



Instituto Nacional de Estadística y Censos
PROGRAMA MECOVI-ARGENTINA
BID-BM-CEPAL

SERIE METODOLOGÍAS

**ESTIMACIÓN DE MEDIDAS DE DESIGUALDAD
DE LOS INGRESOS Y CÁLCULO DE SUS
VARIANZAS**

Gonzalo Marí

Buenos Aires, 2001

INDICE

1.INTRODUCCIÓN	1
2.ESTIMACIÓN DE MEDIDAS DE DESIGUALDAD DEL INGRESO.....	2
2.1. Función de Distribución Acumulada.....	2
2.2. Parámetros de Desigualdad del Ingreso y Medidas de Polarización en Poblaciones Finitas	3
3.ESTIMACIÓN DE VARIANCIA.....	5
4.PROGRAMA DE CÁLCULO DE VARIANCIA	7
5.RESULTADOS	12
BIBLIOGRAFÍA.....	14

1. INTRODUCCIÓN

Es habitual calcular y publicar los errores de muestreo en la estimación de las características centrales de cada encuesta. Incluso cuando se estiman ingresos, es habitual indicar el error de estimación de los ingresos medios. No sucede lo mismo con respecto a parámetros descriptivos de la distribución del ingreso. El presente trabajo considera la construcción de un algoritmo de cálculo que permita estimar distintas medidas de desigualdad y de polarización del ingreso, así como el cálculo de las variancias de dichos estimadores.

Las medidas de desigualdad del ingreso y de polarización son estadísticas no lineales, las cuales dependen de cuantiles y de la función de distribución. Las variancias de las mismas no pueden ser expresadas mediante simples fórmulas. Es por ello que se deben buscar técnicas que brinden estimaciones aproximadas de los errores muestrales.

El problema mencionado en el párrafo anterior se suma al hecho de contar con un diseño complejo, lo cual hace más dificultosa la búsqueda de una metodología que se ajuste a los dos inconvenientes.

Se investigarán tres medidas referidas al ingreso, el Índice de Gini, las ordenadas de la curva de Lorenz, y la proporción de unidades bajo la línea de pobreza.

A lo largo de este trabajo asumiremos que trabajamos con una población finita. Asumimos también que seleccionamos una muestra siguiendo un diseño probabilístico de muestreo $p(s)$. Más específicamente, diremos que la población se encuentra estratificada en L estratos, con N_h unidades primarias de muestreo (UPM) en el estrato h . En la primer etapa muestral, n_h (≥ 2) UPM son seleccionadas del estrato h en forma independiente entre estratos. Asumimos que el submuestreo dentro de las UPM seleccionadas nos asegura estimaciones insesgadas de los totales de las UPM, Y_{hc} , $c=1, \dots, n_h$, $h=1, \dots, L$. Para cada unidad elemental (hci) existirá una observación de la variable de interés, y_{hci} , y el correspondiente peso muestral w_{hci} .

2. ESTIMACIÓN DE MEDIDAS DE DESIGUALDAD DEL INGRESO

Se consideraron para el presente trabajo medidas referentes al ingreso que pueden ser agrupadas en aquellas que dependen de la función de distribución acumulada (FDA), y en las que dependen de un número fijo de cuantiles. Dentro de las primeras, se estudiarán la Curva de Lorenz y el Índice de Gini. Mientras que dentro del segundo grupo será de interés para este trabajo la estimación de la proporción de unidades bajo la línea de pobreza.

2.1. Función de Distribución Acumulada

Sea la población finita $U=\{1,\dots,N\}$. Para una variable y definida en dicha población, la FDA es definida por $F_N(y) = \frac{1}{N} \sum_{i \in U} I\{Y_i \leq y\}$

donde $I\{a\}$ es la función indicadora que toma el valor 1 si el evento a es verdadero, y 0 en otro caso.

Con el fin de estimar θ_N , una muestra s de la población finita U es seleccionada de acuerdo a un diseño $p(s)$ (en nuestro caso particular, estratificado, dos etapas), donde las probabilidades de inclusión son

$$\pi_{hci} = \sum_{s \ni hci} p(s), \quad s \in S(U), \quad hci \in U$$

donde $S(U)$ es el conjunto de todas las muestras posibles de U bajo el diseño probabilístico $p(s)$.

El π -estimador de $F_N(y)$ es

$$\begin{aligned} \hat{F}(y) &= \sum_{hci \in s} I\{y_{hci} \leq y\} \pi_{hci}^{-1} / \sum_{hci \in s} \pi_{hci}^{-1} \\ &= \sum_{hci \in s} I\{y_{hci} \leq y\} \tilde{w}_{hci} \end{aligned}$$

$$\text{donde } \tilde{w}_{hci} = \pi_{hci}^{-1} / \sum_{hci \in s} \pi_{hci}^{-1}, \quad hci \in s.$$

La forma anterior cubre no solo π -estimadores de la FDA, sino otros. Un parámetro de una población finita θ_N puede ser estimado por

$$\hat{\theta} = \sum_s J(y_{hci}) \tilde{w}_{hci} = \sum_U J(y_{hci}) \tilde{w}_{hci}$$

donde $\tilde{w}_{hci} = 0$ para $hci \notin s$, el cual goza de la propiedad de ser consistente en población finita (Sarndal et al., 1992, pag 168).

2.2. *Parámetros de Desigualdad del Ingreso y Medidas de Polarización en Poblaciones Finitas*

Vamos a presentar los parámetros de las medidas antes mencionadas y los estimadores considerados.

La ordenada de la curva de Lorenz brinda el porcentaje de ingreso percibido por cierta porción p acumulada de unidades, y es definido como una función de p , donde p varía entre 0 y 1. Como parámetro de una población finita, es una razón

$$L_N(p) = \mu_N(p) / \mu_N$$

donde $\mu_N = \sum_{i \in U} Y_i / N$ es la media poblacional, y

$$\mu_N(p) = \sum_U I\{F_N(Y_i) \leq p\} Y_i \frac{1}{N}$$

El estimador de la ordenada de la curva de Lorenz depende del estimador de la FDA, y es igual a

$$\hat{L}(p) = \frac{\hat{\mu}_p}{\hat{\mu}} = \frac{\sum_s I\{\hat{F}(y_{hci}) \leq p\} y_{hci} \tilde{w}_{hci}}{\sum_s y_{hci} \tilde{w}_{hci}}$$

El Índice de Gini está definido como el área entre la curva de Lorenz y una línea a 45°, normalizado entre 0 y 1. Puede ser expresado como un parámetro de población finita en términos de una razón,

$$G_N = \mu_N(G) / \mu_N$$

donde

$$\mu_N(G) = \sum_U [2F_N(Y_i) - 1] Y_i \frac{1}{N}$$

con μ_N definido con anterioridad. El estimador viene dado por

$$\hat{G} = \frac{\hat{\mu}_G}{\hat{\mu}} = \sum_s [2\hat{F}(y_{hci}) - 1] y_{hci} \tilde{w}_{hci} / \sum_s y_{hci} \tilde{w}_{hci}$$

Con respecto al segundo grupo de medidas, o sea aquellas que dependen de un número fijo de cuantiles, definiremos en primer término al cuantil de una población finita

$$\xi_N(p) = \inf \{ Y_i \in U | F_N(Y_i) \geq p \} \quad 0 \leq p \leq 1$$

y su respectivo estimador a partir de los cuantiles muestrales

$$\hat{\xi}_p = \inf \{ y_i \in s | \hat{F}(y_i) \geq p \} \quad 0 \leq p \leq 1$$

Si un parámetro es una función de cuantiles $\xi_N = \{\xi_N(p_1), \dots, \xi_N(p_k)\}$, por ejemplo $\theta_N = g(\xi_N)$, luego el estimador de dicho parámetro será $\hat{\theta} = g(\hat{\xi})$ donde $\hat{\xi} = (\hat{\xi}_{p_1}, \dots, \hat{\xi}_{p_k})$.

La proporción de unidades bajo la línea de pobreza relativa es el porcentaje de unidades con ingreso bajo en la población, $\Lambda_\alpha = F_N(\lambda_\alpha)$, donde $\lambda_\alpha = \alpha \xi_N(0.5)$ es la línea de pobreza definida como una fracción α de la mediana, con $0 < \alpha \leq 1$. Para estimar esta proporción se debe estimar la FDA y la línea de pobreza, o sea,

$$\hat{\Lambda}_\alpha = \hat{F}(\hat{\lambda}_\alpha)$$

donde $\hat{\lambda}_\alpha = \alpha \hat{\xi}_{0.5}$

3. ESTIMACIÓN DE VARIANCIA

La estimación de variancia de estadísticas no suaves, como es el caso de funciones basadas en cuantiles, no es de aplicación directa de los estimadores usuales, más aun cuando el diseño muestral utilizado no es el simple al azar, sino que se considera un esquema de selección complejo.

Si bien los métodos de linearización son útiles para estadísticas no lineales, son dificultosos de implementar en los estimadores vistos en el punto 2.2 debido a que los cuantiles involucran la estimación de la densidad. Binder (1991), Binder y Kovacevic (1995) y Kovacevic y Binder (1997) obtuvieron estimadores consistentes de variancia para medidas no suaves de desigualdades del ingreso y de la polarización utilizando el método de linearización de Taylor dentro de la estructura de las ecuaciones de estimación (EE).

En el presente trabajo se estimarán las medidas de error muestral de las estimaciones de los 3 parámetros de interés mediante las EE, quedando como una posible ampliación del presente trabajo la comparación de dicha técnica con la de Bootstrap.

3.1. Linearización de Taylor vía Ecuaciones de Estimación

La aproximación por EE no utiliza en forma intensiva métodos computacionales, como si lo hacen los métodos de replicaciones. Este método, basado en la linearización de Taylor, brinda fórmulas de variancias asintóticas.

Aplicando la metodología EE se obtienen las siguientes expresiones para estimadores de variancia aproximados de las medidas estudiadas

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1} \sum_c (u_{hc}^* - \bar{u}_h^*)^2$$

donde $u_{hc}^* = \sum_i \tilde{w}_{hci} u_{hci}^*$, $\bar{u}_h^* = \sum_c u_{hc}^* / n_h$, y \tilde{w}_{hci} es el peso normalizado. Las u_{hci}^* son las variables transformadas para la estimación de variancia vía EE,

Medida	u_{hci}^*
Indice de Gini	$2[\hat{A}(y_{hci}) y_{hci} + \hat{B}(y_{hci}) - \hat{\mu}(\hat{G} + 1)/2] / \hat{\mu}$
Curva de Lorenz	$[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci} \hat{L}(p)] / \hat{\mu}$
Proporción bajo línea de pobreza	$-\frac{\hat{f}(\hat{\xi}_{0.5}/2)}{2\hat{f}(\hat{\xi}_{0.5})} [I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 1/2] + [I\{y_{hci} \leq \hat{\xi}_{0.5}/2\} - \hat{\Lambda}_{0.5}]$

donde $\hat{A}(y) = \hat{F}(y) - (\hat{G} + 1)/2$, $\hat{B}(y) = \sum_s \tilde{w}_{hcj} y_{hcj} I\{y_{hcj} \geq y\}$, y $\hat{f}(\cdot)$ representa el estimador de la densidad de la población finita.

La expresión de u_{hci}^* para la proporción bajo la línea de pobreza depende del estimador de la función de densidad valuada en la mediana y en la mediana dividido 2. Para hallar ese estimador usaremos un procedimiento sugerido por Francisco y Fuller (1991).

Estimamos la mediana $\xi_{0.5}$ y su variancia, $\text{var}(\hat{\xi}_{0.5})$ utilizando los intervalos de confianza de Woodruff para la mediana, $(\hat{\xi}_l, \hat{\xi}_s)$, los cuales quedan determinados por el siguiente sistema de ecuaciones

$$\begin{cases} \hat{\xi}_l = \inf \left\{ y_{hci}, \hat{F}(y_{hci}) \geq 0.5 - z_{1-\alpha/2} \sqrt{\text{var}(\hat{F}(\hat{\xi}_{0.5}))} \right\} \\ \hat{\xi}_s = \inf \left\{ y_{hci}, \hat{F}(y_{hci}) \geq 0.5 + z_{1-\alpha/2} \sqrt{\text{var}(\hat{F}(\hat{\xi}_{0.5}))} \right\} \end{cases}$$

donde $z_{1-\alpha/2}$ es el percentil $100(1-\alpha/2)$ de una distribución $N(0,1)$, y $\text{var}(\hat{F}(\hat{\xi}_{0.5}))$ es obtenido utilizando el método de EE con $u_{hci}^* = I\{y_{hci} \leq \hat{\xi}_{0.5}\} - \hat{F}(\hat{\xi}_{0.5})$.

Definiendo $D_\alpha(\hat{\xi}_{0.5}) = 1/2(\hat{\xi}_S - \hat{\xi}_I)$ como la mitad de la longitud del intervalo de confianza del $100(1-\alpha)\%$ para $\xi_{0.5}$, podemos obtener un estimador de la función de densidad valorizada en la mediana

$$\hat{f}(\hat{\xi}_{0.5}) \approx \frac{z_{1-\alpha/2} \sqrt{\text{var} \left[\sum_s \tilde{w}_{hci} (I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) \right]}}{D_\alpha(\hat{\xi}_{0.5})}$$

La $\text{var} \left[\sum_s \tilde{w}_{hci} (I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5) \right]$ que aparece dentro de la raíz puede ser calculado mediante EE, tomando $u_{hc}^* = \sum_{i=1}^{n_{hc}} \tilde{w}_{hci} [I\{y_{hci} \leq \hat{\xi}_{0.5}\} - 0.5]$

De igual forma puede ser obtenida la función de densidad valorizada en $\hat{\xi}_{0.5}/2$.

4. PROGRAMA DE CÁLCULO DE VARIANCIA

Se diseñó un programa para realizar el cálculo de variancias mediante la técnica vista en el punto anterior, o sea, mediante EE. Fue realizado en SAS versión 6.12, utilizando lenguaje de macro, y en esta sección se dará una explicación de cómo funciona el mismo, qué parámetros se deben especificar y como se interpretan las salidas.

Consta de 4 etapas:

- a) Seteo de valores
- b) Definición de las macros
- c) Armado de las bases
- d) Llamado de macros

Excepto en la etapa correspondiendo a la definición de las macros, en las restantes se deben definir valores, variables y otras características propias del

funcionamiento del programa. Veremos a continuación el funcionamiento de cada etapa y los valores a definir en las mismas.

Etapa i: Seteo de valores

```
%let adato=x.xxxxxxx;
```

Se debe ingresar cuál es el archivo con la base de los datos original. La misma deberá contener una variable que contenga el identificador de estrato, una que identifique a los conglomerados, la variable con los pesos muestrales, y la o las variables de interés.

```
%let peso=xxxxxx;
```

Se debe ingresar el nombre de la variable de la base de datos definida en adato que contenga los pesos muestrales definitivos.

```
%let estrato=xxxxxx;
```

Se debe ingresar el nombre de la variable de la base de datos definida en adato que contenga el identificador del estrato.

```
%let cluster=xxxxxx;
```

Se debe ingresar el nombre de la variable de la base de datos definida en adato que contenga el identificador del conglomerado.

```
%let printgin=0;
```

```
%let printlor=0;
```

```
%let printlip=0;
```

Estos valores siempre deben permanecer iguales a 0.

Etapa ii: Definición de las macros

Es la parte del programa propiamente dicha, en la cual se realizan los cálculos necesarios para la obtención de las estimaciones de los errores.

Se definen 6 macros, 3 de cálculo (gini, lorenz, lip) y 3 de impresión de resultados (prntgin, prntlor, y prntlip). Tales como los nombres lo indican, la macro gini estima variancias para estimaciones del Índice de Gini, la macro lorenz para las estimaciones de las ordenadas de la curva de Lorenz, y la macro lip para las estimaciones de la proporción bajo la línea de pobreza.

Respecto a las macros de impresión de resultados, la macro prntgin imprime los resultados correspondientes a la macro gini en el caso que la misma haya sido invocada¹, la macro prntlor hace lo mismo con la macro lorenz, mientras que la macro prntlip imprime los resultados a la macro lip en el caso que esta haya sido invocada.

Cabe recordar que en esta etapa del programa no se deben especificar valores o variables para el funcionamiento del programa.

Etapa iii: Armado de las bases

En esta etapa se construye una base de datos llamada file que es la que se utilizará para calcular todas las estimaciones. La misma surge de pegar a la base original los pesos normalizados.

Para que el programa funcione con una base sin las variables que no se vayan a utilizar, es posible especificar que esta base file solamente contenga las variables que sean utilizadas durante la ejecución del programa. Para ello en la definición de

¹ Se entiende por invocar una macro al hecho de ejecutar la misma con el fin de obtener resultados

la base file, se puede especificar mediante un keep las variables que queremos contenga nuestra base de trabajo, no olvidando de incluir las que determinan el estrato (&estrato), el conglomerado (&cluster), el peso (&peso), y las de interés.

Por ejemplo, si deseamos realizar estimaciones sobre la variable ingreso, deberíamos especificar en el programa:

```
data file;  
  
set &adato(keep=&estrato &cluster &peso ingreso);  
  
if _n_=1 then set sal(keep=sump);  
  
peso=&peso/sump;  
  
run;
```

Etapa iv: Llamado de macros

En esta etapa se especifican aquellas macros que serán invocadas y las variables que determinaran las estimaciones. Veremos cómo llamar cada una de ellas.

- % gini(infile,var);

donde:

infile --> archivo con datos (llamada file)

var --> nombre de la variable de interés

Por ejemplo,

```
%gini(file,ingreso);
```

donde file es la base definida en la etapa iii, e ingreso es la variable de interés.

- `%lorenz(infile,var,p);`

donde:

`infile` --> archivo con datos (llamada file)

`var` --> nombre de la variable de interés

`p` --> percentil

Por ejemplo,

```
%lorenz(file,ingreso,.3);
```

donde file es la base definida en la etapa iii, ingreso es la variable de interés, y .3 será la ordenada que determina el valor acumulado de unidades para el cual se requiere el ingreso acumulado por los mismos.

```
%lip(infile,var);
```

donde:

`infile` --> archivo con datos (llamado file)

`var` --> nombre de la variable de interés

Por ejemplo,

```
%lip(file, ingreso);
```

donde file es la base definida en la etapa iii, e ingreso es la variable de interés.

5. RESULTADOS

Los siguientes resultados corresponden a los cálculos efectuados a partir de la información de la Encuesta Nacional de Gastos de los Hogares 1996/1997 para un conjunto de variables a modo de ejemplo del funcionamiento del algoritmo construido.

Los errores standard calculados corresponden al Índice de Gini, a la ordenada de la curva de Lorenz (con p igual a 0.30), y a la proporción de unidades bajo la línea de pobreza, a nivel Región y Total País de las variables seleccionadas.

	Índice de Gini	
	GINI	STD
Ingreso total del hogar		
1- País	0,45795	0,0040098
2- Región Metropolitana	0,45970	0,0068150
3- Región Pampeana	0,44486	0,0056741
4- Región Noroeste	0,43826	0,011159
5- Región Noreste	0,46055	0,0096698
6- Región Cuyo	0,43473	0,0095271
7- Región Patagónica	0,45094	0,0079187
Ingreso total del perceptor		
1- País	0,49486	0,0035414
2- Región Metropolitana	0,49521	0,0055467
3- Región Pampeana	0,48072	0,0056793
4- Región Noroeste	0,47750	0,0062484
5- Región Noreste	0,50714	0,013751
6- Región Cuyo	0,50714	0,013751
7- Región Patagónica	0,49014	0,0057700
Ingreso de la ocupación principal		
1- País	0,4921	0,0035104
2- Región Metropolitana	0,4793	0,0055121
3- Región Pampeana	0,4688	0,0059307
4- Región Noroeste	0,4722	0,0052042
5- Región Noreste	0,4847	0,014149
6- Región Cuyo	0,4943	0,014149
7- Región Patagónica	0,5356	0,0059331

	Proporción de unidades bajo la línea de pobreza	
	PUBLP	STD
Ingreso total del hogar		
1- País	-	-
2- Región Metropolitana	0,20578	0,0068359
3- Región Pampeana	0,19979	0,0064546
4- Región Noroeste	0,20734	0,0119950
5- Región Noreste	0,18977	0,0092963
6- Región Cuyo	0,19056	0,0082363
7- Región Patagónica	0,22072	0,0072160
Ingreso total del perceptor		
1- País	-	-
2- Región Metropolitana	0,23950	0,0047453
3- Región Pampeana	0,22093	0,0054352
4- Región Noroeste	0,22127	0,0062712
5- Región Noreste	0,22320	0,0118170
6- Región Cuyo	0,24540	0,0056521
7- Región Patagónica	0,24903	0,0098018
Ingreso de la ocupación principal		
1- País	-	-
2- Región Metropolitana	0,21917	0,0068489
3- Región Pampeana	0,21668	0,0060627
4- Región Noroeste	0,23516	0,0072726
5- Región Noreste	0,21942	0,010979
6- Región Cuyo	0,26532	0,0074854
7- Región Patagónica	0,25487	0,010414

	Ordenadas de la curva de Lorenz (p=0.30)	
	LORENZ	STD
Ingreso total del hogar		
1- País	-	-
2- Región Metropolitana	0,082906	0,0018243
3- Región Pampeana	0,088587	0,0014443
4- Región Noroeste	0,088933	0,0044186
5- Región Noreste	0,085698	0,0034861
6- Región Cuyo	0,090981	0,0025628
7- Región Patagónica	0,082282	0,0021984
Ingreso total del perceptor		
1- País	-	-
2- Región Metropolitana	0,069478	0,0014684
3- Región Pampeana	0,077164	0,0017141
4- Región Noroeste	0,074378	0,0020438
5- Región Noreste	0,067402	0,0043774
6- Región Cuyo	0,064911	0,0017311
7- Región Patagónica	0,067017	0,0017791

	Ordenadas de la curva de Lorenz ($p=0.30$)	
	LORENZ	STD
Ingreso de la ocupación principal		
1- País	-	-
2- Región Metropolitana	0,073136	0,0018513
3- Región Pampeana	0,075547	0,0019158
4- Región Noroeste	0,071294	0,0018668
5- Región Noreste	0,071884	0,0054708
6- Región Cuyo	0,057618	0,0022113
7- Región Patagónica	0,066340	0,0023913

BIBLIOGRAFÍA

- Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 34-42.
- Binder, D.A. y Kovacevic (1995). Estimating some measures of Income Inequality from Survey Data: An Application of the Estimating Equation Approach. *Survey Methodology*, Vol 21, No. 2, 137-145.
- Encuesta Nacional de Gastos de los Hogares 1996/1997. El Ingreso de los Hogares. Total del País. Regiones. Instituto Nacional de Estadísticas y Censos, Buenos Aires, 2000.
- Francisco, C.A. y Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- Kovacevic, M.S. y Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics*, Vol. 13, No. 1.
- Sarndall, C.E., Swensson, B. y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

- Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.