DIFICULTADES EN ESTUDIANTES UNIVERSITARIOS DEL ESTADISTICO COMO VARIABLE ALEATORIA EN LA DISTRIBUCION DEL MUESTREO DE MEDIAS

José Armando Albert Huerta y Blanca Ruiz Hernández
Tecnológico de Monterrey. México
albert@itesm.mx, bruiz@itesm.mx

Resumen. En esta investigación mostramos cómo el estadístico visto como una variable aleatoria es una idea clave para el desarrollo y fundamento de la estadística inferencial. Abordamos, en particular, el caso de la media por ser el más recurrente en el discurso matemático escolar a nivel universitario y porque resulta ser un buen punto de partida para el desarrollo de conceptos como el de estadístico. Se muestra como a través de hacer interactuar a los estudiantes de ingeniería con elementos constructivos del bloque generador de variables aleatorias $X_1, X_2, \ldots, X_n$ los estudiantes pueden construir funciones $f(X_1, X_2, \ldots, X_n)$ de valor real, en particular, la media, que también son variables aleatorias cuyo valor esperado puede estimar el valor del parámetro. Se reportan aspectos relevantes identificados en este proceso de construcción conceptual de los estudiantes relativos al estadístico como concepto fundamental de las distribuciones muestrales, así como de su naturaleza de variable aleatoria

Palabras clave: educación estadística, distribución muestral, media

Abstract. In this research we show how the statistic seen as a random variable is a key idea for the development of inferential statistics. We approach, in particular, the case of the mean to be a highly recurrent concept in school mathematical discourse at university level and it turns out to be a good starting point for the development of concepts such as statistical. We show how to interact through engineering students with constructive elements $X_1, X_2, \ldots, X_n$ of the random variable generator block, they can build real-valued functions $f(X_1, X_2, \ldots, X_n)$, in particular the average which are also random variables whose expected value can estimate the value of parameter. We report relevant aspects identified in this process students' conceptual construction of the statistic as a fundamental concept, and its nature of random variable

Key words: education statistics, sampling distribution, average

Introducción

Esta investigación hace un acercamiento a una de las ideas fundamentales para el desarrollo del razonamiento alrededor de la estadística inferencial. Sin lugar a dudas el *estadístico* es un concepto clave para el desarrollo de otros como el de distribución del muestreo y estimación. Sabemos que la enseñanza de la estadística inferencial es una tarea compleja. A este respecto Chance, delMas y Garfield (2004) afirman que la dificultad en su comprensión se debe a que los estudiantes requieren de la integración y combinación de muchas ideas de estadística, no sencillas, a su vez, de aprender. En el caso del *estadístico*, podemos observar la conjunción de ideas que se involucra sólo en su definición:

Sea $X_1, \ldots, X_n$ una muestra aleatoria de tamaño *n* de una población y sea $T(x_1, \ldots, x_n)$ una función evaluada en valores reales o valores vectoriales cuyo dominio incluye el espacio muestral de $(X_1, \ldots, X_n)$. Entonces la variable aleatoria o vector aleatorio

$Y = T(X_1, \ldots, X_n)$ es llamado un estadístico. La distribución de probabilidad del estadístico Y es llamada la distribución muestral de Y. (Casela y Berger, 2001, p. 211).

La complejidad del concepto de estadístico, desde una perspectiva didáctica, es notable porque integra muchos otros conceptos tales como: variable aleatoria, aleatoriedad, muestra aleatoria, función de varias variables y distribución del muestreo, entre otros. Es por eso que en esta investigación nos hemos dado a la tarea de estudiarlo más de cerca desde una perspectiva de la compresión del estudiante universitario.

Antecedentes

Algunos estudios realizados con estudiantes universitarios reportan dificultades didácticas para su aprendizaje. Vallecillos y Batanero (1997) concluyen que una dificultad importante fue la falta de apreciación de la característica de variable aleatoria del estadístico muestral. En su estudio, todos los alumnos cometieron errores que evidencian falta de comprensión referidas a las relaciones entre la distribución del estadístico, las regiones y el nivel de significancia. Reportan, además, que el alumno no considera la media muestral como una variable aleatoria, lo que puede ocasionar la creencia en que la hipótesis pueda referirse indistintamente a la media de la muestra, tanto como a la de la población. Well, Pollatsek y Boyce (2004) sostienen que para el desarrollo de una buena comprensión de la inferencia estadística se necesario entender que cuando las muestras se obtienen de una población de referencia, estas muestras variarán y como consecuencia también el valor numérico de los estadísticos derivados de dichas muestras, conformando un patrón predecible de variación, las distribuciones del muestreo que, según Vallecillos (1995), son el concepto esencial de la Inferencia estadística porque cualquier procedimiento inferencial implica conocer la distribución muestral de algún estadístico. Por otra parte, c pudieron observar que los estudiantes tienden a enfocarse en muestras individuales y resúmenes estadísticos de las mismas, en vez de enfocarse en cómo se distribuyen las colecciones de estadísticos muestrales. De manera similar, Cox y Mouw (1992) señalan que la mayoría de sus sujetos estudiados en su investigación sobre muestreo vieron a una muestra como una representación fija y exacta de una población y no tuvieron claridad para ver a un estadístico como un estimador que resulta de una distribución muestral. Por otra parte, Ruiz (2006) reporta la dificultad de los estudiantes de reconocer la existencia de la variable aleatoria al hacer composición entre variables aleatorias. Esto es especialmente importante para el caso del estadístico por ser una función de valor real de variables aleatorias. Esta investigación se circunscribe en el enfoque sistémico y sólo desarrolla una de sus componentes: la cognitiva.

Metodología y resultados

El contexto de la investigación se sitúa en una universidad del norte de México con estudiantes de tercer semestre que ingeniería que estaban llevando el curso de probabilidad y estadística y que contaban con acceso a computadoras personales y al software Excel. Se implementó una secuencia de actividades con uso de tecnología justo como punto de partida para el desarrollo del tema de distribuciones muestrales. La muestra de estudiantes consistió en dos grupos de estudiantes de 40 alumnos cada uno y divididos en equipos de 2 personas o tres. Se trató de tres secuencias didácticas realizadas una por clase y se le dedicó la última media hora de cada clase de hora y media. Las actividades fueron:

Distribución de Población

Bloque aleatorio

Estadístico

Con el propósito de observar a través del desarrollo constructivo del concepto del estadístico de la media muestral basado en un bloque aleatorio de $X_1, X_2, ..., X_n$ variables aleatorias idénticamente distribuidas, las concepciones y dificultades de los estudiantes.

Distribución de Población

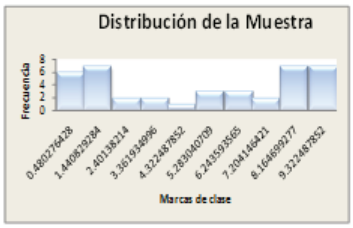
En esta primera fase, se buscó que los estudiantes identificaran una estrategia que les permitiera descubrir aproximadamente qué tipo de distribución tiene la Población. La actividad consistió en tomar muestras aleatoria del tamaño que crean conveniente (de una población infinita de tiempos

de caída de paracaidistas a un punto en tierra fijo) para que al graficar su distribución de frecuencias, y analizar si es posible encontrar algún patrón al aumentar el tamaño de la muestra. También se les pidió que a cada muestra le calcularan su media e identificaran cuál podría ser el valor aproximado de la media de población.

La población era una hoja oculta de Excel con 50 mil datos obtenidos con el software R y distribución uniforme con dominio [0,10]. Los estudiantes tenían acceso a los datos a través de una Urna virtual (otra hoja Excel) que elegía muestras al azar de esos datos del tamaño que ellos eligieran y les daba los datos de la muestra. Podían recurrir al Excel o R para graficar el histograma de la distribución de la muestra obtenida.

La mayoría de los equipos escogieron muestras de no más de 50 datos y no hicieron repeticiones, con una vez les resultó suficiente para concluir. Se puede observar que para ellos "grande"

significa no más de 50 datos. La mayoría no pudo identificar alguna forma de la distribución de frecuencias de la muestra.



"el histograma muestra una distribución muy dispareja". "no encontramos patrón entre sus barras".

Figura 1. Grafico representativo de la mayoría de los grupos de trabajo.

Pero un equipo había ensayado con 1000 datos y concluyeron:

"se nota una tendencia a distribución uniforme, la gráfica asemeja más a un rectángulo con todas las marcas de clase y cada valor de x tiene la misma probabilidad"

Y en la fase de institucionalización, los demás compañeros, al ver este resultado también ensayaron con muestras más grandes y pudieron concluir que se trataba de una distribución uniforme. Pudieron percatarse que a través del muestreo aleatorio es posible conocer características de la población y que ese conocimiento era mejor a medida que la muestra era de mayor tamaño.

Respecto a la media también les bastó el cálculo de la media muestral para proponerla como valor aproximado de la población, pero a nadie se le ocurrió explorar tendencias con varias medias. No mostraron necesidad de repetir el experimento, ni de considerar la variabilidad.

Bloque aleatorio

Con el propósito de familiarizar a los estudiantes con más de una variable aleatoria y algunas de sus operaciones y propiedades, se les propuso la siguiente actividad la cual consistió en que ellos construyeran un bloque de 50 muestras de tamaño 40 obtenidas aleatoriamente con la Urna virtual. Los encabezados de las columnas fueron $X_1, X_2, \ldots, X_n$. De esta matriz resultante se les preguntó sobre si X_1 era una variable, y si la consideraban aleatoria, si podían identificar su distribución y su valor esperado. Y así X_2, hasta X_{40}. Posteriormente se les preguntó sobre si $(X_1, X_2, \ldots, X_n)$ era también aleatorio. Los resultados se muestran en la Tabla 1.

X	Pregunta	Comentarios	Proporción
X1	¿es variable?	"sí, porque varía", "sí porque está tomando Distintos valores…" "sí porque adquiere Valores distintos"	100%
	¿es variable Aleatoria?	"sí porque no sabías Que valores iba a obtener"	100%
		"sí porque los valores fueron Tomados al azar"	
		"sí porque cualquier muestra De la población tiene la misma Oportunidad de ser elegida"	
		"sí porque se tomaron valores Sin seguir un orden"	
	¿distribución?	"más o menos uniforme"	42%
		"no se puede identificar"	33%
		"distribución normal"	25%
	¿valor	"5 idealmente, 5.4 realmente"	8%
	Esperado?	"5.49, 5.4, 4.92, 5.0, 4.9, etc."	92%
X2	¿es variable?	"de la misma manera que x1"	100%
	¿es variable	Sí, por las mismas razones que x1	100%
	Aleatoria?		
	¿distribución?	"más o menos uniforme"	42%
		"no se puede identificar"	42%
		"distribución normal"	17%
	¿valor esperado?	"5 idealmente, 5.8 realmente"	8%
		"5.02, 5.23, 6.4, … con el promedio"	92%
X40	¿es variable?	Sí, por las mismas razones que	100%
		Las otras.	
	¿es variable	Sí, por las mismas razones que	100%
	Aleatoria?	Las otras.	
	¿distribución	"no, porque todos sus valores	75%
	Similar a las	Son diferentes"	
	Otras?	"debe ser muy parecido"	25%
	¿distribución?	"tiende a ser uniforme"	67%
		"no se identifica distribución"	33%
	¿valor esperado?		
		"5 idealmente, 4.9 realmente"	8%

		"5.81, 4.89, 5.3 … es la media"	92%

Tabla 1. Sobre las concepciones de los alumnos en el bloque aleatorio

Se puede observar que, aparentemente, el hecho de repetir varias veces el muestreo y graficar cada vez, les permitió identificar una tendencia en los gráficos. Así, en la actividad 1, sólo el 8.3% había identificado una distribución uniforme, para X_1 identifican el 42% de los alumnos que es uniforme, para X_{40} ya es el 67%.

En cambio, no sucedió lo mismo con el valor esperado pues sólo el 8% del grupo identificó su valor durante las distintas repeticiones del experimento.

Hasta aquí, tenemos que los estudiantes han identificado a $X_1, X_2, \ldots, X_n$ como variables aleatorias. Además, se puede observar que los estudiantes pueden, en su mayoría, identificar una propiedad común a $(X_1, X_2, \ldots, X_n)$ como una unidad aleatoria aunque no siempre con los mismos argumentos. A un porcentaje bajo le costó desprenderse del contexto de la medición de la distancia en la caída entre el punto de aterrizaje del paracaidista y el blanco. Puede observarse que los estudiantes le atribuyen a $(X_1, X_2, \ldots, X_n)$ aleatoreidad porque cada una de las variables que lo componen es aleatoria. Sin embargo, en estudios previos (Ruiz y Albert, 2008) hemos podido identificar su dificultad para aceptar a Y como variable aleatoria si Y es variable dependiente de otras variables aleatorias.

Estadístico

Con los antecedentes sobre variables aleatorias de las actividades anteriores, la siguiente actividad tiene por propósito hacer a los estudiantes interactuar con una variable aleatoria específica: la media muestral que es una función dependiente de variables aleatorias independientes, es decir, $\bar{x} = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$. La actividad consistió en calcular las medias de las anteriores muestras y analizar si se trataba este promedio de una nueva variable y si así fuera, si era aleatoria. Luego, se les pidió identificaran su distribución, su valor esperado y dieran una aproximación de cuál podría ser el verdadero promedio de la población. Finalmente, se les pidió identificar estadísticos y si eran variables aleatorias. Los resultados se muestran en la Tabla 2.

Preguntas	Comentarios tipo de los estudiantes	Proporción
Observe la nueva columna obtenida de promedios de las muestras obtenidas, ¿Será este promedio una variable?	"Sí, ya que al variar las variables también la media lo hará"; "es variable dependiente"	58%
	"No, porque no puede ser sustituido por algún miembro del universo, es dependiente"	8%
	"Sí, porque varía el promedio de una muestra a	33%

	otra"	
¿El promedio será una variable aleatoria?	"Sí, ya que las muestras fueron aleatorias, el promedio depende de las muestras, este será aleatorio también"	33%
	"No, deja de ser aleatorio ya que es un número calculado"	67%
Grafique los valores de los promedios obtenidos ¿Cómo qué tipo de distribución identifica?	"Normal"	42%
	"No se puede identificar con el gráfico"	58%
Calcule el promedio de los promedios. ¿Cuál sería de todos los promedios el más adecuado para proponerlo como una mejor aproximación del promedio general de la población?	"5, ya que es el valor esperado"	20%
	"El promedio de promedios, ya que es un valor en función de todas las medias de las muestras realizadas"	80%
Si definimos un estadístico como cualquier cantidad cuyo valor puede ser calculado a partir de datos muestrales, ¿qué estadísticos pueden hallarse en las muestras antes encontradas?	"Media, mediana, moda, varianza, desviación estándar, rango,…"	84%
	"Los que producen una gráfica con distribución normal"	8%
	Otras respuestas	8%
¿Cuáles estadísticos hallados pueden considerarse variables aleatorias y cuáles no?	"todos porque dependen de los valores aleatorios que tome la muestra" "todos porque vienen de variables aleatorias".	30%
	Ninguno, ya que son calculados a partir de los datos obtenidos" Ninguno porque no pueden tomar el lugar cualquier miembro del conjunto el cual crea el valor de estos estadísticos"	50%
	"Algunas como la Media, mediana y moda"	20%

Tabla 2. Sobre las concepciones de los alumnos en torno al estadístico.

Aunque rápidamente la mayoría de los estudiantes (84%) identificaron lo que es un estadístico, no así que se trataba de una variable aleatoria pues sólo un 30% lo pudo ver. Influyó en esto su creencia errónea de que al ser un estadístico variable dependiente, no hereda el ser también variable aleatoria. Sin embargo, esto es importante, para poder asociarla a una distribución de probabilidad que más adelante se usaría para la inferencia estadística. Por otra parte, aunque su intuición de valor esperado de la media poblacional es acertada, no para todos fue claro que la distribución del estadístico es aproximadamente normal. En parte se debió a un error de interpretación de los gráficos pues algunos no consideraron igualar las escalas de los ejes para luego comparar las formas. Se pudo constatar que a los estudiantes les resultó de mucha utilidad

observar el contraste entre la forma aproximada de la distribución de población obtenida a partir de muestras grandes (en nuestro caso fue uniforme) con la distribución obtenida de las medias de las muestras (aproximadamente normal).

Conclusiones

Esta investigación pudo mostrar algunas dificultades de los estudiantes para la construcción de una nueva variable aleatoria, el estadístico, a partir de otras variables aleatorias independientes. Esto sugiere hacer diseños de actividades más finos que permitan describir mejor este fenómeno didáctico para construir una posible forma de superarlo, desde un punto de vista didáctico. Sin embargo, también se pudo comprobar la riqueza conceptual que se gana en favor de la variable aleatoria cuando los estudiantes se ven involucrados en situaciones de varias variables que son el punto de partida para el desarrollo del concepto de estadístico como variable aleatoria y su distribución.

Abre otras posibles vías de investigación sobre las dificultades de los estudiantes en el razonamiento interpretativo de gráficos, particularmente en lo referido a la consideración de las escalas.

La complejidad de la inferencia estadística queda manifiesta por los múltiples conceptos que intervienen cada vez que se usa. Es por eso que pareciera más oportuno investigar, no sobre conceptos aislados, como tal vez podría hacerse en otras áreas de las matemáticas, sino en complejos formados por redes de conceptos cercanos es el caso del *estadístico*.

Referencias bibiográficas

Casella, G. & Berger, R. L. (2001). *Statistical Inference*. (2nd ed.). Duxbury Press.

Chance, B., del Mas, R. y Garfield, J. (2004). Reasoning about sampling distributions. *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Kluwer Academic Publishers.

Cox, C., & Mouw, J.T. (1992). Disruption of the representativeness heuristic: Can we be perturbed into using correct probabilistic reasoning. *Educational Studies in Mathematics*, 23, 163-178.

Ruiz, B. (2006). *Un acercamiento cognitivo y epistemológico a la didáctica del concepto de variable aleatoria*. Tesis de Maestría. CICATA. México.

Ruiz, B. R., & Albert, J. A. (2008). Obstáculos epistemológicos sobre la variable aleatoria. *Encuentro Latinoamericano De Educación Estadística*. México.

Vallecillos, A. y Batanero, C. (1997). Análisis del aprendizaje de conceptos clave en el contraste de hipótesis estadísticas mediante el estudio de casos. *Recherches en Didactique des Mathématiques*, 17(1), 29-48.

Vallecillos, A. (1995). Sugerencias metodológicas para la introducción del teorema central del límite en la enseñanza secundaria. *Actas del I congreso nacional de bachillerato*. Universidad de Granada.Well, A., Pollatsek, A. & Boyce, S. (2004). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, 47(2), December 1990, 289-312.