



Diseño estadístico

**Encuesta para la
Medición
Multidimensional
del Desarrollo y
Bienestar Local
del Municipio de
San Juan
Sacatepéquez**

Departamento de muestreo INE

Índice

1. Diseño estadístico	1
1.1. Población objetivo	1
1.2. Dominios de estudio	1
1.3. Marco de muestreo	3
1.4. Esquema de muestreo	3
1.4.1. Probabilístico.....	3
1.4.2. Bietápico	4
1.4.3. Estratificado	4
1.4.4. Por conglomerados	4
1.5. Tamaño de la muestra	4
1.6. Afijación de la muestra	5
1.7. Estructura del diseño estadístico	5
1.8. Selección de la muestra	5
1.8.1. Selección de las unidades primarias de muestreo.....	6
1.8.2. Selección de las unidades secundarias de muestreo.....	6
1.9. Ponderación.....	7
1.9.1. Ajuste de los factores de expansión	7
1.10. Estimadores puntuales.....	8
1.11. Estimación por intervalo	9
1.11.1. Cálculo de la varianza, el error estándar, la precisión y el intervalo de confianza	9

1. Diseño estadístico

Un diseño estadístico comprende el conjunto de actividades relacionadas con la selección de la muestra, el tamaño suficiente para la generación de estimaciones respecto a la población objeto de estudio, el marco de muestreo, y la construcción y evaluación de las estimaciones planteadas para la expansión de la información, a partir de los datos obtenidos en el operativo de campo.

Básicamente lo que se busca al realizar una estimación es garantizar que, en la construcción del intervalo de confianza, se cumplan dos características importantes: confiabilidad y precisión. El término confiable, se refiere a que una gran proporción de las muestras posibles a seleccionar, el 95% o 99% según se desee, con un método de selección y un tamaño de muestra específicos, contengan dentro de su intervalo de confianza, el valor real del parámetro de la población. Y la precisión hace referencia a que la longitud de este intervalo sea pequeña, de tal manera que le proporcione al usuario elementos de juicio para poder tomar decisiones.

Por lo anterior, el presente documento describe el conjunto de procedimientos técnicos necesarios para garantizar una muestra probabilística, significativa y estadísticamente representativa de las unidades de análisis.

1.1. Población objetivo

La población objeto de análisis para la encuesta lo conforman los hogares y las personas que residen permanentemente en las viviendas particulares ubicadas dentro del municipio de San Juan Sacatepéquez del Departamento de Guatemala.

1.2. Dominios de estudio

Un dominio de estudio denota la población para la cual son válidas las mediciones (censales o muestrales). Para el caso particular de la encuesta se tiene contemplado el municipio de San Juan Sacatepéquez como único dominio de estudio y cuatro campos de tabulación (4 microregiones). Los campos de tabulación generalmente se definen como los subgrupos de análisis para los que se quieren datos con igual fiabilidad. A continuación, se detallan los lugares poblados correspondientes a cada microregión.

Micro-región I: Área Sur

1. Lo de Mejía
2. Carranza
3. La Económica
4. San José lo de Ortega
5. Villa Verde
6. San Juaneritos
7. Villas de Quetzal
8. Los Encerros
9. Ciudad Gótica
10. Colonia los Robles I, II, III, IV
11. Colonia Las Margaritas
12. San Francisco El Bosque I y II
13. Colonia Las Vistas

14. Lomas de San Pedro
15. Colonia Colinas I, II y III y
16. Ciudad Quetzal (centro)

Micro-región II: Área Central

17. Chitol
18. Cruz Verde
19. Sajcavilla
20. Comunidad de Zet
21. Camino a San Pedro
22. Pachali Chiquito
23. Comunidad de Ruiz
24. Pacayaj
25. Barrio Chitun
26. Sector Uno Ojo de Agua (Cruz B.)
27. Cruz de Ayapan
28. La Viña
29. Colonia Cerro Candelaria

Micro-región III: Área Occidente

1. Cruz Blanca
2. Asunción Chivoc
3. Joya de las Flores
4. Lo de Ramos
5. Pajoques
6. Pilar II
7. Pilar I
8. San Antonio las Trojes II
9. Santa Fe Ocaña
10. Los Guamuches
11. Los Pajoc
12. Loma Alta
13. Los Quequezque

Micro-región IV: Área Norte

14. Estancia Grande
15. Los Guates
16. Colonia San Jorge
17. La Soledad
18. Santa Rosa
19. Las Palmas
20. Patanil
21. San Matías
22. Suacite
23. Montufar

24. Caserío Los Sineyes
25. Caserío los Chajones
26. Sacsuy
27. Caserío Los Cux
28. Candelaria
29. Cerro Alto
30. RealGuit
31. Ajvix
32. Los Patzanes II
33. Los Chajones (cerro alto)
34. San Francisco Las Animas y
35. Llanos de San Juan
36. Joya Linda
37. Pachum
38. Los Caneles
39. Los Caneles II
40. Los Pirires
41. Sector la Buena Esperanza
42. Estancia El Rosario
43. San Jerónimo Chuaxan

1.3. Marco de muestreo

El marco de muestreo (MM) es el conjunto de unidades con probabilidades conocidas y diferentes de cero del cual se seleccionan las viviendas que forman las muestras de hogares. Conforme al Sistema Integrado de Encuestas de Hogares (SIEH) del Instituto Nacional de Estadística, la encuesta tiene como marco general el Marco Maestro de Muestro de 182 unidades primarias de muestreo (26,665 viviendas ocupadas) que son los sectores cartográficos o conglomerados de viviendas del municipio de San Juan Sacatepéquez.

1.4. Esquema de muestreo

El esquema de muestreo se refiere a una combinación de opciones técnicas en cuanto al tipo y modalidad del muestreo. El esquema de muestreo que se adoptará para la encuesta es de tipo probabilístico, bietápico, estratificado y por conglomerados. A continuación, se mencionan las características principales de estos tipos de muestreo.

1.4.1. Probabilístico

La naturaleza matemática del muestreo probabilístico es lo que permite fundamentar científicamente las estimaciones que se desean realizar en la encuesta. El muestreo probabilístico en el contexto de una encuesta de hogares abarca los métodos mediante los cuales se seleccionan los elementos de la población objetivo —unidades geográficas, hogares y personas— para su inclusión en la muestra. El muestreo probabilístico requiere que: a) cada elemento tenga una probabilidad matemática conocida de ser seleccionado, b) dicha probabilidad sea mayor que cero, y c) dicha probabilidad sea calculable numéricamente. Es importante destacar que la probabilidad de cada elemento de ser seleccionado no tiene que ser igual, sino que puede variar en función de los objetivos de la encuesta.

1.4.2. Bietápico

Seleccionar la muestra en etapas tiene ventajas prácticas para el proceso de selección mismo. Para el caso de la encuesta, se considera la selección de unidades primarias de muestreo equivalentes a sectores cartográficos y unidades secundarias de muestreo equivalentes a 30 viviendas ocupadas.

1.4.3. Estratificado

Es estratificado porque las unidades de muestreo se clasifican de acuerdo con el área geográfica (urbana o rural) y la microregión de pertenencia.

1.4.4. Por conglomerados

El muestreo de conglomerados denota los métodos de selección en los cuales la unidad de muestreo, que es la unidad de selección, contiene más de un elemento de la población (unidad compuesta). Para el caso particular de la encuesta, en las etapas de selección de la muestra 1 y 2, se seleccionan conglomerados como unidades de muestreo.

1.5. Tamaño de la muestra

Para el cálculo del tamaño de la muestra del dominio se estableció como variable principal de diseño muestral la tasa de pobreza extrema del departamento de Guatemala obtenida de la ENCOVI 2014; se consideró esta proporción debido a que ésta presenta un alto grado de correlación con las variables incluidas en el instrumento de recolección de datos.

El tamaño de muestra para cometer un error relativo de muestreo en presencia del coeficiente de confianza adicional fue calculado con el siguiente algoritmo:

$$n = \frac{Z_{1-\alpha/2}^2 (1 - \hat{p}) \text{Deff}}{e_r^2 \hat{p}(1 - \text{tnr}) \bar{X}}$$

Donde:

- | | |
|-----------|--|
| Z | Valor en la escala de Z de una distribución normal de probabilidad correspondiente al nivel de confiabilidad asumido (95% de confianza). |
| p | Tasa de pobreza extrema del departamento de Guatemala (5.4%). |
| Deff | Efecto de diseño o efecto de conglomeración asumido, que es la pérdida o ganancia en la eficiencia del diseño por efecto de estratificación y/o conglomerar elementos de la población, para formar unidades muestrales. Generalmente se obtiene como el cociente de dividir la varianza de la variable en el diseño complejo entre la varianza de la variable considerando un MAS (1.8). |
| e | Error relativo de muestreo (10%). |
| $\bar{X}$ | Promedio de personas por hogar (5.45). |
| tnr | Tasa de no respuesta (10% de rechazo). |

El total de muestra calculado a nivel del municipio de San Juan Sacatepéquez correspondió a 2,469 viviendas. Este resultado fue ajustado con el fin de obtener un número fijo de muestra de segunda etapa por Unidad Primaria de Muestreo de 30 viviendas ocupadas, por lo que el tamaño de muestra resultante correspondió a 2,490 viviendas ocupadas. De esta cuenta, y dado el número prefijado de viviendas por

UPM, el total de UPMS a seleccionar como muestra de primera etapa corresponde a 83 sectores cartográficos. Los tamaños de muestra para cada microregión se detallan a continuación.

Tamaños de muestra			
Microregión	Sectores marco de muestreo	Sectores muestra	Viviendas muestra
1	78	33	990
2	24	12	360
3	35	18	540
4	45	20	600
Total general	182	83	2,490

1.6. Afijación de la muestra

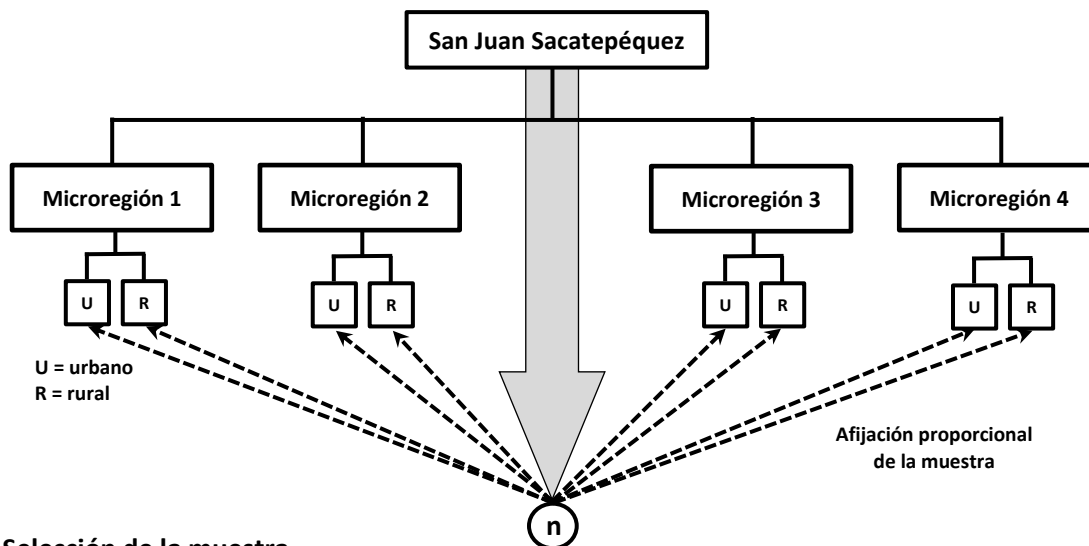
Con la finalidad de que la muestra sea representativa del dominio de estudio, se realizó la distribución de ésta de manera proporcional al tamaño de las cuatro microregiones y sus respectivas áreas urbana y rural. La distribución se llevó a cabo a través de la siguiente expresión:

$$n_h = \frac{N_h}{N} n$$

Donde:

- n_h Tamaño de muestra en sectores para la h-ésimo estrato en el dominio de estudio.
- n Tamaño de la muestra en sectores.
- N_h Total de viviendas en el h-ésimo estrato del dominio de estudio.
- N Total de viviendas en el dominio de estudio según marco de muestreo.

1.7. Estructura del diseño estadístico



1.8. Selección de la muestra

La selección de la muestra se efectúa de manera independiente en cada estrato perteneciente a un campo de tabulación. El procedimiento de selección se describe a continuación.

1.8.1. Selección de las unidades primarias de muestreo

Al interior de cada h-ésimo estrato, de un campo de tabulación, se seleccionaron sistemáticamente " k_n " UPM con probabilidad proporcional al tamaño (PPT) como muestra de primera etapa (la variable de ponderación será número de viviendas por sector cartográfico).

1.8.2. Selección de las unidades secundarias de muestreo

La selección de la muestra de segunda etapa, 30 viviendas al interior de cada sector cartográfico, se llevará a cabo de manera independiente en cada unidad primaria de muestreo a través de un muestreo inverso. El muestreo inverso fue descrito por Haldane's en el año 1945 y William Cochran lo retomó en su libro Técnicas de Muestreo, Tercera Edición, en el año 1977.

El muestreo inverso puede ser aplicado a través de algún algoritmo coordinado negativo. Básicamente un algoritmo coordinado negativo ordena la población de forma aleatoria. De hecho, se trata de una permutación aleatoria de los elementos de la población. De esta forma, si la población tiene N individuos, y se ordenan aleatoriamente, entonces los primeros (o los últimos) k elementos conforman una muestra aleatoria simple de tamaño k .

Las razones por las que se aplicará el muestreo inverso en la selección de las unidades secundarias de muestreo de la encuesta son las siguientes:

1. Es probable que durante el operativo de campo se observe una tasa de no respuesta mayor a la esperada en el modelo teórico propuesto en el diseño de la muestra. Esto debido a las situaciones de violencia y confrontación que existe entre la población del municipio de San Juan Sacatepéquez y el Ministerio de Energía y Minas (MEM).
2. Es importante contar con datos suficientes para la generación de resultados confiables a nivel de campo de tabulación.

La aplicación del muestreo inverso demanda el establecimiento y actualización del total de viviendas ocupadas al interior de cada UPM seleccionada. Para ello, en cada UPM seleccionada se procederá a contabilizar el número total de viviendas ocupadas mediante el uso del formulario FC01 del INE (actualización cartográfica). El listado resultante del proceso de actualización permitirá conocer las probabilidades de inclusión de las unidades secundarias y su posterior selección.

Una vez actualizado un sector cartográfico, se procederá a digitar el total de viviendas actualizadas en un archivo Excel (proporcionado por el INE) y a dar click en la opción "generar selección". La macro en Excel devolverá el marco aleatorio y el detalle de las viviendas que deben medirse, junto con las viviendas de reemplazo (este mecanismo debe realizarse en cada UPM de la encuesta).

Si alguna de las 30 viviendas visitadas presenta algún tipo rechazo (independientemente de la naturaleza del mismo), se procederá a reemplazar la unidad de no respuesta con la estructura inmediata superior del ordenamiento aleatorio. Este procedimiento debe repetirse hasta cumplir la cuota de 30 viviendas por sector.

1.9. Ponderación

Definidos los procedimientos de selección para las unidades primarias y secundarias de muestreo, la probabilidad de seleccionar una vivienda en un campo de tabulación está dada por:

$$P\{V_{CThi}\} = \frac{k_{CTh} m_{CThi}}{m_{CTh}} \times \frac{30}{m_{CThi}^*} = \frac{30k_{CTh} m_{CThi}}{m_{CTh} m_{CThi}^*}$$

Su factor de expansión está dado por:

$$F_{CThi} = \frac{m_{CTh} m_{CThi}^*}{30k_{CTh} m_{CThi}}$$

Donde:

- F_{CThi} Factor de expansión de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación.
- K_{CTh} Número de UPMS seleccionadas en el h-ésimo estrato para el marco maestro de muestreo del campo de tabulación.
- m_{CThi} Número de viviendas en la i-ésima UPM, del h-ésimo estrato, según marco de muestreo del campo de tabulación.
- m_{CTh} Número de viviendas en la en el h-ésimo estrato según marco de muestreo del campo de tabulación.
- m_{CThi}^* Número de viviendas en la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación según actualización cartográfica.

1.9.1. Ajuste de los factores de expansión

Los factores de expansión se ajustarán con base a los siguientes criterios:

a. Ajuste por no respuesta

El ajuste por no respuesta atribuida al informante se realizará a nivel de UPM, en cada grupo de tabulación, mediante la siguiente expresión:

$$F_{CThi}^{nr} = F_{CThi} \frac{m_{CThi}}{m_{CThi}^r}$$

Donde:

- F_{CThi}^{nr} Factor de expansión corregido por no respuesta para viviendas de la i-ésima UPM, del estrato h-ésimo, en el campo de tabulación.
- F_{CThi} Factor de expansión de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación (factor natural).
- m_{CThi} Número de viviendas seleccionadas habitadas en la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación.
- m_{CThi}^r Número de viviendas seleccionadas habitadas con respuesta en la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación.

b. Ajuste por proyección al momento central del empadronamiento

Los factores de expansión ajustados por la no respuesta se corrigen, a fin de asegurar que en el dominio de interés de la encuesta se obtenga la población total determinada por la proyección de población generada por INE referida al punto medio del levantamiento, mediante la siguiente expresión.

$$FE_{hi}^{CT} = F_{CThi}^{nr} \frac{PP_D}{PE_D}$$

Donde:

FE_{hi}^{CT}	Factor de expansión corregido por proyección para viviendas de la i-ésima UPM, del estrato h-ésimo, en el campo de tabulación.
F_{CThi}^{nr}	Factor de expansión corregido por no respuesta para viviendas de la i-ésima UPM, del estrato h-ésimo, en el campo de tabulación.
PP_D	Proyección de población para el dominio de estudio.
PE_D	Población total a la que expande la encuesta en el dominio de estudio.

1.10. Estimadores puntuales

Para la encuesta se podrán realizar dos tipos de estimaciones: estimaciones para totales y estimaciones para proporciones, tasas y promedios, para estos últimos se utilizan estimadores de razón.

El estimador del total de la característica X en el dominio de estudio es:

$$\hat{X} = \sum_h \sum_i FE_{hi}^{CT1} \left(\sum_s \sum_j X_{sj}^{CT1} \right) + \sum_h \sum_i FE_{hi}^{CT2} \left(\sum_s \sum_j X_{sj}^{CT2} \right) + \sum_h \sum_i FE_{hi}^{CT3} \left(\sum_s \sum_j X_{sj}^{CT3} \right) + \sum_h \sum_i FE_{hi}^{CT4} \left(\sum_s \sum_j X_{sj}^{CT4} \right)$$

El estimador del total de la característica X en un campo de tabulación es:

$$\hat{X}_{CT} = \sum_h \sum_i FE_{hi}^{CT} \left(\sum_s \sum_j X_{sj}^{CT} \right)$$

Donde:

FE_{hi}^{CT1}	Factor de expansión final de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación 1.
FE_{hi}^{CT2}	Factor de expansión final de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación 2.
FE_{hi}^{CT3}	Factor de expansión final de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación 3.
FE_{hi}^{CT4}	Factor de expansión final de la i-ésima UPM, del h-ésimo estrato, en el campo de tabulación 4.
X_{sj}^{CT1}	Valor observado de la característica de interés X de la j-ésima persona, en la s-ésima vivienda, en el campo de tabulación 1.
X_{sj}^{CT2}	Valor observado de la característica de interés X de la j-ésima persona, en la s-ésima vivienda, en el campo de tabulación 2.
X_{sj}^{CT3}	Valor observado de la característica de interés X de la j-ésima persona, en la s-ésima vivienda, en el campo de tabulación 3.
X_{sj}^{CT4}	Valor observado de la característica de interés X de la j-ésima persona, en la s-ésima vivienda, en el campo de tabulación 4.

Para la estimación de proporciones, tasas y promedios se utilizará el estimador de razón:

Donde $\hat{Y}$ (el conjunto del que X es un subconjunto) se define en forma análoga a $\hat{X}$. $R = \frac{X}{Y}$

1.11. Estimación por intervalo

Para determinar los intervalos de confianza de las principales estimaciones de la encuesta, se procederá a calcular la varianza del estimador, el error estándar de estimación, el margen de error y la precisión estadística.

1.11.1. Cálculo de la varianza, el error estándar, la precisión y el intervalo de confianza

Para la evaluación de los errores de muestreo de las principales estimaciones se utilizará el método de Conglomerados Últimos, basado en que la mayor contribución a la varianza de un estimado, en un diseño bietápico, es la que se presenta entre las unidades primarias de muestreo.

Para obtener las precisiones de los estimadores de razón, conjuntamente al método de Conglomerados Últimos se aplicará el método de las series de Taylor, obteniéndose la siguiente fórmula para estimar la varianza de R en el dominio de estudio:

$$V(R) = \frac{1}{Y^2} \sum_{CT}^4 \left\{ \sum_h \frac{n_{CTh}}{n_{CTh} - 1} \sum_i^{n_{CTh}} \left[\left(X_{CThi} - \frac{1}{n_{CTh}} X_{CTh} \right) - R \left(Y_{CThi} - \frac{1}{n_{CTh}} Y_{CTh} \right) \right]^2 \right\}$$

Donde:

X_{CThi} Total ponderado de la variable de estudio X para la i-ésima UPM, en el estrato h-ésimo, en el campo de tabulación.

X_{CTh} Total ponderado de la variable de estudio X para el h-ésimo estrato, en el campo de tabulación.

Y_{CThi} Total ponderado de la variable de estudio Y para la i-ésima UPM, en el estrato h-ésimo en el campo de tabulación.

Y_{CTh} Total ponderado de la variable de estudio Y para el h-ésimo estrato, en el campo de tabulación.

n_{CTh} Número de UPM en el h-ésimo estrato, en el campo de tabulación.

La varianza de R en un campo de tabulación:

$$V(R) = \frac{1}{Y^2} \left\{ \sum_h^L \frac{n_h}{n_h - 1} \sum_i^{n_h} \left[\left(X_{hi} - \frac{1}{n_h} X_h \right) - R \left(Y_{hi} - \frac{1}{n_h} Y_h \right) \right]^2 \right\}$$

Donde:

X_{hi} Total ponderado de la variable de estudio X para la i-ésima UPM, en el estrato h-ésimo.

- X_h Total ponderado de la variable de estudio X para el h-ésimo estrato.
- Y_{hi} Total ponderado de la variable de estudio Y para la i-ésima UPM, en el estrato h-ésimo.
- Y_h Total ponderado de la variable de estudio Y para el h-ésimo estrato.
- n_h Número de UPM en el h-ésimo estrato.

La estimación de la varianza del estimador de un total en el dominio de estudio, se calcula con la siguiente expresión:

$$V(X) = \sum_{CT=1}^4 \sum_{h=1}^{L_{CT}} \frac{n_{CTh}}{n_{CTh} - 1} \sum_{i=1}^{n_{CTh}} \left(X_{CThi} - \frac{1}{n_{CTh}} X_{CTh} \right)^2$$

La estimación de la varianza del estimador de un total, se calculará con la siguiente expresión:

$$V(X) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(X_{hi} - \frac{1}{n_h} X_h \right)^2$$

Las estimaciones del error estándar de estimación (EE), el efecto de diseño (Deff) y coeficiente de variación (CV) se calculan mediante las siguientes expresiones:

$$EE = \sqrt{V(\theta)} \quad Deff = \frac{V(\theta)}{V(\theta)_{MAS}} \quad CV = \frac{\sqrt{V(\theta)}}{\theta}$$

Finalmente, el intervalo de confianza al 95%, se calcula de la siguiente forma:

$$I_{1-\varepsilon} = \left(\theta - 1.96\sqrt{V(\theta)}, \theta + 1.96\sqrt{V(\theta)} \right)$$