

METODOLOGIA DEL DISEÑO MUESTRAL

ENCUESTA DE HOGARES - 2009

Contenido

1.	INTRODUCCION	2
2.	ANTECEDENTES.....	3
3.	OBJETIVOS.....	4
3.1	Objetivos de la Encuesta.....	4
3.1.1	Objetivo General	4
3.1.2	Objetivo específicos.....	4
3.2	Objetivos del Diseño Muestral	4
3.2.1	Objetivo General	4
3.2.2	Objetivo específicos.....	4
4.	EL DISEÑO MUESTRAL.....	5
4.1	Cobertura del estudio	5
4.2	Marco Muestral.....	5
4.3	Método estadístico de muestreo.....	6
4.4	Estratificación.....	6
4.5	Unidades de muestreo	8
4.6	Tamaño de la muestra	9
4.7	Asignación de la muestra	13
4.8	Probabilidades de selección.....	15
4.8.1	Probabilidad de primera etapa.....	16
4.8.2	Probabilidad de segunda etapa	16
4.8.3	Probabilidad de tercera etapa	16
4.8.4	Probabilidad de cuarta etapa	17
4.9	Actualización cartográfica de las unidades de muestreo	17
4.10	Factores de expansión	18
4.10.1	Factor de expansión base	18
4.10.2	Ajustes a los factores de expansión.....	19
4.11.1	Determinación de los estimadores de un total poblacional.....	21
4.12	Cálculo de errores muestrales	22
4.12.1	Estimación de la varianza	24
4.13	Encuesta Piloto	29
5.	EVALUACIÓN DE LA CALIDAD DE LOS DATOS.....	29
6.	BIBLIOGRAFÍA	30

1. INTRODUCCION

El presente documento establece la construcción del diseño muestral para la Encuesta a Hogares 2009, esta encuesta es desarrollada por el Instituto Nacional de Estadística de Bolivia y se la realiza durante los meses de octubre y noviembre de cada año, con el fin de obtener información de la población respecto a aspectos socioeconómicos, por esta razón se hace sumamente necesario considerar la construcción de un diseño muestral que se ajuste a estos requerimientos, la encuesta se ejecuta mediante una muestra representativa a nivel nacional, realizando una distinción en cuanto al área urbano rural, además de utilizar un estrato que fue construido en base al NBI (Índice de necesidades insatisfechas) para mejorar la precisión de la encuesta,

El diseño muestral engloba todos los aspectos teóricos dentro la teoría del muestreo estadístico, se utiliza un muestreo complejo utilizando los principios, de la estratificación por conglomerados y el muestreo polietápico, incluyendo también al muestreo proporcional al tamaño y el muestreo sistemático, el marco muestral utilizado en esta encuesta es el denominado “Marco Maestro” que fue construido en base a la Información generada del Censo Nacional de Población y Vivienda 2001, este es un marco de áreas que enmarca todo el territorio nacional, disgregado en unidades compuestas que dan el escenario ideal para la ejecución de un muestreo complejo, dentro de los denominativos estadísticos este es un “Marco en sentido amplio”, porque contiene variables que permiten la inmersión en la planeación de un muestreo complejo.

El diseño muestral de esta encuesta se convierte en un marco referencial para otras instituciones que acogen la utilización del muestreo estadístico, probabilístico, dentro del seno de sus labores de recolección de información, dirigidas a hogares a cualquier nivel de dominio de estudio, además de ser un aporte para los especialistas inmersos en el ámbito del muestreo y para las personas que estén interesadas en profundizar su conocimiento, así todos empecemos con el reconocimiento de esta ciencia que es la estadística y que en esta ocasión mostramos un ámbito de esta, que es la teoría del muestreo.

2. ANTECEDENTES

Desde el año 1978, el Instituto Nacional de Estadísticas ha efectuado una serie de rondas de encuestas por muestreo a hogares: Encuesta Permanente de Hogares (EPH) de 1980 - 1988, Encuesta Integrada de Hogares (EIH) de 1989 - 1995, Encuesta Nacional de Empleo (ENE) de 1996 - 1998, Encuesta de Hogares del Programa MECOVI de 1999 - 2004.

En 1999 a iniciativa del Banco Mundial en la región, se inició el Programa de Mejoramiento de las Encuestas y Medición sobre Condiciones de Vida en América Latina y el Caribe (MECOVI). Como parte de este programa, desde el año 1999 se realiza cada año la Encuesta a Hogares del programa MECOVI, ya en los años 2003 - 2004, dentro del mismo programa se realiza la Encuesta Continua de Hogares y el restante Encuesta de Hogares Puntuales periódicas a partir de la implementación de un cuestionario multitemático, que consta de los siguientes módulos: Información general de los miembros del hogar, migración, salud, educación, empleo, ingresos no laborales, gastos en consumo, vivienda, contingencias y préstamos del hogar e ingreso del productor agropecuario independiente, etc. El objetivo de la misma es generar información sobre las condiciones de vida de la población del país en términos de alcance, cobertura, confiabilidad y, la relevancia para fines de políticas dirigidas hacia la reducción de la pobreza y el aumento de la equidad social y la consecuente formulación de políticas y programas que contribuyan a mejorar las condiciones de bienestar de los hogares.

A partir del año 2005 se plantea presentar un panorama completo sobre las condiciones de vida de la población, como un aporte para el estudio de la realidad nacional y apoyo para instancias de toma de decisiones en materia de política tendientes a la reducción de la pobreza, como usuarios organismos internacionales, comunidad académica, investigadores privados y estudiantes.

Cada Encuesta de Hogares tuvo enlazado un diseño muestral, que busca siempre lograr la mayor representatividad de la población con la implementación de la teoría de muestreo de una manera parsimoniosa.

3. OBJETIVOS

3.1 Objetivos de la Encuesta

3.1.1 Objetivo General

El objetivo general de la Encuesta de Hogares 2009 es obtener información sobre las condiciones de vida de los hogares, a partir de la recopilación de información de variables socioeconómicas y demográficas de la población boliviana, necesarias para la formulación, evaluación, seguimiento de políticas y diseño de programas de acción en el área social.

3.1.2 Objetivo específicos

- Generar indicadores que permitan conocer la evolución de la pobreza, del bienestar y las condiciones de vida de los hogares.
- Medir el alcance de los programas sociales en la mejora de las condiciones de vida de la población.
- Servir de fuente de información a instituciones públicas y privadas, así como a investigadores.
- Permitir la comparabilidad con investigaciones similares, en relación a las variables investigadas.

3.2 Objetivos del Diseño Muestral

3.2.1 Objetivo General

Presentar el Diseño Muestral ejecutado en la Encuesta a Hogares 2009

3.2.2 Objetivo específicos

- Mostrar la técnica de muestreo estadístico asociado al Diseño Muestral.
- Mostrar los elementos involucrados en el desarrollo de la construcción del diseño muestral.
- Especificar el cálculo del tamaño de la muestra y el nivel de la desagregación.
- Mostrar la forma de las probabilidades de selección y los factores de expansión.

4. EL DISEÑO MUESTRAL

En este apartado se dará todo el tratamiento respecto a los componentes involucrados dentro la construcción de un diseño muestral, iniciando con un reconocimiento en cuanto a la cobertura del estudio, los ámbitos espaciales y temporales de la encuesta y el marco muestral utilizado, para luego enfocarnos en el método estadístico, la estratificación, las unidades muestrales, la determinación del tamaño de la muestra y la asignación de la muestra, para finalizar con mostrar la estructura de las probabilidades de selección, los factores de expansión y los ajustes considerados sobre los factores de expansión.

4.1 Cobertura del estudio

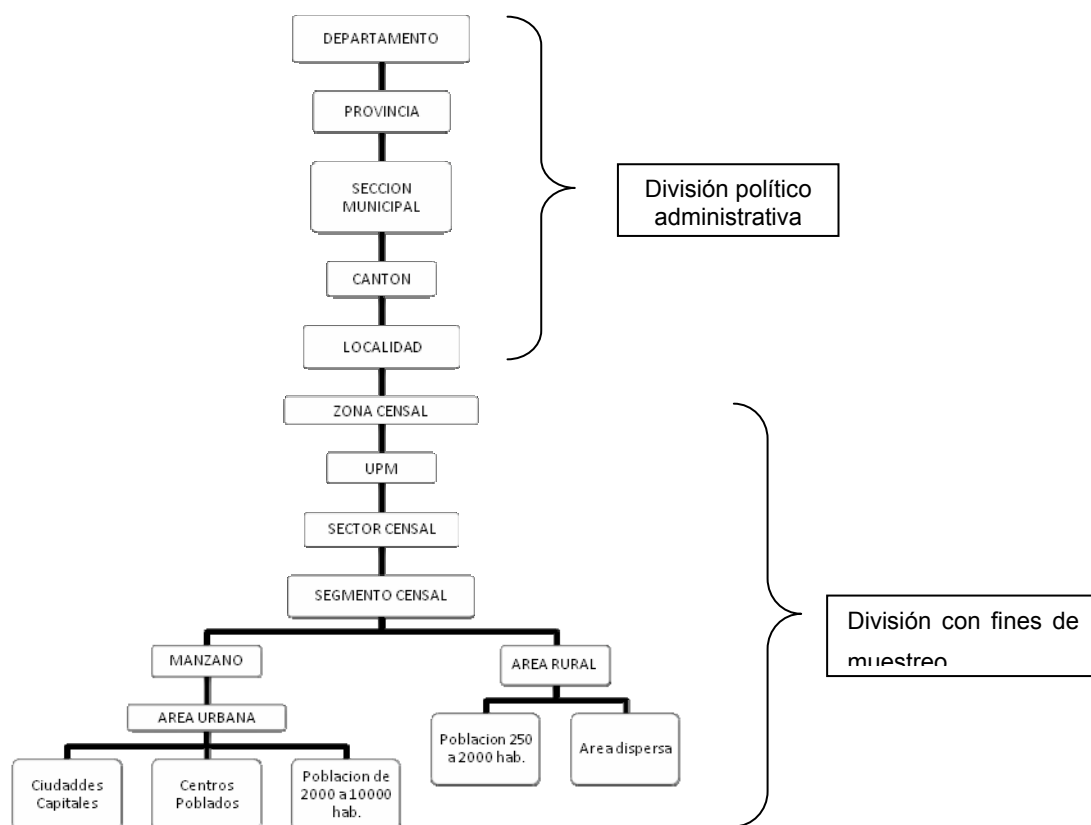
La investigación está dirigida al conjunto de los hogares establecidos en viviendas particulares ocupadas de las ciudades capitales, resto urbano y área rural de Bolivia, en el periodo del año 2009. Excluye a los hogares constituidos en viviendas colectivas, de esta manera el ámbito espacial es el territorio nacional, y el ámbito temporal nos conduce al periodo 2009.

4.2 Marco Muestral

El marco muestral con el que se trabaja es el que proporciono el Censo Nacional de Población y Viviendas de 2001 (CNPV-2001) denominado “Marco Maestro”, es un marco en sentido amplio, es decir; tiene variables que permiten la estratificación, la conglomeración, además es un marco que contiene unidades de muestreo compuestas, entre UPMs, sectores censales, segmentos censales, y otras variables auxiliares que permiten la aplicación del muestreo complejo en todas sus dimensiones, al tener una conformación por viviendas contenidas dentro de cada unidad compuesta, una característica importante de este marco es que es un “Marco de áreas”, esto implica que toda Bolivia está representada en el marco.

Un esquema de la estructura del “Marco Muestral Maestro” se presenta a continuación:

Esquema N° 1, Estructura del Marco Maestro



4.3 Método estadístico de muestreo

El método de muestreo se enmarca en los muestreos probabilísticos, es un muestreo de cuatro etapas, estratificado, por conglomerados, utilizando un muestreo proporcional al tamaño (PPT) en sus tres primeras etapas, utilizando como peso el número de viviendas dentro las UPMs, USMs y UTMs, ya en última etapa se utiliza al muestreo sistemático simple con arranque aleatorio e intervalo fijo en las UUMs que son las viviendas.

4.4 Estratificación

Debido a que se cuenta con un marco muestral compuesto, este nos permite la utilización de un muestreo complejo, y como se describió anteriormente, este Diseño Muestral contiene cuatro etapas, además involucra una estratificación en primera etapa que se describen a continuación.

La estratificación siempre debe responder a un hecho de dividir a la población de estudio en grupos independientes que muestren diferencias notorias en cuanto a las respuestas de las unidades de información respecto a variables que estén

dentro la temática de interés, para así garantizar una ganancia en precisión. En este entendido la estratificación de la Encuesta a Hogares 2009 considera la unión de dos variables que forman grupos que entran dentro la condición de estratos, la primera variable identifica el área urbana o área rural de la Unidad primaria de muestreo, y la segunda, denominada estrato estadístico (NBI), que fue construida en base a los niveles del NBI, estos son:

Cuadro N° 1, Niveles del NBI

Niveles del NBI
Necesidades Básicas Satisfechas (NBS)
Umbral de Pobreza
Pobreza Moderada
Indigencia
Marginalidad

Los NBIs de las áreas temáticas para el mapa de pobreza se clasificaron en cuatro subestratos denominados estratos estadísticos mediante el método estadístico de Clúster de k medias para los fines y objetivos establecidos en la Encuesta de Hogares, de acuerdo al siguiente criterio aproximado:

- 1: Estrato Alto**, son aquellas unidades muestrales que se encuentran con las necesidades básicas satisfechas.
- 2: Estrato Medio Alto**, son unidades muestrales que están en el umbral de pobreza.
- 3: Estrato Medio Bajo**, son unidades muestrales que están en la clasificación pobreza moderada.
- 4: Estrato Bajo**, son unidades muestrales que están entre la indigencia y marginalidad de pobreza.

A continuación se muestra la estructura de las UPMs en primera etapa con la estratificación Área Urbano Rural y Estrato estadístico NBI, obteniendo 8 estratos independientes que asocian a las 16790 UPMs del Marco Maestro.

Cuadro N° 2 UPMs por estrato estadístico y área

		Estrato estadístico NBI				Total
		1	2	3	4	
Área	Urbano	1216	4173	4937	1843	12169
	Rural	23	151	1135	3312	4621
Total		1239	4324	6172	5155	16790

4.5 Unidades de muestreo

Unidad de primera etapa

El marco muestral contiene 16790 unidades compuestas de muestreo denominadas UPMs que engloban o enmarcan a todo el territorio nacional, que no son más que la asociación de uno a mas sectores censales, estas denominadas UPMs las utilizaremos como las unidades primarias de muestreo.

Unidad de segunda etapa

Por la estructura del marco muestral la unidad inmediatamente inferior a la UPMs es el sector censal, la razón de elegir al sector como unidad secundaria de muestreo (USM) recae en el hecho de que si bien es cierto una UPMs es la asociación de uno o más sectores como se muestra a continuación.

Cuadro N° 3, Frecuencia y porcentaje de sectores censales por UPMs

Numero de sectores por UPM		Frecuencia	Porcentaje
1		11270	67,1
2		5406	32,2
3		103	,6
4		9	,1
5		2	,0
Total		16790	100,0

Existe un 32.9% de UPMs que tienen más de un sector asociado, el manejo de esta situación en campo es dificultosa ya que en muchos casos los sectores dentro la UPM no son continuos.

Unidad de tercera etapa

La denominamos UTM (Unidad terciaria de muestreo), esta unidad es lo que se denomina “segmento censal” dentro el Marco Maestro, que es la unidad inmediatamente menor del sector censal, el numero de segmentos por sector oscila entre lo más común de entre 5 a 7 segmentos, esta etapa será aplicada exclusivamente en el área rural dispersa, puesto que su utilidad es mayor en esta, debido al hecho de distancias y manejo en el operativo de campo.

Unidad de cuarta Etapa

En esta etapa entramos a las viviendas dentro del segmento (área rural dispersa), y dentro el sector censal (área urbana, resto rural), así la UUM (unidad ultima de muestreo) es la vivienda, aclarando en este punto que serán excluidas del estudio las viviendas colectivas.

4.6 Tamaño de la muestra

Debido al uso de un diseño de muestral trietapico en el área urbana y tetra-etápico en el área rural, respectivamente la determinación del tamaño de la muestra se convierte en un trabajo complejo, ya que es necesario incorporar un análisis de precisión de las variables de estudio, tanto en las Unidades Primarias de Muestreo (UPM) como las Unidades Secundarias de Muestreo (USM), las Unidades Terciarias de Muestreo (UTM), y las unidades cuartas de muestreo (UCM) respectivamente. En este caso, se suma la dificultad asociada al hecho de que la infraestructura del diseño muestral es de propósitos múltiples y, por lo tanto, eso hace necesario considerar la posibilidad de que otras variables estén bien medidas.

En este contexto, la determinación del tamaño de la muestra, trata, en primer lugar, de ser compatible con los niveles de estimación o dominios de estudio y la importancia de los estratos definidos, y con las principales variables que son objeto de estimación.

En ese entendido, el problema consiste en determinar el tamaño de muestra óptimo considerando las recomendaciones sobre tópicos de tamaño de muestra se dan en el manual sobre Encuestas de Población Económicamente Activa de la Organización Internacional de Trabajo (OIT) y el Mapa de Pobreza generado en base a la información Censal de 1992 y 2001, respectivamente.

El tamaño de muestra óptimo para estimar las proporciones multinomiales en este caso la Condición o Estrato de Pobreza (1. Necesidades Básicas Satisfechas; 2. Umbral de pobreza; 3. Pobreza Moderada; 4. Indigencia; 5. Marginalidad), que es equivalente a la construcción de intervalos de confianza simultáneos para la variable pobreza que se ajusta a una distribución multinomial debido que está compuesto por 5 categorías.

Distintos autores (Quesenberry y Hurts, 1964; Goodman, 1965; Tortora, 1978; Medina, 1998) han analizado procedimiento para determinar el tamaño de muestra necesarios para proporciones multinomiales, que consiste en dividir una muestra de tamaño n en k (5) categorías exhaustivas y mutuamente excluyentes, como sigue.

Cuadro N° 5, Población por NBI

Condición de pobreza	Ni	Pi	Qi	piqi
Necesidades Básicas Satisfechas (NBS)	1328873	0,16581108	0,83418892	0,138317765
Umbral de Pobreza	1990043	0,24830904	0,75169096	0,18665166
Pobreza Moderada	2742319	0,34217482	0,657825184	0,225091211
Indigencia	1738130	0,21687641	0,783123585	0,169841035
Marginalidad	215015	0,02682865	0,973171349	0,026108874
Total	8014380	1		

Donde

P_i : representa la proporción de la población que se encuentra en una de las cinco categorías, que satisface la siguiente condición $\sum_{i=1}^5 p_i = 1$ (por ejemplo, $p_1 = N_1/N = 1328873/8014380 = 0.16581108$).

N_i : Representa la frecuencia absoluta en la i – ésima categoría de la población y

$$\sum_{i=1}^5 N_i = N$$

Por tanto, para un nivel de significancia determinado α se establece un conjunto de intervalos de confianza que contenga al verdadero valor del parámetro P_i con un nivel de confianza igual a $1 - \alpha$, y tiene siguiente expresión matemática (propuesto por Goodman, 1965):

$$P\left(p_i - \sqrt{\frac{Bp_iq_i}{n}} \leq \pi_i \leq p_i + \sqrt{\frac{Bp_iq_i}{n}}\right) = 1 - \alpha$$

Donde

$1 - \alpha$: Nivel de confianza de la encuesta.

α : Nivel de significancia.

B : Es el percentil del límite superior $1 - \frac{\alpha}{2}$ de una distribución chi – cuadrada (χ^2)

con un grado de libertad.

$\sqrt{\frac{p_iq_i}{n}}$: Es la desviación estándar de la i – ésima categoría de una proporción multinomial.

Si se desea para cada categoría un ε_i error permisible, se forma el siguiente sistema de ecuaciones:

$$\varepsilon_i = \sqrt{\frac{Bp_iq_i}{n}}$$

Despejando n se tiene:

$$n = \max\left\{\frac{Bp_iq_i}{\varepsilon_i^2}\right\}, \text{ por ejemplo; para la condición de pobreza NBS: } n_1 =$$

$$\frac{6,6348913 * 0.138317765}{0.00936^2} = 10475.15 \quad \text{y} \quad \text{así} \quad \text{sucesivamente} \quad \text{para} \quad n_4 =$$

$$\frac{6,6348913 * 0.026108874}{0.00936^2} = 1977.289486. \text{ Y procedimiento tradicional de determinación del}$$

$$\text{tamaño de la muestra de la forma: } n^* = \frac{Z^2_{1-\frac{\alpha}{2}} p_i q_i}{\varepsilon^2}. \text{ Ejemplo: } n^*_2 = \frac{1.96^2 * 0.18665166}{0.00936^2} =$$

8184.502806, etc.

Fijando el error máximo deseado por $\varepsilon = 0,00936$ para la incidencia verdadera de pobreza, con un nivel de confianza del 95% de confianza y la notación utilizada por Tortora ($1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow Z_{1-\frac{\alpha}{2}} = Z_{1-0.05} = Z_{0.975} = 1.96$; $\varepsilon_i = \varepsilon = 0,00936$

$\forall i$), por lo tanto, el tamaño de muestra preliminar está dado:

Condición de pobreza	Ni	Pi	Qi	Pili	ni (Tortora)	ni* (mas)
NBS	1328873	0,16581108	0,83418892	0,138317765	6064,865342	10475,14586
Umbral de Pobreza	1990043	0,24830904	0,75169096	0,18665166	8184,177805	14135,59104
Pobreza Moderada	2742319	0,34217482	0,657825184	0,225091211	9869,649651	17046,71313
Indigencia	1738130	0,21687641	0,783123585	0,169841035	7447,076706	12862,48091
Marginalidad	215015	0,02682865	0,973171349	0,026108874	1144,804535	1977,289486
Total	8014380	1			9869,649651	17046,71313

ni/ni*	0,578977
Viviendas particulares con NBI	2258162
α : nivel de significancia	0,05
$Z_{1-\alpha/2}$: Valor normal del nivel de confianza	1,9599611
ε : error permisible	0.00936
B: valor acumulado de $\chi^2_{1-\frac{\alpha}{k}, 1}$	6,6348913
$\frac{\alpha}{k}=0.05/5$, donde k es el N° de estratos de pobreza (5)	0,01
n personas	9869.65
n _o viviendas = n/ M	2780.91
Tamaño de muestra: n	4260.564 \cong 4260
M = Número de miembros/ hogar según NBI	3,54907
Deff (encuesta continua de hogares 2003 – 2004)	1,421
TNR: (encuesta continua de hogares 2003 – 2004)	0,0725

Conforme a los cálculos habituales, se aconsejaría elegir $9869.65 \cong 9870$ personas, situación que representa el 58% del número de selecciones sugeridas a partir del procedimiento propuesto por Tortora.

En una primera aproximación se trató de fijar un tamaño de muestra, en términos del total de viviendas, que resultase compatible con la obtención de un error de muestreo de baja magnitud en las estimaciones. El tamaño se fijó suponiendo un Muestreo Aleatorio Simple (MAS) de viviendas, el cual se ajustó por un factor que da cuenta del “efecto de diseño (Deff)” en el tamaño, al muestrear por conglomerados (sectores censales) en lugar de viviendas.

Otro de los aspectos necesarios es considerar en la determinación del tamaño de la muestra la cobertura de la encuesta, con el propósito de identificar viviendas donde no se hizo la encuesta continua de hogares de 2003 - 2004, se tiene:

$$n = \frac{n_0 * Deff}{(1 - TNR)} = 4260$$

Donde:

TNR: Tasa de No Respuesta Máxima, 7.25%.

Deff: Es el cociente de la varianza del diseño utilizado entre la varianza que resulta de utilizar el muestreo aleatorio simple, con un mismo tamaño de muestra. El cual es de 1.421.

4.7 Asignación de la muestra

El tamaño de muestra es de 4260 viviendas a nivel nacional, se está utilizando un muestreo estratificado considerando 8 estratos, ahora nos resta distribuir estas 4260 viviendas en los ocho estratos. Por un aspecto estrictamente operativo, el número de viviendas entrevistadas por UPM será de 12, de esta forma tenemos un total de 355 UPMs, que debemos asignar a los estratos.

Dentro los requisitos deseables de la encuesta está el hecho de recabar mayor información del área rural, notemos, que si bien es cierto, el área rural nacional en el marco muestral, ocupa un 28%; dentro la muestra se asignara un 40% aproximado al área rural y el 60% restante al área urbano.

La técnica de asignación utilizada será la denominada “asignación de poder”, que se muestra a continuación:

$$n_h = n * \frac{N_h^\alpha}{\sum_h N_h^\alpha}; \quad \text{Con } \alpha \in [0,1]$$

Utilizando un $\alpha=0.5$ obtenemos la siguiente distribución en los ocho estratos considerados.

Cuadro N° 6, Distribución de la muestra con la asignación de poder por estratos

POR UPM		Estrato estadístico NBI				Total
		1	2	3	4	
Área Urbano Rural	Urbano	36	67	73	44	220
	Rural	5	15	41	74	135
Total		41	82	114	118	355
POR VIVIENDAS		Estrato estadístico NBI				Total
		1	2	3	4	
Área Urbano Rural	Urbano	432	804	876	528	2640
	Rural	60	180	492	888	1620
Total		492	984	1368	1416	4260

Realizando el sorteo tenemos la siguiente distribución por departamento, área de las UPMs y las viviendas.

Cuadro N°7, Distribución por departamentos y área, UPMs

		Área		Total
		Urbano	Rural	
DEPARTAMENTO	Chuquisaca	8	8	16
	La Paz	73	38	111
	Cochabamba	35	23	58
	Oruro	11	9	20
	Potosí	13	16	29
	Tarija	9	5	14
	Santa Cruz	61	27	88
	Beni	8	6	14
	Pando	2	3	5

		Área		Total
		Urbano	Rural	
DEPARTAMENTO	Chuquisaca	8	8	16
	La Paz	73	38	111
	Cochabamba	35	23	58
	Oruro	11	9	20
	Potosí	13	16	29
	Tarija	9	5	14
	Santa Cruz	61	27	88
	Beni	8	6	14
	Pando	2	3	5
Total		220	135	355

Cuadro N° 8, Distribución por departamentos y área, Viviendas

		Área		Total
		Urbano	Rural	
DEPARTAMENTO	Chuquisaca	96	96	192
	La Paz	876	456	1332
	Cochabamba	420	276	696
	Oruro	132	108	240
	Potosí	156	192	348
	Tarija	108	60	168
	Santa Cruz	732	324	1056
	Beni	96	72	168
	Pando	24	36	60
Total		2640	1620	4260

4.8 Probabilidades de selección

El muestreo en cuanto a la selección, utiliza un muestreo PPT (Proporcional al tamaño) en sus tres primeras etapas, y en la última etapa una selección sistemática simple con arranque aleatorio.

4.8.1 Probabilidad de primera etapa

La selección de la i – ésima UPM se realizó mediante probabilidad proporcional al tamaño (PPT), medido el tamaño por el número de viviendas particulares a la fecha del Censo de Población y Vivienda de 2001 que se encuentra dentro el marco maestro, entonces la probabilidad de selección de la i – ésima UPM en el estrato h , se denota por $P(UPM_{hi})$ está dada por la siguiente expresión matemática:

$$P(UPM_{hi}) = n_h \frac{M_{hi}}{M_h}$$

Donde:

n_h = Tamaño de muestra de primera etapa para el estrato h .

M_{hi} = Numero de viviendas en la i -ésima UPM del estrato h .

M_h = Numero de viviendas en el estrato h .

4.8.2 Probabilidad de segunda etapa

En segunda etapa $P(USM_{hj/i}) = P_{j/i}$ denota la probabilidad condicional de selección del j – ésimo sector censal en la i – ésima UPM, dado que la i – ésima UPM ha sido seleccionado en la primera etapa. Entonces, la ecuación matemática de la probabilidad de selección $P(USM_{hij}) = P_{ij}$ de la j – ésimo sector censal en la i – ésima UPM estrato h es como sigue:

$$P(USM_{hij}) = P_{ij} = P_i * P_{j/i} = n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}}$$

Donde:

n_{hi} = Tamaño de muestra de segunda etapa dentro la i -ésima UPM del estrato h .

M_{hij} = Numero de viviendas en la j -ésima USM de la i -ésima UPM del estrato h .

4.8.3 Probabilidad de tercera etapa

En el área rural $P(UTM_{hk/i,j}) = P_{k/i,j}$ denota la probabilidad condicional de selección del k – ésimo segmento en el j – ésimo sector censal en la i – ésima UPM en el estrato h , dado que la j – ésimo USM y la i – ésima UPM han sido seleccionado en las primeras dos etapas. Entonces, la ecuación matemática de la probabilidad de selección $P(UTM_{hijk}) = P_{ijk}$ del k – ésimo segmento en el j – ésimo sector censal en la i – ésima UPM en el estrato h es como sigue:

$$P(UTM_{hijk}) = P_i * P_{j/i} * P_{k/i,j} = n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}} n_{hij} \frac{M_{hijk}}{M_{hij}}$$

Donde:

n_{hij} = Tamaño de muestra de tercera etapa dentro la j-esima USM de la i-esima UPM del estrato h.

M_{hijk} = Numero de viviendas en la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h.

4.8.4 Probabilidad de cuarta etapa

Sea $P(USM_{hl/ijk}) = P_{l/ijk}$ denota la probabilidad condicional de selección de la l-esima vivienda dado que han salido el k - ésimo segmento en el j - ésimo sector de la i - ésima UPM dentro el estrato h, en las tres primeras etapas. Entonces, la probabilidad conjunta de selección de la l-esima vivienda del k - ésimo segmento en el j - ésimo sector de la i - ésima UPM dentro del estrato h es $P(USM_{hijkl}) = P_{ijkl}$ y está dado por siguiente expresión matemática:

$$P(UUM_{hijkl}) = P_i * P_{j/i} * P_{k/i,j} * P_{l/i,j,k} = n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}} n_{hij} \frac{M_{hijk}}{M_{hij}} \frac{n_{hijk}^L}{M_{hijk}^L}$$

Donde:

n_{hijk} = Tamaño de muestra de cuarta etapa dentro la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h.

M_{hijk}^L = Numero de viviendas en la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h, obtenido en el listado de viviendas.

4.9 Actualización cartográfica de las unidades de muestreo

En las unidades muestrales superiores a la UUM (sector area urbana, segmento área rural) seleccionadas se realizará la actividad de la actualización cartográfica antes de la realización de la encuesta, para actualizar el número de viviendas, con la finalidad de detectar los cambios experimentados entre el 5 de septiembre de 2001 y el período de la encuesta y determinar, así, las viviendas particulares ocupadas al momento de la encuesta, las que constituyeran el marco muestral

actualizado para seleccionar las viviendas, a esta Etapa la denominamos el “listado de viviendas” que posteriormente será el insumo para la utilización de la tabla de selección del muestreo sistemático, que se utiliza en última etapa.

4.10 Factores de expansión

El factor de expansión es la inversa de la probabilidad de selección de las unidades muestrales en sus diferentes etapas, y es el componente trascendental que permite generar el proceso de la inferencia estadística, y el insumo mas importante para la construcción de los estimadores y sus errores de muestreo asociados, la interpretación de este valor indica la magnitud de representación de la unidades respecto a la población objetivo de estudio, en nuestro caso al ser la unidad ultima de muestreo la vivienda, el factor de expansión nos indicara, cuántas viviendas representa dentro del territorio nacional una viviendas seleccionada en la muestra.

4.10.1 Factor de expansión base

Como ya se menciona el factor de expansión es la inversa de la probabilidad, de esta manera formamos los factores de expansión base, que los denominamos de esta manera, puesto que sobre estos realizaremos posteriormente ajustes debidos a situaciones que se presentan el operativo de campo, que escapan al control previo de un diseño muestral.

Factor de expansión base:

$$W_{viviendas} = W_{hijkl} = \frac{1}{P(UUM_{hijkl})} = \frac{1}{P_i * P_{j/i} * P_{k/i,j} * P_{l/i,j,k}} = \left(n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}} n_{hij} \frac{M_{hijk}}{M_{hij}} \frac{n_{hijk}}{M_{hijk}^L} \right)^{-1}$$

Donde

n_h = Tamaño de muestra de primera etapa para el estrato h.

M_{hi} = Numero de viviendas en la i-ésima UPM del estrato h.

M_h = Numero de viviendas en el estrato h.

n_{hi} = Tamaño de muestra de segunda etapa dentro la i-ésima UPM del estrato h.

M_{hij} = Numero de viviendas en la j-ésima USM de la i-ésima UPM del estrato h.

n_{hij} = Tamaño de muestra de tercera etapa dentro la j-esima USM de la i-esima UPM del estrato h.

M_{hijk} = Numero de viviendas en la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h.

n_{hijk} = Tamaño de muestra de cuarta etapa dentro la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h.

M_{hijk}^L = Numero de viviendas en la k-esima UTM de la j-esima USM de la i-esima UPM del estrato h, obtenido en el listado de viviendas.

La anterior expresión es el factor de expansión para cada vivienda dentro la muestra, puesto que la vivienda es la unidad ultima de muestreo y la de la que nos interesa la expansión, además nos interesa una expansión a nivel de personas dentro de cada hogar, ahora bien, debido a que dentro del hogar se realiza un relevamiento completo de información a todos los miembros, el factor de expansión para las personas dentro del hogar no sufre alteraciones respecto al presentado anteriormente.

Ahora presentaremos los ajustes que se realizaran a los factores de expansión base, estos ajustes responden a situaciones presentadas en el operativo de campo y debidas a la utilización de un marco muestral anticuado.

4.10.2 Ajustes a los factores de expansión

Debido a la no respuesta

Ocorre la no - respuesta cuando las viviendas u otras unidades de observación que se han seleccionado para la inclusión en una encuesta no reportan todos o algunos de los datos. Esta pérdida para obtener los resultados completos de todas las unidades seleccionados que puede levantarse de diferentes fuentes, mientras dependiendo en la situación de la encuesta. Hay dos tipos de no - respuesta: ítem de no - respuesta y la unidad de no - respuesta completa. La unidad de no - respuesta completa se refiere a la perdida de recolectar cualquier dato en la encuesta de una unidad muestral. Por ejemplo, no pueden obtenerse los datos de una vivienda elegible en la encuesta debido a la ausencia del informante, los caminos intransitables, la negativa para participar en la entrevista, o indisponibilidad del informante por otras razones. Este tipo de no - respuesta en la

Encuesta de Hogares se llama Tipo A de no - entrevista. Históricamente, entre 4 y 5 por ciento de las unidades elegibles en un año dado de Tipo A de no - entrevistadas. Recientemente, la proporción de tipo A ha subido a entre 6 y 7 por ciento. El ítem de no - respuesta ocurre cuando una unidad encuestado tiene inexactitudes o se niega a proporcionar algunos ítems (preguntas) específicos de información.

Cuadro N° 9, Grupos de no respuesta

Grupo A	Grupo B
Temporalmente Ausente	Entrevista Incompleta
Informante no Calificado	
Falta de contacto	
Rechazo	
Vivienda Desocupada	

La corrección a la no respuesta la realizamos sobre el factor de expansión base, considerando el número de viviendas planificadas y el numero de viviendas que no ingresan en ninguno de los grupos descritos en el cuadro N° 9

$$W_{hijkl} = \left(n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}} n_{hij} \frac{M_{hijk}}{M_{hij}} \frac{n_{hijk}}{M_{hijk}^L} \right)^{-1} * \frac{n_{hijk}}{n_{hijk}^r}$$

n_{hijk} = Numero de viviendas planificadas

n_{hijk}^r = Numero de viviendas ejecutadas efectivamente (reales)

Debido a la utilización de un marco anticuado

La utilización de un marco muestral que data del Censo Nacional de Población y Viviendas del año 2001, hace necesario la utilización de un ajuste que calibre dos aspectos, el primero, que se compense la antigüedad del marco del cual fue seleccionado la muestra, y segundo que se utilice la información trabajada a partir de los censos de los diferentes años y las encuestas que se realizan, este último se lo realiza mediante las proyecciones de población que realiza la institución.

De esta manera utilizamos a las proyecciones de población, y los factores obtenidos por muestreo en la encuesta, para obtener un factor de expansión que contiene dentro de si, información pertinente que da una mayor perspectiva de la información. Este ajuste lo realizamos a un nivel que nos permita las proyecciones de población que tiene la

institución, en este sentido se utiliza las proyecciones de población departamentales y área urbana rural de la siguiente forma:

$$W_{hijkl} = \left(n_h \frac{M_{hi}}{M_h} n_{hi} \frac{M_{hij}}{M_{hi}} n_{hij} \frac{M_{hijk}}{M_{hij}} \frac{n_{hijk}}{M_{hijk}^L} \right)^{-1} * \frac{n_{hijk}}{n^r_{hijk}} * \frac{P_d^{proy}}{P_d^{Enc}}$$

Donde

P_d^{proy} = Población proyectada de la institución, para el momento de la encuesta en el departamento d

P_d^{Enc} = Población proyectada debida a la encuesta en el departamento d, obtenida mediante F_{hijkl} .

4.11 ESTIMACIONES

Un estimador se define como una cantidad calculada en base a las observaciones muestrales de una o más variables de interés, con la finalidad de realizar algunas inferencias de la población objetivo.

4.11.1 Determinación de los estimadores de un total poblacional

a) Estimador del total poblacional (\hat{Y}_h)

Ahora surge el problema de analizar la forma de los estimadores puntuales óptimos $\hat{\theta} = \hat{\theta}(Y_1, Y_2, Y_3, \dots, Y_n)$ para los parámetros poblacionales típicos. Resulta que las mejores propiedades suelen presentarlas los estimadores lineales insesgados de la forma:

$$\hat{\theta} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

Estimador del total de estrato h, tiene la siguiente expresión matemática:

$$\hat{Y}_h = \frac{\sum_{i=1}^{n_h} W''_{hij} y_{hij}}{\sum_{i=1}^{n_h} W''_{hij}}$$

Por ejemplo, la edad promedio de la población en la Encuesta de Hogares 2007, es igual a:

$$\hat{Y}_h = \frac{\sum_{i=1}^{n_h} W''_{hij} y_{hij}}{\sum_{i=1}^{n_h} W''_{hij}} = \frac{262399926}{9902633} = 26.49799563$$

Por ejemplo, el porcentaje de hombres de la población en la Encuesta de Hogares 2007, es igual a:

$$\hat{P}_h = \frac{\sum_{i=1}^{n_h} W''_{hij} y_{hij}}{\sum_{i=1}^{n_h} W''_{hij}} = \frac{4786686}{9902633} = 0.483375078 \text{ (48,34\%)}$$

4.12 Cálculo de errores muestrales

Una de las medidas más importantes de precisión en las encuestas por muestreo es el error muestral, un indicador de variabilidad introducido al seleccionar una muestra en lugar de enumerar en toda la población, asumiendo que la información recolectado en la encuesta satisface con la propiedad de insesgado. Para cualquier encuesta dada, un error muestral de un estimador puede ser evaluado y utilizado para indicar la exactitud o acuracidad de los estimadores. Para diseños muestrales de encuestas de hogares que a menudo involucran estratificación, conglomeración, y la probabilidad de selección desigual, las formas de estos estimadores son a menudo complejos y muy difíciles evaluar. El cálculo de errores muestrales para los datos de encuestas de hogares requiere procedimientos que tengan en cuenta

la complejidad del diseño muestral de los datos generados, y el empleo apropiado de un software computacional.

Para el cálculo de errores muestrales existen habitualmente ocho paquetes estadísticos: CENVAR, Epi Info, PC CARP, SAS, STATA, SUDAAN, SPSS y WesVarPC. La mayoría de estos paquetes utilizan las aproximaciones de serie de Taylor para determinar las estimaciones de los errores muestrales. En este marco, el cálculo de errores muestrales del programa se utilizará el módulo Muestras Complejas de SPSS versión 13 o más Statistical Package for Social Science produce una tabla de salida o reporte con los indicadores:

- a) El valor estimado del parámetro (Valor estimado de la variable o indicador estadístico)
- b) Error estándar (Es el error muestral expresado en unidades de la variable que se está analizando)
- c) Coeficiente de variación (Es el error muestral expresado en términos relativos y mide los niveles de precisión de las estimaciones de los parámetros).
- d) Intervalo de confianza con un nivel de confianza del 95% (Intervalos con un nivel de confianza establecido (95%), incluyen el valor poblacional).
- e) El efecto del diseño (Deff) (Mide la eficiencia del diseño muestral multietápico, con respecto a un diseño muestral aleatorio simple).
- f) Número de observaciones

Las estimaciones de las variables de interés obtenidas de la muestra pueden tener un alta o poca confiabilidad estadística. Es decir, pueden tener una muy buena precisión o servir solamente como información referencial, tal como muestra la tabla siguiente:

CV: Coeficiente de variación	Precisión obtenida
Menor o igual 0.05	Muy buena
0.06 – 0.10	Buena
0.11 – 0.16	Aceptable
0.17 - 0.20	Aceptable (usar con precaución)

Más de 0.20	Sólo referencial
-------------	------------------

4.12.1 Estimación de la varianza

Algunas encuestas por muestreo utilizan un mecanismo sistemático para la selección de las unidades primarias de muestreo, lo cual complica la estimación de la varianza. En efecto, no existe un estimador insesgado de la varianza bajo muestreo sistemático y no se puede medir la variabilidad muestral de los estimadores puntuales presentados en la sección anterior -ya que en efecto se realiza una sola selección aleatoria por estrato y se necesitan como mínimo dos selecciones aleatorias para estimar la varianza-. En este caso, se dice que la varianza no es medible y no se puede calcular únicamente a partir de los datos de la muestra.

Existen dos enfoques para tratar el problema: (1) utilizar modelos para la estimación de la varianza, y (2) utilizar múltiples selecciones aleatorias. En esta sección presentamos brevemente ambas alternativas. La decisión sobre que enfoque utilizar dependerá de los supuestos que asuma el usuario y de la capacidad computacional disponible.

4.12.2 El método de serie de Taylor (Linearización)

La expansión de serie de Taylor ha sido utilizada en una variedad de situaciones en matemática y estadística. La puesta en práctica de la expansión de serie fue obtener una aproximación del valor de las funciones que son a calcular, por ejemplo, la e^x exponencial o la función de logarítmica $[\log(x)]$. Esta aplicación en los anteriores años era las calculadoras tuvieran teclas de función especiales y caso contrario se tenía acceso a las tablas apropiadas. La expansión de serie de Taylor para e^x supone tomar las derivadas de primer orden y de orden más alto de e^x con respecto a x ; evaluando las derivadas para algún valor, generalmente cero; y construir una series de términos sobre la base de las derivadas. La expansión para e^x es:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots + \frac{x^n}{n!} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Este es una aplicación específica de la siguiente formula general expandida:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + \frac{f^{iv}(a)(x-a)^4}{4!} + \dots + \frac{f^n(a)(x-a)^n}{n!} + \frac{f^{n+1}(a)(x-a)^{n+1}}{(n+1)!}$$

En la estadística, la serie de Taylor se utiliza para obtener una aproximación para alguna función no – lineal, y entonces la varianza de la función está basada sobre la aproximación de serie de Taylor de la función. A menudo, la aproximación proporciona un cálculo aproximado razonable de la función, y a veces la aproximación es incluso una función lineal. Esta idea de la estimación de la varianza tiene diferentes nombres en la literatura, incluyendo el método de linearización, el método de delta (Kalton, 1983, Pág. 44), y la propagación de la varianza (Kish, 1965, Pág. 583).

En la estadística aplicada, la expansión es evaluado en la media o el valor esperado de x , escrito como $E(x)$. Si se utilizará $E(x)$ para una expansión en el anterior fórmula general, se tiene

$$f(x) = f(E[x]) + f'(E[x])(x-E[x]) + \frac{f''(E[x])(x-E[x])^2}{2!} + \frac{f'''(E[x])(x-E[x])^3}{3!} + \frac{f^{iv}(E[x])(x-E[x])^4}{4!} + \dots + \frac{f^n(E[x])(x-E[x])^n}{n!} + \frac{f^{n+1}(E[x])(x-E[x])^{n+1}}{(n+1)!}$$

La varianza de $f(x)$ es $V[f(x)] = E[f^2(x)] - E^2[f(x)]$ por definición y utilizando la expansión de serie de Taylor, se tiene

$$V[f(x)] = V[f(E[x])] + \{f'(E[x])\}^2 V(x) + \dots = 0 + \{f'(E[x])\}^2 V(x) + \dots = \{f'(E[x])\}^2 V(x) + \dots$$

Las mismas ideas se manejan para las funciones de más de una variable aleatoria. En el caso de una función de dos varianzas, la expansión de serie de Taylor está dada

$$V[f(x_1, x_2)] \cong \left(\frac{\partial f(x_1, x_2)}{\partial x_1} \right) \left(\frac{\partial f(x_1, x_2)}{\partial x_2} \right) \text{cov}(x_1, x_2)$$

Aplicando para la anterior ecuación para el cociente de dos variables x e y o vale decir $r = \frac{y}{x}$, obteniendo la fórmula del estimador de la varianza del cociente (razón)

$$V(r) = \frac{V(y) - 2r \text{cov}(x, y) + r^2 V(x)}{x^2} + \dots = r^2 \left(\frac{V(y)}{y^2} + \frac{V(x)}{x^2} - 2 \frac{\text{cov}(x, y)}{xy} \right) + \dots$$

Generalizando la ecuación $V[f(x_1, x_2)] \cong \left(\frac{\partial f(x_1, x_2)}{\partial x_1} \right) \left(\frac{\partial f(x_1, x_2)}{\partial x_2} \right) \text{cov}(x_1, x_2)$ para el caso de c variables aleatorias, la varianza aproximada de $\theta = f(x_1, x_2, x_3, \dots, x_c)$ es

$$V(\theta) \cong \sum \sum \left(\frac{\partial f}{\partial x_i} \right) \left(\frac{\partial f}{\partial x_j} \right) \text{cov}(x_i, x_j)$$

Aplicando la anterior ecuación para un estimador ponderado

$$f(y) = \hat{Y}_i = \sum w_i y_{ij}, j = 1, 2, 3, \dots, c$$

Involucrando las c variables en una muestra de n observaciones, Woodruff (1971) ha demostrado que

$$V(\theta) \cong V \left[\sum w_i \sum \left(\frac{\partial f}{\partial y_j} \right) y_{ij} \right]$$

Esta alternativa de la forma de la varianza linearizado de un estimador no lineal brinda las ventajas computacionales porque evita el cálculo de la matriz de covarianza de $c \times c$ en la ecuación $V(\theta) \cong \sum \sum \left(\frac{\partial f}{\partial x_i} \right) \left(\frac{\partial f}{\partial x_j} \right) \text{cov}(x_i, x_j)$. Esta conveniencia de convertir un problema de estimación multi-etápico en un problema univariante es realizado por un simple intercambio de sumatorias. Este procedimiento computacional general puede ser aplicado para un estimador

univariante no lineal, incluyendo coeficientes de regresión (Fuller, 1975; Tepping, 1968).

Para una encuesta compleja, este método de aproximación es aplicado para el total de UPMs dentro del estrato. Es decir, la estimación de la varianza es una combinación ponderada de la variación en la ecuación $V(\theta) \equiv V\left[\sum w_i \sum \left(\frac{\partial f}{\partial y_j}\right) y_{ij}\right]$ a través de las UPMs dentro del mismo estrato.

Estas fórmulas son complicadas pero pueden requerir mucho menor tiempo de computación que los métodos de replicación discutidos más arriba. Este método puede ser aplicado a cualquier estadística que está expresada matemáticamente - por ejemplo, la media o el coeficiente de regresión - pero no a tales estadísticas no funcionales (no centrada) como la mediana y otros percentiles denominados como medidas de tendencia no central.

En esta sección, se ha presentado diferentes métodos de estimación estadística para encuestas complejas (para la discusión adicional, ver Rust y Rao, [1996]). Los ejemplos de GSS y otras encuestas cuidan mostrar que el efecto del diseño es mayor a 1 en la mayoría de las encuestas complejas. Los ejemplos adicionales pueden ser encontrados en E. S. Lee, Forthofer, y Lorimor (1986) y Eltinge, Parsons, y Jang (1997). Ejemplos en el Capítulo 6 demostrarán la importancia de usar uno de los métodos examinados arriba en el análisis de los datos de encuesta complejas.

Utilizando la expansión de Taylor de una función $f(x, y)$ alrededor (x_0, y_0)

$$f(x, y) = f(x_0, y_0) + (x - x_0) \frac{\partial f(x, y)}{\partial x} \Big|_{(x_0, y_0)} + (y - y_0) \frac{\partial f(x, y)}{\partial y} \Big|_{(x_0, y_0)} + \dots$$

Donde los puntos suspensivos indican el término cuadrático y términos de orden superior la cual haremos la suavización para los propósitos de la aproximación, llamado Linearización de Taylor. En la expresión anterior substituyendo $x \rightarrow \bar{x}$, y

$\rightarrow \bar{y}, x_0 \rightarrow \bar{X}, y_0 \rightarrow \bar{Y}$, y se define $f(x, y) \equiv x_0 \frac{y}{x}$ así que $f(x_0, y_0) = y_0$, $\frac{\partial f(x, y)}{\partial x} = -x_0 \frac{y}{x^2}$

y $\frac{\partial f(x, y)}{\partial y} = \frac{x_0}{x}$ que dado por siguiente expresión matemática

$$\bar{Y}_R = \bar{Y} - (\bar{x} - \bar{X}) \bar{X} \frac{\bar{Y}}{\bar{X}^2} + (\bar{y} - \bar{Y}) \frac{\bar{X}}{\bar{X}} + \dots \quad o \quad \bar{Y}_R - \bar{Y} = \frac{1}{\bar{X}} \left[-\frac{\bar{x}\bar{X}\bar{Y}}{\bar{X}} + \frac{\bar{X}\bar{X}\bar{Y}}{\bar{X}} + \bar{y}\bar{X} - \bar{X}\bar{Y} \right] + \dots \quad o$$

$$\bar{Y}_R - \bar{Y} \approx \frac{1}{\bar{X}} \left[-\bar{x}\bar{Y} + \bar{X}\bar{Y} + \bar{y}\bar{X} - \bar{X}\bar{Y} \right] \approx \frac{(\bar{y}\bar{X} - \bar{X}\bar{Y})}{\bar{X}} \approx \bar{y} - \frac{\bar{Y}\bar{X}}{\bar{X}} \approx \bar{y} - R\bar{X} \approx \bar{y} - R\bar{x}$$

Por lo tanto, $E(\bar{Y}_R) - \bar{Y} \approx E(\bar{y} - R\bar{x}) = \bar{Y} - R\bar{X} = 0$, es aproximadamente insesgado. Puede también utilizarse la linealización para determinar una aproximación para la varianza:

$$V(\bar{Y}_R) \approx V(\bar{y} - R\bar{x}) = V(\bar{y}) - 2R \text{cov}(\bar{x}, \bar{y}) + R^2 V(\bar{x})$$

La cual da la fórmula de varianza requerida (segunda versión) usando el teorema 2.1 para \bar{x} y \bar{y} , respectivamente, y también proporciona la covarianza de las medias, para demostrar se utiliza el mismo método como el en la prueba del teorema 2.1 como sigue:

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{N-n}{N} \text{Cov}(x, y)$$

v_0 es el estimador de la varianza sustituyendo los momentos de la muestra por los momentos de la población pero la prueba de que él, al mismo tiempo que v_1 y v_2 (y efectivamente cualquier v_k para cualquier k fijo), es aproximadamente insesgado para la fórmula de la varianza (y por lo tanto también la varianza es verdadera) está más allá del alcance de este curso

Otras técnicas de replicación utilizadas son el Bootstrap, Jackknife y las Semi-Muestras Balanceadas (Balance Half Samples, BHS). Para una revisión de técnicas

de estimación de varianza ver Kovar, Rao y Wu (1988). Un resumen comparativo de técnicas de estimación de varianza se encuentra también en Rust (1985).

4.13 Encuesta Piloto

La aplicación de la encuesta piloto fue realizada a principios de la implementación de las encuestas por muestreo del Programa MECOVI, con el objeto de evaluar y subsanar los instrumentos de recolección de información (Cuestionarios, Manual de encuestadores, Supervisores, etc.) y la logística de trabajo de campo (determinación de carga de trabajo, recorrido, rutas críticas y no críticas (prioridades), distancias, el cálculo del presupuesto final, etc.).

5. EVALUACIÓN DE LA CALIDAD DE LOS DATOS

Los errores que afectan a toda encuesta pueden agruparse en dos grandes grupos:

Errores de muestreo, que se originan por la obtención de resultados sobre las características de una población, a partir de la información recogida en una muestra de la misma.

Errores ajenos al muestreo, que son comunes a toda investigación estadística, tanto si la información es recogida por muestreo como si se realiza un censo. Estos errores se presentan en cualquier fase del proceso estadístico:

- **Antes de la recolección de datos:** por deficiencias del marco muestral e insuficiencias en las definiciones y/ conceptos y cuestionarios.
- **Durante la recolección de datos:** por defectos en la labor de los entrevistadores y/o supervisores del trabajo de campo e incorrecta declaración por parte de los informantes.
- **Después de la recogida de los datos:** errores en la depuración, crítica-codificación, grabación, tabulación e impresión de los resultados.

6. BIBLIOGRAFÍA

1. Bushery, J. M., Dewey, J. A., and Weller, G. (1995), "Reinterview's Effect on Survey Response," Proceedings of the 1995 Annual Research Conference, U.S. Census Bureau, pp. 475-485.
2. Lohr, Sharon L, (2000), "Muestreo: Diseño y Análisis", International Thomson Editores.
3. U.S. Census Bureau (1996), Current Population Survey Office Manual, Form CPS-256, Washington, DC: Government Printing Office, chapter. 6.
4. U.S. Census Bureau (1996), Current Population Survey SFR Manual, Form CPS-251, Washington, DC: Government Printing Office, chapter 4.
5. U.S. Census Bureau (1993), "Falsification by Field Representatives 1982-1992," memorandum from Preston Jay Waite to Paula Schneider, May 10, 1993.
6. U.S. Census Bureau (1997), "Falsification Study Results for 1990-1997," memorandum from Preston Jay Waite to Richard L. Bitzer, May 8, 1997.
7. Cesar Pérez López, "Técnicas de muestreo Estadístico", Alfaomega grupo editor, S.A. (2000)
8. Sarndal, Swensson, Wretman, "Model Assisted Survey Sampling", Springer-Verlang.