2.3.11. Construcción del concepto de estimación de parámetros a través de simulación con software R

Félix Amadeo Canales Conce
Ubaldo Cayllahua Yarasca
Universidad Nacional de Huancavelica, Huancavelica, Perú

Resumen

Esta experiencia se realizó en la asignatura de estadística inferencial con estudiantes del VII ciclo del programa de estudios de Matemática, Computación e Informática de la Escuela Profesional de Educación Secundaria de la Facultad de Educación de la Universidad Nacional de Huancavelica, con el propósito de construir y comprender la estimación de parámetros por el método de estimación puntual empleando la simulación como recurso didáctico mediado con el software R. Para este efecto, previamente se hizo la consideración del contexto de trabajo, luego se diseñó las actividades de simulación para el objeto estadístico y finalmente la ejecución de las secuencias didácticas y su evaluación. La mayoría de los estudiantes lograron comprender que la estimación de parámetros, es obtener información sobre una población, para el cual no se requiere estudiar a todos y cada uno de los individuos de la población, sino es suficiente aplicar técnicas de muestreo y hacer la generalización del análisis realizado. El uso de la simulación como recurso didáctico mediado con el software R, facilita la construcción y comprensión del concepto y sus propiedades de estimación de parámetros por el método de estimación puntual.

Introducción

El objetivo principal de la estadística inferencial es la *estimación* de los parámetros de una población de estudio. Para calcular necesitaríamos tener todos los posibles valores de la población, lo cual no suele ser posible en la realidad. No obstante, la estadística emplea la teoría de muestreo para obtener una muestra aleatoria de esta población y calcular los valores estadísticos y generalizar esta información a la población; es decir, la estimación son valores estadísticos calculados a partir de lo observado en una muestra, las que se generalizan dicho resultado muestral a la población total, de modo que lo estimado es el valor generalizado a la población. La notación general para un parámetro poblacional es θ y para un estadístico de la muestra es $\hat{\Theta}$.

Existen dos formas de estimar parámetros, la estimación puntual y la estimación por intervalo de confianza. En la primera se busca, con base en los datos muestrales, un único valor estimado para el parámetro. Para la segunda, se determina un intervalo dentro del cual se encuentra el valor del parámetro, con una probabilidad determinada.

Para esta experiencia didáctica se ha considerado la estimación de parámetros a través de la estimación puntual, cuyas propiedades fundamentales (Walpole, Myers y Myers, 1999), son las siguientes:

a) *Insesgamiento.* Significa que su media o valor esperado coincide con el parámetro poblacional.

$$E(\hat{\theta}) = \theta$$

La diferencia entre el valor esperado del estimador de un parámetro y el parámetro a estimar se llama sesgo del estimador que se escribe como:

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Cuando el sesgo del estimador es positivo, este tenderá a sobrestimar el valor del parámetro, mientras si el sesgo es negativo, el estimador tenderá a infravalorar. Finalmente, si el sesgo es cero, el estimador es insesgado.

b) *Eficiencia.* Dos estimadores $\hat{\theta}_1$ y $\hat{\theta}_2$ de una parámetro poblacional obtenidos de la misma muestra, el estimador $\hat{\theta}_1$ será más eficiente que el estimador $\hat{\theta}_2$ si:

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

Esta propiedad emplea dos o más estimadores para comparar las varianzas y determinar como el estimador más eficiente aquel que tenga menor variabilidad.

Las propiedades de insesgamiento y eficiencia, son independientes del tamaño muestral. Sin embargo, el tamaño muestral influye en la precisión del estimador.

c) *Consistencia.* Es cuando el valor del estimador puntual tiende a estar más cerca del parámetro poblacional a medida que el tamaño de la muestra aumenta.

$$\lim_{n\to\infty} P(\left|\hat{\theta} - \theta\right| > \varepsilon) = 0$$

En otras palabras, una muestra grande tiende a proporcionar mejor estimación puntual que una pequeña.

La simulación a través del software R

El diccionario de la Real Academia Española (1992, p. 1883), define la simulación como una acción de simular, que significa representar una cosa, fingiendo o imitando lo que no es. Shannon y Johannes (1976), citado en Wikipedia, enciclopedia libre, la simulación se define, como el proceso de diseñar un modelo de un sistema real y llevar a término experiencias con él, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias (…) para el funcionamiento del sistema (párr. 2). Es decir, la simulación debe entenderse como la reproducción o representación simplificada de un fenómeno real mediante un modelo más sencillo y más adecuado para ser estudiado, a fin de obtener de ella algunas propiedades o características de su funcionamiento.

La modelación de algún fenómeno, puede realizarse por sistemas físicos, matemáticos o computacionales. Girard (1997, citado en Batanero, 2003), menciona que "al trabajar mediante simulación estamos ya modelizando, porque debemos no solo simplificar la realidad, sino fijar los aspectos de la misma, que queremos simular y especificar unas hipótesis matemáticas sobre el fenómeno estudiado" (p. 45). Esta referencia permite considerar que la simulación es un instrumento de modelización de fenómenos aleatorios dentro de la estadística y probabilidades.

Para este estudio se considera la simulación como una herramienta para el proceso de enseñanza y aprendizaje de la estadística. Katz y Medina (2010), nos indica que "la simulación no constituye solamente un medio propicio para obtener respuestas aproximadas a una variedad de problemas. También es un recurso didáctico para facilitar la comprensión de algunas propiedades que resultan complejas por el nivel de abstracción que requieren".

Elousa, López, Artamendi, Yenes y Mujica (2013), concuerdan con esta posición de recurso didáctico, además refieren que el uso de la simulación en la enseñanza del análisis de datos y de la psicometría, debe realizarse con apoyo de Tecnologías de Información y Comunicación, herramienta fundamental en el proceso de enseñanza de las probabilidades.

Desde estas consideraciones cobra sentido considerar a la simulación además de ser un modelo que representa la realidad, una herramienta de apoyo para la enseñanza del docente y para el aprendizaje del estudiante de objetos estadísticos que tienen alto grado de abstracción, empleando ordenadores y software. Asimismo, facilita acercamientos formales en la enseñanza y a la vez permite generar datos y contextos a modo de experimentos controlados por el propio estudiante de acuerdo a su requerimiento.

El software R, de acuerdo a Paradis (2002), es un lenguaje orientado a objetos, con una simplicidad y flexibilidad en su programación. Es un lenguaje interpretado (como Java) y no compilado (como C, C++, Fortran, Pascal, ...), lo cual significa que los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir ejecutables. Su sintaxis es muy simple e intuitiva, el cual hace que un principiante adquiera las nociones básicas de programación y avanzar progresivamente.

Este software tiene la potencialidad por su funcionalidad al brindar dos posibilidades: de programación a través de una sintaxis computacional y por el uso de sus funciones propias de manera directa de la librería base y paquetes especializados, en vista que es un programa basado en comandos. Por el aspecto pedagógico, es un recurso didáctico de excelente mediación en el proceso de construcción de aprendizaje de objetos estadísticos.

La simulación de cualquier objeto o fenómeno empleando como recurso software R, se realiza a través de un conjunto de funciones que componen un paquete implementado, por ejemplo la función *tosscoin* del paquete de probabilidades (*prob*) permite la simulación del lanzamiento de monedas; también pueden ser funciones elaboradas o creadas por el propio usuario de acuerdo a las necesidades, las que realizan tareas que no estaban definidas en el paquete base u otro, el cual es una característica de la potencialidad y comodidad del software R (Elosua, 2011; Jay Kerns, 2013).

En esta experiencia, se hizo uso las funciones de la librería base y la función *sample* del paquete de probabilidades, que permite tomar una muestra aleatoria simple de tamaño *n* a partir de un vector de valores con o sin reemplazamiento. Su sintaxis es *sample (x, size, replace=TRUE, prob=NULLL)*. Donde *x* es un vector de datos de tipo integer (números enteros) o character (cadena de caracteres) del cual se eligen aleatoriamente los elementos, *size* es un entero positivo que indica el número de elementos que se quieren elegir, *replace=FALSE* es un valor lógico que indica que el muestreo se hace sin reemplazamiento, mientras que *replace=TRUE* indica con reemplazamiento. Por último, *prob* es un vector

de probabilidades en el que cada elemento será la probabilidad con la que se elegirá el elemento que corresponde del vector *x* que va a ser muestreado (Jay Kerns, 2013).

Las ventajas fundamentales de este software R, tanto para el usuario particular como para las empresas o universidades, es la completa gratuidad, que es libre y de código abierto. Este hecho es clave especialmente en entornos universitarios (donde cada vez su penetración es mayor), dado que es un software que, aparte del ahorro económico, permite evitar la dependencia por parte del estudiante de lugares físicos, como centros de cómputo, para poder realizar sus tareas y prácticas académicas. Otra ventaja es una herramienta de análisis estadístico permanentemente actualizada gracias a la contribución de una comunidad de desarrolladores a nivel mundial, mediante la incorporación de las últimas técnicas estadísticas a través de la creación de librerías, o pequeños programas que se pueden instalar dentro del entorno (Méndez, 2018).

Metodología

Contexto de trabajo. La experiencia se ha realizado en la asignatura de estadística inferencial con un grupo de 12 estudiantes del VII ciclo del programa de estudios de Matemática, Computación e Informática de la Escuela Profesional de Educación Secundaria de la Facultad de Educación de la Universidad Nacional de Huancavelica, quienes previamente han llevado la asignatura de estadística descriptiva como prerrequisito para estadística inferencial y los mismos que cuentan con saberes previos de uso de software de entorno gráfico y de programación que es software R.

Planificación de medios. Previo al desarrollo de la sesión de aprendizaje, se hace llegar a través del aula virtual el código de comandos escrito en lenguaje R y el resumen del tema de la clase, a fin de que se revise los propósitos que tiene cada uno de las actividades y cuáles serán los elementos de entrada y salida respectivamente en cada actividad.

Ejecución de actividades y evaluación. El tiempo de desarrollo de la sesión de aprendizaje fue de tres horas. Planteado la situación problemática, los estudiantes trabajan en grupos de dos ejecutando los códigos y analizando los resultados para diferentes valores y condiciones en cada actividad, luego intercambian resultados y formalizan en un lenguaje verbal los resultados. La evaluación se realiza durante el proceso de desarrollo de las actividades en base a una lista de cotejo.

Resultados

Se presenta por cada uno de las actividades de simulación desarrolladas en la experiencia.

Actividad 1. Definición de una población de datos

Con el código en R (anexo), se genera una población de datos sobre el peso de 1000 estudiantes, que se encuentran entre 40 y 80 kilogramos. En referencia a esta información se grafica el histograma y se hace el cálculo de algunos valores de los parámetros de la población, por ejemplo, la media poblacional, como se ve en la siguiente figura.

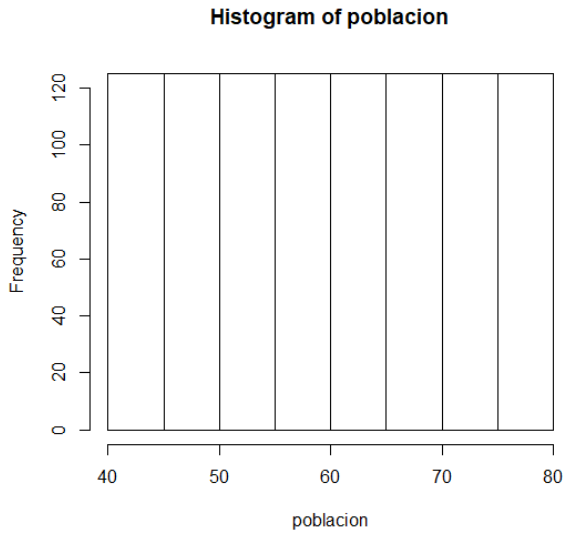

Figura 1. Distribución de los datos de la población.

De esta actividad, se evidencia que la totalidad de los estudiantes lograron identificar la forma de la distribución que tienen los datos de la población, como una distribución uniforme, observando que los datos no tienen una tendencia a concentrarse alrededor de la media poblacional de 60 kilogramos.

Actividad 2. Obtención de muestras aleatorias de diferentes tamaños con y sin reemplazo

Con esta actividad se realiza la simulación de sacar muestras aleatorias de cualquier tamaño con reemplazo o sin reemplazo y analizar la representación que hacen a toda la población.

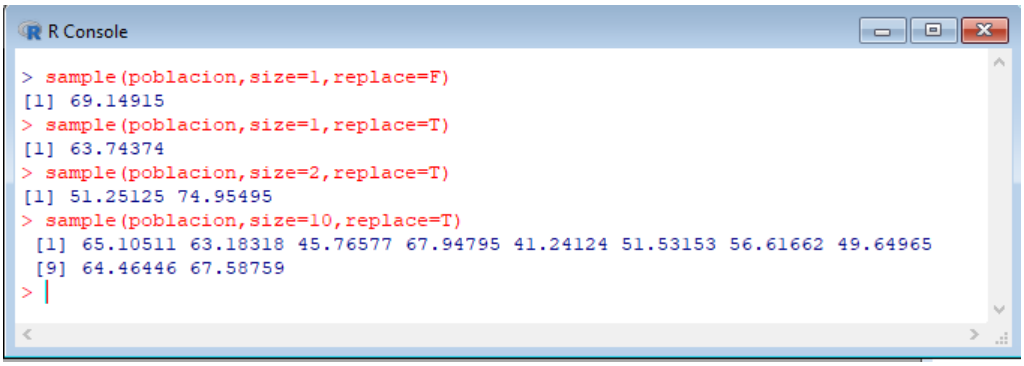

Figura 2. Extracción de muestras con y sin reemplazo.

Los resultados nos evidencian que la mayoría de los estudiantes diferencia entre una muestra con y sin reemplazo, haciendo una analogía con bolas numeradas en una urna; indicando que en una muestra sin reemplazo nunca un individuo puede salir dos veces. Respecto a la representación que hace una muestra a toda la información de la población, solo algunos de los estudiantes logran explicar que muestras pequeñas no hacen una buena representación a toda la información de la población.

Actividad 3. Representación del histograma y el valor del estimador

Se realiza la simulación de sacar muestras aleatorias de cualquier tamaño con reemplazo, luego representar la distribución de la muestra y del valor de la estadística (círculo rojo).

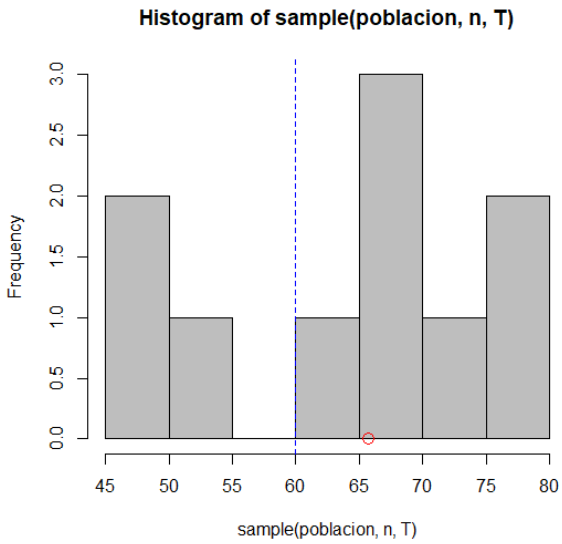

Figura 3. Distribución de los datos de una muestra y representación del valor del estimador con respecto al parámetro.

Todos los estudiantes reconocen que el histograma es la distribución de la muestra y el círculo rojo es la que representa al valor del estimador. Pero no todos evidencian que lo normal es que sea muy raro que se aleje muchísimo del parámetro, y lo más frecuente es que se obtengan valores más o menos cercanos a la verdadera media de la población.

Actividad 4. Representación de la distribución y del estimador para diferentes tamaños de muestra

Se realiza la simulación de extraer muestras aleatorias de cualquier tamaño con reemplazo, cuyas representaciones se presentan en la siguiente figura.

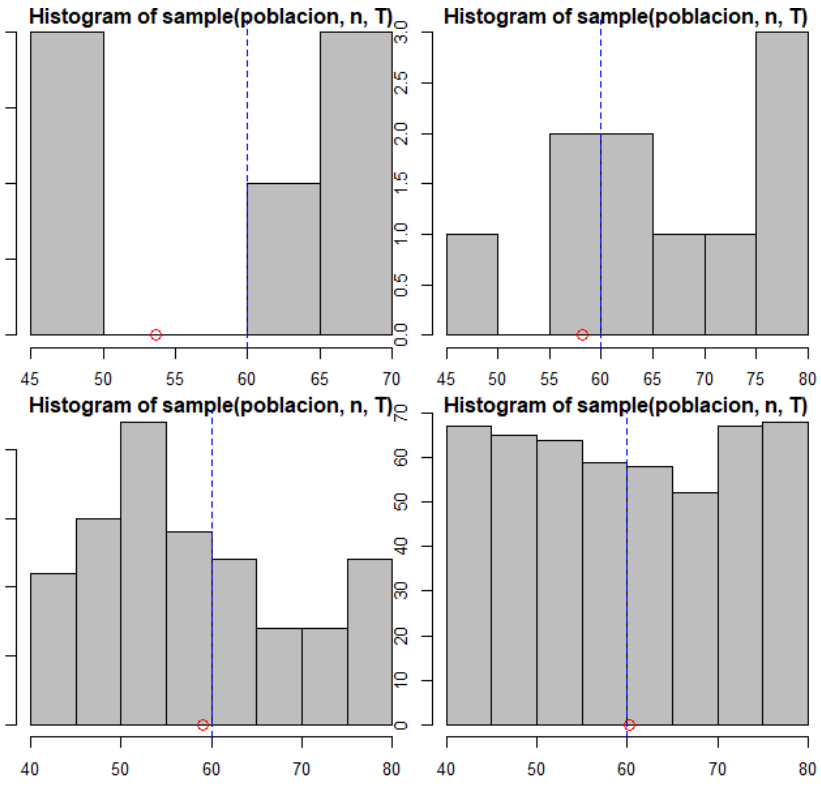

Figura 4. Distribución de una muestra de tamaño 5, 10, 100 y 500, y representación del valor del estimador respecto al parámetro de 60.

A partir de la figura, la mayoría de los estudiantes reconocen que a mayor tamaño de la muestra la distribución muestral tiende a tomar la forma de la distribución poblacional que una distribución uniforme. De igual manera, los valores de los estimadores también tienden a ser igual al valor del parámetro a medida que el tamaño de la muestra se hace grande.

Conclusiones

La experiencia didáctica con estudiantes de matemática, ha permitido arribar a las siguientes conclusiones:

- El uso de la simulación como recurso didáctico mediado con el software R, permitió en los estudiantes realizar la construcción y comprensión del concepto de estimación de parámetros poblacionales por el método de estimación puntual, así como de sus propiedades fundamentales que deben tener un buen estimador puntual.

- Los estudiantes muestran interés para investigar y desarrollar modelación y simulación con software R en otros temas de la estadística y probabilidades para procesos de enseñanza y aprendizaje en los diferentes niveles educativos.

- Fue una oportunidad en los estudiantes para fortalecer sus conocimientos y habilidades de programación utilizando códigos del software R.

Referencias

Batanero, C. (2003). La simulación como instrumento de modelización en probabilidad. *Revista Educación y Pedagogía*, 35(15), 37-54.

Elousa, P. (2011). *Introducción al entorno R*. Bilbao: Universidad del País Vasco / Euskal Herriko Unibertsitatea.

Elousa, P., López, A., Artamendi, J., Yenes, F. y Mujika, J. (2013). Simulación con R y aprendizaje cooperativo: dos niveles de innovación educativa en la enseñanza/aprendizaje de conceptos cuantitativos en las ciencias sociales. En *Educar para transformar, X Jornadas Internacionales de Innovación Universitaria* (pp. 904-910). Universidad Europea: Facultad de Psicología. Recuperado de http://www.ehu.es/gip/publicaciones/articulos/2013/4.pdf

Jay Kerns, G. (2013). *Elementary probability and the prob package*. http://127.0.0.1:24148/library/prob/html/prob.html. Consultado 17/03/15.

Katz, R. D., y Medina, M. A. (2010). La simulación: un recurso didáctico para facilitar el aprendizaje de la Probabilidad y la estadística. Presentado en la *III Jornadas de experiencias Innovadas en educación en la FCEIA*, Universidad Nacional del Rosario: EPEC-FCEIA. Recuperado de http://www.fceia.unr.edu.ar/fceia/3jexpinnov/

Méndez, M. (2018). *Análisis de datos con R: una aplicación a la investigación de mercados*. Madrid, España: ESIC.

Paradis, E. (2002). *R para principiantes.* France: Institut des Sciences de l'Évolution Universit Montpellier II.

Real Academia Española (1992). *Diccionario de la lengua española*. Tomo II, (21ª ed.). Madrid España: Espasa Calpe, S. A.

Shannon, R. y Johannes, J. D. (1976). Systems simulation: the art and science. *IEEE Transactions on Systems, Man and Cybernetics*, 6(10), pp. 723-724. Recuperado de http://es.wikipedia.org/wiki/Simulaci%C3%B3n

Walpole, R., Myers, R. y Myers, S. (1999). *Probabilidad y estadística para ingenieros,* (6ª ed.). Mexico: Prentice Hall, Hispanoamericana, S.A.

Volver al índice de autores

ANEXO

CÓDIGOS EN SOFTWARE R

```
##############################################################################
# Actividad 1: Generar una población de 1000 personas cuyos pesos varíen
# entre 40 y 80 kilogramos.
##############################################################################
poblacion <- seq(40, 80, length=1000)
poblacion
hist(poblacion)
mean(poblacion)
median(poblacion)
##############################################################################
# Actividad 2: Extraer de la población una muestra aleatoria de tamaño "n"
# con reemplazo.
##############################################################################
sample(poblacion, size=1, replace=T)
sample(poblacion, size=2, replace=T)
sample(poblacion, size=2, replace=T)
# Para una muestra de tamaño 10, calcular la media y graficar histograma.
muestra1=sample(poblacion, size=10, replace=T)
muestra1
mean(muestra1)
hist(muestra1, col=8)
##############################################################################
```

```
# Actividad 3: Función para sacar una muestra de tamaño "n" de la población,
# luego representar su histograma y la media.
#############################################################################
dibujaMuestra <- function(poblacion, n){
        poblacion.media <- mean(poblacion)
        muestra.media <- mean(sample(poblacion, n, T))
        hist(sample(poblacion, n, T), col="gray")
        abline(v=poblacion.media,col="blue", lty=2)
        points(muestra.media, 0, col ="red", bg="black", cex=1.5)
        print(muestra.media)
        }
dibujaMuestra(poblacion, 20)
# Sacar 6 muestras para diferentes tamaños
par(mfrow=c(3, 2), mai=c(0.4, 0.1, 0.1, 0.1))
dibujaMuestra(poblacion, 5)
dibujaMuestra(poblacion, 10)
dibujaMuestra(poblacion, 50)
dibujaMuestra(poblacion, 100)
dibujaMuestra(poblacion, 500)
dibujaMuestra(poblacion, 800)
# Sacar cuatro muestras de tamaño 20 de la población y representar histograma
par(mfrow=c(2, 2), mai=c(0.5, 0.1, 0.1, 0.1))
for(i in 1:4) {dibujaMuestra(poblacion, 20)}
```