ASESORÍA ESTADÍSTICA: DATA EXPLORER

Índice

1. INTRODUCCIÓN

La Encuesta Nacional de Protección y Seguridad Social (ENAPROSS) es un estudio orientado a la evaluación de la situación socioeconómica y de protección social de los hogares localizados en el territorio de la República Argentina y de los individuos que residen en los mismos.

El estudio se basa en una muestra probabilística en la que se seleccionan viviendas de localidades urbanas que, de acuerdo al Censo Nacional de Población 2001 realizado por el Instituto Nacional de Estadística y Censos (INDEC), poseían en ese momento, 5.000 o más habitantes[1].

La fecha de referencia de la información corresponde al período de la ejecución del operativo de campo.

La principal utilidad de la ENAPROSS es la de suministrar la información necesaria e imprescindible para la formulación de eficientes políticas de Protección y Seguridad Social a nivel del país, regiones y provincias.

2.OBJETIVOS

El relevamiento apuntó a conocer los obstáculos y limitaciones que tiene la población en general y las familias vulnerables, en particular, para acceder a los derechos sociales tales como educación, salud, vivienda, oportunidades de trabajo y protección y seguridad social, considerando que los mismos pueden darse por dificultades en las condiciones de accesibilidad, falta de información, distancia, bajo nivel educativo, entre muchos otros factores.

Concretamente la ENAPROSS caracteriza las posibilidades y grado de acceso de la población a un hábitat saludable, a ingresos monetarios adecuados, al sistema de salud (de gestión pública o privada), al sistema educativo (de gestión pública o privada), a la protección y a la seguridad social.

3. POBLACIÓN Y DOMINIOS

3.1. POBLACIÓN OBJETIVO

La población objetivo en la presente investigación está compuesta por los hogares y personas que durante el año 2011 habitaban en viviendas particulares ubicadas en localidades del país y que, de acuerdo con el Censo Nacional de Población y Viviendas del 2001, tenían 5.000 o más habitantes. El censo 2001 era la única información disponible en el momento de realizar el diseño de la muestra.

Algunas definiciones generales relacionadas con el presente proyecto son:

Hogar particular: es el constituido por toda persona o personas que comparten una misma vivienda bajo el régimen de tipo familiar y consumen alimentos con cargo al mismo presupuesto, así como también comparten los gastos de alquiler, pagos de los servicios, etc., independientemente de que sean parientes o no. Cada persona integrante de un hogar particular es un miembro del hogar.

1. El Censo 2001 era la única información disponible en el momento de realizar el diseño de la muestra.

Vivienda: es el recinto fijo o móvil que ha sido construido o adaptado para alojar personas.

También se consideran viviendas aquellos locales no destinados originalmente para alojar personas, pero que son utilizados para ese fin. Una vivienda puede contener uno o más hogares.

Se excluyó de la población en estudio a las personas que viven en hogares colectivos, como ser asilos, guarderías, orfanatos, instituciones religiosas, hospitales, instituciones militares, etc. En cambio sí se incluyeron los hogares particulares constituidos en hoteles, pensiones e inquilinatos.

3.2. DOMINIO

Es un concepto de carácter técnico, definido en el año 1950 por la Organización de las Naciones Unidas en relación a la teoría del muestreo. Se lo define como “cualquier subdivisión de la población acerca de la cual se puede dar información numérica con una precisión posible de ser estimada”

Dominios básicos: La mayor desagregación prevista para el suministro de información, correspondió a jurisdicciones completas, incluyendo entre ellas a la Ciudad Autónoma de Buenos Aires y, cinco dominios correspondientes a provincias subdivididas en dos partes, estas últimas son:

- La Provincia de Buenos Aires, que se subdividió en 24 Partidos del Gran Buenos Aires y resto de la provincia.
- La Provincia de Santa Fe, que se subdividió en Gran Rosario y resto de la provincia.
- La Provincia de Córdoba, que se subdividió en Gran Córdoba y resto de la provincia.
- La Provincia de Tucumán, que se subdividió en Gran Tucumán y resto de la provincia.
- La Provincia de Mendoza, que se subdividió en Gran Mendoza y resto de la provincia.

De acuerdo con esa división los dominios resultantes[2] fueron:

a) Un dominio que corresponde a la Ciudad Autónoma de Buenos Aires (CABA).

b) Un dominio integrado por los 24 Partidos de la Provincia de Buenos Aires que forman el llamado Conurbano Bonaerense.

c) Un dominio compuesto por los restantes Partidos de la Provincia de Buenos Aires.

d) Cuatro dominios que corresponden a cada uno de los aglomerados urbanos de 700.000 o más habitantes, estos son: Gran Rosario, Gran Córdoba, Gran Tucumán y Gran Mendoza.

e) Cuatro dominios compuestos por las restantes localidades de cada una de las provincias donde se ubican los aglomerados urbanos indicados en el punto anterior, éstos se definen como el “resto” de las Provincias de Santa Fe, Córdoba, Tucumán y Mendoza.

2. El diseño de muestra inicial incluyó también dieciocho dominios correspondientes a cada una de las restantes Provincias del país, es decir: Catamarca, Corrientes, Chaco, Chubut, Entre Ríos, Formosa, Jujuy, La Pampa, La Rioja, Rio Negro, Misiones, Neuquén, Salta, San Luis, San Juan, Santa Cruz, Santiago del Estero y Tierra del Fuego. Finalmente, el relevamiento se llevó a cabo en los dominios (a), (b), (c), (d) y (e).

Formación de dominios por agregación: adicionando los dominios básicos enunciados anteriormente se pueden construir nuevos dominios de información. En particular son de interés las provincias completas de Santa Fe, Córdoba, Tucumán y Mendoza y la región GBA conformada por CABA y los 24 Partidos del Gran Buenos Aires, y el total conformado por CABA y las localidades de 5000 y más habitantes de las 5 provincias anteriormente mencionadas.

4. DISEÑO DE LA MUESTRA

El diseño de la muestra fue probabilístico, esto es que a cada unidad de la población se le asoció una probabilidad distinta de cero de ser seleccionada para integrar la muestra. Las probabilidades fueron asignadas a las unidades de cada uno de los marcos de muestreo definidos.

El Censo Nacional de Población 2001 fue el marco general disponible a nivel de provincia. No suministraba datos sobre la localización de cada uno de los hogares, ni de las viviendas de cada localidad, pero sí la información correspondiente a la cantidad de hogares y de población que contenía cada radio censal definido en el año 2001. La desventaja fue la desactualización lógica ocasionada por el transcurso del tiempo, pero en el momento del diseño muestral no se pudo contar con los datos, a nivel de radio, del más reciente censo de población del año 2010.

En consecuencia un objetivo técnico fue un diseño de la muestra que posibilitara lograr un marco apropiado a efectos de reducir al máximo el sesgo por el tiempo transcurrido.

Se consideró que el diseño probabilístico más eficiente era un **"Muestreo estratificado de conglomerados a tres etapas"**.

En la primera etapa de la investigación en cada dominio se eligió una **muestra de localidades** que en el censo tenían como mínimo 5.000 habitantes; en una segunda etapa, dentro de cada localidad elegida se seleccionó una **submuestra de radios censales;** en cada uno de éstos se listaron todas las viviendas ubicadas dentro de sus límites; en la tercera y última etapa se seleccionó la **submuestra de viviendas** y se entrevistaron a todos los habitantes de las viviendas así elegidas.

Previo a la selección de los radios, en cada localidad incluida en la muestra, se detectó el posible surgimiento de nuevas zonas edificadas en espacios que en el año 2001 estaban despobladas, como así también la desaparición de viviendas.

Para ello se estimó para cada radio la densidad de viviendas por kilómetro cuadrado sobre la base de imágenes satelitales provistas por el programa Google Earth, las que fueron comparadas con los correspondientes datos del Censo 2001. Aquellos radios donde la densidad estuvo a dos desvíos estándar por debajo o por arriba del promedio de densidad de la localidad, fueron investigados y en caso de comprobarse una diferencia significativa, se aseguró disponer de la cantidad correcta de viviendas antes de proceder a la selección de la muestra.

Finalmente se seleccionaron los radios y en cada uno de ellos se procedió a realizar el listado y conteo de las viviendas para disponer de un marco actualizado de Unidades de Tercera Etapa.

En resumen, en el presente diseño se definieron localidades como unidades de muestreo de primera etapa (UPE), radios censales como de segunda etapa (USE), y las viviendas como las de tercera o última etapa (UTE).

5. LOCALIDADES: UNIDADES DE MUESTREO DE PRIMERA ETAPA (UPE)

Las UPE se seleccionaron en todas las provincias, usando una estrategia mixta de inclusión forzosa y selección en forma aleatoria.

- **UPE de inclusión en forma forzosa,** fueron todas las Capitales de provincia más las ciudades que, al momento del censo 2001, contaban con al menos 140.000 habitantes. Este conjunto de grandes ciudades no participó en la selección aleatoria sino que entró en la muestra con probabilidad igual a la unidad.

- **UPE seleccionadas en forma aleatoria,** quedó constituido por el conjunto formado por las localidades de 5000 y más habitantes según el Censo de Población de 2001, excluidas las que integraron la muestra por inclusión forzosa. Las unidades fueron seleccionadas con probabilidad proporcional al tamaño, utilizando como medida del tamaño la cantidad de habitantes que cada localidad tenía en ese Censo.

Para lograr una mejor dispersión geográfica de las localidades sobre la superficie de la provincia, se preparó un marco especial a partir de ubicar las localidades de cada provincia arbitrariamente en dos o tres zonas al sólo efecto de selección de la muestra.

Existe una gran heterogeneidad entre las provincias argentinas con relación a la cantidad de localidades con más de 5.000 habitantes. Esto implicó que la cantidad de localidades en la muestra difiera significativamente entre provincias.

El área finalmente relevada en campo comprendió a 35 localidades de las cuales 11 fueron de inclusión forzosa y 24 de selección aleatoria.

6. RADIOS: UNIDADES DE MUESTREO DE SEGUNDA ETAPA (USE)

Las USE fueron los radios censales definidos por el INDEC para el censo 2001. En general los radios censales en ese momento contenían, en promedio 300 viviendas.

Dentro de cada localidad se seleccionaron las USE con probabilidad proporcional al tamaño, medido éste por la cantidad de hogares. El número de USE a seleccionar dentro de cada localidad dependió de: a) la cantidad población, b) el efecto diseño y c) indicadores del nivel de pobreza.

Con respecto al efecto diseño la teoría estadística indica que es necesario tener un adecuado balance entre la cantidad de encuestas a realizar en cada radio y el número de radios seleccionados, dado que al aumentar la cantidad de encuestas por radio el efecto diseño hace que disminuya la eficiencia de los estimadores.

El objetivo de la ENAPROSS está orientado a la evaluación de la situación socioeconómica y de protección social, por lo tanto la población de mayor interés para esta investigación es la de menores recursos, ya que hacia ellos van dirigidas las políticas de protección social. Por este motivo el diseño ha contemplado una mayor cantidad de unidades de segunda etapa de selección en zonas de bajos recursos.

Para asegurar esta mayor participación en la muestra, en cada localidad se estudió, a nivel de radio, el rango de variación del Índice de Privación Material de los Hogares (IPMH) elaborado por el INDEC.

El Índice de Privación Material de los Hogares -IPMH- es una variable que identifica a los hogares según su situación respecto a la privación material en cuanto a dos dimensiones: recursos corrientes y patrimoniales. La combinación de estas dimensiones define cuatro grupos de hogares: aquellos que no tienen ningún tipo de privación y tres grupos diferenciados según el tipo de privación que presentan: sólo de recursos corrientes, sólo patrimonial y convergente. Por lo tanto, la identificación de las personas en cierta categoría se establece a partir de las características del hogar al que pertenecen[3] .

Se conformaron dos estratos, uno con los radios cuya proporción de hogares con al menos una de las privaciones era menor al 40%, y otro con los radios de 40% o más de los hogares con privación. Esta estrategia sólo se utilizó en las ciudades más grandes, dado que no representaba una mejora en la distribución de la muestra para las más pequeñas.

Las ciudades con representación diferenciada en la muestra son: CABA, 24 Partidos del Gran Buenos Aires, Gran Rosario, Gran Córdoba, Gran Tucumán, Gran Mendoza, La Plata, Mar del Plata y la Ciudad de Santa Fe.

En las otras ciudades, si bien no se aumentó el tamaño de la muestra, en los radios con hogares de mayor privación, se utilizó el IPMH para el ordenamiento de los radios, lo que aseguraba, al utilizar el método de selección sistemático, que hubiera una mejor representación de la población en relación a esa variable. Si bien este ordenamiento no es estrictamente una estratificación, las estimaciones tienden a tener menores variancias.

Se determinó la cantidad de unidades de segunda etapa para cada estrato por afijación óptima, logrando así el objetivo de tener mayor proporcionalidad de radios en el estrato de más alta proporción de hogares con alguna privación.

El método de afijación óptima se refiere a la meta de asignar razones de muestreo a los estratos, de manera que se obtenga la mínima variancia para la media y el total para un costo fijo. Cuando el costo es el mismo entre estratos se lo conoce como la asignación de Neyman (Leslie Kish 1979)- (Scheaffer et al. 1987)

La determinación de tamaños de muestra no proporcionales a la estructura poblacional implica construir los factores de expansión de modo tal que permitan, en el momento de la expansión, reproducir la estructura poblacional del área bajo estudio.

Los datos disponibles del IPMH corresponden al censo 2001, por lo tanto eran de esperar variaciones, no sólo en la distribución de hogares con privaciones, sino también en la cantidad de hogares residentes en cada radio.

Por tal motivo se previó, en el diseño de la muestra, el listado de viviendas en los radios seleccionados, de esta forma se puede estimar el crecimiento ocurrido entre el 2001 y el momento de la encuesta.

Las variaciones detectadas en el momento de la encuesta en cuanto a la estratificación no afectan a las estimaciones de las variables, solamente aumentan sus variancias en la medida que se alejen de la homogeneidad prevista para el estrato.

El diseño no planteó utilizar los estratos como dominios de información, simplemente se trató de asegurar mayor tamaño de muestra en áreas con probabilidad de tener, en el momento de la encuesta, una proporción más grande de hogares con privación material. Cabe recordar que esa mayor probabilidad está referida a la información brindada por el Censo 2001.

3. Para obtener mayores detalles sobre la metodología ver INDEC (2004) El estudio de la pobreza según el Censo Nacional de Población, Hogares y Viviendas 2001. Indice de Privación Material de los Hogares (IPMH), DNESyP/DEP/P5/PID, Serie Pobreza, Documento de Trabajo Metodológico, Buenos Aires (mimeo)

La estrategia de distribución no proporcional de la muestra implica una mejor calidad de las estimaciones en aquellas variables que se presentan con mayor frecuencia en hogares de menores recursos, como es la Asignación Universal por Hijo (AUH)

Para los análisis que se realicen con los datos de la encuesta se utilizarán variables estratificadoras (post estratificación) extraídas de la misma encuesta, como por ejemplo quintiles de ingreso per cápita familiar.

Como ya se dijo, la información sobre la cantidad de viviendas por radio proviene del Censo del 2001, por lo tanto estaba desactualizada ya sea por el surgimiento o por la desaparición de viviendas. Para reducir este sesgo se recorrieron las manzanas que componían cada radio elegido y se listaron todas las viviendas particulares existentes en ese momento, obteniendo así marcos de viviendas actualizados a la fecha del listado.

La construcción del listado se facilitó debido a que los listadores utilizaron el dispositivo “Personal Digital Assistant” (PDA), de esa forma, al finalizar el recorrido ya se dispuso del marco de selección con el consecuente ahorro de tiempo. No obstante se destaca que en algunas zonas no fue conveniente utilizar PDA, y se reemplazó por una Hoja de Ruta, que luego de completada se grabó en la base datos.

Las tareas de listado fueron supervisadas a efectos de validar su cantidad y su calidad.

7. VIVIENDAS: UNIDADES DE MUESTREO DE TERCERA ETAPA (UTE)

Las UTE fueron las viviendas. Los listados realizados en cada uno de los radios constituyeron los marcos para selección de viviendas, que, como ya se dijo, estuvieron disponibles en una base de datos, lo que posibilitó el desarrollo de una aplicación informática que realiza la selección sistemática y que, además, elabora automáticamente la Hoja de Ruta del Encuestador.

Las viviendas fueron elegidas con probabilidad igual dentro de cada radio integrante de la muestra.

La no respuesta: Uno de los inconvenientes de las encuestas dirigidas a hogares es que la tasa de no respuesta puede ser altamente significativa y, además, variable según localidad y barrio. En general hay una correlación positiva entre la tasa de no respuesta y el nivel socioeconómico de los hogares.

Para poder lograr una muestra efectiva acorde a lo esperado en el diseño y evitar tener que seleccionar a posteriori nuevas viviendas para completarla sin afectar las probabilidades, se extrajo una muestra mayor de viviendas en cada radio elegido.

8. TAMAÑO DE LA MUESTRA

El tamaño total final de la muestra fue de 11.600 viviendas.

El objetivo de obtener estimaciones que tengan una determinada precisión y nivel de confianza similar para cada uno de los dominios, hizo necesario establecer una cantidad mínima de UTE a ser seleccionada en cada uno de ellos. Además, con fines prácticos para el trabajo de campo, se adoptaron cantidades iguales de viviendas en los radios de cada dominio.

En las provincias que incluyen aglomerados urbanos como dominios de información se incrementó el tamaño de muestra total provincial para obtener estimaciones para ambos

dominios (el aglomerado y total provincia). El incremento se adjudicó al resto de la provincia para lograr un nivel de error muestral razonable.

9. ADJUDICACIÓN DE RADIOS POR LOCALIDAD

Los 995 radios y las 11.600 viviendas debieron ser adjudicados en cada una de las 35 localidades elegidas.

El método de adjudicación fue proporcional a la cantidad de habitantes que tenía la localidad con respecto al dominio según los datos del Censo Nacional de Población del 2001, con la restricción, por problemas operativos, de elegir un mínimo de 3 radios en cada localidad.

La regla de adjudicación de los radios de cada dominio entre las localidades fue:

$$c_{dl} = \frac{Pob_{dl}}{\sum_{l} Pob_{dl}} \times c_d$$

Si $C_{dl} < 3$ se hace $C_{dl} = 3$ Ajustando los restantes valores

Donde:

C_d : Cantidad de radios en la muestra del dominio “d”

C_{dl} : Cantidad de radios a adjudicar en la localidad l-ésima del dominio “d”

Pob_{dl} : Cantidad de habitantes de la localidad l-ésima del dominio “d” en el censo de población 2001.

10. ESTIMADORES

Los **estimadores** se definen como las expresiones matemáticas construidas a partir de los datos de la muestra y que tienen como objetivo estimar los **valores poblacionales** o **parámetros del estudio**. La estructura de estas fórmulas depende de la forma en que fueron seleccionadas las diferentes unidades de muestreo.

En la presente investigación el diseño de muestra es complejo: en cada dominio la selección fue en tres etapas, donde las UPE se seleccionaron con probabilidad variable, las USE se estratificaron y seleccionaron también con probabilidades variables y las UTE con probabilidades iguales. Los dominios de información requieren el suministro de estimaciones por separado.

Se trabajó con dos tipos de estimadores:

- Insesgados de expansión simple: para estimar parámetros totales y promedios de una variable cuantitativa o dicotómica como así también los errores debidos al muestreo. Se utilizan en forma directa o se incorporan a los estimadores por razón.

- Sesgado de razón: en muestras suficientemente grandes, como la presente, el sesgo es despreciable y la ventaja es que suelen ser más precisos. Permiten estimar el parámetro razón de dos variables, totales y promedios de una variable cuantitativa o dicotómica y los estimadores del error de muestreo.

El desarrollo teórico de los estimadores utilizados de presentan en el Anexo I.

11. COMPARACIÓN DE ESTRUCTURAS CON OTRAS FUENTES

Algunas estructuras básicas elaboradas con los resultados de la ENAPROSS se compararon con los correspondientes datos de otras fuentes disponibles.

Los 11 dominios relevados fueron comparados con la Encuesta Anual de Hogares Urbanos (EAHU) de INDEC, 3° trimestre 2010. Para los dominios CABA y 24 partidos del Gran Buenos Aires también se utilizaron datos del Censo 2010.

Las variables comparadas fueron: estructuras por tramos de edad, índice de masculinidad, promedio de hogares por vivienda, promedio de personas por hogar y cantidad de hogares según la cantidad de miembros por hogar.

12. ERRORES DE MUESTREO

La base de datos de la ENAPROSS en formato SPSS o Stata, permite estimar el error típico, el coeficiente de variación y el efecto diseño para cada estadístico, utilizando el modulo de muestras complejas.

Para ello es necesario construir el archivo csplan de SPSS o la sintaxis en Stata para indicar el diseño muestral utilizado. Los estratos deben ser indicados por las variables identificadas en la base como “COD_PROV” y “ESTRATO” y los conglomerados por “ID_RADIO”.

Se recomienda tener en cuenta la siguiente sugerencia para evaluar la confianza de las estimaciones:

Estadísticos cuyos coeficientes de variación sean inferiores al 10% se consideran muy aceptables, los que superan el 10% y hasta un 20%, se los considera razonables, cuando superan el 20%, tomar con precaución el dato estimado.

ANEXO I

Estimadores

Se definen los siguientes símbolos

v: Representa la v-ésima UTE (vivienda) seleccionada en forma sistemática dentro de un radio.

En la población, v: 1, 2, …, M_{dlhr}

En la muestra, v: 1, 2, …, m_{dlhr}

r: Representa la r-ésima USE (radio censal) seleccionado con probabilidad proporcional al tamaño dentro de una localidad.

En la población, r: 1, 2, …, C_{dlh}

En la muestra, r: 1, 2, …, c_{dlh}

h: Representa al estrato formado por la agrupación de USEs de cada la localidad en dos niveles de Necesidades Básicas Insatisfechas (NBI).

Donde h: 1, 2

l: Representa la l-ésima UPE (localidad) dentro de un dominio d.

En la población l: 1,2, …, N_d

En la muestra l: 1, 2, …, n_d

d: Representa el d-ésimo dominio.

Donde, d: 1, 2, …, 29

La variable en estudio se simboliza:

y_{dlhrv}: Es el valor de la variable en estudio de la v-ésima UTE, de la r-ésima USE del estrato h, de la l-ésima UPE ubicada en el dominio d.

1. EXPANSIÓN DE UNA VARIABLE A NIVEL DE UTE

La expresión que corresponde a la expansión de una variable al radio completo viene dada por la siguiente expresión

$$\hat{Y}_{dlhr} = \frac{M_{dlhr}}{m_{dlhr}} \sum_{v}^{m_{dlh}} y_{dlhrv} \quad (1)$$

Donde:

y_{dlhrv} Representa el valor de la variable para la v-ésima UTE, de la r-ésima USE del estrato h, y la l-ésima UPE del dominio d.

$\hat{Y}_{dlhr}$ Representa el total de la variable expandida a la r-ésima USE del estrato h, y de la l-ésima UPE del dominio d.

Mdlhr Es el total de UTEs que fueron listadas en la r-ésima USE del estrato h, y de la l-ésima UPE del dominio d.

mdlhr Es el total de UTEs que fueron seleccionadas en forma sistemática del listado, de la r-ésima USE del estrato h, y de la l-ésima UPE del dominio d.

2. EXPANSIÓN DE UNA VARIABLE A NIVEL DE UPE

Es la expansión de una variable al conjunto total de radios de una localidad

$$\hat{Y}_{dl} = \sum_{h}^{2} \frac{1}{c_{dlh}} \sum_{r}^{c_{dlh}} \frac{\hat{Y}_{dlhr}}{Q_{dlhr}} \quad (2)$$

Donde: C_{dlh} Representa la cantidad USEs seleccionadas en el estrato h del dominio d.

$\hat{Y}_{dl}$ Representa el total de una variable expandida a la l-ésima UPE del dominio d.

Q_{dlhr} Son probabilidades variables de selección de la r-ésima USE del estrato h y la l-ésima UPE del dominio d.

Las expresiones (1) y (2) se definen solo como expansiones y no como estimadores, debido a que no es el objetivo del presente diseño de muestra la estimación de los parámetros a nivel de localidad, excepto que la localidad haya sido definida como dominio, en el caso de ser así la expresión (2) se convierte en un estimador.

3. ESTIMADORES A NIVEL DOMINIO

Estimador del total de una variable, por simple expansión, a nivel dominio

El dominio constituye la menor agregación para la que se suministra información. Dentro de cada dominio se eligen las UPEs.

En el presente diseño se distinguen tres casos en relación a la forma en que las UPE fueron seleccionadas.

a) La UPE es el dominio, es decir que está compuesto por una única localidad que por su importancia fue específicamente definida como tal, es decir que no existe selección de localidad.

b) La UPE fue seleccionada aleatoriamente. La localidad fue elegida con probabilidad proporcional a la cantidad de habitantes que tenía en el Censo de Población 2001.

c) La UPE fue incluida forzosamente por ser la capital de una provincia y/o arrojaba al menos 140.000 habitantes en el Censo. Las localidades que cumplen esta condición fueron incluidas en la muestra sin que exista selección aleatoria, como ejemplos la ciudad de Río IV no es capital provincial pero se incluyó por tener más de 140.000 habitantes.

Los tres casos anteriores, desde un punto de vista metodológico, conducen a dos tipos de dominios que, a los efectos del presente diseño de muestra y con el fin de diferenciarlos, se designan como Dominios de tipo I y Dominios de tipo II:

Dominios tipo I: son los integrados por una sola UPE (caso a). Son los siguientes seis: Ciudad Autónoma de Buenos Aires, Gran Córdoba, Gran Mendoza, Gran Rosario, Gran Tucumán y el Conurbano Bonaerense integrado por el conjunto de 24 Partidos del Gran Buenos Aires.

El siguiente estimador de un dominio tipo I, es igual a la expansión dada en (2) excepto que corresponde a una sola localidad.

$$\hat{Y}_d = \sum_{h}^{2} \frac{1}{c_{dh}} \sum_{r}^{c_{dh}} \frac{\hat{Y}_{dhr}}{Q_{dhr}} \quad (3)$$

Dominios tipo II: Son 5 dominios integrados por UPEs, algunas de las cuales fueron seleccionadas en forma aleatoria y otras incluidas forzosamente en la muestra. Los 5 dominios son: resto de Provincia de Buenos Aires, resto de Provincia de Santa Fe, resto de Provincia de Córdoba, resto de Provincia de Tucumán y resto de Provincia de Mendoza.

Las UPEs aleatorias fueron elegidas con probabilidad de selección “Pdl” (caso b) y las UPEs forzosas con probabilidad igual a la unidad (caso c).

El siguiente es el estimador del total de una variable para un dominio que presenta ambos tipos de selección (caso b y c).

$$\hat{Y}_d = \frac{1}{n_d} \sum_{l}^{n_d} \frac{\hat{Y}_{dl}^{AL}}{P_{dl}} + \sum_{l}^{N'_d} \hat{Y}_{dl}^{IF} \quad (4)$$

Donde:

$\hat{Y}d$ Es el estimador del total de una variable para el dominio d. .

$\hat{Y}_{dl}^{AL}$ Representa la expansión de una variable correspondiente a la l-ésima UPE del dominio d, que fue seleccionada en forma aleatoria, se utiliza la siguiente fórmula

$$\hat{Y}_{dl}^{AL} = \sum_{h}^{2} \frac{1}{c_{dlh}} \sum_{r}^{c_{dlh}} \frac{\hat{Y}_{dlhr}}{Q_{dlhr}}$$

$\hat{Y}_{dl}^{IF}$ Representa la expansión de una variable correspondiente a la l-ésima UPE del dominio d, que fue incluida en forma forzosa en la muestra, se utiliza la siguiente fórmula

$$\hat{Y}_{dl}^{IF} = \sum_{h}^{2} \frac{1}{c_{dlh}} \sum_{r}^{c_{dlh}} \frac{\hat{Y}_{dlhr}}{Q_{dlhr}}$$

P_{dl} Son las probabilidades de selección de la l-ésima UPE del dominio d, se calcularon en forma proporcional a la cantidad de hogares.

n_d Es el total de UPEs que fueron seleccionadas aleatoriamente en el dominio d.

N'_d Total de localidades de inclusión forzosa en el dominio d

Se observa que en el primer sumando de la fórmula (4) la variable aparece multiplicada por el factor de expansión de localidad “1/Pdl” en cambio en el segundo no aparece en forma explícita por valer la unidad.

4. ESTIMADOR DEL PROMEDIO POR ELEMENTO (POR VIVIENDA), POR SIMPLE EXPANSIÓN, A NIVEL DOMINIO

El estimador del promedio por elemento, para el caso de un dominio tipo I

$$\hat{\bar{Y}}_d = \frac{1}{(M_d)} \left(\sum_{h}^{2} \frac{1}{c_{dh}} \sum_{r}^{c_{dhl}} \frac{\hat{Y}_{dhr}}{Q_{dhr}} \right) \quad (5)$$

El estimador del promedio por elemento, para el caso de un dominio tipo II

$$\hat{\bar{\bar{Y}}}_d = \frac{1}{(M_d)}\left(\frac{1}{n_d}\sum_{l}^{n_d}\frac{\hat{Y}_{dl}^{AL}}{P_{dl}} + \sum_{l}^{N'_d}\hat{Y}_{dl}^{IF}\right) \quad (6)$$

Donde

$\hat{\bar{\bar{Y}}}_d$ Es el estimador promedio por elemento del dominio d.

Md Es el total poblacional de UTEs del dominio d.

Por ser un valor poblacional los Md no son conocidos y deberán ser estimados a partir de los resultados de los listados de la muestra de localidades. Para ello se utilizan las fórmulas (3) o (4). Se debe tener en cuenta que Md = MdAL + MdIF

5. ESTIMADOR DE UNA RAZÓN, A NIVEL DE DOMINIO

El parámetro razón entre dos variables cualesquiera se estima a partir del cociente de dos estimadores de simple expansión que utilizan la fórmula (3) o la fórmula (4) según sea dominios de tipo I o de tipo II.

$$\hat{R}_d = \frac{\hat{Y}_d}{\hat{X}_d} \quad (7)$$

6. ESTIMADOR DE UN TOTAL, POR EL MÉTODO DE RAZÓN, A NIVEL DOMINIO

Para estimar el total de una variable en un dominio por el método de razón, se necesita conocer el valor poblacional de una variable auxiliar “W” que debe estar correlacionada significativamente con las variables que se quieren estimar. En este caso el estimador por razón presentará ganancias en precisión con relación a uno de simple expansión.

El estimador del total de una variable por el método de razón, correspondiente al dominio d

$$\hat{\bar{\bar{Y}}}_d^{Razón} = \frac{\hat{Y}_d}{\hat{W}_d}\bar{\bar{W}}_d \quad (8)$$

Donde: W_d es el total poblacional de la variable auxiliar.

$\hat{W}_d$ es la estimación con la muestra de la variable auxiliar.

7. ESTIMADOR DEL PROMEDIO POR ELEMENTO (POR VIVIENDA), POR EL MÉTODO DE RAZÓN, A NIVEL DOMINIO

$$\hat{Y}_d^{Razón} = \frac{\hat{Y}_d}{\hat{W}_d}W_d \quad (9)$$

Donde: $\bar{\bar{W}}_d$ es el promedio de la variable auxiliar.

Estimadores de la variancia

No es suficiente con obtener los estimadores de los parámetros sino que es fundamental establecer las expresiones de los estimadores de la variancia de los propios estimadores.

Se debe tener en cuenta que se refiere a la variancia debida al proceso de muestreo y que se expresa en unidades al cuadrado. En consecuencia una vez que se obtienen es necesario extraerles la raíz cuadrada a efectos de obtenerlos en las mismas unidades de medida de la variable.

1. Estimador de la variancia de un total, por simple expansión, a nivel de dominio.

El estimador de la variancia del estimador de un total por simple expansión dado en (3), correspondiente a un dominio definido como de tipo I

$$\hat{\sigma}^2_{(\hat{Y}_d^{Raz})} = \sum_{h}^{2}\frac{1}{c_{dh}(c_{dh}-1)}\sum_{r}^{c_{dh}}\left(\frac{\hat{Y}_{dhr} - \hat{R}_{dh}\hat{W}_{dhr}}{Q_{dhr}}\right)^2 \quad (10)$$

En el caso de un dominio de tipo II, el estimador de la variancia del estimador de un total por simple expansión dado en (4)

$$\hat{\sigma}^2_{(\hat{Y}_d^{Raz})} = \frac{1}{n_d(n_d-1)}\sum_{l}^{n_d}\left(\frac{\hat{Y}_{dl}^{AL} - \hat{R}_{dl}\hat{W}_d}{P_{dl}}\right)^2 + \sum_{l}^{N'_d}\sum_{h}^{2}\frac{1}{c_{dh}(c_{dh}-1)}\sum_{r}^{c_{dh}}\left(\frac{\hat{Y}_{dhlr}^{IF} - \hat{R}_{dhl}\hat{W}_{dhlr}}{Q_{dhlr}}\right)^2 \quad (11)$$

N´d: corresponde al total de localidades seleccionadas como de Inclusión Forzosa.

2. Estimador de la variancia de un promedio por elemento, por simple expansión, a nivel de dominio.

Es estimador de la variancia de un promedio por elemento dado en (5) o en (6) para un dominio de tipo I o II

$$\hat{\sigma}^2_{(\hat{\bar{\bar{Y}}}_d)} = \frac{1}{M_d^2}\sigma^2_{(\hat{Y}d)} \quad (12)$$

3. Estimadores de la variancia de un total, por razón, a nivel de dominio

Un estimador de la variancia de un total por razón correspondiente a un dominio definido como de tipo I, dado en (8), el que es ligeramente sesgado.

$$\hat{\sigma}^2_{(\hat{Y}_d)} = \sum_h^2 \frac{1}{c_{dh}(c_{dh}-1)} \sum_r^{c_{dh}} \left(\frac{\hat{Y}_{dhr}}{Q_{dhr}} - \hat{Y}_{dh} \right)^2 \qquad (13)$$

En el caso de un dominio tipo II el estimador de la variancia de un total por razón, dado en (9)

$$\hat{\sigma}^2_{(\hat{Y}_d)} = \frac{1}{n_d(n_d-1)} \sum_l^{n_d} \left(\frac{\hat{Y}^{AL}_{dl}}{P_{dl}} - \hat{Y}_d \right)^2 + \sum_l^{N'_d} \sum_h^2 \frac{1}{c_{dh}(c_{dh}-1)} \sum_r^{c_{dh}} \left(\frac{\hat{Y}^{IF}_{dhlr}}{Q_{dhlr}} - \hat{Y}^{IF}_{dhl} \right)^2 \qquad (14)$$

4. Estimador de la variancia de un promedio por elemento, por el método de razón, a nivel de dominio

Un estimador de la variancia del estimador por razón de un promedio por elemento para un dominio de tipo I o II

$$\hat{\sigma}^2_{(\hat{\bar{Y}}^{Raz}_d)} = \frac{1}{M_d^2} \sigma^2_{(\hat{Y}^{Raz}_d)} \qquad (15)$$

5. Estimador de la variancia de una razón a nivel de dominio

Un estimador de la variancia del estimador por razón de una razón, correspondiente a un dominio definido como de tipo I o tipo II es

$$\hat{\sigma}^2_{(\hat{R}_d)} = \frac{1}{\hat{X}_d^2} \sigma^2_{(\hat{Y}^{Raz}_d)} \qquad (16)$$

<u>Estimadores del Error Estándar y el Coeficiente de Variación</u>

Ya se mencionó que a las variancias que se obtienen se les debe extraer la raíz cuadrada y así se obtienen los Errores Estándar que en muestreo son las verdaderas medidas del error debido al proceso de muestreo es decir de “observar” solo a la muestra y no la población completa.

$$\hat{\sigma}_{(\hat{\theta})} = \sqrt{\hat{\sigma}^2_{(\hat{\theta})}}$$

El símbolo anterior representa que el error estándar se obtiene como la raíz cuadrada de la variancia.

Además, a efectos de facilitar su interpretación, se acostumbra utilizar una segunda medida de tipo relativa, que se denomina Coeficiente de Variación que corresponde al cociente entre el Error Estándar y su estimador, usualmente expresado en porcentaje.

$$\text{CV\%} = \frac{\text{Error Estándar del Estimador}}{\text{Estimasor del parámetro}} . 100$$