



Formas de razonamiento en la generación de distribuciones muestrales mediante simulación computacional

Greivin **Ramírez** Arce
Escuela de Matemática, Instituto Tecnológico de Costa Rica
Costa Rica
gramirez@itcr.ac.cr

Resumen

En el presente artículo se reportan resultados de una investigación con estudiantes universitarios avanzados que estaban tomando un curso de estadística, acerca de la comprensión que poseen sobre la generación de distribuciones muestrales con el software Fathom, expresadas por medio de histogramas al obtener muestreos repetidos. Los resultados muestran niveles intermedios de comprensión y razonamiento, pues permanecieron con dificultades al definir los estadísticos en la continua búsqueda de generación de las distribuciones.

Palabras clave: distribuciones muestrales, estadística, simulación computacional.

Antecedentes y planteamiento del problema

La mediación instrumental, brindada por una calculadora, un paquete computacional (Excell, Fathom, Probability Explorer), entre otros; pueden permitir al individuo aumentar sus conocimientos sobre los procesos estocásticos. Desde el análisis de información que provee los medios para la toma de decisiones, hasta la comprensión de conceptos complejos que hace que se vuelven accesibles y palpables.

La National Council of Teachers of Mathematics (2000) con los Principios y Estándares para la Educación Matemática pone de relieve en el principio tecnológico cómo las calculadoras y computadoras son herramientas que proporcionan imágenes matemáticas visuales que facilitan la organización y análisis de datos. Además, permite al estudiante tomar decisiones, reflexionar, razonar y resolver problemas (funciones psicológicas de nivel superior, según Vigotsky en Kozulin, 2000). Agrega en el estándar de Análisis de Datos y Probabilidad que:

Se recomienda que los alumnos aprendan a recoger sus propios datos y organizar los ajenos, representarlos en gráficos y diagramas y que con estos hagan inferencias y obtengan conclusiones. El estudiante aprenderá que las soluciones a algunos problemas estadísticos dependen de las hipótesis iniciales y tienen cierto grado de incertidumbre. (p. 51)

En esta investigación los estudiantes buscan datos reales de indicadores de países del mundo tales como: densidad de población (hab/km²), esperanza de vida (años), alfabetización (% en edades superior a 15 años) y mortalidad infantil (núm de muertes por cada 1000 nacidos), con el fin de que estos datos tengan un contexto significativo para ellos. Luego hacen un análisis descriptivo inicial de los datos, para después dar principal importancia a la creación de distribuciones muestrales a partir del estudio de estadísticos en la extracción repetitiva de muestras. Curcio (1989) lo llama primero “leer los datos” y después “leer dentro de los datos”.

Las distribuciones muestrales son la piedra angular de la inferencia estadística (estimación de intervalos de confianza, pruebas de hipótesis, entre otros), pues se pueden deducir conclusiones de una población a partir del estudio de muestras concentrándose en un estadístico (por ejemplo la media en las distribuciones de los promedios muestrales, la proporción en la distribución de las proporciones muestrales, entre otros) (Garfield, et al, 2004, p. 295; Wild & Seber, 2000, p. 277; Ramírez, 2007)

Sin embargo, muchas veces se pone atención a los resultados finales más que a los procesos iniciales e intermedios de creación de las distribuciones (Inzunsa, 2006; Lipson, 2002). Las distribuciones de los promedios, proporciones y varianzas muestrales, son los primeros ejemplos que no se escapan de ser abarcadas como distribuciones teóricas que se reducen a consultar tablas preelaboradas sin que el estudiante tenga idea de dónde surgen y sea capaz de construirlas (Garfield, delMas & Chance, 2004, p. 295).

El paquete computacional Fathom permite evaluar los procesos subyacentes mediante la simulación en la extracción de muestras (sugerida por Shaughnessy, 1992; Burrill, 2002; Sánchez, 2002; Lipson 2002; Inzunsa, 2006; Mills, 2002; Ramírez, 2007); y con esto resolver problemas, extraer conclusiones y generalizar resultados desde la creación de las distribuciones.

La presente investigación busca evidenciar el impacto positivo de la simulación computacional en la generación de distribuciones muestrales de tal manera que las formas de razonamiento que muestran los estudiantes sean de procesos integrados.

Pregunta de investigación

Específicamente la pregunta de investigación es la siguiente:

- ¿Cuáles son las formas de razonamiento que muestran estudiantes de ingeniería sobre la generación de distribuciones muestrales con el uso de simulación computacional em Fathom?

Las actividades están intencionadas con el fin de que los estudiantes obtengan muestras repetitivas de estadísticos y vayan creando distintas distribuciones cada vez más complejas, de tal manera que comprenda y caracterice la distribución (media, desviación, variabilidad, forma y relación con la población). Se evalúa la evolución de los niveles de razonamiento que presentan los estudiantes a través del tiempo.

Fathom permiten crear diversas simulaciones con la variación de tan solo un dato, ya sea mediante la creación de diagramas de muestras, manipulación de parámetros, construcción de tablas, entre otros. Los estudiantes reciben retroalimentación inmediata bajo cualquier movimiento, con el fin de que pueda establecer comparaciones entre la distribución muestral obtenida y la distribución poblacional.

Objetivos de la investigación

El objetivo general es investigar las formas de razonamiento que muestran estudiantes de ingeniería en la generación de distribuciones muestrales basado en un enfoque frecuencial mediado con el recurso de la computadora, de tal manera que permita un desarrollo empírico de las distribuciones muestrales con la simulación que se puede realizar con Fathom.

Los objetivos específicos son: determinar las virtudes de Fathom, a través de la simulación, para desarrollar funciones cognitivas de nivel superior, además de validar el uso de actividades intencionadas con datos reales como estrategia de enseñanza que le permite al estudiante percibir o producir significado de las distribuciones muestrales que genera.

Marco conceptual

En las últimas tendencias de la investigación del razonamiento, pensamiento y cultura estadística, el desarrollo de niveles o jerarquías para describir el avance cognitivo de los sujetos de estudio, ha llegado a ser un objetivo importante de investigación (Reading & Reid, 2006; Sánchez et. al, 2008). Se considera un marco base propuesto por Garfield, delMas & Chance (1999) y adaptado por Ramírez (2007), que establece categorías de razonamiento acerca de las distribuciones muestrales.

El modelo consta de cinco categorías o niveles, las cuales se describen a continuación:

Nivel 1. Razonamiento ideosincrático (RI): Los estudiantes conocen palabras y símbolos relacionados con distribuciones muestrales, pero los usan incorrectamente y sin un entendimiento completo. Además, frecuentemente usan esas palabras y símbolos con información no relacionada.

Nivel 2. Razonamiento verbal (RV): Los estudiantes tienen un entendimiento verbal de las distribuciones muestrales, pero no pueden aplicar esto al comportamiento de los promedios, proporciones y varianzas muestrales en muestras repetidas. Los estudiantes pueden seleccionar una definición correcta y centrarse en ella, pero no entienden por ejemplo conceptos claves de cómo la variabilidad y la forma están relacionados.

Nivel 3. Razonamiento de transición (RT): Los estudiantes son capaces de identificar una o dos características del proceso de muestreo que genera las distribuciones muestrales. En particular, para la distribución de los promedios muestrales, esas características se refieren a cuatro aspectos del teorema del límite central: entender que el promedio de la distribución de los promedios muestrales es igual al promedio poblacional, que la desviación estándar de la distribución de los promedios muestrales es igual a la desviación estándar poblacional entre la raíz cuadrada del tamaño de la muestra, que la variabilidad de la distribución de los promedios muestrales disminuye conforme aumente el tamaño de la muestra y que la forma de la distribución de los promedios muestrales tiende a la distribución normal conforme aumente el tamaño de la muestra. Ya sea relacionar el promedio, o bien la tendencia a la normal, o bien la variabilidad cuando crece el tamaño de la muestra.

Nivel 4. Razonamiento de procesos (RP): Los estudiantes son capaces de identificar las características del proceso de muestreo, pero no hacen una integración total de ellas. Por ejemplo los estudiantes pueden predecir correctamente cuales distribuciones muestrales corresponden a los parámetros dados, pero no pueden explicar el proceso, y no tienen confianza completa de cuando predecir una distribución de una población dada para un tamaño de muestra dado.

Nivel 5. Razonamiento de procesos integrados (RPI): Los estudiantes tienen un entendimiento completo del proceso de muestreo y distribuciones muestrales, las reglas y los conocimientos de estocástica son coordinados. Por ejemplo, los estudiantes pueden explicar el proceso con sus propias palabras, describir por qué la distribución de los promedios muestrales puede llegar a ser más normal y tener menor variabilidad con forma aumente el tamaño de la muestra.

Se toman muestras de datos poblaciones sobre indicadores de países del mundo, de tal manera que se calcula los estadísticos promedio, proporción y variancia de cada una para formar distribuciones muestrales con el fin de obtener comparaciones con respecto a la población. Se utilizan estos estadísticos para construir otros más complejos de tal manera que permitan formar otras distribuciones. Por ejemplo, partir de las variables independientes $Z \sim N(0,1)$ y $V \sim \chi^2(v)$ para estudiar el estadístico $T = \frac{Z}{\sqrt{V/v}}$ y generar la distribución t con v grados de libertad.

Metodología

En la investigación participaron doce estudiantes de ingeniería en computación que estaban matriculados en un curso semestral de estadística. Éstos habían recibido el semestre anterior un curso de probabilidades en el que abarcaron en forma teórica distintas distribuciones de probabilidad discretas y continuas (Distribución Uniforme, Binomial, Geométrica, Hipergeométrica, Normal, T-student, Chi-cuadrado y Gamma), analizaron sus funciones de probabilidad, dedujeron de manera formal sus medias, varianzas y distribuciones acumuladas.

En la investigación se desarrollaron dos actividades de dos horas cada una sobre conocimiento de Fathom. Luego se desarrollaron dos actividades, de dos horas cada una, de instrucción en los temas: manejo de datos, construcción y análisis de graficas, medidas de tendencia central y variabilidad (este último tema que debe ser abordado según Garfield, et al., 2004, como requisito fundamental antes de abordar el tema de distribuciones muestrales). Luego se llevó a cabo dos actividades-problema, en forma individual, de dos horas cada una, sobre creación de distribuciones muestrales con el uso de Fathom, para luego institucionalizar a través de la discusión las conclusiones obtenidas con todo el grupo.

La primera actividad-problema se concentró en la generación de distribuciones muestrales de estadísticos básicos tales como la media, proporción y varianza. En la segunda se utilizaban estas distribuciones para generar otras con estadísticos más complejos.

Los contenidos que están involucrados en la investigación son: muestreo individual, efecto del tamaño de la muestra, dispersión, forma y variabilidad de la distribución poblacional, las distribuciones muestrales y de las distribuciones: uniforme, distribución binomial, distribución normal, distribución t y distribución χ^2 cuadrado,

Como material de análisis, se tiene el reporte escrito de cada una de las actividades que se realizaron, se cuenta con bitácoras, hojas de campo y los archivos computacionales construidos por los estudiantes para responder a las actividades.

La investigación es de tipo cualitativo en la que se pretende con el desarrollo de las actividades que los estudiantes establezcan las propiedades de las distribuciones muestrales creadas a partir de la simulación computacional en la toma repetitiva de muestras.

Resultados

En la segunda actividad-problema se les pide a los estudiantes que seleccionen una variable aleatoria para hacer el estudio completo de la generación de distribuciones. Se presenta el análisis de dos casos, según el trabajo desarrollado en la actividad, con el fin de explicar la clasificación en los diversos niveles.

En el primer caso Julio selecciona la variable aleatoria mortalidad infantil y presenta la distribución de sus datos:

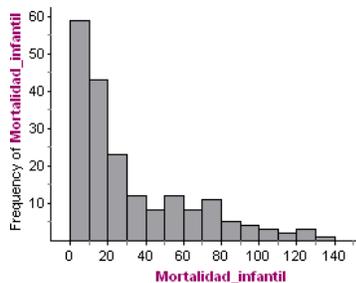


Figura 2. Histograma poblacional de la variable mortalidad infantil

Toma una muestra de cualquier tamaño (40 en este caso) y construye su gráfico:

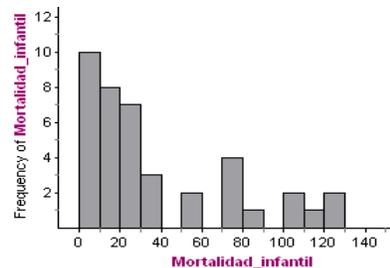


Figura 3. Histograma de una muestra de la variable mortalidad infantil

Luego calcula los estadísticos básicos de la muestra (promedio, proporción y varianza) y se obtienen 1000 muestras de tamaño 40 para formar las distribuciones de los promedios, proporciones y varianzas muestrales. Construye los histogramas respectivos y se le sugiere que construya las distribuciones teóricas sobre las mismas gráficas con el fin de que determine la distribución que siguen y les permita además, analizar la variabilidad en la extracción aleatoria de las muestras:

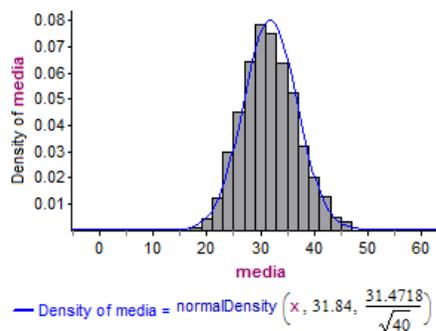


Figura 4. Distribución de promedios muestrales

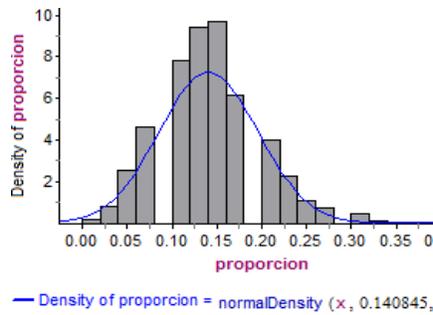


Figura 5. Distribución de proporciones muestrales

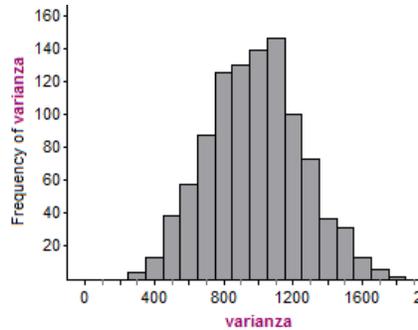


Figura 6. Distribución de varianzas muestrales

Hasta aquí Julio demuestra que entiende el proceso de creación de las distribuciones muestrales de los estadísticos básicas, intuyendo la distribución que siguen y analizando la variabilidad de los mismos, ubicándolo en el nivel de razonamiento de procesos integrados.

Luego utiliza la distribución de promedios muestrales, cuya distribución es normal (figura 4) con media $\mu = 31.84$ y desviación estándar $\frac{\sigma}{\sqrt{n}} = \frac{31.4}{\sqrt{40}}$, para crear otras distribuciones.

Toma una muestra de tamaño 15 (podría ser cualquier tamaño) y se le sugiere calcular el estadístico $X = \frac{s^2(n-1)}{\sigma^2}$. Luego se toman 1000 muestras con el fin de que se observe el comportamiento de su gráfica y deduce que sigue una distribución χ^2 con $\nu = n - 1$ grados de libertad. Gráficamente:

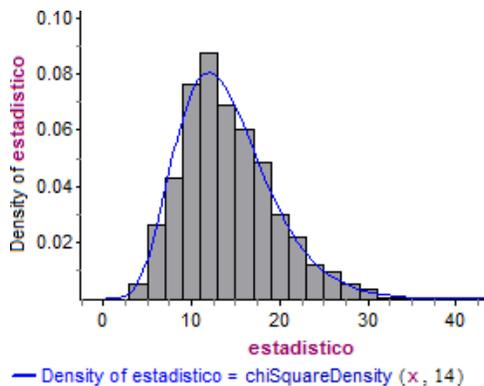


Figura 7. Distribución de varianzas muestrales

Con esta última distribución y la distribución normal (figura 4), que se estandariza a través del estadístico $Z = \frac{x - \mu}{\sigma}$ para obtener una distribución normal con media cero y desviación

estándar uno, se le sugiere obtener el estadístico $T = \frac{Z}{\sqrt{\frac{X}{n}}}$, para el cuál empieza a construir

gráficas hasta que deduce que la distribución teórica que mejor se ajusta a sus muestreos es la *t-Student* con $\nu = n - 1$ grados de libertad. Gráficamente:

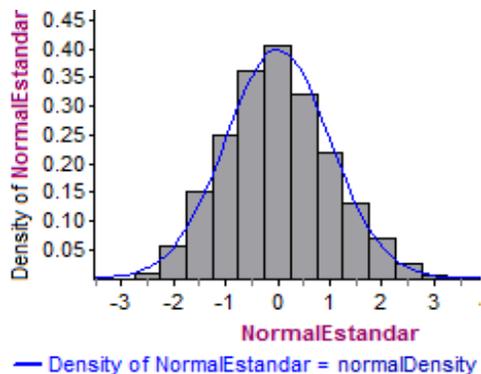


Figura 8. Distribución normal estándar

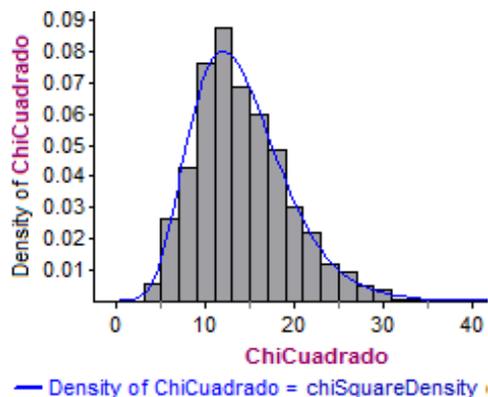


Figura 9. Distribución Chi-Cuadrado

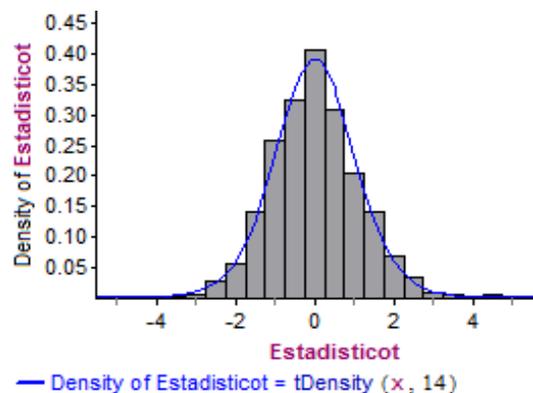


Figura 10. Distribución t-Student que surge a partir de las dos anteriores

Julio fue el único que presenta un nivel de procesos integrados en todo el proceso, pues logra construir diversas distribuciones muestrales y determinar la relación que tienen con la distribución poblacional. Además, construye distribuciones teóricas en el mismo gráfico de la distribución empírica, deduciendo el tipo de distribución que siguen los datos con sus respectivos parámetros.

En el caso de Cristian genera correctamente la distribución de promedios muestrales, la estandariza y construye la distribución teórica en forma correcta sobre su mismo gráfico:

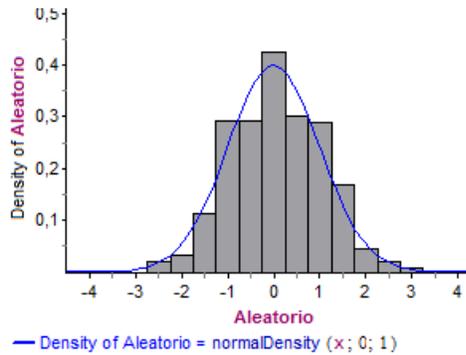


Figura 11. Distribución de promedios muestrales

Toma una muestra de tamaño 16 y se le sugiere calcular el estadístico $X = \frac{s^2(n-1)}{\sigma^2}$, sin embargo, él sólo obtiene la varianza como estadístico, luego toma 1000 muestras y cree haber construido la distribución χ^2 con $\nu = n - 1$ grados de libertad. Gráficamente:

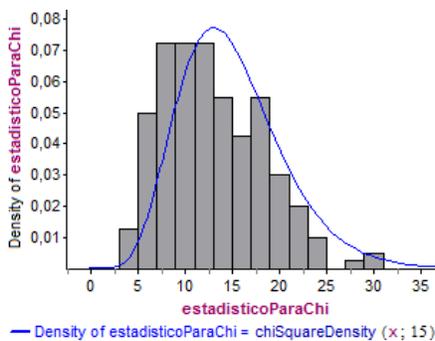


Figura 12. Distribución de varianzas muestrales

Sin embargo, cuando se le cuestiona que la distribución teórica no se ajusta completamente a la distribución empírica obtenida, resta importancia a esto y menciona que aproximadamente son iguales, la diferencia se debe a la variabilidad en la toma de muestras.

Cuando se le sugiere utilizar las dos distribuciones anteriores para construir la distribución t-Student insiste en calcular estadísticos básicos (media, proporción y varianza) en cada una y no utiliza la sugerencia del estadístico $T = \frac{Z}{\sqrt{\frac{X}{n}}}$ para construir la distribución. Construye gráficas

$$T = \frac{Z}{\sqrt{\frac{X}{n}}}$$

como la siguiente:

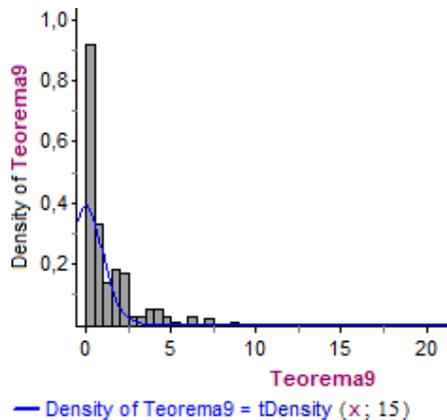


Figura 13. Distribución t-Student que construye a partir de las dos anteriores

Cristian se ubica en el nivel de transición pues cree que todas las distribuciones muestrales surgen a partir de los estadísticos básicos y utiliza la variabilidad como una opción extrema de diferencia entre las distribuciones teóricas y las distribuciones empíricas que construye.

En la siguiente tabla se muestra los niveles de razonamiento donde se ubican los estudiantes. La mayoría se ubican en el nivel de transición.

Tabla 1

Estudiante	Nivel de Razonamiento	Justificación
Michael y Johan	Razonamiento de procesos	Problemas iniciales de creación de distribuciones al generar estadísticos.
David, Carlos, Walter, Mónica, Liev, Daniel y Pablo	Razonamiento de transición	Dificultades iniciales de generación de distribuciones al reconocer estadísticos. Grafica distribuciones teóricas que no corresponden con las empíricas que obtuvo en las distribuciones muestrales. No relaciona completamente la variabilidad de los datos con los parámetros obtenidos en las distribuciones empíricas.
Adrián	Razonamiento verbal	Muestra las dificultades anteriores y además, cree que debe generarse datos a partir de la distribución uniforme. Al final menciona que todas las distribuciones deben ser normales.

Fuente: datos de actividades. 2010.

Conclusiones

La mayoría de estudiantes se ubican en los niveles intermedios de razonamiento, de transición y de procesos. Sólo uno muestra un nivel de procesos integrados y sólo uno un nivel de razonamiento verbal.

La simulación computacional a través de la extracción repetitiva de muestras, vuelve accesible, casi palpable conceptos difíciles como el de generación de distribuciones muestrales, pues los estudiantes puedan crear las distribuciones muestrales más que concentrarse en resultados finales. Pueden obtener el cálculo de probabilidades sin consultar tablas preelaboradas

que no saben de dónde surgen ni la razón de su uso.

Los datos reales, bajados de internet por los mismos estudiantes, permiten mantener el contexto en todo momento de su manipulación, para llegar a conclusiones con sentido para el estudiante.

Los estudiantes mostraron algunas dificultades que no les permitieron alcanzar a los niveles más altos de razonamiento como los siguientes:

- Muestras dificultades al crear estadísticos diferentes a los básicos como lo son las medias, proporciones y varianzas. Creen que las distribuciones más complejas se generan a partir de múltiplos de esos estadísticos básicos.
- Al sugerirles construir la distribución teórica en el mismo gráfico de la distribución obtenida en la extracción repetitiva de muestras, buscan casos particulares en los parámetros de las distribuciones. Más tarde definen sliders para ir variando esos parámetros.
- La generación de distribuciones complejas es muy difícil si el estudiante no tiene conocimientos fuertes sobre distribuciones básicas en conceptos como forma, promedio, varianza y distribución acumulada.

Referencias y bibliografía

- Burrill, G. (2002). Simulation as a tool to develop statistical understanding. En B. Phillips (Ed). *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town South Africa.
- Curcio, F. (1989). *Developing Graph Comprehension*. Reston, VA: NCTM.
- Garfield, J.; delMas, R. & Chance, B. (2004). Reasoning about sampling distributions. En D. Ben-Zvi & J. Garfield (Eds), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295-323). Netherlands: Kluwer Academic Publishers.
- Inzunza, S. (2006). Significados que estudiantes universitarios atribuyen a las distribuciones muestrales en un ambiente de simulación computacional y estadística dinámica. Tesis doctoral no publicada. CINVESTAV-IPN. México.
- Kozulin, A. (2000). *Instrumentos Psicológicos*. Primera edición. Barcelona: Paidós.
- Lipson, K. (2000). The Role of the Sampling Distribution in Developing Understanding of Statistical Inference. Tesis doctoral no publicada. University of Technology of Swinburne, Australia.
- Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. En B. Phillips (Ed.). *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town South Africa.
- Mills, J. D. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistics Education* 10(1). [en línea] Recuperable en <http://www.amstat.org/publications/jse/v10n1/mills.html>.
- National Council of Teachers of Mathematics. (2000). *Principios y Estándares para la Educación Matemática*. Sevilla: Sociedad Andaluza de Educación Matemática Thales.
- Ramírez, G. (2007). Formas de razonamiento que muestran estudiantes de maestría de Educación Matemática sobre Distribuciones Muestrales mediante problemas de simulación en Fathom. En E. Marcera y C. Pérez (Eds.). *Memorias de la XII Conferencia Interamericana de Educación Matemática*. Querétaro, México.

- Reading, Ch. & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: from a variation perspective. *Statistics Education Research Journal* 5(2), 46-48. [Disponible en línea]. <http://www.stat.auckland.ac.nz/serj>. International Association for Statistical Education (IASE/ISI).
- Sánchez, E. (2002). Teacher's beliefs about usefulness of simulation with the educational software Fathom for developing probability concepts statistics classroom. En B. Phillips (Ed.). *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town South Africa.
- Sánchez, E., Inzunza, S., Yáñez, G., Behar, R., Salcedo, A., Ramírez, G. (2008). Comprensión de las implicaciones del teorema del límite central a través de histogramas: un estudio con estudiantes y profesores de América Latina. En R. Luengo (Ed.). *Memorias del XII Simposio de la Sociedad Española de Investigación en Educación Matemática*. Badajoz, España.
- Shaughnessy, M. (1992). Research in Probability and Statistics: Reflections and Directions. En Grouws, D. A.(Ed.). *Handbook of Research on Mathematics Teaching and Learning*. New York. Macmillan Publishing Company, 465-494.
- Wild, C., Seber, G. (2000). *Chance Encounters: a first course in data analysis and inference*. Primera edición. United States: John Wiley & Sons, Inc.