DISEÑO DE MUESTREO Y APLICACIONES

MUESTREO ALEATORIO (10 MINUTOS)

Como vimos en la primera clase, la estadística que estamos aprendiendo en este curso se basa en hacer inferencias de una muestra para sacar conclusiones sobre una población. Ahora la pregunta es, ¿por qué usar una muestra cuando se puede hacer un censo de toda la población? La respuesta depende de la situación, pero usualmente es más fácil y económico seleccionar una muestra. Tomar una muestra de una población es algo que hacemos todos los días. Por ejemplo, al decidir si van a ir o no ir a una fiesta, no le preguntan a todos sus amigos si ellos irán, sino solamente a unos pocos. Algo similar pasa cuando están decidiendo qué película ver en el cine: le preguntan a un grupo de amigos y no a todos los que la han visto para inferir si la película es entretenida o no.

Tanto en el ejemplo de la fiesta como de la película, asumimos que uno le pregunta a sus amigos. Es decir, la elección de la muestra no es al azar o aleatoria, sino que está restringida a nuestras amistades. Pero esta no es la única forma de hacer un muestreo. Como veremos en esta clase, existen distintas formas de selección una muestra de la población con el fin de lograr que la represente de la mejor manera y que sea conducente a los objetivos de la investigación. Por ejemplo, aunque suene obvio, si solo queremos ver el ingreso de personas en edad de jubilar, en la muestra no deberían entrar personas con menores de 65 años.

Para la inferencia estadística el ideal es que el muestreo sea completamente aleatorio, sin estar condicionado por preferencias conscientes o subconscientes, limitaciones de presupuesto, conveniencia u cualquier otro factor distinto al azar. Esto no es siempre posible, pero afortunadamente existen alternativas con ventajas y desventajas bien descritas. En términos generales, existen dos errores típicos que tenemos que evitar: un error de muestreo en que hacemos conclusiones demasiado generales acerca de una población a partir de una muestra demasiado específica, o un error de inferencia en donde hacemos conclusiones de una población mucho más grande de la que se tomó la muestra originalmente.

TIPOS DE MUESTREO (45 MINUTOS)

Para dar una buena pincelada de lo que es el muestreo, dividiremos en dos grandes grupos los diferentes tipos de clasificación de muestras. Esto no significa que existan diferentes criterios de clasificación. Los dos grande grupos que veremos serán los métodos de muestreo probabilístico y métodos de muestreo no probabilístico.

Muestreo probabilístico

Este tipo de muestreo se basa en que cada individuo de la población tiene la misma probabilidad de ser elegido en la muestra “n”, y de la misma manera toda muestra “n” tiene la misma posibilidad de ser seleccionada. Un ejemplo muy simple es una rifa dentro del curso: si a cada uno le asignamos un numero distinto del 1 al 39, todos tienen la misma probabilidad de ser elegidos.

- *Muestreo aleatorio simple*

Este tipo de muestreo en específico lo estudiaremos con mayor profundidad más adelante. Básicamente se trata de elegir los elementos de la población de uno en uno, de forma que cada individuo tenga la misma posibilidad de ser elegido. Si bien este modelo se ve muy fácil y cómodo de implementar, no es muy útil con poblaciones muy grandes. Las formas de aplicar este tipo de muestreo aleatorio simple, incluyen por ejemplo: sacar números de una bolsa, utilizar una tabla de números aleatorios o recurrir a números generados por algún programa.

¿Qué ejemplos de la vida cotidiana se les ocurre en donde existe un muestreo aleatorio simple?

- *Muestreo sistemático*

Para muestreos probabilísticos con poblaciones grande se necesita mucho tiempo para tomar una muestra aleatoria simple. Una alternativa para este tipo de poblaciones es usar un muestreo sistemático. El muestreo sistemático consiste en hacer la elección de los números de una muestra en base al total de la población y el tamaño que se requiere.

$$k = N/n$$

N es el tamaño de la población, y n es el tamaño de la muestra. El número k representa cada cuántos elementos de la población se elegirá un agente para la muestra. En la práctica, lo que hacemos es elegir un parámetro de partida, el cual tiene que ser un número al azar entre 1 y k. Este parámetro de partida lo denotaremos con la letra i. Por lo tanto, los elementos que integraran la muestra serán:

$$i\ , i+k\ , i+2k\ , i+3k\ , \dots , i+(n-1)k$$

Este tipo de muestreo puede tener ciertas falencias, y es el hecho de que al estar determinado por una periodicidad constante (k), este puede replicar ciertos patrones en la población, introduciendo homogeneidad en la muestra que no necesariamente se da en la población. Por ejemplo, si seleccionamos personas sobre listas de 10 personas, donde las primeras 5 son mujeres y los últimos 5 son hombres en cada lista, con k=10, estaríamos solo eligiendo mujeres o solo hombres dependiendo de nuestro punto de partida, sin haber una representación lógica de la población al no tener una muestra de ambos sexos.

Otro ejemplo podría ser elegir a 10 compañeros para que hagan un trabajo para la próxima clase, dónde $k = \frac{40}{10} = 4$. Si el número i fuera 3, los números 3, 7, 11, 15, 19, 23, 27, 31, 33 y 37 de la lista formarían parte de la muestra y podríamos sacar el promedio del curso, la estatura y muchas otras características.

¿En qué situación podríamos querer aplicar un muestreo aleatorio sistemático?

- *Muestreo aleatorio estratificado*

Lo que se pretende con el muestreo aleatorio estratificado es considerar distintas categorías que posean gran homogeneidad a su interior. Los estratos pueden ser según profesión, sexo, edad, región, estado civil, etc. Lo importante es que todos los estratos de interés estarán bien representados en la muestra. Elaborar este tipo de muestreo puede tener ciertas complicaciones porque para elaborar una estratificación acertada requerimos de un conocimiento bastante profundo de la población.

La distribución de la muestra se puede hacer tomando una muestra aleatoria simple de cada estrato. A la vez existen varias formas de combinar los resultados de la muestra, tales como las siguientes:

  - Afijación simple: cada estrato reparte igual número de observaciones a la muestra.
  - Afijación proporcional: la distribución se hace de acuerdo al peso que tiene el estrato con respecto a la población.
  - Afijación óptima: se toma en cuenta la dispersión de los datos en cada estrato, pudiendo considerar la proporción y la desviación estándar. Dado que generalmente no se conoce la desviación estándar, este caso tiene poca aplicación.

El muestreo aleatorio estratificado tiende a mostrar mejores resultados mientras más homogéneos sean los elementos del estrato. Cuando los estratos son homogéneos a su interior, estos tienen varianza pequeña y permiten

buenas estimaciones con muestras relativamente pequeñas. Los resultados pueden ser casi tan precisos como los elaborados a partir de un muestreo aleatorio simple, pero con una muestra más pequeña.

Un buen ejemplo es el Estudio Nacional de Opinión Pública, elaborado por el Centro de Estudios Públicos (CEP), donde se tratan temas como los siguientes: percepción económica, visión del país, principales problemas, identificación política, evaluación de las coaliciones políticas, coyuntura política, reformas en curso, temas laborales, temas valóricos, evaluación de personajes políticos y evaluación del gobierno, el congreso y los jueces. El universo es la población urbana y rural de 18 años y más, residente a lo largo de todo el país, excluyendo Isla de Pascua. El muestreo aleatorio estratificado incluyó en julio del 2014 a 1.442 persona entrevistadas en sus hogares, que vivían en 149 comunas del país. Se estratifico por región y zona urbana/rural con un muestreo aleatorio y probabilístico en cada una de sus tres etapas (manzana-hogar-entrevistado). El error muestral obtenido fue de $\pm 3\%$, considerando varianza máxima y un 95% de confianza.

¿Qué tipo de afijación creen que se puedo haber utilizado, o cual creen que es la más correcta?

- *Muestreo por conglomerado*

En este el muestreo por conglomerado se dividen los elementos de la muestra en grupos separados, llamados conglomerados. Generalmente se dividen en comunas, barrios, manzanas, para muestreo de áreas bien definidas. Este muestreo requiere un mayor tamaño de muestra que el muestreo aleatorio simple o el estratificado. Sin embargo, el muestreo por conglomerado es muy útil a la hora de reducir costos, ya que al recolectar los datos en una manzana, por ejemplo, se obtienen muchos datos en poco tiempo. En la práctica lo que se hace es elegir aleatoriamente un cierto número de conglomerados que permite alcanzar el tamaño muestral establecido.

¿Qué otros ejemplos de muestreos por conglomerados se les ocurren?

Muestreo no probabilístico

Muchas veces a la hora de elaborar una muestra uno puede encontrarse con barreras de costos. Una alternativa a métodos de muestreo probabilístico es el muestreo no probabilístico. Este tipo de muestreo no es útil para hacer inferencia sobre la población, ya que no tenemos la certeza de que la muestra sea representativa. A diferencia de la clasificación anterior, los sujetos de la población no tienen la misma posibilidad de ser elegidos en la muestra. De todas maneras se sigue intentando que los seleccionados, dentro de lo posible, conformen una muestra representativa de las características de la población (pero por probabilidades no es estadísticamente representativa).

- *Muestreo por cuotas*

El muestreo por cuotas se asemeja al muestreo aleatorio estratificado, pero no tiene el carácter de aleatorio o probabilístico. El muestreo por cuotas se basa en el conocimiento de los estratos de la población o de los individuos más representativos para los fines de la investigación.

Es muy común que este tipo de muestreo se aplique en encuestas de opinión públicas, donde se establecen "cuotas" de los integrantes que debe tener la muestra, es decir, que tengan determinadas características. Por ejemplo, una encuesta sobre una nueva marca de ropa juvenil podría realizarse mediante una encuesta a 100 personas entre 18 y 25 años, en Santiago, de las cuales 50 tienen que ser hombres y 50 mujeres. Al determinar esta cuota, alguien podría venir a la UDP y cumplir con estas cuotas que no necesitan ser aleatorias.

¿Qué otros ejemplos de encuestas por cutas conocen?

- *Muestreo de conveniencia*

El muestro por conveniencia se orienta por criterios de comodidad y no probabilidades. Pensemos en encuestas que se les hace a personas de fácil acceso. El ejemplo más típico es cuando un profesor está haciendo una investigación y emplea encuestas a sus propios alumnos simplemente porque están a la mano.

¿Qué otros ejemplos de encuestas de fácil acceso conocen?

- *Bola de nieve*

El muestreo de bola de nieve va armando la muestra a partir de cada encuestado. Primero se encuesta a un individuo y este conduce a otros individuos, de modo que la "bola" va creciendo hasta conseguir la muestra necesaria. Este tipo de muestreo típicamente se aplica a delincuentes, sectas, enfermedades con alto estigma u otras poblaciones "marginales."

- *Muestreo discrecional o subjetivo*

En el muestreo direccional o subjetivo, el investigador elabora la muestra y elige a los sujetos que la integraran. Esto se puede deber a que tiene un conocimiento profundo del tema o que considera que esos agentes son representativos de la población.

EJERCICIO VENTAJAS Y DESVENTAJAS DE LOS DISTINTOS MUESTREOS PROBABILISTICOS (20 MINUTOS)

	Ventajas	**Desventajas**
Muestreo aleatorio simple	• • • •	• • • •
Muestreo sistemático	• •	•
Muestreo aleatorio estratificado	• • •	•
Muestreo por conglomerado	• •	• •

Calculadores de muestra

En este curso también estudiamos el tamaño que debe tener una muestra y cómo calcularlo. Hoy en día existen muchas herramientas que nos ayudan a calcular el tamaño necesario de una muestra. En internet encontramos diversas páginas que pueden resultar de ayuda para esto. Por ejemplo, Creative Research System es una página orientada a la elaboración de softwares para estudios de mercado, recursos humanos, ciencias sociales, encuestas políticas y cualquier otro que use cuestionarios. En su página tienen un calculador de muestras, donde explican los conceptos y operaciones que hay detrás del cálculo: http://www.surveysystem.com/sscalc.htm.

Para calcular el tamaño muestral necesitamos definir el intervalo de confianza, nivel de confianza y tamaño de la población. El intervalo de confianza es el margen de error que puede tener cierta información recogida de la muestra. El nivel de confianza es un porcentaje que representa qué tan seguro podemos estar que la respuesta escogida por toda la población está dentro del intervalo de confianza. Por ejemplo, en la primera clase vimos un ejemplo sobre aprobación presidencial cuyo margen de error fue de +/- 3,7%. Esto quiere decir que si la desaprobación presidencial alcanzó un 70% en la muestra de 716 personas, si le preguntáramos a toda la población el resultado podría estar entre 66,3% o 73,7%. Esto último es el intervalo de confianza del 95%.

Probemos en la calculadora estos datos, con una población en edad de votar de 10.000.000. ¿Qué pasa si solo contamos la cantidad de habitantes que fue a votar (6.000.000)? ¿Cómo cambia el resultado si aumentamos el intervalo de confianza? ¿Es una muestra representativa de la población?