DISEÑO ESTADÍSTICO
ENCUESTA DE CULTURA AMBIENTAL 2026

1 Introducción

La encuesta de Cultura Ambiental, la cual hace parte de la Encuesta Bienal de Cultura, y realizada por la Dirección Observatorio y Gestión del Conocimiento Cultural - DOGCC, desde su creación se ha consolidado como una de las principales fuentes de información respecto de las prácticas y percepciones relacionadas con el comportamiento ambiental de los ciudadanos, el uso eficiente del agua y otras prácticas sostenibles. Esta información busca identificar áreas de oportunidad para fortalecer la educación y la sensibilización ambiental en Bogotá, contribuyendo a la creación de políticas públicas orientadas a la sostenibilidad y la preservación de los recursos naturales.

En el año 2001 se comenzó a aplicar una de las principales encuestas con las que cuenta el distrito que fue la Encuesta de Cultura Ciudadana y desde entonces, aunque ha venido cambiando su nombre, se ha mantenido y se aplica cada dos años, contando a la fecha con doce aplicaciones.

Generalmente las encuestas diseñadas desde la Dirección consultan a personas de mínimo trece años, que residen en las zonas urbana y rural de las 20 localidades de Bogotá.

La información recolectada permite estimaciones de totales con representatividad a nivel distrital, local, por sexo y rangos de edad.

2 Población objetivo

La población objetivo del estudio está conformada por las personas de 13 años o más, residentes habituales en viviendas particulares ubicadas en las veinte localidades de Bogotá, D.C., al momento de la recolección de la información.

3 Cobertura geográfica

La cobertura geográfica del estudio comprende las veinte localidades de Bogotá, D.C., las cuales constituyen el ámbito geográfico para el cual se diseñó la muestra y sobre el que se producirán las estimaciones estadísticas. La encuesta permitirá obtener resultados

representativos para el total de Bogotá, para cada una de sus localidades, y para zona urbana y rural, de acuerdo con los escenarios de diseño definidos.

La cobertura del estudio se restringe a las viviendas particulares y a la población residente habitual ubicada dentro de los límites político-administrativos del Distrito Capital. En consecuencia, no hacen parte del universo de estudio las personas que residen en viviendas colectivas ni aquellas sin residencia habitual en Bogotá.

Las localidades incluidas en el diseño son:

- Usaquén
- Chapinero
- Santa Fe
- San Cristóbal
- Usme
- Tunjuelito
- Bosa
- Kennedy
- Fontibón
- Engativá
- Suba
- Barrios Unidos
- Teusaquillo
- Los Mártires
- Antonio Nariño
- Puente Aranda
- La Candelaria
- Rafael Uribe Uribe
- Ciudad Bolívar

- Sumapaz

4 Marco de muestreo

El marco de muestreo del presente estudio se define a partir del Marco Geoestadístico Nacional (MGN) del Departamento Administrativo Nacional de Estadística (DANE), empleando la cartografía oficial de manzanas urbanas para las diecinueve localidades urbanas de Bogotá, D.C. (código DANE 11001). Para la localidad de Sumapaz, dada su naturaleza predominantemente rural, se utilizó la cartografía rural del MGN, conformada por los centros poblados y el área rural dispersa.

El marco de muestreo se complementa con las proyecciones oficiales de población del DANE para el año 2026, desagregadas por localidad y zonas (urbano – rural). Esta información permite identificar la distribución de la población objetivo en cada dominio de estudio.

El marco de muestreo está conformado por el conjunto de unidades cartográficas del Distrito Capital, manzanas urbanas para las localidades urbanas y unidades cartográficas rurales para Sumapaz junto con la información asociada de viviendas y población residente. Estas unidades constituyen las Unidades Primarias de Muestreo (UPM) del diseño probabilístico y sirven como base para la selección de la muestra.

La información demográfica proveniente de las proyecciones oficiales del DANE para 2026 se utilizará para la determinación del tamaño de muestra, la asignación de la muestra entre localidades y grupos de edad, así como para el cálculo de las probabilidades de selección y los factores de expansión del estudio.

5 Esquema de muestreo utilizado

El diseño muestral corresponde a un muestreo probabilístico, estratificado y trietápico. Las veinte (20) localidades de Bogotá constituyen los estratos de muestreo y los principales dominios de estudio para la producción de estimaciones a nivel local. Adicionalmente, el diseño permite obtener estimaciones representativas para el total de Bogotá, diferenciando sexo. Este diseño garantiza que todas las unidades de la población objetivo tengan una probabilidad de selección conocida y diferente de cero.

En la primera etapa se seleccionan las Unidades Primarias de Muestreo (UPM), correspondientes a las manzanas urbanas en las diecinueve localidades urbanas y a las unidades cartográficas rurales en la localidad de Sumapaz. La selección se realiza mediante Muestreo Aleatorio Simple sin reemplazo (MAS), utilizando el algoritmo de coordinado

negativo, con el fin de preservar las probabilidades de inclusión y facilitar la coordinación con futuras operaciones estadísticas.

En la segunda etapa se seleccionan las viviendas particulares ocupadas pertenecientes a las UPM previamente seleccionadas. La selección se realiza mediante muestreo sistemático con arranque aleatorio, garantizando la selección de un número fijo de viviendas por cada UPM seleccionada.

En la tercera etapa se selecciona una única persona elegible dentro de cada vivienda mediante el método de Kish, garantizando que todos los residentes habituales pertenecientes al grupo de edad objetivo tengan la misma probabilidad de ser seleccionados.

6 Definición de unidades

Para la implementación del diseño probabilístico se definen las siguientes unidades de muestreo, observación y análisis:

6.1 Unidad Primaria de Muestreo (UPM)

La Unidad Primaria de Muestreo (UPM) está constituida por las manzanas urbanas de las diecinueve localidades urbanas de Bogotá y por las unidades cartográficas rurales (centros poblados y área rural dispersa) de la localidad de Sumapaz, definidas a partir del Marco Geoestadístico Nacional (MGN) del DANE. Estas unidades corresponden a las seleccionadas en la primera etapa del diseño muestral.

6.2 Unidad Secundaria de Muestreo (USM)

La Unidad Secundaria de Muestreo (USM) corresponde a las viviendas particulares ocupadas ubicadas en las UPM seleccionadas. La selección de las viviendas se realiza mediante un procedimiento de muestreo sistemático con arranque aleatorio, garantizando que todas las viviendas tengan una probabilidad conocida de ser seleccionadas.

6.3 Unidad Terciaria de Muestreo (UTM)

La Unidad Terciaria de Muestreo (UTM) corresponde a las personas residentes habituales de las viviendas seleccionadas que tengan 13 años o más, y que residan de manera habitual en la ciudad de Bogotá. En cada vivienda se selecciona una única persona mediante algún método aleatorio, quien constituye la unidad final de selección.

6.4 Unidad de observación

La unidad de observación corresponde a la persona seleccionada dentro de la vivienda, de quien se recolecta la información mediante la aplicación del instrumento de recolección de datos.

6.5 Unidad de análisis

La unidad de análisis corresponde a las personas de 13 años o más, residentes habituales en viviendas particulares ubicadas en las veinte localidades de Bogotá, D.C. Sobre esta unidad se realizan las estimaciones e inferencias estadísticas de las variables relacionadas con el uso y la percepción de la inteligencia artificial.

7 Estratos de muestreo y dominios de estudio

Con el propósito de garantizar la representatividad territorial y producir estimaciones independientes, el diseño muestral se estructuró mediante una estratificación geográfica, en la cual cada una de las veinte localidades de Bogotá, D.C. constituye un estrato de muestreo.

La selección de la muestra se realiza de manera independiente dentro de cada estrato, lo que permite asignar tamaños de muestra acordes con la población objetivo de cada localidad y con los niveles de precisión establecidos para los diferentes escenarios de inferencia. Los estratos de muestreo corresponden a las localidades descritas en la Sección 3. Cobertura geográfica.

Los dominios de estudio comprenden dos niveles de inferencia. El primero corresponde al total de Bogotá, D.C., para las zonas urbana y rural de Bogotá, distribuidos proporcionalmente en hombres y mujeres. El segundo corresponde a cada una de las veinte localidades, para las cuales se producirán estimaciones independientes utilizando las probabilidades de selección y los factores de expansión definidos en el diseño muestral.

8 Algoritmos de selección

La selección de la muestra se realiza de manera independiente dentro de cada una de las veinte localidades de Bogotá, D.C. Para cada etapa del diseño muestral se emplea un algoritmo de selección específico, el cual garantiza la aleatoriedad del proceso y el cumplimiento de las probabilidades de selección definidas en el diseño.

8.1 Primera etapa: selección de las Unidades Primarias de Muestreo (UPM)

La selección de las Unidades Primarias de Muestreo (UPM), correspondientes a las manzanas urbanas de las diecinueve localidades urbanas y a las unidades cartográficas rurales de la

localidad de Sumapaz, se realiza mediante Muestreo Aleatorio Simple sin reemplazo (MAS) utilizando el algoritmo de coordinado negativo.

El procedimiento comprende las siguientes etapas:

1. Se identifica el conjunto de UPM pertenecientes a cada localidad.
2. A cada UPM se le asigna un número aleatorio permanente (*Permanent Random Number*, PRN), generado a partir de una distribución uniforme en el intervalo $(0,1)$.
3. Los números aleatorios se utilizan como coordenadas permanentes para la selección de la muestra.
4. Las UPM se ordenan de acuerdo con el valor de su coordenada.
5. Se seleccionan las primeras m_h UPM requeridas para la localidad correspondiente.

El uso de coordenadas permanentes permite coordinar la muestra con futuras operaciones estadísticas, reduciendo el traslape entre muestras sin alterar el diseño probabilístico.

8.2 Segunda etapa: selección de viviendas

Una vez seleccionadas las UPM, se realiza el listado completo de las viviendas particulares ocupadas existentes en cada una de ellas. Sea V_{hi} el número total de viviendas particulares ocupadas de la i-ésima UPM de la localidad h y sea v_{hi} el número de viviendas a seleccionar en dicha unidad.

El procedimiento de selección es el siguiente:

1. Se enumeran consecutivamente las V_{hi} viviendas de la UPM, desde 1 hasta V_{hi}.
2. Se calcula el intervalo de selección como:

$$k_{hi} = \frac{V_{hi}}{v_{hi}}$$

3. Se selecciona un arranque aleatorio r_{hi}, tal que:

$$r_{hi} \sim U(0, k_{hi})$$

4. Las viviendas seleccionadas corresponden a las posiciones:

$$r_{hi},\ r_{hi} + k_{hi},\ r_{hi} + 2k_{hi},\ \ \ldots,\ r_{hi} + (v_{hi} - 1)k_{hi}$$

Cuando las posiciones obtenidas no son enteras, se aproximan al entero superior correspondiente para identificar la vivienda seleccionada dentro del listado. Este procedimiento garantiza una distribución aproximadamente uniforme de las viviendas seleccionadas a lo largo de la UPM y asegura que todas las viviendas tengan la misma probabilidad de ser seleccionadas.

8.3 Tercera etapa: selección de personas

En cada vivienda seleccionada se elabora el listado de las personas residentes habituales pertenecientes a la población objetivo del estudio (13 años o más). Posteriormente, se selecciona una única persona elegible.

9 Determinación del tamaño de muestra

El tamaño de muestra se determinó con el propósito de obtener estimaciones de proporciones con los niveles de precisión establecidos para los diferentes dominios de estudio. Para ello se utilizaron las proyecciones oficiales de población del DANE para el año 2026.

El diseño contempla dos escenarios de representatividad. El primero corresponde a estimaciones para el total de Bogotá, mientras que el segundo permite obtener estimaciones independientes para cada una de las veinte localidades.

Para la determinación del tamaño de muestra se empleó la expresión propuesta por Kish (1965) para la estimación de proporciones en poblaciones finitas:

$$n = \frac{z_{1-\alpha/2}^2 PQDeff}{\varepsilon^2 + \frac{z_{1-\alpha/2}^2 PQDeff}{N}}$$

donde:

- n: tamaño de muestra requerido.
- N: tamaño de la población objetivo.
- P es el parámetro trazador del indicador de interés, fijado en P=0.5 para asumir máxima variabilidad y garantizar la estimación de cualquier proporción con los niveles de precisión establecidos;
- $Q = 1 - P$.
- ε es el margen de error máximo esperado, fijado en ±3%, ±5%.
- $z_{1-\alpha/2}$: valor de la distribución normal para un nivel de confianza del 95%.

- $Deff$ representa el efecto de diseño asociado al esquema de muestreo, el cual se establece en 1.5 siguiendo la recomendación de la ONU para encuestas en hogares en diseños multietápicos (UN 2009)

Los cálculos fueron implementados en el software R mediante la función *ss4p()* del paquete *samplesize4surveys*.

Tabla 1. Tamaño de muestra con base a la población de 13 años en adelante proyectada a 2026 – Zona Urbana.

Nombre Localidad	Hombres 13 años y más	Mujeres 13 años y más	Total 13 años y más	Muestra Hombres	Muestra Mujeres	Muestra Total
Usaquén	226.788	271.011	497.799	123	147	270
Chapinero	65.342	75.210	140.552	99	115	214
Santa Fe	46.854	48.359	95.213	88	90	178
San Cristóbal	159.413	175.029	334.442	119	131	250
Usme	155.434	166.491	321.925	121	129	250
Tunjuelito	72.166	77.904	150.070	96	104	200
Bosa	303.817	333.565	637.382	129	141	270
Kennedy	435.391	481.057	916.448	128	142	270
Fontibón	157.357	181.203	338.560	116	134	250
Engativá	333.756	382.091	715.847	126	144	270
Suba	501.308	561.852	1.063.160	132	148	280
Barrios Unidos	57.285	64.983	122.268	93	106	199
Teusaquillo	63.751	76.960	140.711	97	117	214
Los Mártires	32.534	38.998	71.532	81	97	178
Antonio Nariño	31.341	36.788	68.129	117	138	255
Puente Aranda	106.808	113.263	220.071	87	93	180
La Candelaria	7.716	7.171	14.887	78	72	150
Rafael Uribe Uribe	154.913	163.848	318.761	121	129	250
Ciudad Bolivar	265.588	280.863	546.451	131	139	270

Tabla 2. Tamaño de muestra con base a la población de 13 años en adelante proyectada a 2026 – Zona Rural.

Nombre Localidad	Hombres 13 años y más	Mujeres 13 años y más	Total 13 años y más	Muestra Hombres	Muestra Mujeres	Muestra Total
Usaquén	200	261	461	52	68	120
Chapinero	426	459	885	58	62	120
Santa Fe	475	389	864	66	54	120
San Cristóbal	21	43	64	7	13	20
Usme	4.171	3.912	8.083	93	87	180

Suba	1.737	1.590	3.327	94	86	180
Ciudad Bolivar	4.388	4.347	8.735	90	90	180
Sumapaz	1.470	1.404	2.874	92	88	180

Con el anterior, esquema, se tendrán resultados con representatividad por localidad con los parámetros indicados en la tabla 1, alcanzando una muestra para total Bogotá de 5.500 encuestas efectivas aproximadamente; tamaño que para resultados a nivel distrital se establece a partir de un error del 3% para eventos con una proporción del 50% y un ajuste por no respuesta del 2.24%.

Por su parte, la distribución por urbano y rural, representan tamaños de muestra de 4.398 y 1.100 encuestas efectivas respectivamente. En el caso del área urbana el tamaño se establece para eventos con frecuencia relativa del 50% y errores de 1.83%, mientras que para el caso del área rural, para la misma proporción el error admisible no supera 3%

10 Probabilidades de inclusión y factores de expansión

La probabilidad de inclusión de una persona corresponde al producto de las probabilidades de selección en cada una de las tres etapas del diseño muestral.

- **Primera etapa: selección de las Unidades Primarias de Muestreo (UPM)**

Sea:

- M_h: número total de UPM de la localidad h.
- m_h: número de UPM seleccionadas en la localidad h.

La probabilidad de selección de la UPM i es:

$$\pi_{Ihi} = \frac{m_h}{M_h}$$

- **Segunda etapa: selección de viviendas**

Sea:

- V_{hi}: número total de viviendas particulares ocupadas de la UPM i.
- v_{hi}: número de viviendas seleccionadas en la UPM i.

La probabilidad de selección de la vivienda j es:

$$\pi_{IIhij} = \frac{v_{hi}}{V_{hi}}$$

- **Tercera etapa: selección de personas**

Sea E_{hij} el número de personas elegibles pertenecientes al grupo de edad objeto de estudio residentes en la vivienda j. La probabilidad de seleccionar la persona k mediante el método de Kish es:

$$\pi_{IIIhijk} = \frac{1}{E_{hij}}$$

- **Probabilidad final de inclusión**

La probabilidad final de inclusión de una persona corresponde al producto de las probabilidades de selección de las tres etapas:

$$\pi_{hijk} = \pi_{Ihi} \times \pi_{IIhij} \times \pi_{IIIhijk}$$

Sustituyendo las probabilidades de cada etapa:

$$\pi_{hijk} = \left(\frac{m_h}{M_h}\right)\left(\frac{v_{hi}}{V_{hi}}\right)\left(\frac{1}{E_{hij}}\right)$$

- **Factores de expansión**

El peso básico de diseño corresponde al inverso de la probabilidad final de inclusión de cada unidad seleccionada. En un diseño probabilístico estratificado y trietápico, este peso se obtiene como el producto de los pesos de diseño asociados a cada una de las etapas de selección.

- **Factor de expansión de primera etapa**

$$d_{Ihi} = \frac{1}{\pi_{Ihi}} = \frac{M_h}{m_h}$$

- **Factor de expansión de segunda etapa**

$$d_{IIhij} = \frac{1}{\pi_{IIhij}} = \frac{V_{hi}}{v_{hi}}$$

- **Factor de expansión de tercera etapa**

$$d_{IIIhijk} = \frac{1}{\pi_{IIIhijk}} = E_{hij}$$

- **Factor de expansión básico**

El factor de expansión asociado a cada persona seleccionada está dado por:

$$d_{hijk} = d_{Ihi} \times d_{IIhij} \times d_{IIIhijk}$$

o, de manera equivalente,

$$d_{hijk} = \frac{1}{\pi_{hijk}} = \frac{M_h}{m_h} \times \frac{V_{hi}}{v_{hi}} \times E_{hij}$$

El peso básico de diseño representa el número de personas de la población que son representadas por cada persona entrevistada y constituye el peso inicial utilizado para la estimación de los parámetros poblacionales mediante el estimador de Horvitz-Thompson.

11 Parámetros de interés y estimadores

Los principales parámetros de interés del estudio corresponden a totales, promedios, proporciones y razones asociados a las variables investigadas sobre el uso y la percepción de la inteligencia artificial en la población residente de Bogotá, D.C. Las estimaciones se obtendrán mediante estimadores basados en el diseño de muestreo, utilizando las probabilidades de inclusión derivadas del diseño probabilístico estratificado en tres etapas.

11.1 Parámetros poblacionales

- **Total poblacional**

El total poblacional de una variable de interés Y está definido por

$$t_Y = \sum_U y_i$$

donde U representa la población objetivo y y_i corresponde al valor observado de la variable de interés para la unidad i.

- **Razón poblacional**

La razón poblacional se define como

$$R = \frac{t_Y}{t_Z}$$

donde t_Y y t_Z representan los totales poblacionales de las variables de interés Y y Z, respectivamente.

- **Promedio poblacional**

El promedio poblacional está dado por

$$\bar{Y} = \frac{\sum_U y_i}{N}$$

donde N representa el tamaño de la población.

- **Proporción poblacional**

La proporción poblacional corresponde al promedio de una variable dicotómica y_i, definida como

$$P = \frac{\sum_U y_i}{N}$$

donde

$$y_i = \begin{cases} 1, & \text{si la unidad presenta el atributo de interés} \\ 0, & \text{en otro caso} \end{cases}$$

11.2 Estimadores

Para la estimación de los parámetros poblacionales se utilizará el estimador de Horvitz-Thompson, el cual produce estimaciones insesgadas bajo diseños probabilísticos de muestreo.

- **Estimador insesgado del total**

Para estimar el total poblacional se emplea el estimador de Horvitz-Thompson (Horvitz y Thompson, 1952; Särndal et al., 1992). Para una variable de interés y, el estimador del total poblacional está dado por:

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \hat{t}_{y\pi,h} = \sum_{h=1}^{H} \sum_{i \in S_{Ih}} \frac{\hat{t}_{hi}}{\pi_{Ihi}}$$

Con

$$\hat{t}_{hi} = \sum_{\mathrm{j} \in S_{\mathrm{IIhi}}} \frac{\hat{t}_{hij}}{\pi_{IIhij}}$$

Y

$$\hat{t}_{hij} = \sum_{\mathrm{k} \in S_{\mathrm{IIIhij}}} \frac{y_{\mathrm{hijk}}}{\pi_{\mathrm{III\ hijk}}}$$

donde:

- h: índice asociado al estrato de muestreo (localidad).
- H: número total de localidades ($H = 20$).
- S_{Ih}: muestra de Unidades Primarias de Muestreo (UPM) seleccionadas en la localidad h.
- π_{Ihi}: probabilidad de selección de la UPM i en la localidad h.
- S_{IIhi}: muestra de viviendas seleccionadas en la i-ésima UPM de la localidad h.
- π_{IIhij}: probabilidad de selección de la vivienda j, ubicada en la i-ésima UPM de la localidad h.
- S_{IIIhij}: muestra de personas seleccionadas en la vivienda j de la i-ésima UPM de la localidad h.
- $\pi_{IIIhijk}$: probabilidad de selección de la persona k, perteneciente a la vivienda j de la i-ésima UPM de la localidad h.
- y_{hijk}: valor observado de la variable de interés en la persona k, perteneciente a la vivienda j de la i-ésima UPM de la localidad h.
- $\hat{t}_{y\pi}$: estimador del total poblacional de la variable de interés para el universo de estudio (Bogotá, D.C.).
- $\hat{t}_{y\pi,h}$: estimador del total de la variable de interés en la localidad h.
- $\hat{t}_{hi}$: estimador del total de la variable de interés en la UPM $i$de la localidad h.
- $\hat{t}_{hij}$: estimador del total de la variable de interés en la vivienda $j$de la UPM i de la localidad h.

- **Estimador insesgado de la media poblacional**

A partir de los estimadores del total poblacional se construyen los estimadores de la media poblacional para cada nivel de desagregación, definidos de la siguiente manera:

Desagregación	Estimador
General	$\hat{\bar{y}}_{U,\pi} = \frac{\hat{t}_{y\pi}}{N}$
Estrato	$\hat{\bar{y}}_{U_h,\pi} = \frac{\hat{t}_{y\pi,h}}{N_h}$

donde:

- $\hat{t}_{y\pi}$, $\hat{t}_{y\pi,h}$ corresponden a los estimadores del total poblacional definidos previamente;
- N y N_h representan los tamaños poblacionales del universo y de la localidad h respectivamente.
- $\hat{\bar{y}}_{U,\pi}$ corresponde al estimador de la media poblacional de la variable de interés y para el universo completo U.
- $\hat{\bar{y}}_{U_h,\pi}$ corresponde al estimador de la media poblacional de la variable de interés y en la localidad h.

- **Estimador insesgado de la proporción poblacional**

Cuando la variable de interés es dicotómica, tomando valores ($y = 1$) si la unidad presenta la característica de interés y ($y = 0$) en caso contrario, la media poblacional coincide con la proporción poblacional. En consecuencia, el estimador de la proporción está dado por:

Desagregación	Estimador
General	$\hat{\mathrm{p}}_{U,\pi} = \frac{\hat{t}_{y\pi}}{N}$
Estrato	$\hat{\mathrm{p}}_{U_h,\pi} = \frac{\hat{t}_{y\pi,h}}{N_h}$

donde:

- $\hat{\mathrm{p}}_{U,\pi}$ es el estimador de la proporción poblacional para el universo de estudio.
- $\hat{\mathrm{p}}_{U_h,\pi}$ es el estimador de la proporción poblacional para la Localidad h.

- **Estimador aproximadamente insesgado de una razón poblacional**

Para la estimación de razones poblacionales se utilizará el cociente entre los estimadores de los totales poblacionales de las variables y y z. Las expresiones correspondientes para cada nivel de desagregación son las siguientes:

Desagregación	Estimador
General	$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$
Estrato	$\hat{R}_h = \frac{\hat{t}_{y\pi,h}}{\hat{t}_{z\pi,h}}$

donde:

- $\hat{R}$ corresponde al estimador de la razón poblacional para el universo U;
- $\hat{R}_h$ corresponde al estimador de la razón poblacional en la Localidad h;
- $\hat{t}_{y\pi}$ y $\hat{t}_{z\pi}$ corresponden a los estimadores de los totales poblacionales de las variables y y z para el universo completo U, respectivamente;
- $\hat{t}_{y\pi,h}$ y $\hat{t}_{z\pi,h}$ corresponden a los estimadores de los totales poblacionales de las variables y y z en la Localidad h, respectivamente.

11.3 Estimación de la varianza

La precisión de las estimaciones se evaluará mediante procedimientos de estimación de varianza consistentes con el diseño probabilístico estratificado en tres etapas. Las estimaciones de varianza, errores estándar, coeficientes de variación e intervalos de confianza se obtendrán mediante técnicas de linealización de Taylor implementadas en software especializado para análisis de encuestas complejas, incorporando la estratificación por Localidad, las Unidades Primarias de Muestreo (UPM) y los factores de expansión derivados del diseño.

- **Varianza del estimador de un total**

La varianza del estimador del total poblacional se estima como la suma de las contribuciones de varianza de cada estrato:

$$\hat{V}(\hat{t}_{y\pi}) = \sum_{h=1}^{H} \hat{V}_h\left(\hat{t}_{y\pi,h}\right)$$

siendo

$$\hat{V}_h\left(\hat{t}_{y\pi,h}\right) = \frac{m_h}{m_h - 1} \sum_{i \in S_{Ih}} (\hat{z}_{hi} - \hat{\bar{z}}_h)^2$$

donde

$$\hat{z}_{hi} = \hat{t}_{hi}/\pi_{Ihi}$$

Y

$$\hat{\bar{z}}_h = \frac{1}{m_h} \sum_{i \in S_{Ih}} \hat{z}_{hi}$$

En consecuencia, la expresión también puede escribirse como:

$$\hat{V}_h(\hat{t}_{y\pi,h}) = \frac{m_h}{m_h - 1} \sum_{i \in S_{Ih}} \left(\frac{\hat{t}_{hi}}{\pi_{Ihi}} - \hat{\bar{z}}_h \right)^2$$

donde:

- m_h: número de manzanas seleccionadas en la Localidad h.
- S_{Ih}: muestra de manzanas seleccionadas en la Localidad h.
- π_{Ihi}: probabilidad de selección de la manzana i en la Localidad h.
- $\hat{t}_{hi}$: estimador del total de la variable de interés en la manzana i de la Localidad h.
- $\hat{z}_{hi}$: total estimado expandido asociado a la manzana i.
- $\hat{\bar{z}}_h$: promedio de los totales estimados expandidos de las manzanas seleccionadas en el estrato h.

- **Estimador de la varianza del estimador de la media**

Dado que la media poblacional se obtiene dividiendo el estimador del total entre el tamaño poblacional conocido, la varianza del estimador de la media para el universo de estudio se estima como:

$$\hat{V}(\hat{\bar{y}}_{U,\pi}) = \frac{1}{N^2} \hat{V}(\hat{t}_{y\pi})$$

Para la localidad h, la varianza se estima como:

$$\hat{V}(\hat{\bar{y}}_{U_h,\pi}) = \frac{1}{N_h^2} \hat{V}(\hat{t}_{y\pi,h})$$

donde:

N: tamaño poblacional del universo de estudio.
N_h: tamaño poblacional de la Localidad h.

- **Estimador de la varianza del estimador de una proporción**

La proporción poblacional corresponde a la media de una variable dicotómica, donde y toma el valor 1 si la persona presenta la característica de interés y 0 en caso contrario. Por tanto, la varianza del estimador de una proporción se estima de forma análoga a la varianza de la media.
Para el universo de estudio:

$$\hat{V}(\hat{P}_{U,\pi}) = \frac{1}{N^2}\hat{V}(\hat{t}_{y\pi})$$

Para la Localidad h, la varianza del estimador de una proporción se estima como:

$$\hat{V}(\hat{P}_{U_h,\pi}) = \frac{1}{N_h^2}\hat{V}(\hat{t}_{y\pi,h})$$

- **Estimador de la varianza del estimador de una razón**

Para estimar la varianza de una razón se utiliza la linealización de Taylor. Sea la razón estimada:

$$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}}$$

Se define la variable linealizada:

$$e_k = y_k - \hat{R}z_k$$

La varianza del estimador de razón se aproxima mediante:

$$\hat{V}(\hat{R}) = \frac{1}{(\hat{t}_{z\pi})^2}\hat{V}(\hat{t}_{e\pi})$$

donde:

- $\hat{R}$: estimador de la razón entre los totales de las variables y y z.
- $\hat{t}_{y\pi}$: estimador del total poblacional de la variable y.
- $\hat{t}_{z\pi}$: estimador del total poblacional de la variable z.
- e_k: variable linealizada asociada a la razón.
- $\hat{t}_{e\pi}$: estimador del total poblacional de la variable linealizada e_k.
- $\hat{V}(\hat{t}_{e\pi})$: varianza estimada del total de la variable linealizada, calculada con la misma expresión utilizada para el estimador del total.

11.4 Implementación computacional

Las expresiones anteriores corresponden a la aproximación por linealización de Taylor de primer orden, la cual constituye el procedimiento estándar para la estimación de varianzas en encuestas por muestreo con diseños probabilísticos complejos. En el presente estudio, la implementación computacional se realizará mediante el paquete *survey* del software R, especificando los estratos correspondientes a las Localidades, las Unidades Primarias de Muestreo (UPM) y los factores de expansión derivados del diseño, obteniendo así estimaciones consistentes de errores estándar, coeficientes de variación e intervalos de confianza para todos los parámetros de interés.

12 Ajuste por no respuesta

En las operaciones estadísticas basadas en muestreo probabilístico es posible que algunas unidades seleccionadas no proporcionen información debido a rechazos, ausencia de los residentes, imposibilidad de contacto u otras causas operativas. Estas situaciones generan no respuesta y pueden introducir sesgos en las estimaciones si no se aplican los ajustes correspondientes. Con el fin de preservar la representatividad de la muestra, los pesos básicos de diseño obtenidos a partir de las probabilidades de inclusión podrán ajustarse para compensar la no respuesta observada durante el operativo de campo.

El ajuste se realizará dentro de grupos homogéneos de respuesta, definidos a partir de variables geográficas o de diseño que presenten comportamientos similares frente a la probabilidad de respuesta. Para cada grupo de ajuste g, el factor de ajuste por no respuesta se define como:

$$A_g = \frac{n_g}{r_g}$$

donde:

- n_g: número de unidades seleccionadas en el grupo de ajuste g.
- r_g: número de unidades que respondieron efectivamente dentro del grupo de ajuste g.

El peso ajustado por no respuesta para la persona k, seleccionada en la vivienda j, de la Unidad Primaria de Muestreo (UPM) i, perteneciente a la localidad h, se obtiene mediante:

$$d_{hijk}^{NR} = d_{hijk} \times A_g$$

donde:

- d_{hijk}: peso básico de diseño de la persona k, seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.
- A_g: factor de ajuste por no respuesta del grupo de ajuste g.
- d_{hijk}^{NR}: peso ajustado por no respuesta de la persona k, seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.

Este procedimiento redistribuye el peso de las unidades no respondientes entre las unidades respondientes pertenecientes al mismo grupo de ajuste, manteniendo la estructura del diseño muestral dentro de cada grupo.

Antes de aplicar el ajuste, el operativo de campo implementará los protocolos establecidos para minimizar la no respuesta, incluyendo revisitas, verificación de direcciones y seguimiento a los casos pendientes, con el propósito de maximizar la tasa de respuesta y reducir los posibles sesgos asociados a la pérdida de unidades seleccionadas.

13 Calibración de los factores de expansión

Una vez ajustados los pesos por no respuesta, estos podrán ser calibrados utilizando información auxiliar proveniente de fuentes oficiales, con el propósito de mejorar la precisión de las estimaciones y garantizar la coherencia entre las estimaciones muestrales y los totales poblacionales conocidos.

La calibración consiste en modificar los pesos ajustados por no respuesta de manera que los estimadores reproduzcan determinados totales auxiliares conocidos, manteniendo al mismo tiempo la mayor cercanía posible respecto a los pesos ajustados por no respuesta.

Sea x_{hijk} el vector de variables auxiliares observado para la persona k, seleccionada en la vivienda j, de la Unidad Primaria de Muestreo (UPM) i, perteneciente a la localidad h, y sea X el vector de totales poblacionales conocidos de dichas variables. Los pesos calibrados w_{hijk} satisfacen la siguiente restricción:

$$X = \sum_{h=1}^{H} \sum_{i \in S_{Ih}} \sum_{j \in S_{IIhi}} \sum_{k \in S_{IIIhij}} w_{hijk}\, x_{hijk}$$

donde:

- w_{hijk}: peso final calibrado de la persona k, seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.
- x_{hijk}: vector de variables auxiliares observado para la persona k, seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.
- X: vector de totales poblacionales conocidos de las variables auxiliares.
- S_{Ih}: muestra de UPM seleccionadas en la localidad h.
- S_{IIhi}: muestra de viviendas seleccionadas en la UPM i, perteneciente a la localidad h.
- S_{IIIhij}: persona seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.

En términos generales, los pesos calibrados pueden expresarse como:

$$w_{hijk} = d_{hijk}^{NR} \times g_{hijk}$$

donde:

- d_{hijk}^{NR}: peso ajustado por no respuesta de la persona k, seleccionada en la vivienda j, de la UPM i, perteneciente a la localidad h.
- g_{hijk}: factor de calibración obtenido mediante el algoritmo de calibración seleccionado.
- w_{hijk}: peso final calibrado utilizado en la estimación.

La implementación de la calibración se realizará mediante procedimientos de calibración generalizada asociados al estimador de regresión generalizada (*Generalized Regression Estimation*, GREG), disponibles en software especializado para el análisis de encuestas complejas. En particular, podrá implementarse mediante el paquete survey del software R.

Los pesos calibrados constituyen los pesos finales utilizados para la estimación de los parámetros poblacionales. Su aplicación permite reducir los posibles sesgos asociados a la no respuesta y mejorar la eficiencia de las estimaciones mediante el aprovechamiento de información auxiliar confiable.

14 Bibliografía

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). John Wiley & Sons.
Deville, J. C., & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference, 128*(2), 569–591. https://doi.org/10.1016/j.jspi.2003.11.011
Gutiérrez-Rojas, J. A. (2016). *Teoría del muestreo: Diseño y análisis de encuestas*. Universidad Nacional de Colombia.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685. https://doi.org/10.2307/2280784
Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.
Lumley, T. (2021). *survey: Analysis of Complex Survey Samples* (R package version 4.x). https://cran.r-project.org/package=survey
Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.
Tillé, Y. (2006). *Sampling Algorithms*. Springer.
United Nations. (2009). *Designing Household Survey Samples: Practical Guidelines*. Department of Economic and Social Affairs, Statistics Division (Studies in Methods, Series F, No. 98).
Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer.