R. (2012). A conceptual pathway to confidence intervals. *ZDM Mathematics Education*, *44*(7), 899-911.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*. [CD-ROM]. Salvador Bahía, Brazil.

Saldanha, L. A., & Thompson, P. W. (2014). Conceptual issues in understanding the inner logic of statistical inference: Insights from two teaching experiments. *Journal of Mathematical Behavior*, *35*, 1–30.

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society Series A*, *174*, 247-295.

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. Recuperado de http://www.stat.auckland.ac.nz/serj

Articles published in the Proceedings are copyrighted by the authors.