El Muestreo sí sirve

Luis Rojas Torres[1]

Resumen

El uso de las muestras aleatorias simples es un tema muy común en los noticieros, no obstante, muchos espectadores desconfían de su efectividad. El objetivo de este taller es presentar una metodología para la enseñanza del muestreo simple aleatorio, enfocada en mostrar la efectividad de esta técnica. El taller consistirá en usar el software Excel para seleccionar muestras aleatorias de una base de datos y verificar cómo el promedio de las muestras se "acerca" al promedio poblacional. También se utilizarán estas muestras aleatorias para introducir los conceptos de margen de error y nivel de confianza. Por último, se analizará el problema de cuál es el tamaño de muestra necesario para algunos estudios.

Palabras claves: Muestreo aleatorio simple, metodología de enseñanza, nivel de confianza, error de muestreo

Resumen

The use of simple random sampling is a very common topic in news broadcasts, however, many viewers are suspicious of its effectiveness. The objective of this workshop is to present a methodology for teaching simple random sampling, focused on showing the effectiveness of this technique. The workshop will consist of using Excel software to select random samples from a database and verify how the average of the samples is "close" to the population average. These random samples will also be used to introduce the concepts of margin of error and confidence level. Finally, the problem of what sample size is necessary for some studies will be discussed.

Key words: simple random sampling, teaching methodology, confidence level, sampling error

Modalidad: Taller [3 horas]

[1] Universidad de Costa Rica, luismiguel.rojas@ucr.ac.cr

Introducción

El auge de las redes sociales ha puesto en evidencia que un grupo de personas tiende a desconfiar de los resultados de las encuestas. En los comentarios de los reportajes de las encuestas es normal encontrar comentarios como: "la casa encuestadora manipuló los datos" o "la persona favorecida financió la encuesta". En [2] se menciona que en dos encuestas de finales de los setentas, un tercio de las personas reportaron que no confiaban en los resultados de las encuestas, ya que consideraban que las empresas manipulaban los resultados o las personas no eran sinceras en el momento de responder.

Otro comentario común sobre las encuestas es el siguiente: "es imposible hacer conclusiones sobre todo el país con solo 1000 personas". A diferencia de otros comentarios, éste tiene la apariencia de ser objetivo, no obstante, dista de la realidad ya que la estadística inferencial permite, por ejemplo, usar la información de una muestra aleatoria para generar un intervalo de confianza de una proporción poblacional de interés.

Para contrarrestar esta situación, el Programa de estudios de Matemática de Costa Rica considera objetivos de aprendizaje como "reconocer la importancia del muestreo en el análisis de datos" e "identificar la importancia del azar en los procesos de muestreo estadístico" [3]. Además, en el Programa se indica que una persona egresada de secundaria debe ser capaz de comprender las ideas básicas de muestreo. No obstante, en las aulas costarricenses el tema de Estadística y Probabilidad ha mostrado menores niveles de aprendizaje que las otras áreas de la matemática [5]. Por otro lado, durante el 2020 esta área fue la más perjudicada del "apagón educativo debido a la pandemia", ya que se dejó de lado en primaria y en secundaria [6]. Estos resultados señalan que la enseñanza de la Estadística y la Probabilidad sigue sin afianzarse en las instituciones costarricenses.

Debido a lo anterior, resulta relevante que las personas docentes conozcan nuevas formas de abordar los distintos temas de Estadística y Probabilidad, ya que para muchas de estas personas estos temas y su didáctica son tópicos bastantes nuevos. En particular, la importancia del muestreo puede ser un tema complejo de enseñar, debido a que se basa en matemáticas avanzadas, es por esto, que es importante crear espacios de discusión para que las personas docentes aclaren sus dudas y puedan dictar sus clases con mayor confianza.

Ahora bien, el objetivo de este trabajo es construir una estrategia didáctica para la enseñanza del muestreo, que permita que el estudiantado valore el potencial de esta técnica estadística.

Descripción de la estrategia didáctica

La estrategia didáctica que se propone en este trabajo consiste en realizar un laboratorio dirigido a que la persona estudiante descubra por sí misma que el muestreo aleatorio con tamaños relativamente pequeños presenta promedios de una variable de interés muy cercanos al promedio en la población. Este laboratorio se desarrollará con el software

Excel y la persona estudiante no necesitará conocer de fórmulas, ya que se le dará una hoja de cálculo con la base de datos de trabajo y los comandos escritos.

El laboratorio presenta los siguientes momentos:

- Acercamiento a una base de datos de una población: En este paso la persona estudiante analizará los estadísticos descriptivos de las variables de una población de interés (mínimo, máximo, promedio y desviación estándar). Dado que se utilizará una hoja de Excel prediseñada, la persona solo debe interpretar los estadísticos seleccionados.

  Se puede utilizar una base de datos de los estudiantes del nivel de interés, la cual puede ser recolectada por medio de un formulario de Google. Esta base puede tener 3 variables: una dicotómica, una cuantitativa discreta y una cuantitativa, por ejemplo, juega fútbol (1=sí o 0=no), número de hermanos y nota en el último examen de matemática.

  La base de datos que se utilizará en este laboratorio es una base ficticia con 200 casos y las variables mencionadas en el ejemplo.

- Selección y estudio de muestras aleatorias: En este momento, se le enseñará a la persona estudiante que la función “aleatorio.entre” de Excel genera números aleatorios en un rango determinado y se le indicará que esta función es útil para generar muestras.

  Luego, se mostrará que en una sección de la hoja asignada hay una columna con 35 espacios con esta función digitada y se les hará ver que cada vez que oprimen el botón *enter* se generan nuevos números aleatorios (la justificación de este tamaño de muestra se presenta en la sección de tamaño de muestra). Esta columna es el conjunto de casos que conforma una muestra aleatoria. Luego de este paso, las personas deberán realizar una leve manipulación de la hoja, la cual consiste en copiar los casos aleatorios y pegarlos con formato *valores* en el espacio correspondiente para la muestra aleatoria definitiva. Esta manipulación deberán hacerla cinco veces, ya que en la hoja de excel hay 5 segmentos dedicados a la construcción de 5 muestras aleatorias de 35 casos.

  Posteriormente, en cada una de estas muestras la persona estudiante observará el promedio de cada una de las tres variables de la base de datos. Estos promedios serán generados automáticamente por la hoja de Excel utilizada.

- Comparación de promedios: En este paso se analizará la diferencia de los promedios obtenidos en las muestras con el verdadero promedio.

  Además, se les pedirá a los estudiantes que determinen cuántas diferencias en la variable 1, 2 y 3 son menores al 10%, a 0,5 y a 5, respectivamente (se considerará este valor como el límite para una diferencia pequeña).

- Socialización de los resultados de la variable 1 y conclusión general sobre variables dicotómicas: En este momento, el docente puede indicarles a sus estudiantes que al estudiar variables dicotómicas, como en las encuestas, la

selección de 35 personas de 200, permite obtener en aproximadamente el 80% intentos diferencias de promedios menores al 10%.

Luego, se procede a contar cuántos estudiantes obtuvieron 0, 1, 2, 3, 4 y 5 muestras con diferencias menores al 10% y se determina el porcentaje de muestras estimadas en las que la diferencia fue menor al límite establecida. Se espera que la mayoría del estudiantado obtengan 4 o 5 muestras con diferencias pequeñas.

Con base en este proceso, se concluye que el muestreo aleatorio permite obtener, con un grado de certeza preestablecido, estimaciones del promedio cercanas al real. Por último, se menciona que las casas encuestadoras, por lo general, escogen un tamaño de muestra para que la diferencia de los promedios en variables dicotómicas sea menor al 3% en 95 de cada 100 intentos. Para fortalecer esta conclusión se puede presentar un artículo como el de [4].

- Análisis del resto de variables: En la última fase de la estrategia se le puede pedir a un estudiante que exponga las diferencias observadas en las otras variables y se analiza en conjunto con el grupo si estas diferencias fueron pequeñas o grandes. En este parte, se debe indicar al estudiantado que el tamaño de muestra utilizado permite obtener diferencias pequeñas de los promedios en gran cantidad de muestras aleatorias, siempre y cuando las variables estudiadas tengan una "variabilidad baja". En el caso de la base de datos de estudio se sabe que habrá diferencias pequeñas (en la sección de tamaño de muestra se indicará por qué se puede hacer esta conclusión).

Para ejemplificar los momentos del laboratorio, en la tabla 1 se presentan los promedios que pudo haber obtenido un estudiante en las cinco muestras aleatorias. Se puede observar que en las 5 muestras estudiadas, las diferencias del promedio de la variable "juega fútbol" con el promedio verdadero fueron inferiores al 10%.

En la variable número de hermanos, la cual presenta un intervalo de variación de 0 a 4, la diferencia de promedios en las muestras seleccionadas fue a lo más de 0,14 unidades, es decir, un 3,5% del intervalo observado. Mientras que en la nota de matemática la mayor diferencia fue de 2,42 unidades, es decir, un 2,42% del rango posible de variación de la nota (de 0 a 100) y un 7,11% del rango real observado (de 51 a 85). Por lo cual, en este caso en particular, este tamaño de muestra parece ayudar a obtener resultados cercanos al promedio real.

Tabla 1

Promedios de las variables de interés en 5 muestras aleatorias de 35 personas

Num. muestra	**Juega fútbol**		**Num. Hermanos**		**Nota en Matem.**	
	Prom	Dif	Prom	Dif	Prom	Dif
1	0,77	-0,09	1,09	0,02	68,45	0,16
2	0,63	0,05	1,23	-0,12	66,19	2,42
3	0,77	-0,09	1,00	0,11	68,02	0,59
4	0,71	-0,03	0,97	0,14	70,24	-1,63
5	0,74	-0,06	1,11	0,00	68,61	0,00
Población	0,68		1,11		68,61	

El uso de Excel en la estrategia didáctica

Para poder implementar la estrategia didáctica expuesta es necesarios describir dos elementos esenciales: la hoja de trabajo y la estrategia para la construcción de la muestra.

La hoja de trabajo

La hoja de trabajo de Excel viene preparada previamente para que la persona estudiante la complete. La estructura es la siguiente:

- Filas 1-201: Base de datos.
- Filas 203-212: Espacio para los análisis descriptivos.

	A	B	C	D	E	F	G
203	**Estadístico**	**Fútbol**	**Hermanos**	**Matem.**	**Dif Fút**	**Dif Her**	**Dif Mat**
204	**Mínimo Pob**	=Min(B2:B201)	=Min(C2:C201)	=Min(D2:D201)			
205	**Máximo Pob**	=Max(B2:B201)	=Max(C2:C201)	=Max(D2:D201)			
206	**Desv. Est. Pob**	=Desvest(B2:B201)	=Desvest(C2:C201)	=Desvest(D2:D201)			
207	**Promedio Pob**	=Promedio(B2:B201)	=Promedio(C2:C201)	=Promedio(D2:D201)			
208	**Prom. Muestra 1**	=Promedio(B216:B250)	=Promedio(C216:C250)	=Promedio(D216:D250)	=B208-B$207	=C208-C$207	=D208-D$207
209	**Prom. Muestra 2**	=Promedio(B256:B290)	=Promedio(C256:C290)	=Promedio(D256:D290)	=B209-B$207	=C209-C$207	=D209-D$207
210	**Prom. Muestra 3**	=Promedio(B296:B230)	=Promedio(C296:C230)	=Promedio(D296:D230)	=B210-B$207	=C210-C$207	=D210-D$207
211	**Prom. Muestra 4**	=Promedio(B336:B370)	=Promedio(C336:C370)	=Promedio(D336:D370)	=B211-B$207	=C211-C$207	=D211-D$207
212	**Prom. Muestra 5**	=Promedio(B376:B410)	=Promedio(C376:C410)	=Promedio(D376:D410)	=B212-B$207	=C212-C$207	=D212-D$207

- Filas 215-250: Espacio para la muestra 1.
- Filas 255-290: Espacio para la muestra 2.
- Filas 295-330: Espacio para la muestra 3.
- Filas 335-370: Espacio para la muestra 4.
- Filas 375-410: Espacio para la muestra 5.

El muestreo con Excel

Para generar los casos de una muestra aleatoria y recuperar los datos de los casos seleccionados se utilizará la siguiente guía de comandos:

	A	B	C
215	**Casos definitivos**	**Futbol**	**Hermanos**
216	Pegar valores	=INDIRECTO(DIRECCION(A216;2;;;))	=INDIRECTO(DIRECCION(A216;3;;;))
217	…	…	…

	D	E
215	**Matemática**	**Casos iniciales**
216	=INDIRECTO(DIRECCION(A216;4;;;))	=aleatorio.entre(2;201)
217		…

En la columna E se utilizará la función "=aleatorio.entre(2;201)" para determinar la fila en la que se encuentran los casos seleccionados en la muestra aleatoria. Los valores 2 y 201 indican el rango de filas en el que se ubican los casos estudiados. Luego, se copiarán estos valores y se pegarán en la columna A en formato "valores", con lo cual se tendrán los números de casos definitivos. Este paso será la única manipulación de la hoja que los estudiantes deberán realizar durante el taller.

La función de la columna B viene escrita en la hoja de Excel, esta llama al valor de la variable "Fútbol" obtenido por el caso seleccionado aleatoriamente. En el caso del dato que iría en la casilla B216, el número de fila se indica en la casilla A216 y el número de columna es la 2 (número asignado a la columna B). De igual manera, se realizará en la columna C y D para llamar a los valores de las variables "Hermanos" y "Matemática".

El tamaño de muestra

Otro aspecto que se debe considerar para el desarrollo de la estrategia es el cálculo del tamaño de muestra, ya que si el docente va a utilizar una base distinta a la propuesta, el tamaño de muestra requerido puede variar.

Para determinar el tamaño de muestra de un estudio se deben considerar cuatro elementos: el error muestral permitido (ε), el nivel de confianza $(1-\alpha)*100\%$, la desviación estándar poblacional de la variable (σ) y el tamaño de la población (N).

El error muestral indica cuál es la máxima diferencia entre el promedio muestral y el poblacional que se desea observar. El nivel de confianza indica la probabilidad de que el promedio de la muestra se ajuste al límite de error establecido.

La fórmula para determinar el tamaño de muestra para obtener un error muestral menor a ε, en aproximadamente un $(1-\alpha)*100\%$, equivale a

$$n=\frac{NZ_{\alpha}^{2}\sigma^{2}}{\varepsilon^{2}(N-1)+Z_{\alpha}^{2}\sigma^{2}}$$

con Z_{α} el cuantil α de la función de densidad acumulada por la derecha de la función normal, cuyos valores más comunes son 1,28; 1,64 y 1,96 para niveles de confianza de 80%, 90% y 95%, respectivamente [1].

Para determinar el tamaño de muestra requerido para la variable "juega fútbol", considerando $\varepsilon=0{,}1$ y una confianza del 80% ($\alpha=0{,}20$), el tamaño requerido es de

$$n=\frac{200(1{,}28)^{2}(0{,}47)^{2}}{(0{,}1)^{2}(200-1)+(1{,}28)^{2}(0{,}47)^{2}}=30{,}77<31$$

Con lo cual se concluiría que el tamaño de muestra debería ser 31, ya que es el número natural mínimo con el que se cumplen las propiedades deseadas.

Para las otras variables estudiadas en el taller y los errores muestrales máximos considerados, el tamaño de muestra de 35 personas fue mucho superior al requerido. En el caso del número de hermanos el tamaño muestral requerido era

$$n = \frac{200(1{,}28)^2(0{,}90)^2}{(0{,}5)^2(200-1) + (1{,}28)^2(0{,}90)^2} = 5{,}19 < 6$$

mientras que en el de nota de matemática era

$$n = \frac{200(1{,}28)^2(9{,}16)^2}{(5)^2(200-1) + (1{,}28)^2(9{,}16)^2} = 5{,}37 < 6$$

El tamaño de muestra sin información a priori

La desventaja con las estimaciones del tamaño de muestra utilizadas previamentes es que demandan información poblacional de la variable (la desviación estándar), la cual es poco esperable que se maneje antes de que se realice un estudio.

Ahora bien, en variable dicotómicas no es necesario conocer la desviación estándar de las variables debido a que tienen un máximo conocido: 0,5 [7]. Por lo cual, cuando se realizan estudios con variables dicotómicas, se utiliza el máximo de la desviación estándar de la variable. En el caso de la variable "juega fútbol", el tamaño de muestra considerando un error de $\varepsilon = 0{,}1$ y una confianza del 80% ($\alpha = 0{,}20$), implica que el tamaño de muestra sea:

$$n = \frac{200(1{,}28)^2(0{,}5)^2}{(0{,}1)^2(200-1) + (1{,}28)^2(0{,}5)^2} = 34{,}14 < 35$$

Con base en el resultado anterior es que en este estudio se decidió utilizar un tamaño de muestra de 35 personas.

Por otro lado, es importante mencionar que si el tamaño de la población es suficientemente grande, la fórmula de cálculo de tamaño de muestra se simplifica a

$$n = \frac{Z_\alpha^2 \sigma^2}{\varepsilon^2}$$

Discusión

Este documento presenta una propuesta de taller para la enseñanza del muestreo en secundaria, la cual demanda de un conocimiento básico de Excel. La razón por la cual se brinda una hoja con todas las funciones escritas es que el objetivo de la estrategia es que la lección se dedique al estudio del muestreo, en lugar de a las funciones. Si alguna persona estudiante presenta problemas con la copia de los casos seleccionados en la muestra, la persona docente puede proporcionarle una hoja prediseñada con los casos seleccionados, para evitar la pérdida de tiempo en un detalle irrelevante.

Se considera que esta estrategia es apropiada para la enseñanza del muestreo, porque permite explorar una base de datos original y observar claramente que las muestras aleatorias son subconjuntos de estas poblaciones. Esta relación de subconjunto, aparentemente tan trivial, es una idea difícil de comprender cuando no hay una verdadera manipulación de datos.

Por otro lado, el contraste de los estadísticos muestrales con los parámetros poblacionales es el principal argumento para justificar la relevancia del uso del muestreo, la observación de estos contrastes le brindará al estudiante una experiencia real con la verdadera importancia del muestreo.

Esta estrategia también brinda una oportunidad para estudiar otros conceptos adicionales, como el margen de error o el nivel de confianza. No obstante, el objetivo principal de la estrategia es que el estudiantado comprenda que el muestreo adecuado permite, "en la mayoría de los casos" (referencia a la confianza), obtener aproximaciones "precisas" (referencia al error muestral) del promedio muestral.

Finalmente, hay que mencionar que esta estrategia es una propuesta que no ha sido experimentada, por lo cual, debe evaluarse su pertinencia. Se espera que la implementación apropiada de esta estrategia provoque que el estudiantado comprenda la relevancia y el funcionamiento del muestreo aleatorio.

Referencias

[1] Kish, L. Muestreo de Encuestas. Editorial Trillas. 1965.

[2] López, R. "Opinión pública y encuestas de opinión en España". Revista mexicana de opinión pública (2020) 28, 149-179.

[3] Ministerio de Educación Pública. Programa de Estudios de Matemática. MEP. 2012.

[4] ODI UCR. "Encuesta: costarricenses muestran opiniones favorables sobre el desempeño del presidente". En https://www.ucr.ac.cr/noticias/2022/08/03/encuesta-costarricenses-muestran-opiniones-favorables-sobre-el-desempeno-del-presidente.html (consultada el 16 de setiembre, 2022).

[5] Programa Estado de la Nación. Séptimo Estado de la Educación Costarricense. PEN. 2019.

[6] Programa Estado de la Nación. Octavo Estado de la Educación Costarricense. PEN. 2021.

[7] Wackerly, D. D., Mendenhall, W. y Scheaffer. R. L. Estadística matemática con aplicaciones. Cengage Learning. 2008.