La post-estratificación
en el análisis de datos de encuestas

Roberto A. Molina Cruz

Introducción

En Guatemala se vienen realizando encuestas nacionales de hogares, dirigidas a investigar o medir variables de diferente tipo como empleo, ingresos y gastos familiares, salud, educación y actividad económica. La mayoría de estas encuestas son realizadas por el Instituto Nacional de Estadística (INE), el Ministerio de Salud Pública y Asistencia Social (MSPAS) y el Ministerio de Educación (MINEDUC).

La información de estas encuestas es inicialmente analizada en forma independiente, pero luego se analiza la información de 2 o más encuestas en forma conjunta. Debemos notar que este análisis conjunto de datos es cada vez más común, por ejemplo ha sido necesario para la elaboración de los Informes de Desarrollo Humano del Programa de Desarrollo de Naciones Unidas (PNUD), y la evaluación de las Metas del Milenio de educación, salud, condiciones de vida (como pobreza, hambre y desarrollo), equidad de género y de medio ambiente.

El análisis conjunto de los datos de varias encuestas presenta dificultades técnicas, debidas principalmente a que cada encuesta es diseñada de acuerdo a objetivos específicos particulares. Por ejemplo, durante el 2008 se estarán realizando las encuestas nacionales de salud materna e infantil (ENSMI), y de ingresos y gastos familiares (ENIGFAM). Aunque las dos encuestas estarán visitando un número similar de hogares, sus muestras incluyen alrededor de 730 y 1,500 sectores cartográficos, respectivamente. De hecho, la ENIGFAM estará encuestando alrededor de 12 hogares por cada sector, mientras que la ENSMI necesita encuestar alrededor de 30 hogares para obtener la información requerida de salud materna e infantil.

Para facilitar el análisis conjunto de los datos de 2 o más encuestas, es necesario realizar algunas acciones en cada encuesta. Estas acciones pueden pertenecer al diseño de cada encuesta, lo que corresponde a la

etapa de pre-muestreo, o bien al análisis de sus datos, esta es la etapa de post-muestreo.

Al nivel del diseño de las encuestas, recientemente el INE ha tomado medidas para facilitar el análisis conjunto de los datos de 2 o más encuestas nacionales. Entre las que está la elaboración de una muestra maestra en base al último censo, realizado en el 2002. La cual está sirviendo como fuente para las muestras de las encuestas nacionales, como la encuesta de condiciones de vida (ENCOVI), la ENIGFAM, la de empleo e ingresos (ENEI), y la ENSMI.

Sin embargo, además de las medidas tomadas en el diseño de las encuestas, es necesario también tomar medidas en el análisis de los datos. Como es la aplicación de métodos modernos para el cálculo de los factores de expansión, pesos o expansores, de las unidades encuestadas. En particular, la aplicación de los ajustes por post-estratificación. Lo que no se hace regularmente en las encuestas del INE, aunque es un procedimiento estándar en otros países.

En este artículo describo primero la estructura general de las encuestas nacionales. Discuto la relevancia de la post-estratificación en el análisis de las encuestas. Para lo cual presento como ejemplos dos encuestas: la ENEI realizada durante el 2002 y la encuesta del Programa Nacional de Evaluación de Rendimiento Educativo (PRONERE) del 2004. Por último describo la teoría relativamente simple del método de post-estratificación.

Por último, deseo agradecer al revisor del artículo por sus observaciones, las cuales sirvieron para una mejora significativa de la versión original. Así, siguiendo una de estas observaciones, al final del artículo incluyo un glosario de los términos de la teoría del muestreo que uso en el artículo.

Las encuestas nacionales

Las encuestas nacionales se basan desde luego en muestras de la población residente en el país, las cuales son obtenidas por medio de mecanismos complejos de selección. Estos mecanismos determinan aparte de la población objetivo de la encuesta, una población que llamamos población en estudio. Esta población comprende a todos los individuos que el mecanismo de selección puede escoger para incluirlos en una muestra.

Los mecanismos de selección se basan regularmente en marcos de muestreo, los cuales son listados con información específica de las unidades —individuos o conglomerados de individuos— que componen la población en estudio de la encuesta. Las mayores unidades de la población que aparecen listadas en los marcos de muestreo se les llaman unidades primarias. Ejemplos de estas unidades son los sectores censales y las escuelas del país. Estas unidades pueden a su vez contener otras unidades de la población, como podrían ser los hogares pertenecientes a los sectores censales y los grados de las escuelas. Las cuales también podrían contener otras unidades de la población, como las personas que componen los hogares, y las secciones y alumnos de cada grado.

Dado lo extenso y complejo de la población del país, los mecanismos de selección de las muestras de las encuestas consideran por lo regular los aspectos siguientes.

a) La estratificación de la población en estudio. Esto es la división de la población en estratos —subpoblaciones o grupos de individuos— más o menos homogéneos, e independientes con respecto al procedimiento de selección.

b) La selección dentro de cada estrato de unidades primarias de muestreo —conglomerados de individuos— e individuos de la población.

c) La asignación a los individuos de la población de probabilidades de selección y expansores posiblemente diferentes.

Así por lo regular, la población del país se estratifica en área rural y urbana, nivel socioeconómico alto, medio y bajo, y en las 8 regiones definidas por el INE.

Dentro de cada estrato se aplican procedimientos de muestreo de por lo menos dos etapas. En la primera etapa se seleccionan como unidades primarias de muestreo sectores censales. En la segunda etapa se seleccionan viviendas de los sectores censales seleccionados, ya sea como unidades secundarias de muestreo o como elementos de estudio. Luego, usualmente se seleccionan a todos los hogares de las viviendas seleccionadas, y a todas las personas de estos hogares.

Prácticamente todas las encuestas hacen uso de un procedimiento de muestreo que debe producir una muestra autoponderada. Así que por lo menos inicialmente, a todos los individuos encuestados en un mismo estrato se les asigna la misma probabilidad de selección y expansor. Sin embargo, es muy común que se deban ajustar este expansor. Por lo que regularmente se resultan teniendo diferentes expansores para los individuos de un mismo estrato.

La práctica de la post-estratificación

Actualmente, los métodos modernos de las encuestas ponen un mayor énfasis en la utilización de información auxiliar en la etapa del cálculo de las estimaciones, esto a través del ajuste de los expansores de las muestras. Este tipo de ajustes tiene ventajas como las siguientes.

a) Las estimaciones pueden hacerse consistentes con algunos totales poblacionales, conocidos por medio de otras fuentes de información.

b) Estos ajustes regularmente hacen uso de post-estratos, lo cual ayuda a que las estimaciones sean más precisas. Esto es, que las estimaciones tengan un menor margen de error.

c) Los ajustes incluso pueden ser usados para compensar alguna posible diferencia entre las poblaciones objetivo y en estudio de la encuesta.

Tradicionalmente, la información auxiliar era utilizada solamente en la etapa del diseño de la encuesta. Principalmente para la definición de los estratos en el marco de muestreo. Sin embargo, es común contar con información auxiliar que no puede ser utilizada para el diseño. Por lo que se emplean métodos que permiten incorporar esta información durante el cálculo de las estimaciones de la encuesta, siendo el más usado el ajuste de los expansores por post-estratificación.

En forma similar a la estratificación, la post-estratificación es más efectiva cuando se da lo siguiente.

a) Los individuos más similares —con respecto a las mismas variables de post-estratificación— pertenecen al mismo post-estrato, y los individuos más disimilares pertenecen a post-estratos diferentes.

b) Las variables de post-estratificación están correlacionadas con las variables de estudio de la encuesta.

La ENEI 2002-1

Durante el año 2002 se realizaron en Guatemala las ENEI. Estas encuestas tuvieron como objetivo principal estudiar en forma longitudinal el fenómeno del empleo en el país. Debido principalmente al poco tiempo que se tenía entre las encuestas, fue necesario trabajar con una muestra mucho más reducida que las demás encuestas nacionales. Por lo que en el marco de muestreo solamente se pudieron considerar las siguientes 2 variables de estratificación: Dominio (3 niveles) y nivel socioeconómico: (3 niveles).

Se vio conveniente ajustar los expansores de sus muestras por medio de una post-estratificación. Basada en hacer coincidir las estimaciones de la Población en Edad de Trabajar (PET), con las proyecciones elaboradas por el Centro Latinoamericano de Demografía (CELADE). Esto en cada uno de los 72 post-estratos definidos por los dominios, el género de las personas, y los 12 grupos de edad: 00 (7- 9 años), 01 (10-14), 02 (15-19), 03 (20-24), 04 (25-29), 05 (30-34), 06 (35-39), 07 (40-44), 08 (45-49), 09 (50-54), 10 (55-64) y 11 (65 y más).

La Tabla 1 presenta las cifras de la PET proyectadas (S) y estimadas por esta encuesta ($\hat{S}$), y el factor de ajuste por post-estratificación (F), correspondientes a los post-estratos de los hombres de los dominios 1 y 3.

En esta tabla notamos que para los post-estratos del dominio 1, la encuesta inicialmente sobreestimó la PET proyectada para los grupos de edad 0 a 6, y la subestimó para los grupos 7 a 11. Mientras que para los post-estratos

Tabla 1

Cifras de la PET proyectadas (S) y estimadas por la encuesta ($\hat{S}$), y el factor de ajuste por post-estratificación (F), correspondientes a los post-estratos de los hombres de los dominios 1 y 3. (PESTRA = Dominio + Género + Grupo de edad.

PESTRA	S	$\hat{S}$	F
1100	314,615	89,275	3.524
1101	142,987	94,870	1.507
1102	140,125	104,380	1.342
1103	122,250	77,875	1.570
1104	98,435	110,485	0.891
1105	76,176	51,365	1.483
1106	57,302	45,280	1.265
1107	45,464	48,110	0.945
1108	37,721	49,645	0.760
1109	31,196	40,125	0.777
1110	45,761	89,885	0.509
1111	42,908	56,270	0.763
3100	1,206,049	298,905	4.035
3101	493,375	436,805	1.130
3102	405,575	307,655	1.318
3103	337,620	303,560	1.112
3104	273,629	153,035	1.788
3105	216,977	187,440	1.158
3106	177,815	94,855	1.875
3107	144,465	173,860	0.831
3108	118,867	76,755	1.549
3109	94,971	96,070	0.989
3110	133,797	103,470	1.293
3111	118,640	150,130	0.790

del dominio 3, la encuesta inicialmente subestimó la PET proyectada para casi todos los post-estratos. Desde luego, las sobreestimaciones se deben a que fueron entrevistados más individuos de los esperados, y las subestimaciones a que fueron entrevistados menos de los esperados, en el correspondiente grupo de edad.

Tabla 2

Estimaciones de varias variables de ocupación, considerando los expansores iniciales de la encuesta.

Variable	Estimación	EE	CV	EDIS
PO	4,769,380	221,465	0.046435	29.2520
PDAA	90,577	10,646	0.117540	1.4775
PSV	737,014	48,955	0.066424	4.1771
PDAP	63,679	15,509	0.243560	4.4452
PEA	4,923,640	226,067	0.045915	30.9640
PEI	3,166,140	155,722	0.049183	14.6920
PDAT	154,256	19,046	0.123470	2.7990
PDTA	809,157	54,685	0.067582	4.7944

Tabla 3

Estimaciones de varias variables de ocupación, considerando los expansores de la encuesta ajustados por post-estratificación.

Variable	Estimación	EE	CV	EDIS
PO	4,834,100	89,554	0.018525	4.8129
PDAA	101,255	12,446	0.122910	1.8087
PSV	771,690	44,616	0.057816	3.3292
PDAP	65,135	14,804	0.227280	3.9600
PEA	5,000,490	87,685	0.017535	4.7009
PEI	3,089,300	87,685	0.028383	4.7009
PDAT	166,390	19,445	0.116860	2.7087
PDTA	821,066	49,006	0.059685	3.8007

En la Tablas 2 presentamos las estimaciones de varias variables de ocupación, considerando los expansores iniciales de la muestra de esta encuesta. Mientras que en la Tabla 3 presentamos las estimaciones de estas mismas variables, pero con los expansores ajustados por post-estratificación. Las variables de ocupación consideradas corresponden a la población ocupada (PO), desocupada abierta activa (PDAA), subocupada visible (PSV), desocupada abierta pasiva (PDAP), económicamente activa (PEA), económicamente inactiva (PEI), desocupada abierta total (PDAT) y desocupada total agregada (PDTA).

En estas tablas notamos que las estimaciones son similares con y sin una post-estratificación. Sin embargo, los errores estándar (S.E.) y coeficientes de

variación (C.V.) de varias variables se reducen considerablemente al aplicar el ajuste por post-estratificación. Similarmente los efectos de diseño, los cuales miden la eficiencia de la encuesta, resultan ser menores empleando los expansores ajustados. Lo que indica que los datos de la encuesta son mejor utilizado aplicando una post-estratificación.

Puede encontrarse una discusión sobre otros aspectos de estas encuestas en otro artículo del autor (1).

La encuesta del PRONERE del 2004

En el 2004 se realizó una de las encuestas de PRONERE para principalmente obtener estimaciones del rendimiento en lectura y matemática, de los alumnos de 1er grado de primaria de niños de los establecimientos de todo el país con las características siguientes: oficiales, con jornada matutina o doble, y del plan regular (diario). Además, por razones de presupuesto fue necesario considerar solamente los establecimientos con una determinada matrícula mínima. La población en estudio de la encuesta incluyó a 13,741 establecimientos.

Estos establecimientos fueron estratificados por medio de las variables: departamento, área (urbana o rural), y el tamaño de los establecimientos. En la encuesta se consideraron a los establecimientos como las unidades primarias de muestreo, las secciones del grado como unidades secundarias, y en cada sección seleccionada se evaluaron a todos los alumnos. La selección de los establecimientos de un mismo estrato se realizó con igual probabilidad de selección, así como la selección de la sección a evaluar en el grado.

Debemos notar que como estas encuestas se están realizando en forma anual, la selección de los establecimientos con probabilidades proporcionales a su tamaño, complica el manejo de la composición de las muestras de cada año. Ya que las muestras de un año podrían contener más establecimientos grandes que la de otros año, o bien se podría terminar encuestando los establecimientos más grandes más frecuentemente de lo deseado.

Por el otro lado, empleando el tamaño de los establecimientos como variable de estratificación, se puede agrupar los establecimientos en paneles para determinar la composición de la muestra de cada año, y el tiempo en que se desea encuestar de nuevo a los establecimientos de cada estrato.

Para el ajuste de los expansores de la encuesta se definieron los post-estratos de la población en estudio por medio de las variables: DEPTO, AREA y SEXO. En la Tabla 4 presentamos las cifras de la población en estudio de alumnos, primero proyectadas en base a registros administrativos (S) y luego estimadas por medio de la encuesta ($\hat{S}$). Además presentamos el factor de ajuste por post-estratificación (F), correspondientes post-estratos con los valores de F que más difieren de 1.

Tabla 4

La población de alumnos de 1er grado en estudio, proyectada en base a registros administrativos (S) y estimadas por la encuesta (Ŝ), y el factor de ajuste por post-estratificación (F), correspondientes a los post-estratos definidos.

		S		Ŝ		F	
DEPTO	SEXO	Urbana	Rural	Urbana	Rural	Urbana	Rural
00	H	5,127	263	3,966	120	1.293	2.192
	M	5,935	233	4,655	197	1.275	1.183
03	H	1,971	1,252	2,363	1,167	0.834	1.073
	M	1,755	1,146	1,059	1,018	1.657	1.126
04	H	2,020	6,470	1,150	5,220	1.757	1.239
	M	2,077	6,295	2,214	5,196	0.938	1.212
05	H	1,999	8,411	1,126	7,227	1.776	1.164
	M	1,925	7,194	1,583	7,243	1.216	0.993
06	H	886	5,937	573	3,948	1.546	1.504
	M	820	5,315	822	4,261	0.997	1.247
07	H	1,217	6,116	1,161	3,977	1.049	1.538
	M	1,212	5,882	1,347	4,379	0.900	1.343
08	H	528	7,351	286	6,138	1.844	1.198
	M	700	7,373	760	5,389	0.921	1.368
14	H	1,534	18,147	1,585	11,235	0.968	1.615
	M	1,490	16,550	1,293	11,282	1.152	1.467
17	H	677	10,214	615	7,212	1.102	1.416
	M	622	9,264	437	7,612	1.425	1.217
20	H	503	5,643	575	3,719	0.875	1.517
	M	641	5,047	420	3,747	1.526	1.347
21	H	472	5,580	544	3,804	0.868	1.467
	M	847	5,003	773	3,637	1.096	1.376
22	H	575	8,690	353	6,028	1.628	1.442
	M	966	8,067	958	5,523	1.008	1.461

Debemos notar que para la mayoría de post-estratos estos factores son aproximadamente iguales a 1, lo que significa que en ellos el ajuste por post-estratificación es prácticamente innecesario. Sin embargo, en varios post-estratos estos factores son considerablemente diferentes a 1. Los factores mayores a 1 significan que en estos post-estratos se terminaron evaluando menos alumnos de los esperados, mientras que los factores menores a 1 significan que se evaluaron más de los esperados.

El objetivo principal del análisis de los datos fue estimar la distribución de los alumnos de la población en estudio, con respecto a su calificación obtenida en las pruebas de lectura y matemática. En la Tabla 5 presentamos como ejemplo la estimación de algunos parámetros de la distribución de la población de alumnos en estudio, con respecto a su calificación cruda -no estandarizada- en lectura.

Tabla 5

Estimaciones de algunos parámetros de la distribución de los alumnos en estudio de 1er grado, con respecto a su calificación cruda en lectura.

2	Variable dependiente: LTOTAL			
2	Media de subpoblación			
2	**Estimación**	**EE**	**CV**	**EDIS**
1	1.3136571D+01	1.2754519D-01	9.70917D-03	1.25851D+01
2	Cuantiles			
2		**Estimación**	**EE**	**Intervalo de confianza (95%)**
1	0.01	1.0667393D+00	2.0050360D-01	(6.21313D-01, 1.42333D+00)
1	0.05	3.5956376D+00	2.0558913D-01	(3.22492D+00, 4.04727D+00)
1	0.10	5.5896773D+00	1.3544908D-01	(5.33210D+00, 5.87389D+00)
1	0.25	9.1905534D+00	1.8869649D-01	(8.81040D+00, 9.56519D+00)
1	0.50	1.3516906D+01	1.7930634D-01	(1.31525D+01, 1.38698D+01)
1	0.75	1.6958119D+01	1.1909375D-01	(1.67173D+01, 1.71937D+01)
1	0.90	1.8773445D+01	7.2083356D-02	(1.86168D+01, 1.89051D+01)
1	0.95	1.9488396D+01	7.4590089D-02	(1.93198D+01, 1.96181D+01)
1	0.99	1.9957300D+01	2.7391028D-02	(1.98885D+01, 1.99980D+01)
2	Rango intercuartil			
2		**Estimación**	**EE**	
2		7.7675660D+00	1.6750384D-01	

Lamentablemente cuando se calcularon estas estimaciones, no se contó con un paquete estadístico que permitiera incorporar la información de la post-estratificación realizada. Sin embargo, dado que se ajustaron los expansores las estimaciones obtenidas son las correctas, pero los errores estándar posiblemente están siendo sobre estimados. Por lo que los intervalos de confianza correctos podrían ser más angostos.

La teoría de la post-estratificación

La post-estratificación es presentada en los textos clásicos de teoría de muestreo, por ejemplo el texto de Cochran (2), como la definición de los post-estratos y el ajuste de los expansores iniciales de la encuesta por medio de la multiplicación del factor

$$F_g = S_g / \hat{S}_g$$

Desde luego, S_g es el valor total de la variable auxiliar en el post-estrato g, y $\hat{S}_g$ es ese mismo valor pero estimado por medio de los expansores iniciales de la encuesta. Es decir,

$$S_g = \sum_{i \varepsilon g} s_i \quad \text{y} \quad \hat{S}_g = \sum_{i \varepsilon g \cap m} w_i \, s_i$$

Donde w_i representa el expansor inicial del individuo *i* perteneciente a la muestra m. Por lo que el expansor ajustado para ese mismo individuo es como sigue.

$$w^*_i = (S_g / \hat{S}_g) \, w_i$$

Sin embargo ahora se sabe que al aplicar una post-estratificación, los estimadores usuales se convierten en estimadores de regresión bajo el modelo de las medias de grupo, o de Análisis de Varianzas, como se explica en el texto de Särndal et al (3).

En este modelo se asume que el valor esperado y la varianza de la variable:

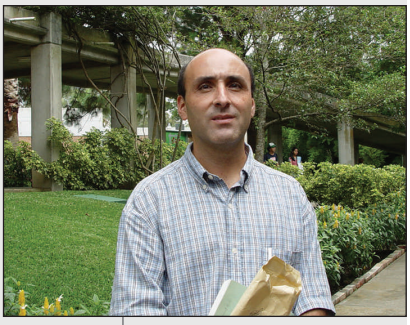
Y_i, son constantes para todos los individuos i del mismo post-estrato g. Esto es,

$$E[Y_i] = \beta_g \quad y \quad V[Y_i] = \sigma^2_g$$

El modelo lineal asociado puede escribirse entonces como $E[Y_i] = \bar{\delta}^t_i \, \bar{\beta}$. Donde los vectores de parámetros del modelo y pertenencia a los post-estratos son como sigue.

$$\beta^t = (\beta_1, \beta_2, \ldots, \beta_G) \;\; y \;\; \underline{\delta}^t_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{Gi})$$

Aquí G representa el número total de post-estratos, y las variables de pertenencia a los post-estratos toman los valores $\delta_{gi} = 1$, si i pertenece al post-estrato g, y $\delta_{gi} = 0$ si no pertenece.



Roberto A. Molina Cruz
rmolina@uvg.edu.gt
Catedrático Departamento de Matemáticas de la Facultad de Ciencias y Humanidades de la Universidad del Valle de Guatemala

Bibliografía

1) Molina-Cruz, R.A. Surveying Employment in a Developing Country. Bulletin of the International Statistical Institute, 54th Session. Berlin, 2003.

2) Cochran, W. G. Técnicas de Muestreo. Compañía Editorial Continental S. A. (CECSA). México, 1987.

3) Särndal, C.E. et al. Model Assisted Survey Sampling. Springer-Verlag. New York, 1992.

Glosario

- **Ajuste de los expansores (en inglés, weighting):** El cálculo de los expansores de una encuesta se realiza regularmente en dos o más etapas. En la primera se calculan los expansores básicos como el inverso de la probabilidad de selección de cada unidad de estudio, y en las demás etapas básicamente se ajustan estos expansores básicos por medio de la multiplicación de factores, estos asociados a una posible no respuesta o una post-estratificación.

- **Coeficiente de variación (CV):** Es conveniente evaluar la precisión de una estimación en forma relativa al mismo valor de la estimación. Para lo que regularmente se emplea el coeficiente de variación, que se calcula como el cociente del error estándar de la estimación y el valor de la estimación.

- **Efecto de diseño (EDIS):** Idealmente toda encuesta debería diseñarse en forma simple, principalmente sin el uso de conglomerados. Sin embargo, esto es regularmente imposible por razones prácticas. Cualquier otro diseño —complejo— usualmente provoca una mayor varianza de los estimadores, por lo que es usual evaluar el diseño de la encuesta comparando estas varianzas con las que pudieron haberse obtenido mediante un diseño simple. Para esto se calcula el efecto del diseño de la encuesta como el cociente de la varianza de una estimación obtenida con el diseño complejo, y la varianza de la misma estimación que pudo haberse obtenido con el diseño simple.

- **Error estándar (EE):** La precisión de toda estimación puede evaluarse en forma absoluta, para lo que regularmente se emplea el error estándar de la estimación. Esta se calcula como una desviación estándar.

- **Estimación:** El objetivo principal de toda encuesta es la obtención de valores que estimen a algunos parámetros poblacionales. Estos valores son obtenidos por medio de estimadores y son llamados estimaciones.

- **Estratificación:** Por esto nos referimos a la división de la población en estudio en estratos, subpoblaciones o grupos de individuos más o menos homogéneos. La incorporación de estratos en el diseño de una encuesta por lo regular mejora la precisión de sus estimadores.
- **Expansor:** Decimos que la muestra de una encuesta representa a la población en estudio, de forma que a cada unidad de la muestra se le asocia un peso, factor de expansión o expansor, el cual indica el número de unidades de la población que esa unidad está representando.
- **Muestra autoponderada:** Bajo un diseño específico, la muestra de una encuesta puede ser obtenida de forma que todas sus unidades tengan asociado un mismo expansor. A estas muestras se les denomina autoponderadas. En la práctica casi todas las encuestas se diseñan para obtener una muestra autoponderada.
- **Margen de error de una estimación:** La precisión de una estimación puede evaluarse en forma absoluta mediante el margen de error de la estimación, el cual es igual al radio del intervalo de confianza de la estimación. Esto desde luego para el nivel de confianza adoptado.
- **Muestreo con probabilidades proporcionales al tamaño (PPT):** En los diseños de encuestas que hacen uso de conglomerados (complejos), la selección de las unidades de estudio se realiza en 2 o más etapas. En la primera de las cuales regularmente se seleccionan los conglomerados asignándoles probabilidades de selección proporcionales a su tamaño --este es el número de unidades que contienen.