Propuesta metodológica para el diseño muestral de una encuesta de hogares multipropósito

Observatorio Laboral de Ñuble

Héctor Garrido Henríquez

Antecedentes

El observatorio Laboral de Ñuble se ha fijado como objetivo investigar tres temáticas relevantes para el desarrollo de la nueva región

- Inmigración internacional
- Informalidad laboral
- Adopción de tecnologías de la información

Para abordar estas temáticas no basta con los instrumentos cuantitativos disponibles (CENSO 2017, Encuesta CASEN 2017, Encuesta Nacional de Empleo, Encuesta Suplementaria de Ingresos, entre otras)

Antecedentes

Esta propuesta metodológica consiste en:

- Un diseño muestral multietápico que sigue los estandares de las encuestas de hogares más importantes del País (Encuesta Nacional de Empleo y Encuesta CASEN).
- Una propuesta de Matching Estadístico para aprovechar el levantamiento de la Encuesta CASEN 2017. Esto implica levantar un número acotado de preguntas en el cuestionario de manera tal de poder combinar ambos instrumentos y aprovechar al máximo el levantamiento de información.

Objetivo de la encuesta.

- El objetivo de la encuesta es caracterizar y cuantificar la población de 15 años y más respecto de tres temáticas: Inmigración internacional, informalidad laboral y adopción de tecnologías de la información.

Población objetivo.

La población objetivo la constituye toda la población de la ciudad de Chillán de la zona urbana residente en viviendas particulares ocupadas.Esta definición excluye a la población que habita en viviendas colectivas como hospitales, cárceles, conventos, cuarteles y otros, pero incluye a las personas que residen en viviendas particulares dentro de dichos centros, como son los porteros, conserjes y otros residentes

Marco Muestral

Cartografía digital del precenso 2016 y el censo de población y vivienda 2017.

A la fecha aún no se libera la base de datos del censo 2017, por lo que en esta presentación se utiliza provisoriamente la información del censo 2002.

Tipos de encuestas

A grandes rasgos existen dos tipos de diseños muestrales:

- Diseños probabilísticos: Para estudios cuantitativos, cuya principal característica es que cada elemento de la muestra tiene una probabilidad conocida de ser seleccionado. **Require necesariamente de un marco muestral exhaustivo**
- Diseños intencionados: Para estudios cualitativos o en aquellos casos en que no se disponga de un marco muestral adecuado. **No se puede realizar inferencia estadística sobre este tipo de muestras. Es decir, las conclusiones no son extrapolables a la población en estudio desde un punto de vista estadístico**

Un error común: Una muestra es probabilística o no lo es ¡No hay intermedios!

Es común cometer errores respecto de los puntos anteriores. Un mal diseño muestral, por ejemplo, es el que utiliza la encuesta *Plaza Pública CADEM*, donde la muestra se recoge principalmente en lugares de alta afluencia de público, como las estaciones de metro, salidas de centros comerciales, etc.

¡La probabilidad de que cada sujeto sea seleccionado es desconocida! ¡No se puede realizar inferencia estadística con ella!

Los resultados desastrosos de esta encuesta (y muchas otras) están a la vista. Fracasaron enormemente en predecir los resultados de la primera vuelta presidencial de 2017.

Estimación del tamaño muestral

Para estimar el tamaño muestral de la encuesta se utilizó la siguiente expresión

$$m_3 = \frac{Z^2_{1-\frac{\alpha}{2}} \cdot S(p)^2}{e_0^2 + Z^2_{1-\frac{\alpha}{2}} \cdot Deff(p) \cdot \frac{S(p)^2}{M}} \cdot \frac{Deff(p)}{(1 - tnr)}$$

Estimación del tamaño muestral

Donde:

- p Prevalencia de la variable cualitativa de interés. Para este estudio, la tasa de desempleo.
- n número de conglomerados o PSU. Para este estudio, 3.595 manzanas.
- $\bar{m}$ número promedio de viviendas a encuestar por PSU. Tomado de CASEN 2015.
- M número de viviendas en la población. 42.916 según CENSO 2002.
- $Deff(p)$ efecto del diseño sobre la varianza de la prevalencia de la variable cualitativa de interés. Es decir, ¿Cuánto aumenta la varianza debido al diseño muestral respecto de un diseño aleatorio simple?
- $SE(p)$ Error de estándar de la tasa de desempleo.
- $S(p)^2$ Cuasivarianza poblacional de la tasa de pobreza
- $Z^2_{1-\frac{\alpha}{2}}$ corresponde al cuantil de la distribución normal asociado a una probabilidad acumulada de $1-\frac{\alpha}{2}$. Con $\alpha = 0.05$

Estimación del tamaño muestral

- Teniendo en cuenta lo anterior, el tamaño muestral de viviendas (SSU) se estima en 845.
- Esto aún no soluciona el problema. Es necesario determinar el número de manzanas a utilizar y subsecuentemente el número de viviendas al interior de dichas manzanas.
- Para estos fines no existe una formula matemática explícita, por lo que el proceso debe realizarse a través de un algoritmo recursivo que se describe a continuación

Algoritmo de selección de conglomerados.

El algoritmo consta de los siguiente pasos.

1. Se agrupan las manzanas en 30 categorías de acuerdo al número de viviendas (Clasificación provista por el INE)
2. Se selecciona al azar un grupo de manzanas con probilidad proporcional al número de viviendas.
3. Dentro del grupo seleccionado se selecciona una manzana con probabilidad igual mediante MAS.
4. Dentro de la manzana escogida en el paso anterior se selecciona un 25% de las viviendas.
5. Se quita la manzana del marco muestral. El algoritmo se repite hasta alcanzar el tamaño muestral objetivo.

Algoritmo de selección de conglomerados.

Ejecutado el algoritmo anterior, se tiene que para alcanzar el tamaño muestral de viviendas se trabajará con 247 manzanas, con un mínimo de 2 viviendas por manzana, un promedio de 3 y un máximo de 5 viviendas.

A pesar de que la encuesta está orientada hacia personas, hasta el momento se ha trabajado con las viviendas como última unidad de muestreo. En el trabajo de campo se debe velar porque al acudir a una vivienda, el jefe de hogar (idealmente), reporte la información de todos los sujetos en edad economicamente activa. Esto significa que el tamaño definitivo de la encuesta se conoce en última instancia una vez realizado el trabajo de campo.

Desarrollo de factores de expansión

En una encuesta de carácter multietápico, como la aquí propuesta, cada observación representa un número diferente de sujetos en la población. Esto no ocurre en las muestras seleccionadas puramente al azar, donde todas las observaciones representan el mismo número de sujetos en la población. Es por esto necesario construir factores de expansión que permitan realizar inferencia de manera adecuada.

La construcción de los factores de expansión no es trivial y consta de varios pasos. En primer lugar se debe construir un factor de expansión para las manzanas, luego para las viviendas y luego para las personas.

Desarrollo de factores de expansión.

El algoritmo de selección de conglomerados descrito previamente no permite determinar formulas explícitas para las probabilidades de selección. Por lo que es necesario recurrir a un procedimiento numérico para la construcción de factores de expansión para manzanas.

Se utiliza un procedimiento descrito en la literatura como "Ponderadores bootstrap". Una vez determinado en el paso anterior la lista de manzanas a las cuales se acudirá se realiza un estudio de simulación con el marco muestral que consiste en tomar 2000 muestras aleatorias y determinar cuántas veces fueron seleccionadas las manzanas inicialmente escogidas. El factor de expansión será entonces el inverso de la probabilidad empírica de selección.

Desarrollo de factores de expansión.

Una vez realizado el procedimiento anterior, se determina que la mediana del factor de expansión es de 14.6, mientras que el mínimo es de 12.35 manzanas y el máximo 17.86 manzanas.

Para determinar los factores de expansión de viviendas y de personas se requiere de las estadísticas de población del CENSO, por lo que se está a la espera de la publicación de la base de datos.

Planificación del trabajo de campo

Mediante los pasos descritos previamente se tiene un conjunto específico de manzanas a encuestar las cuales serán ubicadas mediante la cartografía digital del CENSO utilizando algún sistema de información geográfica (SIG)[1]

Un último paso consiste en seleccionar las viviendas mediante muestreo sistemático[2] al interior de cada manzana.

[1] Para estos fines puede utilizarse QGIS o la librería `ggmaps` de `R`

[2] Este algoritmo permite al igual que el MAS, tener iguales probabilidades de selección

Algunas consideraciones sobre Matching Estadístico.

Para este período en particular hay disponibles varias bases de datos que pueden ser utilizadas para el estudio de las temáticas de interés del observatorio:

- Censo de población y vivienda 2017. Todavía no disponible
- Encuesta de Caracterización socioeconómica Nacional (CASEN) 2017. Todavía no disponible.
- Encuesta Nacional de Empleo.

Existe la posibilidad de 'emparejar' alguna de estas bases de datos con la encuesta que diseñará el OLR, por lo que es necesario no redundar en la información que se recoja. Aún así, debe existir un conjunto de variables en común, las cuales permitirán realizar el matching. Una revisión exhaustiva de los cuestionarios de estos instrumentos es necesaria.