

Ponderación en encuestas

Juan Pablo Ferreira

1. Introducción

Los ponderadores de la muestra son un componente fundamental para producir las estimaciones de los distintos indicadores de la ELPS. Los datos de la muestra son ponderados de forma de obtener estimaciones tanto a nivel nacional como para otros dominios de estudio.

Por ejemplo, la estimación de un total de una variable cualquiera queda definido como

$$\hat{t} = \sum_{i \in s} w_i \times y_i$$

donde y_i y w_i son el ponderador y el valor que toma la variable en el elemento i respectivamente perteneciente a la muestra aleatoria (s).

Debido a que se releva información de un integrante del hogar (seleccionado al azar) así como del hogar (y la vivienda), es necesario disponer de dos sistemas de ponderadores. Uno para el integrante seleccionado y otro para el hogar.

El cálculo de la definición de los ponderadores se realizó en dos etapas. En la primera etapa, se determinaron los ponderadores de los hogares en base a la probabilidad de selección de los mismos en la encuesta, a las tasas de respuestas obtenidas en campo y a un ajuste producto de las proyecciones de población utilizando técnica de calibración. Una vez obtenido los ponderadores de los hogares, se determinaron los ponderadores para la persona sorteada en el hogar, también utilizando técnicas de calibración para ajustar a las proyecciones de población.

2. Ponderadores del hogar

El ponderador muestral teórico depende de la estrategia utilizada para la selección de los hogares en la muestra, es decir, las probabilidades de selección, las cuales provienen del diseño muestral implementado. El ponderador muestral teórico del hogar j perteneciente al estrato de diseño h se define como el inverso de la probabilidad de selección

$$w_{jh} = 1/(n_h/N_h) = N_h/n_h$$

donde N_h es la cantidad de hogares en el estrato h y n_h es el tamaño de muestra teórico en el estrato h . Cada estrato del diseño es auto ponderado, es decir, todos los hogares dentro de un mismo estrato tienen la misma probabilidad de pertenecer a la muestra y por ende el mismo ponderador muestral.

Una vez finalizado el relevamiento de la información de la Ola 1, los ponderadores originales son ajustados en dos etapas:

1. Ajuste por no respuesta.
2. Ajuste en base a información auxiliar acerca de la estructura del universo (calibración).

2.1 Ajuste por no respuesta

La no respuesta es la regla más que la excepción en las encuestas por muestreo. Para un subconjunto de los hogares en la muestra original no es posible relevar la información, debido a la ausencia de sus moradores, rechazos, cambios en el tipo de vivienda, deficiencias en el marco muestral (producto de la desactualización del mismo), etc.

La presencia de no respuesta implica que los ponderadores originales deban ser ajustados para corregir la subestimación provocada por la reducción del tamaño de muestra original. El ajuste por no respuesta se realizó creando clases (o post-estratos) y utilizando información del Censo 2011. La construcción de los post-estratos tiene dos niveles. El primer corte es geográfico y corresponde a los 19 departamentos del Uruguay. Luego, dentro de cada departamento se clasificaron las localidades por tamaño:

- Localidades con menos de 5000 habitantes.
- Localidades con más de 5000 habitantes.
- Áreas rurales.

Para el caso de las localidades que corresponden a la Zona metropolitana (Canelones y San José) se creó un post-estrato aparte.

La definición de los post-estratos se debe a que las tasas de respuestas obtenidas en la primera ola de la ELPS variaba según el tamaño de la localidad y la región geográfica.

El ajuste por no respuesta en el post-estrato G para el hogar j perteneciente al estrato (pre) del diseño queda definido como:

$$w_{jh}^{nr} = (N_G / \hat{N}_G) \times w_{jh}$$

donde N_G es la cantidad de hogares particulares en el post-estrato G según el censo 2011 y \hat{N}_G es la estimación puntual de la cantidad de hogares particulares en el mismo post-estrato utilizando los ponderadores base (originales).

2.2 Ajuste por proyecciones de población

La calibración es un método que consiste en ajustar (calibrar) los ponderadores originales utilizando información auxiliar acerca de la estructura del universo de estudio, con el objetivo que la muestra expandida coincida con la información conocida del universo, lo cual brinda coherencia y comparabilidad a las estimaciones obtenidas.

La metodología de calibración consiste en determinar nuevos ponderadores incorporando información auxiliar de forma de calcular factores de ajustes para los ponderadores provenientes del diseño muestral. Los ajustes producto de la calibración reducen significativamente el sesgo ocasionado por la

no respuesta y minimizan los problemas de cobertura del marco muestral producto del paso del tiempo desde la realización del Censo 2011.

Cualquier usuario que disponga de los microdatos de la ELPS debe poder realizar una primera verificación simplemente sumando los ponderadores de la muestra. Dado que, cada ponderador básicamente equivale al número de hogares que representa cada hogar de la muestra en el universo de estudio, la suma, debe coincidir con el total de hogares del Uruguay. De igual forma, si se trabaja con los microdatos de las personas la suma debe coincidir con la población total del Uruguay. Es decir, la muestra debe estar calibrada, por lo menos, al universo de estudio.

A su vez, se puede estimar alguna otra variable que se conozca su distribución poblacional, por ejemplo, la cantidad de personas por tramo de edad y sexo. Es muy posible que, por el mecanismo aleatorio de selección de la muestra, por la no respuesta obtenida y la desactualización del marco, las estimaciones difieran de los valores poblacionales. Entonces, si la muestra no es capaz de representar la distribución de la población del Uruguay por edad y sexo, tampoco será válida para proporcionar otra información de interés para la ELPS.

Intuitivamente, la calibración se basa en el siguiente supuesto: si la muestra es capaz de estimar sin error los totales de las variables auxiliares utilizadas para la calibración también proporcionará buenas estimaciones de las variables objetivo de la encuesta. La mejora en las estimaciones proporcionada por la calibración dependerá en buena medida de que tanto expliquen las variables auxiliares a las variables objetivo de la encuesta. A su vez, la calibración produce estimadores con menor varianza en comparación con el estimador Horvitz – Thompson, es decir, utilizando únicamente los ponderadores provenientes del diseño muestral y minimizan el sesgo ocasionado por la no respuesta.

El ponderador calibrado para el hogar j queda definido como

$$w_j^* = w_j^{nr} \times \alpha_j$$

donde

$$\alpha_j = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x)' \left(\sum w_j^{nr} \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \mathbf{x}_j$$

donde \mathbf{x}_j el vector de información auxiliar para el hogar j , \mathbf{t}_x es el vector de totales poblacionales de las variables auxiliares utilizadas para la calibración y $\hat{\mathbf{t}}_x$ es el estimador de utilizando los ponderados provenientes del diseño muestral ajustados por la no respuesta (post – estratos). Es importante remarcar que cuando la última etapa del diseño muestral involucra muestreo por conglomerados, como ser el muestreo de todas personas en el hogar, los ponderadores calibrados en el mismo cluster (hogar) deben ser necesariamente iguales. Esto proporciona consistencia entre las estimaciones de los microdatos de la bases de personas y hogares.

La información auxiliar para la calibración de los ponderadores es el número de hogares por departamento provenientes de la Encuesta Continua de Hogares (ECH) del año 2013 y el total de personas desagregado por departamento y sexo provenientes de las proyecciones de población que realiza el INE para el mismo año. Esto implica que la muestra de hogares de la ELPS expandida coincide

con el total de personas por departamento y sexo provenientes de las proyecciones y la muestra expandida coincide con el total de hogares por departamento según la ECH 2013.

3. Ponderadores de personas (informante)

Una vez calculados los ponderadores calibrados para los hogares se determinan los ponderadores para las personas encuestadas por la ELPS también siguiendo técnicas de calibración. Teniendo en cuenta que una única persona elegible (de 14 años o más) era seleccionada al azar en el hogar, el ponderador base para la persona sorteada en el hogar queda definido como:

$$w_{kj} = w_j^* \times E_j$$

donde w_{kj} es el ponderador calibrado del hogar j y E_j es la cantidad de personas elegibles en dicho hogar.

Los ponderadores w_{kj} son también ajustados en base a las proyecciones de población y utilizando post-estratificación incompleta. En la post-estratificación se ajustan los ponderadores de la muestra de acuerdo a los totales poblacionales de una o más variables categóricas, como marginales de una tabla.

El ponderador calibrado para la persona perteneciente al hogar j queda definido como:

$$w_{kj}^* = \beta_{kj} \times w_{kj}$$

Los factores β_{kj} son un proceso de un ajuste iterativo, al que se le denomina raking, el cual consiste en ir ajustando por filas y columnas los totales estimados de las marginales, hasta alcanzar un error fijado a priori entre las estimaciones y los verdaderos valores de los totales poblacionales. La decisión de esta metodología de ajuste se debe a que se requería que los ponderadores replicarían varias características de la estructura de la población, y no se quería a su vez forzar mucho los ajustes de los ponderadores de manera de evitar sesgar las estimaciones.

Las variables auxiliares para la calibración son el total de personas por departamento, sexo y ocho tramos de edad:

1. 14 a 19 años
2. 20 a 29 años
3. 30 a 49 años
4. 50 a 59 años
5. 60 a 64 años
6. 65 a 69 años
7. 70 a 74 años
8. 75 y más años

Esto implica que la muestra de personas expandida coincide con las proyecciones del año 2013 para la estructura de edad y sexo anteriormente mencionada, a nivel nacional y para cada uno de los departamentos.

4. Precisiones de las estimaciones

Las estimaciones de los indicadores que mide la ELPS varían de muestra a muestra, es decir, dependiendo de la muestra que sea seleccionada son las estimaciones que se obtienen para los mismos. Esta variación producto del muestreo describe la incertidumbre de las inferencias acerca de la población en base a una muestra en particular y es medida por el error estándar (SE). Los SE de los indicadores dependen entre otros factores de:

- El diseño muestral (la estrategia de selección de los hogares y las personas en la muestra).
- La variabilidad de los datos para las variables de interés de la encuesta.
- El tamaño de muestra efectivo.
- El método de ajuste de los ponderadores (calibración).

En el siguiente cuadro se presentan las medidas de precisión (SE, CV¹ y DEFF²) para algunos de los Indicadores claves de la ELPS

INDICADORES	SE	CV	DEFF
PROPORCIÓN DE HOGARES UNIPERSONALES INTEGRADOS POR ADULTOS MAYORES	0.85%	1.27%	1.2436
PROPORCIÓN DE PERSONAS DE 14 AÑOS Y MAS CON AL MENOS UNA LIMITACIÓN	0.22%	2.89%	1.2725
PROPORCIÓN DE PERSONAS DE 14 AÑOS Y MÁS CON AL MENOS UNA LIMITACIÓN QUE PRECISAN AYUDA	0.17%	3.93%	1.2616
PROPORCIÓN DE PERSONAS QUE HAN SEGUIDO TRABAJANDO LUEGO DE JUBILARSE	0.74%	3.53%	1.4234
TOTAL DE HOGARES QUE COBRAN ASIGNACIÓN FAMILIAR	3588	1.24%	0.9328
TOTAL DE JUBILADOS	4984	1.14%	0.4548
TOTAL DE PENSIONISTAS	5700	2.03%	0.8669

¹ El coeficiente de variación (CV) es también una medida utilizada para describir la precisión de una estimación. Mientras más pequeño sea el CV, más precisa es la estimación. El CV de una estimación, es el ratio entre el SE de la estimación respecto a la propia estimación.

² El efecto de diseño (DEFF) compara la varianza del estimador del diseño muestral elegido respecto a la varianza del estimador bajo un muestreo aleatorio simple sin reposición (SI). El DEFF es también utilizado para medir la calidad del diseño muestral. Si el efecto del diseño es menor que uno implica que el diseño muestral elegido es más eficiente que el SI. En tanto, si el efecto del diseño es mayor a uno implica que el diseño muestral elegido es menos eficiente que el SI.