DISEÑO MUESTRAL: CULTURA CÍVICA EN MÉXICO, 2009

Guillermo CUEVAS

SUMARIO: I. *Introducción*. II. *México en la década de 1950*. III. *La encuesta de 1959*. IV. *México en 2009*. V. *Fuentes de información*. VI. *Márgenes de error estadístico*. VII. *Referencias*.

I. INTRODUCCIÓN

En 1959 se realiza el primer estudio cuantitativo sobre cultura política del mundo (publicado por primera vez por Almond en 1963). Ese estudio, en una de sus vertientes, utiliza un instrumento cuantitativo, aplicado a una muestra probabilística de cinco países distintos, uno de ellos fue México. A cincuenta años de ese levantamiento, se ha replicado con la intención de encontrar similitudes y diferencias entre las sociedades de ambas épocas.

Debido a que en México no hay un precedente de la réplica, desde el punto de vista probabilístico fue un reto diseñar una muestra que emulara a la anterior, de tal forma que los resultados obtenidos del levantamiento fueran realmente comparables con los de 1959. Sería un exceso decir que se ha logrado la comparabilidad al 100%, sin embargo, creemos que se ha logrado mantener el espíritu de lo que la muestra original pretendía.

En los siguientes párrafos daremos cuenta de los pasos que implicó el diseño de la muestra. Empezando por un breve resumen de lo que se hizo en 1959, se hará una comparación entre el México de aquel entonces y el México de ahora. Paso seguido se detallan las fuentes de información que se utilizaron como referencia para la selección de las diferentes etapas de muestreo. Después, se describen los estratos, conglomerados y métodos de selección de las diferentes unidades de muestreo. A continuación se presenta un análisis sobre los márgenes de error estadístico asociados al levantamiento. Finalmente, se relatan las fórmulas para el cálculo de probabilidad de los individuos, así como su calibración, mismas que son de utilidad para la ponderación de la base de datos.

II. México en la década de 1950

De acuerdo con el VII Censo General de Población (1954), en 1950 México tenía 25,791,017 habitantes, de los cuales, el 30.4% vivía en localidades de más de 10,000 habitantes (7,843,489). El Distrito Federal tenía 3,049,561 habitantes (38.9% de la población urbana). Las otras dos ciudades más importantes del país, Guadalajara y Monterrey tenían 377,016 y 333,422 habitantes, respectivamente (9.1%). Adicionalmente, había otras siete ciudades con más de 100,000 habitantes (Veracruz, Juárez, León, Saltillo, San Luis Potosí, Mérida y Puebla), que en conjunto, tenían 955,428 personas (12.2%). Otras 132 localidades de 10,000 habitantes o más (de las cuales, 33 tenían entre 30,000 y 99,999 habitantes, alcanzando un total de 1,572,832 personas, el 20.1% de la población urbana), tenían 3,127,150 personas entre todas (39.9%), completando el total de 142 localidades de más de 10,000 habitantes que tenía el país al tiempo del levantamiento del censo.

III. La encuesta de 1959

El levantamiento realizado en 1959 incluyó como universo de estudio a todas las personas mayores de 21 años (que era la mayoría de edad) residentes en las zonas urbanas del país (el estudio consideró como localidad urbana a toda aquella localidad con más de 10,000 habitantes). Para garantizar una mejor dispersión geográfica, el país fue dividido en cinco estratos, a saber: la Ciudad de México; las ciudades de Guadalajara y Monterrey; las ciudades de 100,000 a 299,999 habitantes; las ciudades de 30,000 a 99,999 habitantes; y las ciudades de 10,000 a 29,999 habitantes. El tamaño de muestra alcanzada para cada uno de los segmentos entrevistados fue el siguiente:

1. Ciudad de México	191 entrevistas	(18.9%)
2. Guadalajara y Monterrey	117 entrevistas	(11.6%)
3. Ciudades de 100,000 a 299,999 habitantes	270 entrevistas	(26.8%)
4. Ciudades de 30,000 a 99,999 habitantes	230 entrevistas	(22.8%)
5. Ciudades de 10,000 a 29,999 habitantes	200 entrevistas	(19.8%)

Aunque el levantamiento se realizó nueve años después del censo de 1950, este último fue el utilizado como marco de referencia, pues no se contaba con datos actualizados, y aunque existían algunas aproximaciones, éstas no tuvieron en cuenta el disparo demográfico que se dio durante la década.

IV. México en 2009

Desde el levantamiento del VII Censo General de Población, la población del país se ha cuadruplicado, así, de acuerdo con los datos del II Conteo General de Población (2005), en México habitan 103,263,388 personas, de las cuales, el 65.1% (67,187,878) lo hace en localidades de 10,000 habitantes o más. El Distrito Federal cuenta con 8,720,916 personas, mientras que los municipios de Guadalajara y Monterrey ahora albergan a 1,600,894 y 1,133,070 residentes, respectivamente (sin considerar sus zonas metropolitanas). Otros datos relevantes, que fueron considerados al diseñar la muestra, son los siguientes:

La localidad de más de 100,000 habitantes en 1950 que tiene menos habitantes en la actualidad es Veracruz, con 444,438 habitantes.

Las localidades de 10,000 a 399,999 habitantes ahora suman 31,559,084 personas (el 47.0% de la población urbana).

Puesto que en la actualidad rara vez se separa a las localidades de los estratos 4 y 5 contemplados en el estudio de 1959, para el levantamiento actual se han determinado cuatro estratos que repliquen la muestra anterior:

1. Distrito Federal.

2. Guadalajara y Monterrey (incluyendo su zona metropolitana).

3. Localidades de 400,000 habitantes o más.

4. Localidades de 10,000 a 399,999 habitantes.

Además, el levantamiento actual contempla un quinto estrato poblacional:

5. Localidades de menos de 10,000 habitantes.

En el siguiente cuadro (véase cuadro 1) se resumen las diferencias y similitudes entre ambos levantamientos, tanto en su alcance sociodemográfico como en el tamaño de muestra alcanzado en ambos casos (en los siguientes párrafos se dará una explicación detallada sobre el número de muestra en cada estrato).

CUADRO 1*

SIMILITUDES Y DIFERENCIAS ENTRE LOS LEVANTAMIENTOS DE 1959 Y 2009

Muestra 1959			Muestra 2009		
Estrato	*Descripción*	*Tamaño de muestra*	*Estrato*	*Descripción*	*Tamaño de muestra*
1	Ciudad de México	191	1	Distrito Federal	216
2	Guadalajara y Monterrrey	117	2	Guadalajara y Monterrrey	272
3	Ciudades de 100,000 a 299,999 habitantes	270	3	Ciudades de 400,000 habitantes o más	224
4	Ciudades de 30,000 a 99,999 habitantes	230	4	Ciudades de 10,000 a 399,999 habitantes	612
5	Ciudades de 10,000 a 29,999 habitantes	200			
—	—	—	5	Localidades de menos de 10,000 habitantes	504

* Diseño muestral, cultura cívica en México, 2009.

** Fuente: Almond y Verba (1963).

V. FUENTES DE INFORMACIÓN

1. *Marco muestral*

Se utilizó la información oficial disponible, brindada por el Instituto Nacional de Geografía y Estadística (INEGI), y actualizada al 2005 (de acuerdo con el II Conteo Nacional de Población). En los casos en los que no se pudo contar con ese marco de referencia, se construyó un marco *ad hoc* a cada situación.

Respecto a los datos obtenidos a partir del INEGI, se utilizó el sistema de información referenciada geoespacialmente integrada en un sistema (IRIS-

SCINCE), así como los datos de población desagregados a nivel de municipio o localidad.

2. *Población en estudio*

Se consideró a todos los habitantes mayores de edad (18 años) como población objetivo. A cada miembro seleccionado en la muestra se le aplicó un cuestionario de opinión que recopiló diversas opiniones y actitudes sobre cultura política.

3. *Regionalización*

Para fines de selección de la muestra, el país quedará dividido de la siguiente manera:

- El Distrito Federal. Las Ageb's[1] que conforman al Distrito Federal fueron estratificadas a su vez en niveles socioeconómicos alto, medio y bajo.
- La ciudad de Guadalajara. Las Ageb's de la ciudad también fueron estratificadas en tres niveles socioeconómicos alto, medio y bajo.
- La ciudad de Monterrey. Las Ageb's de la ciudad fueron estratificadas en tres niveles socioeconómicos, al igual que en el caso anterior.
- Otras localidades de más de 400,000 habitantes.
- Localidades de 10,000 a 399,999 habitantes.
- Localidades de menos de 10,000 habitantes. Las localidades fueron estratificadas en tres niveles de marginación: alto, medio y bajo.

4. *Esquema de selección*

La selección de las unidades de muestreo se realiza a través de etapas sucesivas, y de manera independiente para cada uno de los estratos definidos.

- De localidades y Ageb's:

  — En el Distrito Federal, las ciudades de Guadalajara y Monterrey: se seleccionaron directamente Ageb's, en el primer caso 27, y

[1] Área Geoestadística Básica. Unidad geográfica delimitada por el INEGI en la que las personas tienen características socioeconómicas parecidas.

18 en el resto. La selección se realizó después de dividir las Ageb's en tres niveles socioeconómicos,[2] a saber, alto, medio y bajo. De cada nivel socioeconómico se seleccionaron 9 y 6 ageb's, respectivamente, con probabilidad proporcional al número de habitantes de 18 años o más (PPT) en cada ageb, y con reemplazo.

— En otras localidades de más de 400,000 habitantes, se seleccionaron 5 localidades, con PPT, con reemplazo. En cada localidad seleccionada se ordenaron las Ageb's de mayor a menor nivel de marginación (conforme a la metodología explicada en el punto anterior), seleccionando 6 de ellas de manera sistemática, con arranque aleatorio.

— En las localidades de 10,000 a 399,999 habitantes, se seleccionaron 27 de ellas, con PPT y reemplazo. Se ordenaron sus Ageb's de acuerdo con su nivel de marginación y se seleccionaron 3 de forma sistemática, con arranque aleatorio.

— Las localidades de menos de 10,000 habitantes fueron divididas según el índice de marginación que Conapo les asignó, y se eligieron 7 con índice de marginación alto o muy alto; 7 con índice de marginación medio; y 7 con índice de marginación bajo o muy bajo. La selección fue PPT, con reemplazo. No hubo selección de Ageb's.

- De manzanas:

— En todas las localidades de 10,000 habitantes o más, adentro de cada Ageb seleccionada se eligieron 2 manzanas. La selección de las mismas fue por PPT, con reemplazo.

— En las localidades de menos de 10,000 habitantes se eligieron 6 manzanas. El método de selección fue sistemático, con arranque

[2] El método que se utilizó para dividir las Ageb's es similar al seguido por Conapo (Consejo Nacional de Población) para la construcción de los índices de marginación: se consideró el porcentaje de población analfabeta de 5 años y más, el porcentaje de población de 15 años y más sin primaria completa, el porcentaje de viviendas particulares sin drenaje ni excusado, el porcentaje de viviendas particulares sin agua entubada en el ámbito de la vivienda, el porcentaje de viviendas con algún nivel de hacinamiento, el porcentaje de viviendas particulares con piso de tierra, y el porcentaje de viviendas particulares sin refrigerador. Los datos fueron obtenidos por parte del SCINCE. Con los datos anteriores, se tomó como índice de marginación a la primera componente principal (que explica el 55% del total de la varianza), y se dividió el *ranking* en tres grupos, de acuerdo con el criterio de Dalenius.

aleatorio, realizando un conteo manual de manzanas. Cuando la localidad estaba conformada por caseríos dispersos (en parte o en su totalidad), se hizo un amanzanamiento virtual, considerando a todo el caserío disperso como una gran manzana, y seleccionando viviendas directamente. La cantidad de viviendas en muestra provenientes de esa gran manzana dependió de la proporción que representara de la localidad: cuando representaba más del 90%, se seleccionaron 24 viviendas; cuando representaba más del 50% se tomaron 16 viviendas y dos manzanas en la parte no dispersa; cuando representaba más del 10% se tomaron 8 viviendas y cuatro manzanas; cuando representaba menos del 10%, la población residente de esos lugares no fue considerada como parte de la muestra.

- De viviendas:

  — Para las localidades con más de 10,000 habitantes, se seleccionaron 4 viviendas por cada manzana seleccionada. El criterio se explica en el siguiente punto.
  — Para las localidades con menos de 10,000 habitantes, por cada manzana en muestra se eligieron 4 viviendas. La selección se realizó de manera sistemática, con arranque aleatorio, después de haber realizado un conteo de las mismas, por parte del personal de campo.
  — De individuos: Para cada vivienda en muestra se obtuvo la opinión de un individuo de 18 años cumplidos o más, el método de selección fue aleatorio simple, con base en una tabla de números aleatorios.
  — El cuadro 2 resume el tamaño de muestra para cada esquema de selección. Derivado de dicho esquema, se consideraron las siguientes unidades de muestreo:

CUADRO 2*

TAMAÑO DE MUESTRA POR ETAPA DE MUESTREO

Estrato	*Descripción*	*Localidades*	*Ageb's*	*Manzanas*	*Viviendas*	*Tamaño de muestra*
I	Distrito Federal	1	27	2	4	216
II	Guadalajara y Monterrey	2	18	2	4	288
III	Ciudades de 400,000 habitantes o más	5	6	2	4	240
IV	Ciudades de 10,000 a 399,999 habitantes	27	3	2	4	648
V	Localidades de menos de 10,000 habitantes	21	—	6	4	504

* Cultura Cívica en México, 2009. Diseño Muestral.

a) Unidades Primarias de Muestreo (UPM): en los estratos I y II son las Ageb's, mientras que en el resto de los estratos, fueron las localidades.
b) Unidades Secundarias de Muestreo (USM): en los estratos I, II y V son las manzanas, y en los otros dos estratos fueron las Ageb's.
c) Unidades Terciarias de Muestreo (UTM): en los estratos I, II y V son las viviendas, d) Unidades Cuaternarias de Muestreo (UCM): para los estratos I, II y V son los individuos seleccionados, siendo ésta la última unidad de muestreo. Para los estratos III y IV fueron las viviendas.
e) Unidades Últimas de Muestreo (UUM): en el caso de los estratos III y IV, las unidades quintas y últimas de muestreo fueron los individuos de cada vivienda seleccionada.

A. *Chiapas*

Cuando ya había iniciado el levantamiento de campo se decidió aumentar el tamaño de muestra en esta entidad federativa para poder alcanzar resultados estadísticamente representativos del estado. Para ello, se realizó un diseño muestral paralelo que no interfiriera con el nacional, y que al final lograra integrarse con el resto de los resultados, al obtener 416 casos adicionales. El muestreo se realizó de la siguiente manera:

- Las localidades de la entidad fueron divididas en tres estratos: de menos de 10,000 habitantes; de 10,000 a 99,999 habitantes; de 100,000

habitantes o más. De cada estrato se extrajeron 6 localidades con PPT, con reemplazo.

- En cada localidad seleccionada se eligieron 3 Ageb's con PPT, con reemplazo. Cuando la localidad no contaba con Ageb's, no fueron seleccionadas.
- En cada Ageb seleccionada se tomaron 4 manzanas de manera sistemática y con arranque aleatorio. Cuando no se seleccionaron Ageb's, se tomaron 5 manzanas, siguiendo el criterio de selección de manzanas en caseríos dispersos explicado con anterioridad.
- En cada manzana seleccionada se consideraron 2 viviendas en muestra, y cuando se trató de caseríos dispersos, por cada manzana se consideraron 4 viviendas. La selección fue sistemática, con arranque aleatorio, y siguiendo el criterio de selección explicado anteriormente.
- En cada vivienda seleccionada, a través de una tabla de números aleatorios se eligió a un individuo mayor de 18 años para que contestara el instrumento.

VI. Márgenes de error estadístico

1. *Tamaño de muestra*

El tamaño de muestra obtenido del estudio fue de 1,896 casos (2,312 si se consideran los casos adicionales de Chiapas. Ese tamaño de muestra, como ya se ha explicado, fue distribuido a diferentes estratos del país, lo cual redunda en la disminución de los márgenes de error estadísticos [para ver en detalle el desglose de la muestra, por etapa de muestreo, véase cuadro 2]).

2. *Estimaciones teóricas de los márgenes de error*

La siguiente fórmula se utilizó para la determinación del tamaño de muestra:

$$n = \frac{(z^2_{1-\frac{\alpha}{2}})(p)(1-p)(Deff)}{s^2(1-TNR)}$$

Donde:

n = Tamaño de muestra considerando que la población a estimar es infinita.

$z^2_{1-\frac{\alpha}{2}}$ = Número con el que se obtiene una probabilidad de éxito superior al (1-α)*100%, suponiendo un modelo de probabilidad normal estándar (en este caso, elevado al cuadrado).

p = Probabilidad de éxito del evento. Se refiere a la probabilidad de éxito esperada. En encuestas complejas, lo mejor es tomar valores de p conservadores. El más conservador de los valores se obtiene cuando $p = 0.5$.

s^2 = Varianza estimada. Nos indica cuánto variarán nuestros resultados, dada una certeza. Es decir, una vez obtenido algún resultado, cuánto se puede mover de acuerdo con la precisión deseada.

TNR = Tasa de no respuesta esperada.

Deff = Efecto de diseño por utilizar un muestreo diferente del muestreo aleatorio simple.

Se consideró una tasa de no respuesta inferior al 5%, y un efecto de diseño de 1.75, el tamaño de muestra. Así, se determinó el tamaño de muestra que lograra estimaciones con márgenes de error máximo de +/- tres puntos porcentuales, con un nivel de confianza del 94%.

3. *Estimaciones muestrales de los márgenes de error*

A partir de los datos recabados, utilizando el módulo de muestras complejas del paquete SPSS 14, el cual considera los estratos (utilizados en la post estratificación y que se detallan más adelante) y las unidades primarias de muestreo, se obtuvieron los siguientes resultados:

1. Considerando las frecuencias simples de cada una de las preguntas del cuestionario, en el 50% de los casos, la amplitud de los márgenes de error asociados (al 95% de confianza) fue menor a tres puntos porcentuales. En el 86% de los casos, dicha amplitud fue menor a seis puntos porcentuales (validando el margen de error máximo de +/- tres puntos porcentuales). Mientras que en el 95% de los casos, cada opción de respuesta de las variables tuvo un margen de error de +/- cuatro puntos porcentuales.
2. Si se estandariza la medida anterior dividiéndola entre el estimador puntual obtenido, en el 76% de las ocasiones, la medida fue inferior a uno (es decir, la diferencia entre el límite superior y el inferior fue menor al estimador puntual). Los casos en que esta medida fue superior a la unidad se debieron a frecuencias cuyo estimador puntual fue pequeño, así, si se des-

cartan los estimadores puntuales inferiores a 0.05, la medida es inferior a la unidad en el 98% de los casos.

3. El promedio del efecto de diseño para las tablas de frecuencia fue de 2.016.

4. El promedio del efecto de diseño para las variables de las que se obtuvieron promedios fue de 1.33.

4. *Cálculo de los ponderadores*

En esta sección se incluye la forma en que fueron calculados los ponderadores utilizados para el cálculo de los estimadores. Para ello, se analizarán las probabilidades de selección de cada una de las unidades de muestreo, conforme a lo establecido en los apartados anteriores.

5. *Probabilidad de selección de la localidad*

Sea X_j una localidad perteneciente a cualquiera de los cinco estratos (I, II, III, IV, V). Entonces, la probabilidad de selección de la localidad es:

$$P\left[X_j\right]=\begin{cases}1, j=\{I, II\} \\ \frac{PT18\left(x_j\right)}{\sum_{x_j} PT18\left(x_j\right)}, j=\{III, IV\} \\ \frac{PT18\left(x_V^i\right)}{\sum_{x_V^i} PT18\left(x_V^i\right)}, i=\{1,2,3\}, j=\{V\}\end{cases}$$

Donde, *PT*18 es la población total de 18 años o más en una localidad *x* del estrato *j*. A su vez, *i* representa el índice de marginación de la localidad, donde 1 es alto o muy alto, 2 es medio, y 3 es bajo o muy bajo (nótese que únicamente es utilizado en el estrato V, pero se puede generalizar al resto de las localidades con la letra *î*, que indica la pertenencia indistinta a cualquier índice de marginación). Por último, se menciona que ésta es la probabilidad de selección de una vez, sin embargo, en los casos donde se hizo selección con reemplazo, se debe considerar que cada localidad pudo salir en más de una ocasión.

6. *Probabilidad de selección de la Ageb*

Sea $X^{i}_{j,k}$ la k – ésima ageb de una localidad perteneciente al j – ésimo estrato, con índice de marginación i. Entonces, la probabilidad de selección de la ageb es:

$$P\left[X^{i}_{j,k}\right]=\begin{cases}\frac{PT18\left(\hat{x}^{j,r}_{j,k}\right)}{\sum_{\hat{x}^{j,r}_{j,k}} PT18\left(\hat{x}^{j,r}_{j,k}\right)}, r=\{1,2,3\}\ j=\{I,II\}\\ \frac{6}{\#AGEB^{\hat{i}}_{x_j}} P\left[x^{\hat{i}}_{j}\right], j=\{III\}\\ \frac{3}{\#AGEB^{\hat{i}}_{x_j}} P\left[x^{\hat{i}}_{j}\right], j=\{IV\}\\ P\left[x^{i}_{j}\right], i=\{1,2,3\}, j=\{V\}\end{cases}$$

Donde $PT18\left(x^{\hat{i},r}_{j,k}\right)$ es la población total de personas de 18 años o más en la k – ésima ageb del j – ésimo estrato, pertenecientes al nivel de marginación r. Y $\#AGEB^{\hat{i}}_{x_j}$ es el total de ageb's de la localidad x_j.

7. *Probabilidad de selección de la manzana*

Sea $X^{i,r}_{j,k,l}$ la l – ésima manzana de la k – ésima ageb con índice de marginación r (de la ageb), de una localidad perteneciente al j – ésimo estrato, con índice de marginación i. Entonces, la probabilidad de selección de la manzana es:

$$P\left[X^{i,r}_{j,k,l}\right]=\begin{cases}\frac{PT18\left(\hat{x}^{\hat{i},r}_{j,k,l}\right)}{\sum_{\hat{x}^{\hat{i},r}_{j,k,l}} PT18\left(\hat{x}^{\hat{i},r}_{j,k,l}\right)} P\left[x^{\hat{i},r}_{j,k}\right], r=\{1,2,3\}\ j=\{I,II\}\\ \frac{PT18\left(\hat{x}^{\hat{i},\hat{r}}_{j,k,l}\right)}{\sum_{\hat{x}^{\hat{i},\hat{r}}_{j,k,l}} PT18\left(\hat{x}^{\hat{i},\hat{r}}_{j,k,l}\right)} P\left[x^{\hat{i},\hat{r}}_{j,k}\right], j=\{III,IV\}\\ \frac{6}{\#MZ^{i,\hat{r}}_{x_{j,k}}} P\left[x^{i,\hat{r}}_{j,k}\right], i=\{1,2,3\}, j=\{V\}\end{cases}$$

Donde $PT18\left(x^{\hat{i},r}_{j,k,l}\right)$ es la población total de personas de 18 años o más en la l – ésima manzana de la k – ésima ageb (con nivel de marginación r)

del *j* – ésimo estrato. A su vez, $\#MZ^{i,\hat{r}}_{x_{j,k}}$ es el total de manzanas de la ageb (localidad) x_j con índice de marginación *i*. Ésta es la probabilidad de selección en una ocasión, sin embargo, en los casos donde se hizo selección con reemplazo, se debe considerar que cada localidad pudo salir en más de una ocasión.

8. *Probabilidad de selección de la vivienda*

Sea $X^{i,r}_{j,k,l,m}$ la *m* – ésima vivienda de la *l* – ésima manzana de la *k* – ésima ageb (con índice de marginación *r)*, de una localidad perteneciente al *j* – ésimo estrato, (con índice de marginación *i)*. Entonces, la probabilidad de selección de la vivienda es:

$$P\left[X^{i,r}_{j,k,l,m}\right]=\begin{cases}\frac{4}{\#VIV^{i,r}_{x_{j,k},l}}P\left[x^{\hat{i},r}_{j,k,l}\right], r=\{1,2,3\}\ j=\{I,II\}\\ \frac{4}{\#VIV^{\hat{i},\hat{r}}_{x_{j,k},l}}P\left[x^{\hat{i},\hat{r}}_{j,k,l}\right], j=\{III,IV\}\\ \frac{2}{\#VIV^{i,\hat{r}}_{x_{j,k},l}}P\left[x^{i,\hat{r}}_{j,k,l}\right], i=\{1,2,3\}, j=\{V\}\end{cases}$$

Donde $\#VIV^{i,r}_{x_{j,k},l}$ es el total de viviendas de la manzana *l*, adentro de la ageb *k* (con índice de marginación *r*) de la localidad *j* (con índice de marginación *i*).

9. *Probabilidad de selección del individuo*

Sea $X^{i,r}_{j,k,l,m,n}$ el n – ésimo individuo de 18 años o más de la vivienda m, que pertenece a la l – ésima manzana de la k – ésima ageb (con índice de marginación r) del j – ésimo estrato (con índice de marginación i). Con esa notación, la probabilidad de selección de ese individuo es:

$$P\left[X^{i,r}_{j,k,l,m,n}\right]=\frac{1}{N^{i,r}_{j,k,l,m}}P\left[X^{i,r}_{j,k,l,m}\right]$$

Donde $N^{i,r}_{j,k,l,m}$ representa el total de individuos de 18 años o más en la vivienda *m* de la manzana *l* que se encuentra en la ageb *k* (con índice de marginación *r*), de la localidad *j* (con índice de marginación *i*).

10. *Probabilidad estimada de selección de la vivienda*

Derivado del uso de un marco de muestreo de información incompleta, no se obtuvo información fidedigna del número de viviendas en cada manzana seleccionada. Por lo anterior, se tuvo que recurrir al siguiente estimador:

$$P\left[X^{i,r}_{j,k,l,m}\right] = \frac{v_j}{\#VIV^{i,r}_{x_{j,k}} \Big/ \#MZ^{i,r}_{j,k}}$$

Donde v_j es el número de viviendas que se seleccionan por manzana (que depende del estrato *j*), y $\#VIV^{i,r}_{x_{j,k}}$ es el número de viviendas en la ageb *k*, información disponible en el marco de muestreo. Dicho valor fue sustituido en lugar del real para el cálculo de la probabilidad de selección del individuo.

11. *Probabilidad de selección en Chiapas*

Sea $X_{j,k,l,m,n}$ la persona *n* – ésima de la vivienda *m* que pertenece a la manzana *l* de la ageb *k* de la localidad *j* del estrato *i*. Entonces, su probabilidad de selección es:

$$P\left[X_{i,j,k,l,m,n}\right] = \left(\frac{PT18(x_i)}{\sum_{x_i} PT18(x_i)}\right)\left(\frac{PT18(x_{i,j})}{\sum_{x_{i,j}} PT18(x_i)}\right)\left(\frac{n_{i,j,k}}{\#MZ_{i,j,k}}\right)\left(\frac{n_{i,j,k,l}}{\#VIV_{i,j,k,l}}\right)\left(\frac{1}{N_{i,j,k,l,m}}\right)$$

Donde $n_{i,j,k}$ es el número de manzanas a seleccionar de la ageb *k* que se encuentra en la *j* – ésima localidad y que depende del estrato *i*. Y $n_{i,j,k,l}$ es el número de viviendas a seleccionar de esa manzana. Cabe señalar que para el cálculo de la probabilidad de la vivienda se recurrió al estimador presentado en el punto anterior, por las mismas razones. Una vez que se tiene la probabilidad de selección, considerando la probabilidad de selección del levantamiento nacional, para cada individuo de Chiapas se tienen dos probabilidades. Por lo anterior, se recurrió al siguiente ajuste:

$$P\left[x\right]=P_1\left[x^{i,r}_{j,k,l,m,n}\right]P_2\left[x_{i,j,k,l,m,n}\right]+P_1\left[x^{i,r}_{j,k,l,m,n}\right]\left(1-P_2\left[x_{i,j,k,l,m,n}\right]\right)+\ 1-P_1\left[x^{i,r}_{j,k,}\right.$$
$$+\left(1-P_1\left[x^{i,r}_{j,k,l,m,n}\right]\right)P_2\left[x_{i,j,k,l,m,n}\right]$$

Donde $P\left[x\right]$ es la probabilidad de seleccionar al individuo x, $P_1\left[x^{i,r}_{j,k,l,m,n}\right]$ es la probabilidad de selección conforme al diseño muestral nacional y $P_2\left[x_{i,j,k,l,m,n}\right]$ representa la probabilidad conforme al diseño muestral de Chiapas.

12. *Probabilidad de selección con reemplazo*

Cuando el diseño muestral contempla la selección con reemplazo para cierta etapa, cada una de las unidades seleccionadas pudo haber salido en más de una ocasión (por ejemplo, si se seleccionan tres ageb's con reemplazo, una ageb puede ser seleccionada cero, una, dos o tres veces). Así, la probabilidad de selección debe considerar esta situación, y no basarse en la probabilidad de selección de una ocasión.

Sea x_j una unidad de muestreo que pertenece a un subconjunto *j*, del cual se harán *m* – extracciones independientes, con reemplazo, y sea $0<p_x<1$ la probabilidad de seleccionar a la unidad. Sin pérdida de generalidad, supongamos que cada extracción se realiza en un instante de tiempo distinto. Así, la probabilidad de que x_j haya salido en una ocasión, es la probabilidad de que hubiera salido en la primera extracción, pero no en el resto; más, la probabilidad de que no hubiera salido en la primera extracción, hubiera salido en la segunda, y no volviera a salir; y así, sucesivamente. En fórmula, se tiene:

$$P\left[x_j=1\right]=\underbrace{p_x\underbrace{\left(1-p_x\right)\cdots\left(1-p_x\right)}_{m-1\ veces}+\left(1-p_x\right)p_x\underbrace{\left(1-p_x\right)\cdots\left(1-p_x\right)}_{m-2\ veces}+\cdots+\underbrace{\left(1-p_x\right)\cdots\left(1-p_x\right)}_{m-1\ veces}p_x}_{m-veces}$$
$$=\sum_{i=1}^{m}p_x\left(1-p_x\right)^{m-1}=m\left[p_x\left(1-p_x\right)^{m-1}\right]$$

Siguiendo este caso, la probabilidad de que x_j hubiera salido en dos ocasiones tiene que considerar la probabilidad de que sea extraído en la primera ocasión, luego sea extraído en la segunda, pero no en el resto; o bien,

que sea extraído en la primera ocasión, no en la segunda, sí en la tercera, y no en el resto. Después, habrá que considerar los casos en los que no fue seleccionada en la primera ocasión, sí en la segunda, sí en la tercera, pero no en el resto; o bien, no en la primera, sí en la segunda, no en la tercera, pero sí en el resto. Siguiendo ese proceso en etapas sucesivas, se llega a la siguiente fórmula:

$$
\begin{aligned}
P\left[x_j=2\right] &= \underbrace{p_x p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-2-veces} + p_x(1-p_x)\,p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-3-veces} + \cdots + p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-2-veces} p_x}_{m-sumandos} \\
&\quad + \underbrace{(1-p_x)\,p_x p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-3-veces} + (1-p_x)\,p_x(1-p_x)\,p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-4-veces} + \cdots + (1-p_x)\,p_x \underbrace{(1-p_x)\cdots(1-p_x)}_{m-3-veces} p_x}_{m-1-sumandos} \\
&\quad + \underbrace{(1-p_x)\cdots(1-p_x)}_{m-2-veces} p_x p_x \\
&= \sum_{i=1}^{m} p_x^2 (1-p_x)^{m-2} + (1-p_x)\sum_{i=1}^{m-1} p_x^2 (1-p_x)^{m-3} + \cdots + (1-p_x)^{m-2} p_x^2 \\
&= p_x^2 (1-p_x)^{m-2}\left[m+m-1+\cdots+1\right] \\
&= \tfrac{m(m+1)}{2} p_x^2 (1-p_x)^{m-2} \\
&= \frac{m!}{2!(m-2)!} p_x^2 (1-p_x)^{m-2} \\
&= \binom{m}{2} p_x^2 (1-p_x)^{m-2}
\end{aligned}
$$

Generalizando la fórmula, se puede demostrar que sigue una distribución binomial con parámetros *m* (el número de repeticiones) y p_x la probabilidad de selección. Así, la probabilidad de que x_j sea seleccionada es:

$$
\begin{aligned}
P\left[x_j \geq 1\right] &= \sum_{i=1}^{m} \binom{m}{i} p_x^i (1-p_x)^{m-i} \\
&= 1-(1-p_x)^m = 1-P\left[x_j=0\right]
\end{aligned}
$$

13. *Post estratificación*

Para que los datos puedan expandir a la población en las proporciones que se presentan en el país, es necesario seguir un proceso de calibración, consistente en multiplicar el factor de expansión original por una constante que permita alcanzar ese objetivo.

Sea $f_1\left(X_{r,s,t}\right)=\sum_x \frac{1}{P\left[x_{r,s,t}\right]}$ el factor de expansión obtenido como el inverso multiplicativo de la probabilidad de selección del individuo *x*, que pertenece al estrato *r*, al sexo *s* y al grupo quinquenal de edad *t,* sumado para cada individuo perteneciente a ese conjunto (es decir, la expansión total de ese conjunto).

Los 19 estratos considerados en esta parte fueron los tres niveles de marginación de las ageb's del Distrito Federal, las ciudades de Guadalajara y Monterrey; las localidades de más de 400,000 habitantes (sin considerar las tres zonas metropolitanas); las localidades de 10,000 a 399,999 habitantes (sin considerar Chiapas); los tres niveles de marginación de las localidades de menos de 10,000 habitantes (sin considerar Chiapas); las localidades de Chiapas de menos de 10,000 habitantes; las localidades de Chiapas de 10,000 a 99,999 habitantes; cada una de las tres localidades de Chiapas de 100,000 habitantes o más.

A su vez, los grupos quinquenales fueron: 18 a 19 años de edad; 20 a 24 años de edad, en quinquenios hasta el grupo de 60 a 64 años de edad; y, 65 años o más.

Sea $g_1\left(X_{r,s,t}\right)=PT_{r,s,t}$ la población total perteneciente al estrato *r*, sexo *s* y grupo de edad *t*. Entonces, el factor de expansión ya corregido por las variables poblacionales es:

$$f_2\left(x_{r,s,t}\right)=\frac{g\left(X_{r,s,t}\right)}{f_1\left(X_{r,s,t}\right)}\frac{1}{P\left[x_{r,s,t}\right]}$$

Pues, como se demuestra a continuación, logra la expansión deseada.

$$\sum_x f_2\left(x_{r,s,t}\right)=\sum_x \frac{g\left(X_{r,s,t}\right)}{f_1\left(X_{r,s,t}\right)}\frac{1}{P\left[x_{r,s,t}\right]}=\frac{g\left(X_{r,s,t}\right)}{f_1\left(X_{r,s,t}\right)}\sum_x \frac{1}{P\left[x_{r,s,t}\right]}=\frac{g\left(X_{r,s,t}\right)}{f_1\left(X_{r,s,t}\right)}f_1\left(X_{r,s,t}\right)=g\left(X_{r,s,t}\right)$$

VII. Referencias

ALMOND, Gabriel y VERBA, Sidney, *The Civic Culture, Political Attitudes and Democracy in Five Nations*, USA, Sage Publications, 1989.

VII Censo General de Población, México, Dirección General de Estadística (ahora INEGI), 1953.

HENRY, Gary T., *Practical Sampling*, USA, Sage Publications (Applied Social Research Methods Series Vol. 21), 1990.

KALTON, Graham, *Introduction to Survey Sampling*, USA, Sage Publications (Quantitative Applications in the Social Sciences Vol. 35), 1983.

LÉVY MANGIN, Jean-Pierre, *Análisis multivariable para las ciencias sociales*, Madrid, Pearson Educación, 2003.