Guía de Diseño de la Muestra para Encuestas

Fase de Diseño de la Norma Técnica del Proceso de Producción de Información Estadística y Geográfica

ÍNDICE

PREFACIO

Con fundamento en el Artículo 14, fracción IV y V de la Norma Técnica del Proceso de Producción de Información Estadística y Geográfica (NTPPIEG) para el Instituto Nacional de Estadística y Geografía (INEGI), se presenta a las Unidades Administrativas, la Guía de Diseño de la Muestra para Encuestas, cuyos propósitos principales son: proporcionar los aspectos generales[1] de las actividades para elaborar un diseño muestral y los principales criterios para observar la calidad de la información.

[1] Artículo 58. Ley del Sistema Nacional de Información Estadística y Geográfica (LSNIEG).

INTRODUCCIÓN

A. Presentación

El INEGI, como coordinador del SNIEG, juega un papel muy importante, ya que al suministrar información de calidad, pertinente, veraz y oportuna[2], se permite que las políticas públicas sean evaluables y sus resultados comparables, considerando los principios constitucionales: accesibilidad, transparencia, y objetividad. La Ley del Sistema Nacional de Información Estadística y Geográfica (LSNIEG) le da la atribución de realizar las acciones tendientes a lograr la adecuación conceptual de la Información de Interés Nacional a las necesidades que el desarrollo económico y social del país impongan, que la Información sea comparable en el tiempo y en el espacio, y la adecuación de los procedimientos estadísticos y geográficos a estándares internacionales para facilitar su comparación.

Para cumplir con el mandato constitucional y la ley del SNIEG, el Instituto genera estadística básica, la cual obtiene de tres tipos de fuentes: censos, encuestas y registros administrativos, así como estadística derivada, mediante la cual produce indicadores demográficos, sociales y económicos, además de contabilidad nacional. Para la generación de esta información ha adoptado el Modelo del Proceso Estadístico y Geográfico (MPEG) el cual se encuentra regulado por la NTPPIEG[3]. Dentro de este modelo, en la Fase de Diseño se encuentra el subproceso de Diseño de la Muestra, el cual comprende la determinación de la población objeto de estudio, el marco muestral y el tipo de muestreo, cuando este aplique, mediante el cual se verifica la cobertura del marco a usar con respecto a la población objeto de estudio y el alcance temático, es decir, que sea completo y actualizado. Además de determinar el diseño de muestreo en el caso de datos recabados por esta vía[4], este conjunto de actividades impacta tanto en la calidad como en la confiabilidad y oportunidad de los resultados.

Con el fin de facilitar la realización de las actividades y la integración de las evidencias a que hace referencia el Capítulo III de la NTPPIEG, se pone a disposición de las Unidades Administrativas que desarrollan actividades para producir información estadística y geográfica, la **Guía de Diseño de la Muestra para Encuestas**. Esta guía tiene el propósito de ofrecer un marco de referencia para las encuestas que se generan en el Instituto, describe los principales elementos técnicos a considerar durante la toma de decisiones que involucra este subproceso, las actividades generales para la conformación del marco de muestreo (y su implementación con los marcos múltiples y maestros), las características para la definición del tipo de muestreo a implementar, así como la distribución de la muestra en los estratos. También incluye una sección sobre la definición de estimadores habituales. Todos estos son elementos de apoyo y herramientas para la realización de la actividad, según las necesidades o particularidades de cada encuesta.

---

2 Ley del Sistema Nacional de Información Estadística y Geográfica, México. 16 de abril de 2008. https://snieg.mx/contenidos/espanol/normatividad/marcojuridico/LSNIEG.pdf

3 La Norma Técnica del Proceso de Producción de Información Estadística y Geográfica (NTPPIEG) para el INEGI fue aprobada mediante el Acuerdo No. 8ª/IX/2018, de la Octava Sesión de la Junta de Gobierno, celebrada el 29 de agosto de 2018 y publicada en la Normateca Institucional el 5 de septiembre del mismo año; sus modificaciones más recientes fueron publicadas en la referida Normateca el 19 de noviembre de 2020.

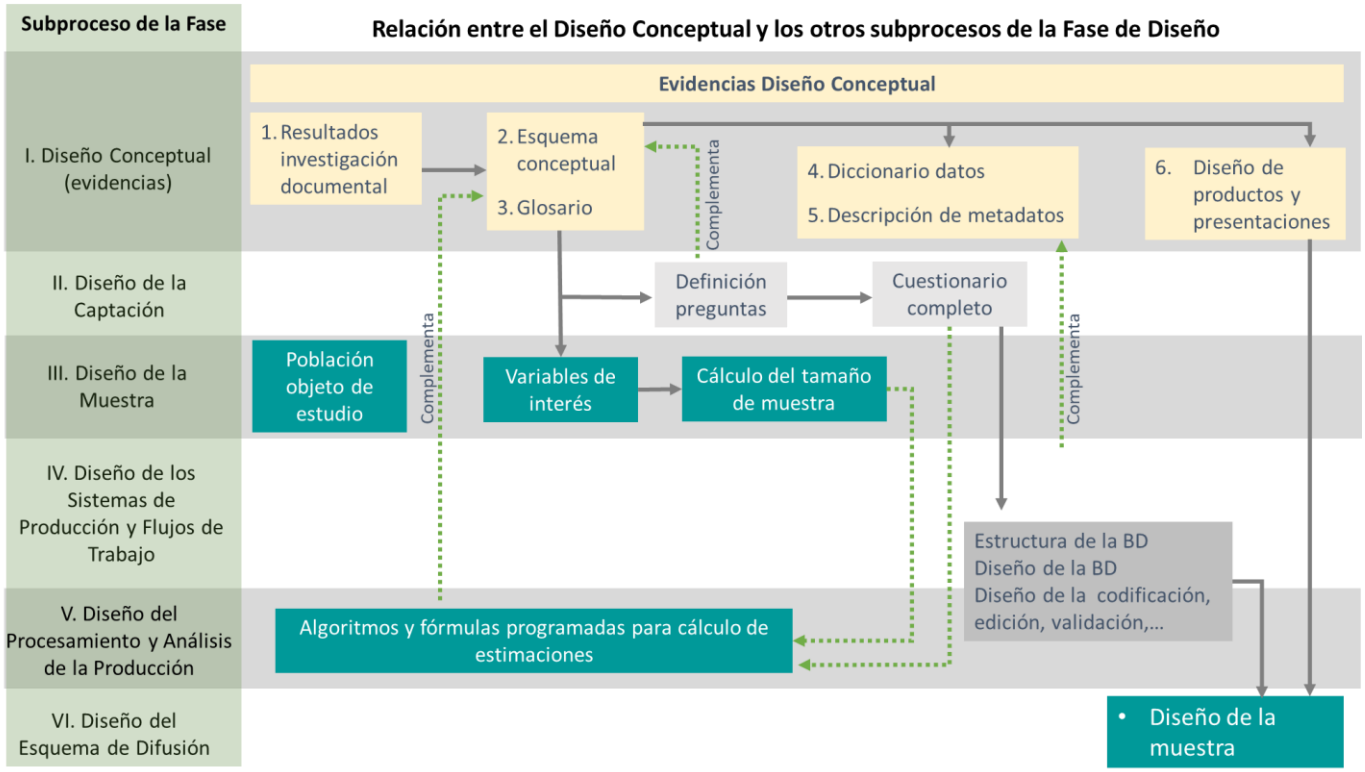
4 Artículo 14. Fracción IV y V. NTPPIEG.

B. Aspectos generales sobre el subproceso de Diseño de la Muestra

El Diseño de la Muestra es un subproceso de la Fase de Diseño en la que se define el esquema de muestreo a utilizar, se determina el tamaño y procedimiento de selección y distribución de la muestra y, en el caso del muestreo probabilístico, se determinan los ponderadores y estimadores que se requieren para la generación de resultados en un Programa de Información basado en encuestas.

Como se detalla en la Guía de Diseño Conceptual para Encuestas, el Diseño de la Muestra interactúa con todos los subprocesos de la Fase de Diseño; particularmente, en el Diseño Conceptual se determinan los indicadores objetivo que se requieren representar o medir, los cuales, junto con la determinación de la población objeto de estudio, el marco muestral y tipo de muestreo, permiten estimar el tamaño de la muestra que se empleará en el Programa de información.

Figura B.1. Relación entre las evidencias del Diseño de la Muestra y otras actividades de la Fase de Diseño



Los subprocesos de Diseño Conceptual y Diseño de la Muestra interactúan en doble sentido, debido a que las coberturas conceptual y geográfica influyen en las decisiones sobre el esquema de muestreo, y éste a su vez puede implicar el ajuste del desglose conceptual y geográfico, debido al límite del presupuesto. Además, es en el Diseño Conceptual en donde se define, en su caso, el diseño de los indicadores necesarios para inferir los resultados de la población objeto de estudio. Asimismo, la interacción con las demás fases de la NTPPIEG se da, en lo general, de la siguiente forma:

La Fase de Documentación de las Necesidades interactúa como condicionante del Diseño de la Muestra debido a que define el alcance de los objetivos del Programa en cuanto a cobertura temática y geográfica, así como las restricciones impuestas por el presupuesto disponible.

La Fase de Captación también interactúa en dos sentidos con el Diseño de la Muestra, dado que la estrategia para el Diseño de Captación implica considerar la distribución y dispersión geográfica de la muestra, en tanto que, en el sentido inverso, el Diseño de la Muestra requiere considerar las opciones idóneas para la captación de los datos, dadas las características del contexto donde se ejecutará el Programa.

Finalmente, el Diseño de la Muestra se relaciona con las Fases de Procesamiento y Análisis de la Producción mediante la entrega de ponderadores, estimaciones e indicadores de precisión estadística correspondientes a los resultados que se obtengan de la Fase de Captación.

Debido a que el subproceso de Diseño de la Muestra de la Fase de Diseño en un Programa específico debe responder a decisiones tomadas en las otras fases de la NTPPIEG, es necesario identificar aspectos que actúan como condicionantes que influyen en las decisiones sobre la definición del esquema de muestreo, los cuales se describen a continuación:

- De la Fase de Documentación de las Necesidades: la temática de la información a generar; la población objeto de estudio; los dominios de estudio; la cobertura y desglose geográfico; referencia temporal; periodicidad de la captación de datos y otras desagregaciones relevantes; así como la unidad de muestreo ya que son fundamentales en el Diseño de la Muestra.

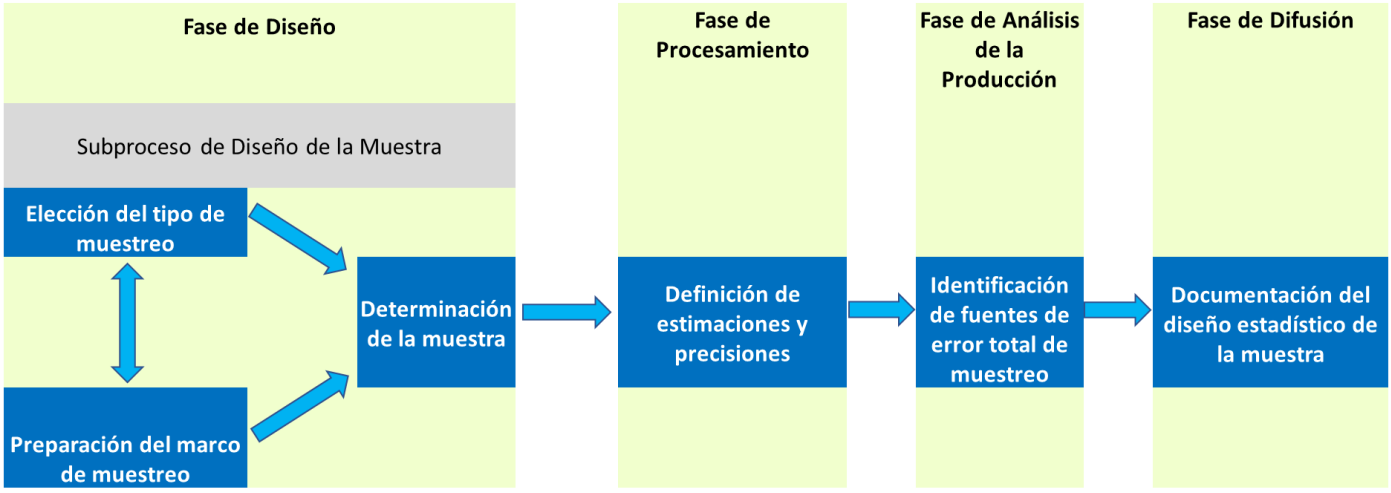
- De la disponibilidad o estructura del marco de muestreo: disponer o no de un marco muestral, así como de su estructura, influye en la determinación del esquema de muestreo, pues de optarse por el muestreo probabilístico es necesario realizar la construcción o actualización del marco; en tanto que de no ser factible disponer de él, o bien, si las unidades muestrales son desiguales, es decir, algunas de ellas contienen la mayor parte de la información a estudiar, entonces se debe utilizar un esquema no probabilístico. Asimismo, determinadas características del marco muestral en un contexto geográfico amplio pueden implicar un esquema de selección de la muestra de varias etapas, partiendo de divisiones territoriales grandes hasta pequeñas áreas de selección.

- De los recursos disponibles: en toda encuesta por muestreo se enfrenta el reto de obtener estadísticas con un aceptable grado de precisión en combinación con un presupuesto razonable, en la perspectiva del costo-beneficio. Existen aspectos asociados con el presupuesto y duración del evento que influyen en el cálculo del tamaño de la muestra derivado del esquema de muestreo elegido. Por ejemplo, una muestra no probabilística puede ser menos costosa, pero no se puede evaluar el error de muestreo ni tampoco se puede generalizar a toda la población objeto de estudio; por otra parte, con una muestra probabilística se requerirá de un presupuesto mayor, pero será posible conocer el grado de precisión de las estimaciones y obtener conclusiones que se generalicen hacia toda la población objeto de estudio.

- Del subproceso de Diseño Conceptual de la Fase de Diseño: la cantidad de variables a estudiar, la frecuencia con la que la característica o fenómeno se presenta en la población objeto de estudio y la cobertura de los valores, son factores que inciden en la determinación del tamaño de la muestra que permiten obtener la representatividad deseada en las estimaciones correspondientes.

El Diseño de la Muestra se desarrolla en una secuencia de actividades en las cuales es necesario optar por alternativas idóneas conforme al análisis de condicionantes correspondientes. Estas actividades se planean y definen para que en las fases subsecuentes estén completamente precisadas para su implementación. En la figura B.2 se describen las fases de la NTPPIEG involucradas en el Diseño de la Muestra.

Figura B.2. Fases de la NTPPIEG involucradas en el Diseño de la Muestra



ACTIVIDAD 1: PREPARACIÓN DEL MARCO DE MUESTREO PARA MUESTREO PROBABILÍSTICO

Esta actividad es esencial y se refiere a la detección, evaluación, organización y elaboración de los listados, directorios o materiales cartográficos para la identificación de todas las unidades de la población objeto de estudio.

La población definida a través del marco de muestreo debe contener a toda la población objeto de estudio, de lo contrario la encuesta aportará resultados poco confiables. Un marco de muestreo debe, entonces, tener propiedades relacionadas con la calidad, con la eficiencia y el costo.

En la práctica, se dan distintas situaciones en relación con la disponibilidad de un marco de muestreo, cada una con implicaciones en cuanto a recursos a utilizarse. Se destacan los siguientes estados de un marco de muestreo:

- Se dispone de todo el marco al inicio del estudio
- El marco es parcial y debe actualizarse
- No existe el marco y debe construirse

El estado ideal de un marco muestral se tiene cuando está actualizado y las imperfecciones se han eliminado o reducido. En las dos siguientes secciones se describen los componentes involucrados en un marco de muestreo y las imperfecciones que puede contener.

1.1 El marco de muestreo

El marco de muestreo denotado por M es el listado en el cual se busca identificar a todos los elementos de una población objeto de estudio y que permite seleccionar una muestra de esta, con fines de estimación estadística. Se deben identificar los elementos que se involucran en el marco de muestreo. La población objeto de estudio se denota como U. La población objeto de estudio listada en el marco de muestreo se etiqueta como U_M. La unión de estas dos poblaciones define el conjunto completo de elementos C, el cual se compone de tres partes:

I. Los elementos de la población objeto de estudio que están listados en el marco de muestreo,

$$U_L = U \cap U_M$$

II. Los elementos de la población objeto de estudio que no están listados en el marco de muestreo,

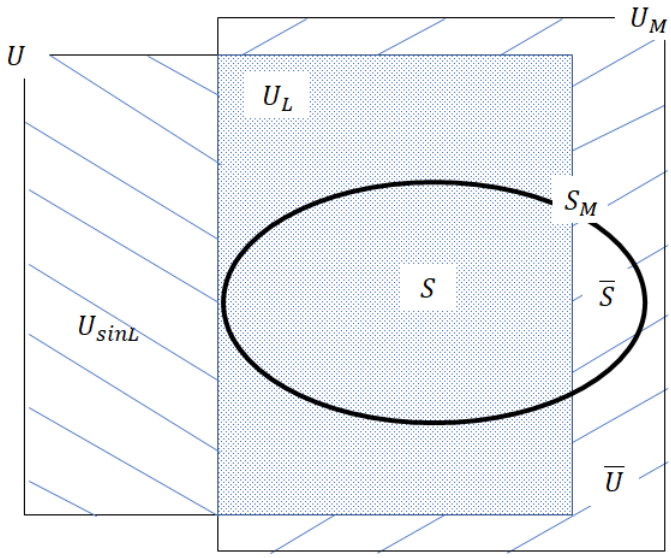
$$U_{sinL} = U - U_M$$

III. Los elementos que no pertenecen a la población objeto de estudio, pero están listados en el marco de muestreo,

$$\overline{U} = U_M - U_L$$

La muestra se selecciona del marco de muestreo M con un diseño muestral específico. Las unidades en la muestra están listadas en s_M, donde s_M es un subconjunto de U_M. Algunos elementos en s_M son parte de la población objeto de estudio, es decir, el conjunto $s = s_M \cap U$. La diferencia, $\overline{s} = s_M - s$, contiene a los elementos que no pertenecen a la población objeto de estudio; dichos elementos en $\bar{s}$ pueden ya no existir y en caso de existir, no son de interés para la encuesta (ver figura 1.1).

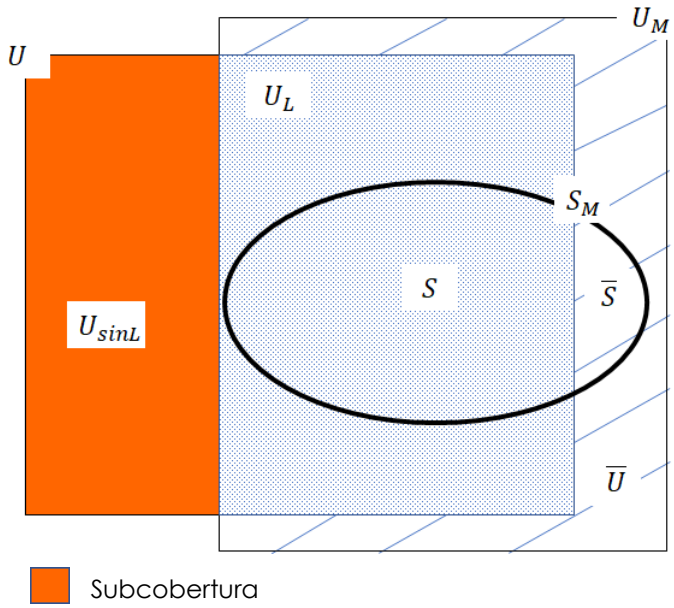
Figura 1.1. Población *objeto de estudio y marco de muestreo*



Existen 3 tipos de imperfecciones del marco de muestreo:

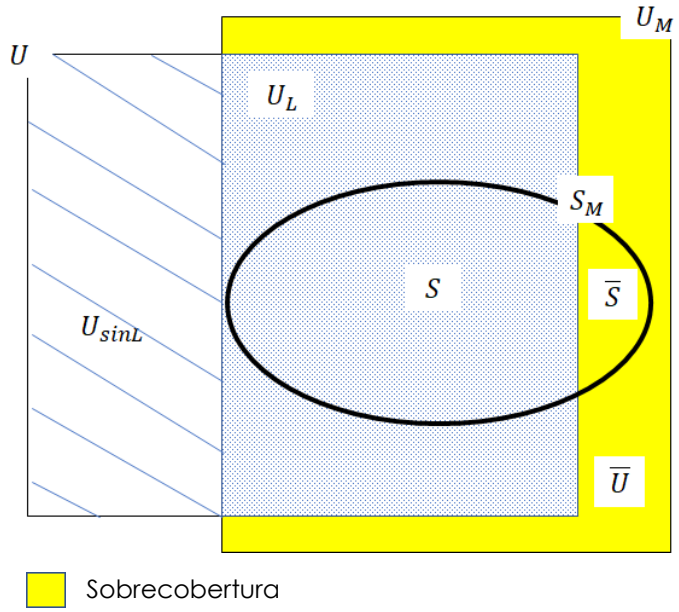
a. Algunos elementos de la población objeto de estudio no están listados y por lo tanto no pueden ser seleccionados en la muestra s. Esto es, U_{sinL} no está vacío (subcobertura en figura 1.2).

Figura 1.2. Subcobertura en el marco de muestreo



b. El conjunto $\overline{U}$ no está vacío, algunos elementos en el marco de muestreo no pertenecen a la población objeto de estudio (sobrecobertura en figura 1.3).

Figura 1.3. Sobrecobertura en el marco de muestreo



c. En el marco de muestreo hay elementos que están listados en más de una ocasión (registros duplicados).

EL Comité de Aseguramiento de la Calidad (CoAC) del INEGI aprobó el indicador Tasa de Sobrecobertura a nivel unidad de observación (TSC)[5] que se define como el porcentaje de unidades de observación que contiene el marco de referencia que no corresponde a las unidades de la población objeto de estudio. Para el cálculo de este porcentaje se asume que se puede distinguir en el marco de referencia cuando hay población en éste que no pertenece a la población objeto de estudio, por ejemplo, después de la Fase de Captación y de la Fase de Procesamiento se puede definir este aspecto sobre las unidades del marco que fueron observadas (todo el marco o bien la muestra que se seleccionó de éste). En el caso de muestreo probabilístico o de un muestreo no probabilístico basado en el tamaño de alguna variable de diseño también se deberá calcular la versión ponderada (ver sección 4.2) de este indicador.

1.2 Construcción del marco de muestreo y mantenimiento

La siguiente lista describe las etapas para la construcción y mantenimiento de un marco muestral:

Etapa 1. Identificación de los elementos: deben tomarse en cuenta los siguientes factores, (i) el costo que involucra el establecimiento y mantenimiento, incluyendo la información auxiliar, (ii) la disponibilidad de la información requerida para cada elemento en el marco, (iii) la estabilidad de los elementos del marco en un periodo y (iv) el tiempo necesario para construir el marco.

[5] https://extranet.inegi.org.mx/calidad/indicadores-de-calidad-y-evaluaciones/

Etapa 2. Desarrollo: consiste en la construcción de una base de datos después de un proceso de investigación, recolección, estandarización y organización de la información requerida para los elementos del marco.

Etapa 3. Validación: consiste en la evaluación de la calidad y la cobertura alcanzada en el marco obtenido después de las Etapas 1 y 2.

Etapa 4. Administración: debe definir procedimientos para preservar la calidad del marco de muestreo. Si se utilizará para futuras encuestas, también se debe considerar las necesidades de otros usuarios del marco.

Etapa 5. Mantenimiento: involucra las modificaciones y actualizaciones requeridas. Duplicados y "muertes" deben ser removidos, los "nacimientos" incorporados y la información auxiliar actualizada. Mantener actualizado un marco de muestreo puede involucrar un alto costo.

La información que debe contener un marco de muestreo son los datos relacionados con:

a. Identificación*: se emplea para identificar de manera única a cada elemento en el marco.
b. Contacto*: sirve para localizar a los elementos seleccionados en la muestra durante el proceso de captación de datos.
c. Clasificación: esta información es útil para la inclusión en la muestra y posiblemente para hacer estimaciones para la muestra, por ejemplo, una medida de tamaño (número de empleados en una empresa, ingresos monetarios totales de una empresa), divisiones geográficas, estratificación, entre otros.
d. Mantenimiento: esta información se necesita en caso de que la encuesta se repita en el futuro, por ejemplo, fechas de agregación de registro o cambios en el marco

*Datos mínimos requeridos en el marco de muestreo.

1.3 Marcos muestrales múltiples

Las encuestas de marcos múltiples pueden mejorar en gran medida la eficiencia y reducir el sesgo de cobertura insuficiente.

Los marcos múltiples deben adoptar un diseño de encuesta muestral que combine el marco de área con un marco lista como apoyo para evitar la inestabilidad de las estimaciones. El papel del marco de área en el enfoque de marcos múltiples es esencialmente resolver los problemas relacionados con la falta de completez del marco.

Dentro de las recomendaciones internacionales se comenta la importancia de tener clasificadas las unidades muestrales en todos los marcos e identificar cuales se localizan en la intersección de ellos, ya que este método de marcos múltiples es sensible a una mala clasificación provocando sesgos en la estimación. Así también, cuidar que las muestras en los diferentes marcos tengan un mismo cuestionario y una misma administración, ya que esto puede provocar sesgos en la estimación.

Se asume que existen dos o más marcos de muestreo, cada uno de ellos con un cierto nivel de subcobertura, pero su unión puede dar una cobertura más completa de la población objeto de estudio. Se asume lo siguiente:

- Cada elemento k en la población objeto de estudio U está en al menos uno de los dos marcos U_A y U_B.
- Una muestra probabilística se toma de cada marco s_A de U_A y s_B de U_B.
- La pertenencia a cada marco puede definirse para cada elemento muestreado.

Se distinguen tres dominios no traslapados de la población objeto de estudio U:

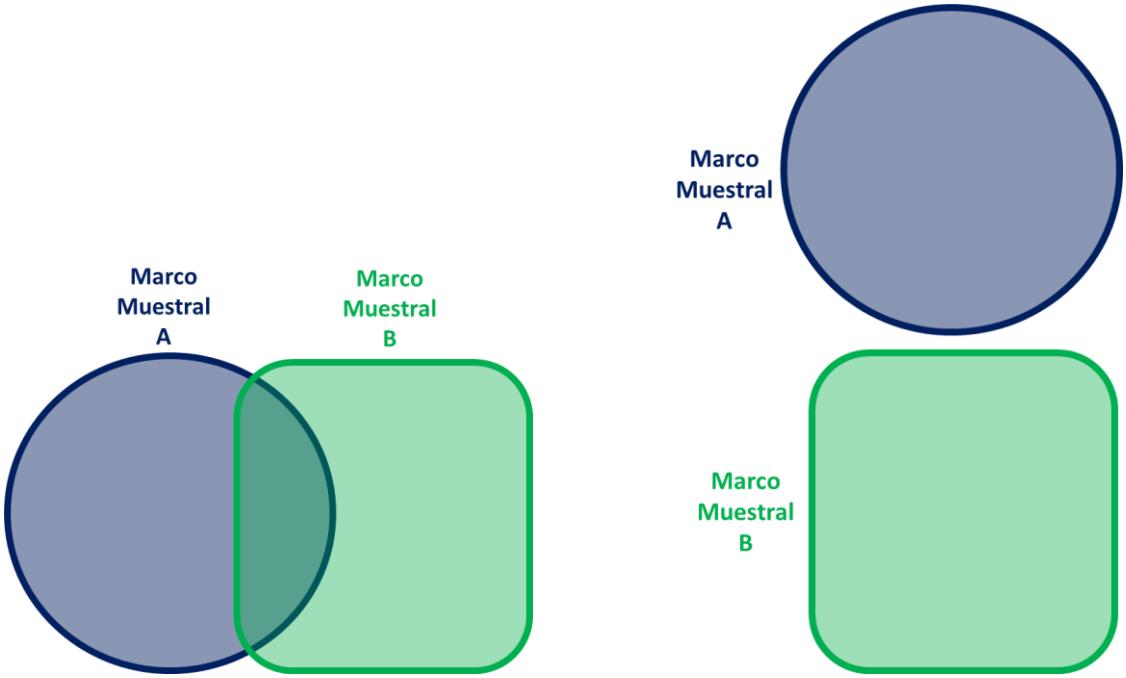
- Dominio A, lo componen los elementos solamente en U_A, es decir, $\mathrm{A} = \mathrm{U} - U_B$
- Dominio B, lo componen los elementos solamente en U_B, es decir, $\mathrm{B} = \mathrm{U} - U_A$
- Dominio AB, lo componen los elementos tanto en U_A como en U_B, es decir, $\mathrm{AB} = U_A \cap U_B$

Bajo estos supuestos, se tiene que $U = A \cup B \cup AB$, y el estimador de, por ejemplo, un total t se obtiene como

$$\hat{t} = \hat{t}_A + \hat{t}_B + \hat{t}_{AB}$$

donde $\hat{t}, \hat{t}_A, \hat{t}_B, \hat{t}_{AB}$ corresponden a los estimadores del total para la población objeto de estudio, domino A, dominio B y dominio AB, respectivamente (ver figura 1.4). Cuando los marcos muestrales son ajenos, el Dominio AB es vacío y por lo tanto el total $\hat{t}_{AB}$ es cero.

Figura 1.4. Marcos muestrales múltiples



a) Marcos muestrales con intersección b) Marco muestrales ajenos

En la figura 1.5 se presenta como ejemplo de marcos de muestreo múltiples a la Encuesta Nacional Agropecuaria (ENA). De acuerdo con las síntesis metodológicas, los marcos contienen dos registros distintos de unidades económicas, empresas o unidades de producción.

Figura 1.5. Ejemplos de Marcos de muestreo múltiples

Programa de Información	Método captación de datos	Marcos de muestreo	Elementos en el marco de muestreo	Tamaño
Encuesta Nacional Agropecuaria (ENA) 2019	Muestreo probabilístico	• Productos agrícolas: el marco derivado de la Actualización del Marco Censal Agropecuario (AMCA) 2016 actualizado con los resultados de la ENA 2017. • Productos pecuarios: el Censo Agrícola, Ganadero y Forestal 2007, actualizado con los resulta-dos de la ENA 2017	Unidades de Producción	3 783 588

1.4 Marcos de áreas

En un marco de área las unidades son áreas geográficas. La población objeto de estudio de la encuesta se encuentra dentro de estas áreas geográficas. Los marcos de área se pueden usar cuando una encuesta es de naturaleza geográfica o cuando no se dispone de un marco muestral adecuado, en cuyo caso el marco de área puede ser utilizado como vehículo para crearlo. Un marco de área puede ser una lista, mapas, fotografías áreas, imágenes de satélite o cualquier otra colección de unidades terrestres.

Hay dos tipos de marcos de área: cuadrícula y terreno. La diferencia entre una cuadrícula y terreno depende del objetivo analítico de la encuesta en lugar de la estructura del marco. Los marcos de terrenos contienen unidades de muestreo finales que se observan en su totalidad, mientras que los marcos de cuadrícula contienen unidades terrestres que serán divididos y muestreados en más etapas. Los marcos de terrenos son usados regularmente en encuestas agropecuarias y de medio ambiente. En algunas ocasiones se debe seleccionar una muestra de unidades dentro de áreas geográficas. Los conglomerados geográficos de unidades muestrales componen el marco de cuadrícula. Primero se muestrean los conglomerados geográficos, después se selecciona una muestra de unidades dentro de los conglomerados muestreados.

Los marcos de área deben cubrir toda la población objeto de estudio y dividirlo en unidades geográficas mutuamente excluyentes. Para encuestas que hacen estimaciones basadas en delimitaciones políticas, como municipios o estados, por lo general se debe hacer alguna compensación entre los límites geográficos visibles y los "invisibles" (por ejemplo, fronteras políticas). Cambios en la geografía política como anexiones, así como cambios en la geografía física, como cauce de ríos, cambio en la vegetación, árboles o carreteras, deben reflejarse en el marco de área. Si no se actualiza el marco pueden causar confusión en la recolección de datos, aumento de sesgo de cobertura y variación de cobertura.

La temporalidad está fuertemente influenciada por el cambio en el nivel de cobertura de un año a otro, provocando errores de muestreo. Un marco de área siempre está completo y permanece útil durante mucho tiempo.

1.5 Marcos de muestreo maestros (MMM)

El MMM se presenta como una lista organizada en forma de base de datos que contiene a las unidades de observación registradas en un censo que participarán en cada una de las fases de diseño, distribución y selección de la muestra de una encuesta. El MMM principalmente se emplea para identificar y seleccionar las unidades de muestreo y como base para realizar estimaciones basadas en los datos de la muestra; esto implica que la población objeto de estudio a ser seleccionada para la muestra debe estar representada de forma física, es decir, el MMM también está formado por todos los mapas y planos a diferentes escalas que permiten identificar en forma precisa y clara los límites físicos que tienen las diferentes unidades de selección, considerándose como parte principal de éste los registros y listados en los que se detalla las referencias que faciliten identificar en forma exacta las unidades de observación seleccionadas.

El Marco contiene información sobre la división político-administrativa y geográfica del país (subdivisiones políticas o zonificación estadística definida para efectuar la enumeración del censo; como también sobre los volúmenes de viviendas (u otras unidades de observación) y de población objeto de estudio total, por grupos de edad y sexo, entre otras variables necesarias para clasificar a los hogares de acuerdo con determinadas características según los objetivos específicos de cada encuesta. Todos y cada uno de los elementos de los que está compuesto el Marco tienen una probabilidad conocida y diferente de cero de ser seleccionados de alguna de las muestras que se puedan extraer del mismo.

La depuración y actualización del marco de muestreo puede realizarse con apoyo en diversos materiales disponibles y en procesos de digitalización de listados y croquis, sistemas para conformar unidades de muestreo y sistemas para la selección. Si con los materiales existentes no se puede integrar el marco, puede ser necesario

construir uno específico, sin embargo, los costos de este pueden hacer inviable el Programa de Información. En este caso se recomienda formar unidades de muestreo intermedias entre las distintas etapas de selección.

Para mejorar la eficiencia del diseño es necesario que el marco contenga información que permita separar las unidades de muestreo en estratos, formados preferentemente en función de variables correlacionadas con los indicadores objetivo. La estratificación óptima logra que los elementos que se incluyan dentro de cada estrato sean homogéneos y que haya heterogeneidad entre los estratos, así mismo, deben incluir a toda la población objeto de estudio de modo que cada unidad de muestreo pertenezca exactamente a un solo estrato.

Por otra parte, hay que considerar la existencia y posibilidad de uso de marcos de muestreo maestros sobre viviendas, unidades de producción y establecimientos, integrados y actualizados por el INEGI.

Estos tipos de marcos permiten:

- La planeación operativa de las encuestas.
- Reducir los costos operativos de las encuestas por concepto de marcos.
- Tener un mejor control de los errores de muestreo y de los ajenos al muestreo.
- Mejorar la congruencia entre las estimaciones de las encuestas.

1.6 Muestras maestras

El uso de muestras maestras tiene como propósito efectuar selecciones múltiples de unidades que permiten atender las demandas de información de encuestas. Las principales ventajas de trabajar con un diseño de muestra maestra son:

- Se dispone de una muestra maestra seleccionada a partir de criterios uniformes.
- Se usa el mismo número de etapas.
- Se utilizan las mismas unidades del marco.
- Se aplica el mismo procedimiento de cálculo para determinar las probabilidades de selección.

Idealmente, el marco de muestreo debe mantenerse actualizado, lo que implica un alto costo operativo y económico, sin embargo, es más viable actualizar la muestra maestra como una actividad previa al diseño de una encuesta. Usando esta estrategia se asume que el crecimiento observado de las viviendas seleccionadas en la muestra maestra será el mismo que las viviendas no incluidas en esta. El costo periódico de un marco de muestreo maestro o muestra maestra es muy alto. Es una práctica generalizada recurrir a las proyecciones de población objeto de estudio para generar estimaciones de los totales de personas en viviendas adicionales y omitidas (CEPAL, p.28, 2002).

El MMM 2012, está conformado por conglomerados de viviendas llamados Unidades Primarias de Muestreo (UPM) construidos a partir de la información cartográfica y demográfica que se obtuvo del Censo de Población y Vivienda 2010. La Muestra Maestra permite la selección de submuestras para todas las encuestas en viviendas que realiza el INEGI. Su diseño es probabilístico, estratificado, unietápico y por conglomerados, pues es en ellos donde se seleccionaron, en una segunda etapa, las viviendas que integran las submuestras de las diferentes encuestas; por ejemplo, la Encuesta Nacional del Uso de Tiempo (ENUT) 2019 utilizó el MMM en su diseño de la muestra.

El uso de muestras maestras también implica desventajas, por ejemplo, el agotamiento de las UPM, es decir, quedarse sin viviendas si la muestra maestra se utiliza en varias ocasiones en un periodo corto. Esta desventaja puede prevenirse con la planeación de submuestras que podrían usarse para cada encuesta, aunque depende de que el total de viviendas en cada UPM de la muestra maestra sea lo suficientemente grande. Otra desventaja, es la acumulación de sesgo cuando la actualización de la muestra maestra no se lleva a cabo.

1.7 Marcos "ab initio"

En ocasiones, un marco puede requerir que se construya "ab initio", en estos casos, debe describirse el método de enmienda o de construcción (Kish, p. 77 – 84, 1975). Una vez definido el universo, se debe recabar información, lo más exacta posible, de sus dimensiones y distribución espacial y temporal, para con ello poder construir el marco muestral, que es la base para hacer el diseño de muestreo. El marco muestral es la información que ubica y dimensiona al universo.

Figura 1.6. Ejemplos de Marcos "ab initio"

Programa de información	Método de captación de datos	Marcos de muestreo	Elementos en el marco de muestreo	Tamaño
Encuesta de Viajeros Internacionales	Muestreo probabilístico	• Viajeros fronterizos: construcción mensual del marco con una metodología definida basada en un conteo intensivo de viajeros que cruzan los puntos fronterizos (días y horarios establecidos).	• Puntos y viajeros fronterizos	Tamaño variable por mes

1.8 Consideraciones adicionales en el uso de marcos muestrales

Complejidad de las unidades de observación (unidades de producción) en la Encuesta Nacional Agropecuaria

En la ENA 2019 se define el siguiente concepto:

Unidad de Producción. Es la unidad económica conformada por uno o más terrenos ubicados en un mismo municipio, en donde al menos en alguno de ellos se realizan actividades agropecuarias o forestales, bajo el control de una misma administración. Si la administración tiene terrenos ubicados en otro municipio, se considera como otra unidad de producción; esto es, habrá tantas unidades de producción como municipios ocupen sus terrenos.

En el párrafo previo se mencionan los términos *terreno* y *productor*, definidos como:

Terreno. Es la superficie continua de tierra, con límites reconocidos por el productor, perteneciente a un solo régimen de tenencia y a un mismo tipo de derechos. Los conceptos predio, parcela, lote y predio rústico, para efectos de las estadísticas agropecuarias, se consideran sinónimos de terreno. Para los casos de superficies de uso común en ejidos y comunidades agrarias con productores a su interior, se considera como terreno, la parte que cada productor tiene bajo su manejo o responsabilidad, aun cuando no sea posible el reconocimiento de los límites.

Productor. Es la persona o el conjunto de personas responsables del manejo y de la toma de decisiones de la unidad de producción.

Por tanto, la complejidad de la unidad de producción es que están definidas por los siguientes puntos que pueden ser muy volátiles en el paso del tiempo:

1) La limitación de la Unidad de Producción por el productor (Terreno).

2) El régimen de tenencia y derechos de la Unidad de Producción (Terreno).

3) La persona o personas responsables del manejo de la Unidad de Producción (Productor).

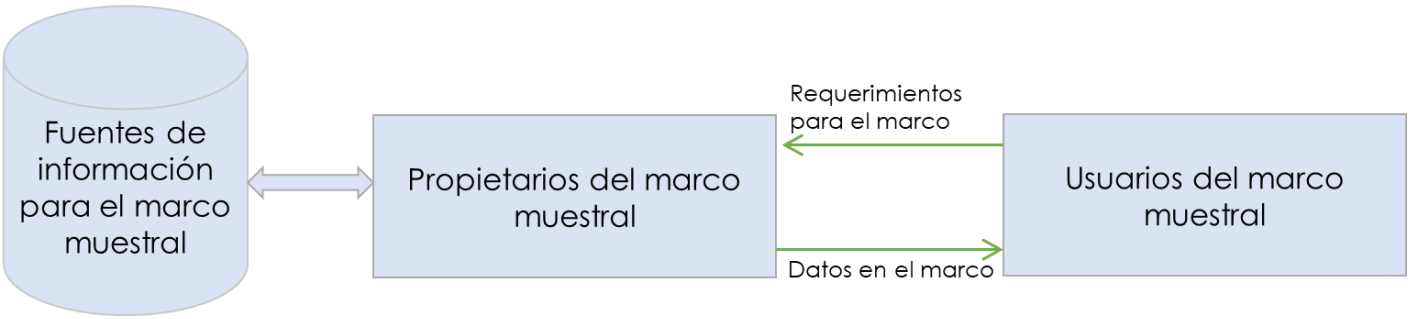
La actualización del marco de muestreo de la ENA debe considerar estas fluctuaciones en la composición de las unidades de producción, ya que, en caso de no preverlas, el marco inevitablemente estará desactualizado y se presentarán las imperfecciones ya mencionadas de subcobertura y sobrecobertura.

Entidades propietarias y usuarias distintas de un marco muestral

En algunos Programas de Información, dos o más entidades construyen, organizan y usan un marco muestral. Una entidad central debe crear, mantener y actualizar el marco, la cual se considera la propietaria del marco muestral.

La entidad usuaria de la información genera los requerimientos de la información que contendrá el marco muestral. La entidad propietaria de la información tiene que organizar y coordinar todas las actividades relacionadas con el marco. La entidad usuaria deberá enviar las solicitudes, sugerencias y nuevas necesidades de información a incorporar en el marco. Por esta razón, es muy importante la coordinación entre la entidad propietaria y la entidad usuaria del marco ya que una comunicación deficiente puede provocar que las necesidades de información no sean cubiertas y como consecuencia los procedimientos de muestreo no se implementen correctamente y los errores no muestrales se incrementen. En la figura 1.7 se muestra el esquema de coordinación entre las entidades propietaria y usuaria del marco muestral.

Figura 1.7. Coordinación entre propietarios y usuarios de un marco muestral



En el caso particular de las encuestas de gobierno, seguridad pública y justicia, para algunos de sus programas de información, la unidad de observación forma parte de registros o listados externos al INEGI, es por ello que previo a la realización de programas en dominios distintos a viviendas y unidades económicas, donde el INEGI es quien coordina los distintos Marcos Muestrales, se realizan ejercicios de revisión, depuración, integración y actualización de los referidos listados de las unidades de observación. No obstante, se reconoce que, en estos casos especiales, las distintas poblaciones o unidades observación por su definición pueden fluctuar en cortos periodos de tiempo, por lo que, puede conducir a la actualización del marco de muestreo durante la ejecución del programa de información, o bien es posible realizar sustituciones de las unidades de observación.

En la figura 1.8 se describen 3 ejemplos en los que la entidad propietaria del marco muestral es proveedora de la información de encuestas realizadas por el INEGI. La información de las unidades de análisis de las tres encuestas

requiere un tratamiento especial debido a la confidencialidad y dificultad de acceso al informante. La información contenida en el marco muestral de estos tres ejemplos contiene datos de identificación tales como sociodemográficos y ubicación del informante, además de variables para la creación de dominios de estudio y estratificación.

Figura 1.8. Ejemplos de encuestas donde la entidad propietaria del marco muestral suministra la información a la entidad usuaria

Nombre	Tipo de muestreo/ unidad de observación	Entidad Propietaria	Entidad Usuaria	Unidad de análisis	Cobertura y desglose geográfico	Descripción del marco
Encuesta Nacional de Población Privada de la Libertad (ENPOL)	Encuesta probabilística en establecimientos	Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP)	DGEGSPyJ	Personas privadas de la libertad (PPL)	Nacional, por entidad federativa y centros penitenciarios de interés	Listados de la población privada de la libertad (PPL)
Encuesta Nacional de Estándares y Capacitación Profesional Policial (ENECAP)	Encuesta probabilística en hogares	Comisión Nacional de Seguridad Procuraduría General de la República	DGEGSPyJ	Servidor público destinado a funciones de seguridad pública	Nacional para el caso de Policía Federal y la PGR Nacional y Estatal para Policía Estatal, Municipal y PGJ/FGE	Registro Nacional de Personal de Seguridad Pública Directorio del personal de la PGR
Encuesta Nacional de Adolescentes en el Sistema de Justicia Penal (ENASJUP)	Encuesta probabilística en establecimientos	Dirección General de Prevención y Tratamiento de menores de la CNS	DGEGSPyJ	Población de adolescentes que estén siendo procesados y que tengan al menos una medida cautelar o que se les haya dictado una medida de sanción por la comisión de un delito del fuero común o del fuero federal	Nacional y por entidad federativa	Listados de la población interna y externa de cada uno de los centros de internamiento y/o autoridad responsable de los adolescentes en externación Directorio de los centros de internamiento del país

ACTIVIDAD 2: ELECCIÓN DEL TIPO DE MUESTREO

El esquema de muestreo se refiere a una combinación de opciones técnicas en cuanto al tipo y modalidad del muestreo, así como al número de etapas de selección, respecto al cual se toman las decisiones para un Programa de Información específico.

Las alternativas metodológicas son diversas y su determinación implica tanto el análisis de los condicionantes como el conocimiento de las ventajas y desventajas de cada una de ellas.

Figura 2.1. Insumos y evidencias de la determinación de la población objeto de estudio, el marco muestral y el tipo de muestreo

INSUMOS	EVIDENCIAS

INSUMOS	EVIDENCIAS
• Esquema conceptual incluye asociación y jerarquización de temas, universo de análisis, variables y clasificaciones • Documentación de marco muestral • Metodologías anteriores si están disponibles	Documentación del Diseño de la Muestra • Determinación del marco muestral y tipo de muestreo • Justificación del marco muestral • Diseño de muestreo cuando corresponda

2.1. Opciones para la determinación del esquema de muestreo

Para elegir el tipo de muestreo idóneo en un Programa de Información, la decisión debe basarse en los objetivos del estudio, el esquema de la investigación y el alcance de sus contribuciones. Los métodos de muestreo se agrupan en dos tipos: probabilístico y no probabilístico; la combinación de ambos se considera como muestreo mixto. En la figura 2.2 se presentan ejemplos de estos tres tipos de muestreo. Las etapas se refieren a los niveles jerárquicos donde se seleccionan unidades de muestreo. En cada etapa el tipo de unidad de muestreo es distinta.

Figura 2.2. Tipos de muestreo y etapas de selección

Tipo de muestreo/Modalidad de muestreo	Número de etapas de selección	
	Unietápico	Bietápico o más etapas
Probabilístico • Aleatorio simple (con reemplazo y sin reemplazo) • Sistemático • Estratificado • Por conglomerados • Mediciones repetidas (longitudinal) **Determinístico (no probabilístico)** • Convencional o accidental • Por cuotas • Cadena o bola de nieve (snow ball) • Intencional o por juicio **Mixto** • Muestreo *cutoff*	Una sola elección del tipo y modalidad de muestreo	Combinación específica del tipo de muestreo, la modalidad y el número de etapas de selección

A. Muestreo probabilístico

Es útil para realizar estimaciones de indicadores objetivo en la población objeto de estudio, donde todos sus elementos tienen una probabilidad conocida, mayor a cero, de ser elegidos. Algunas de las principales ventajas de este tipo de muestreo son, dado que cada unidad es seleccionada aleatoriamente, el sesgo de selección se elimina (Särndal et al, 1992, sec. 1.3) y sus probabilidades de inclusión permiten calcular estimaciones confiables y estimar los errores estándares asociados, por lo que se pueden hacer inferencias acerca de la población objeto de estudio (Fellegi, sección 6.2, 2010). Algunas de las modalidades por considerar se describen a continuación:

Aleatorio simple (MAS)

Se enumeran a todos los individuos de la población objeto de estudio cuyo tamaño es N, estos números se asignan para servir como identificador único. Se eligen tantos sujetos como sea necesario para completar el tamaño de muestra n, con $n \leq N$, por medio de un mecanismo de selección aleatoria, en el que se extrae un conjunto de n números entre 1 y N con probabilidades iguales. Este procedimiento, tiene poca o nula utilidad práctica cuando la población objeto de estudio es muy grande; además, tiene un costo muy alto e involucra un arduo proceso operativo, sin embargo, es una referencia para comparar a otras modalidades de muestreo. Existen dos tipos de selección: con reemplazo y sin reemplazo. El muestreo aleatorio simple tiene una alta entropía, es decir, es muy difícil predecir el tipo de muestra que se obtendrá. El concepto de entropía es útil para la estimación de la varianza de los estimadores muestrales. Cando un diseño muestral tiene una alta entropía es posible obtener una aproximación de las probabilidades de segundo orden en términos de las probabilidades de inclusión de primer orden. (Tillé y Haziza, 2010).

Sistemático

Se numeran todos los elementos de la población objeto de estudio de forma ascendente, pero en lugar de extraer **n** números aleatorios sólo se extrae uno, entre 1 y k, k=N/n donde N es el tamaño de la población objeto de estudio. Se parte de ese número aleatorio para elegir, a intervalos constantes, todos los demás hasta completar la muestra. Cuando la población objeto de estudio está en un orden aleatorio con respecto al Indicador Objetivo que estima a la población objeto de estudio, este método es equivalente al muestreo aleatorio simple (MAS), aunque su uso facilita la extracción de la muestra. Por otro lado, cuando el orden de la población objeto de estudio presenta tendencia a cambios paulatinos de dicho Indicador Objetivo, esta modalidad produce varianzas estimadas menores que el MAS, esto se debe a que la muestra queda más dispersa sobre la población objeto de estudio, es decir, que la muestra es más representativa; aunque en este caso no hay expresiones válidas para obtener los estimadores y las varianzas respectivas, se pueden utilizar las expresiones dadas por el MAS como una aproximación conservadora; en caso contrario, es decir, cuando se tiene un orden que se refleja en cambios periódicos, este método puede producir varianzas estimadas mayores a las generadas mediante el MAS. El muestreo sistemático es el diseño muestral de menor entropía[6].

$$\underbrace{|1|2|3|\ldots|k|}\ldots|120|121|\ldots|N-1|N|$$

Intervalos de tamaño k

Estratificado

Se divide la población objeto de estudio en clases homogéneas y mutuamente excluyentes, llamadas estratos (por grupos de edades, por sexo, entre otros). Hecho esto, la muestra se distribuye de acuerdo con distintos métodos de afijación o distribución (igual en todos los estratos, proporcional al tamaño de cada estrato, distribución de Neyman y distribución óptima). Con esto se asegura que todos los estratos de interés estén representados adecuadamente en la muestra. Cada estrato funciona independientemente, en cada estrato se selecciona una muestra probabilística con esquema no necesariamente iguales en todos los estratos.

Por conglomerados

La población objeto de estudio está dividida en subpoblaciones llamadas conglomerados, en los cuales los elementos que los componen poseen cierta característica que les hace ser propios de cierta cualidad o atributo. Por ejemplo, agrupaciones de viviendas ubicadas en dos o más manzanas de un área geográfica delimitada. El muestreo por conglomerados se utiliza cuando no se cuenta con un marco de muestreo completo de la población

[6] Es una medida del nivel de incertidumbre o de sorpresa en una muestra que será seleccionada. Si un diseño muestral tiene una alta entropía, es muy complicado predecir la muestra que se obtendría (Tillé & Haziza, 2010).

objeto de estudio o es muy caro producirlo, y cuando se requiere reducir costo de captación de datos, principalmente cuando la población objeto de estudio está muy dispersa en una región. El muestreo por conglomerados de una etapa consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y posteriormente todos los elementos de cada conglomerado elegido formaran parte de la muestra. Cuando los conglomerados son áreas geográficas suele hablarse de "muestreo por áreas". El muestreo por conglomerados es menos preciso que el muestreo aleatorio simple o estratificado debido a la homogeneidad inherente de las unidades de muestreo dentro de los conglomerados seleccionados, ya que las unidades son físicamente cercanas y tienden a tener características similares, es decir, la selección de dos o más unidades dentro del mismo conglomerado puede generar información redundante lo cual incrementa los errores muestrales de los indicadores objetivo.

Muestreo multietápico

Cuando en el muestreo por conglomerados se prosigue en el análisis y dentro de cada conglomerado se vuelven a seleccionar, también de forma aleatoria, nuevos subconglomerados, y así sucesivamente hasta seleccionar las unidades últimas, al muestreo se le denomina muestreo por etapas o multietápico.

Varias condiciones pueden propiciar a decidir que la selección de la muestra se realice en más de una etapa: como el no disponer de la localización exacta de cada una de las unidades de observación, o si el tamaño de la población objeto de estudio es demasiado grande; entonces resulta conveniente conformar conglomerados mediante la agrupación de unidades de observación que deberán reunir o compartir algunas características.

Cuando la población objeto de estudio se ha dividido en conglomerados, pero hay gran variabilidad entre sus tamaños, se pueden conformar nuevos conglomerados al interior y elevar el número de etapas de selección a dos o más.

Una alternativa para la situación anterior consiste en combinar a los conglomerados con estratos. En este caso los conglomerados se estratifican, seleccionándose cierto número en cada estrato y de los grupos elegidos se toman las unidades de análisis que integrarán la muestra.

El más frecuente de los muestreos por etapas es el bietápico, en el que se seleccionan, en primer término y de forma aleatoria, los conglomerados o áreas, y en una segunda etapa, las unidades últimas o más elementales del conjunto poblacional, sin necesidad de tener que seleccionar ningún otro tipo de unidad intermedia.

El hecho de usar muestreo por etapas conlleva a una reducción en costos, pero conduce también a un incremento en las varianzas de las estimaciones. El cálculo de un estimador insesgado de la varianza puede complicarse demasiado ya que involucra la estimación de varios componentes de la varianza asociados con cada etapa, por lo que en algunas ocasiones se utilizan estimadores de varianza simplificados con un cierto nivel de sesgo.

Muestreo multifase

El muestreo multifase es bastante diferente del muestreo multiétapico. Aunque el muestreo multifase también implica tomar dos o más muestras, éstas se extraen del mismo marco y las unidades seleccionadas tienen la misma estructura en cada fase. Una muestra multifase recopila información de una muestra grande de unidades y luego, para una submuestra, recopila información más detallada. La forma más común de muestreo multifásico es el muestreo de dos fases (o muestreo doble), pero también son posibles tres o más fases. Sin embargo, al igual que con el muestreo multiétapico, el diseño y las estimaciones de la muestra se vuelven más complicados cuando se incrementa el número de fases más complejo.

Para el muestreo doble se consideran las siguientes fases:

a) En la primera fase, se selecciona una muestra s_a grande de unidades usando un diseño muestral simple. Para estas unidades se recolecta información auxiliar cuya obtención es sencilla o de bajo costo.

b) Con la ayuda de la información auxiliar obtenida en la primera fase, se selecciona una muestra de segunda fase s proveniente de s_a con un diseño muestral condicional a la primera muestra. Los indicadores objetivo se registran para las unidades de la segunda muestra.

La creación de un marco con información útil y confiable para la muestra s_a es la base para el éxito de un muestreo doble. El muestreo doble se ha aplicado para la estimación en encuestas con presencia de no respuesta.

Muestreo con mediciones repetidas

En este tipo de estudios destaca la encuesta longitudinal en la cual se hace un seguimiento de una muestra probabilística de unidades de observación y se obtienen varias mediciones en un periodo. El propósito de una encuesta longitudinal es recopilar y analizar datos sobre el crecimiento, el cambio o tendencia a lo largo del tiempo de uno o varios indicadores objetivo. Este tipo de muestreo generalmente mide con mayor precisión los cambios de una característica de la población objeto de estudio, en comparación de utilizar una serie de muestras independientes.

Un diseño intermedio entre muestras independentes seleccionadas sucesivamente y una muestra longitudinal es el muestreo con panel rotante, en el cual una fracción de la muestra se remplaza en cada ocasión que se lleva a cabo la encuesta. El objetivo principal de una muestra de panel rotantes es obtener estimaciones a niveles agregados. El esquema de rotación se diseña para controlar una muestra en el tiempo, para garantizar que las estimaciones transversales sean insesgadas y para reducir costos y la varianza muestral.

Muestreo con probabilidades proporcionales al tamaño

Este tipo de diseños de muestreo son un caso particular de los diseños con probabilidades de selección desiguales. El muestreo con probabilidades proporcionales al tamaño es un método que usa información auxiliar y genera probabilidades de selección distintas. Si las unidades en la población objeto de estudio varían en tamaños y se conocen los tamaños, estos pueden ser usados en la selección de la muestra para mejorar la eficiencia estadística. a pesar de que el sesgo de selección no se elimina, el muestreo proporcional al tamaño puede incrementar la precisión de las estimaciones si las medidas de tamaño están correlacionadas los indicadores objetivo, es decir, la varianza de las estimaciones puede reducirse en relación con un diseño muestral con probabilidades iguales (Chromy, 2008). Ejemplo de medidas de tamaño son ingresos, número de empleados, número de pacientes atendidos. En este esquema, las probabilidades de selección son proporcionales al tamaño (PPT). Las variantes del muestreo PPT son con reemplazo, sin reemplazo, de Poisson y sistemático. Este último, permite obtener muestras PPT grandes pero no permite obtener estimaciones de varianza insesgadas dado que la mayoría de sus probabilidades de inclusión de segundo orden son cero.

Muestreo con marcos “ab initio”

Cuando no se puede crear un marco de referencia completo debido a que la población objeto de estudio fluctúa derivado de su dinámica y geográficamente no tiene una ubicación fija se pueden utilizar encuestas de flujo (entrada / salida). Este tipo de encuestas se aplican a las poblaciones que cruzan una frontera, por ejemplo, el flujo de personas o mercancías que entran (o salen) de un país, los usuarios de una carretera de peaje, entre otros (Fellegi, sección, 6.3.2, 2010).

Para estos casos, se utiliza un marco de referencia “ab initio” y un muestreo sistemático, o un muestreo de conglomerados en varias etapas con muestreo sistemático dentro de los conglomerados muestreados, para encuestar a dichas poblaciones. El marco muestral puede ser una lista de las unidades de la población enumeradas dentro de un cierto intervalo de tiempo en ubicaciones particulares. Para que el marco tenga una cobertura completa, estas ubicaciones deben ser áreas donde se concentra la población objetivo. A menudo, se utilizan áreas de entrada o salida.

Diseños muestrales complejos

En la práctica es difícil o muy costoso usar el muestreo aleatorio simple. Por tal razón se utilizan diseños muestrales complejos que involucran el uso de estratificación, conglomeración, varias etapas o fases de muestreo, probabilidades de selección distintas y uso de varios marcos de muestreo. El objetivo del uso de diseños muestrales complejos es la minimización de la varianza y disminuir la dificultad de calcular las estimaciones muestrales; además de obtener estimaciones puntuales y de varianza para los parámetros poblacionales de interés.

B. Muestreo determinístico (no probabilístico)

En el muestreo determinístico, también conocido como no probabilístico, el cálculo del tamaño y selección de la muestra se basan en juicios y criterios subjetivos, por lo tanto, se desconoce la probabilidad de selección de las unidades de la población objeto de estudio y no es posible establecer la precisión respecto a niveles de confianza predefinidos. No obstante, el muestreo determinístico representa una alternativa viable, ya sea cuando la aplicación del muestreo probabilístico resulta demasiado costosa cuando no es posible disponer de un marco de muestreo o cuando existe seguridad en que la información recabada bajo este tipo de muestreo es suficientemente útil para los fines de la investigación.

En el contexto del muestreo no probabilístico pueden identificarse las siguientes modalidades:

Muestreo convencional o accidental

Consiste en recopilar datos acerca de los sujetos de estudio que resulten más accesibles. Es un esquema de muestreo rápido y de bajo costo, pero con deficiencias en términos de representatividad. Es útil como parte del proceso de estudios exploratorios con propósitos de orientar la definición de una investigación y no para la caracterización de estructuras o del comportamiento de una población objeto de estudio

Muestreo por cuotas

En este tipo de muestreo se utilizan los datos de subconjuntos o determinados estratos de población objeto de estudio, tales como: sexo, edad o religión, entre otros, para seleccionar miembros que se consideren típicos según los propios fines de la investigación. El muestreo por cuotas recibe su nombre en función de la práctica que consiste en asignar ciertas proporciones de la muestra a determinados estratos de la población objeto de estudio. En el proceso de selección, el entrevistador decide a quien aplicar el cuestionario, bajo determinados criterios generales establecidos previamente. Éstos últimos resultan insuficientes para evitar en la práctica la intervención de elementos subjetivos, por lo cual no es utilizado para fines de estadísticas que requieren buen nivel de confiabilidad. Sin embargo, en la práctica se han obtenido buenos resultados combinado el muestreo probabilístico multietápico con el muestreo por cuotas, en las unidades últimas de muestreo para mejorar la representatividad por grupos de edad y sexo.

Muestreo en cadena o bola de nieve

El muestreo en cadena o de bola de nieve se basa en el supuesto de que los miembros de una población rara se conocen entre sí. Para tomar una muestra de bola de nieve, se identifican a unas cuantas personas con la característica rara. Se pide a cada una de esas personas que identifique a otras personas con la misma característica rara para su muestra, se pide a las nuevas personas en la muestra que identifiquen más personas con la característica rara, y así sucesivamente, hasta que se alcance el tamaño de muestra deseado. El muestreo de bola de nieve puede crear una muestra bastante grande de una población rara. Se necesitan suposiciones fuertes para generalizar los resultados de una muestra de bola de nieve a la población objeto de estudio. Aunque el muestreo de bola de nieve puede identificar miembros de una población rara que sería difícil de encontrar con

otros diseños, la muestra resultante no puede considerarse como MAS. Personas con muchas conexiones en la población objeto de estudio es más probable que se incluya en la muestra que las personas con pocas conexiones; es probable que las personas aisladas no pueden contactarse.

Muestreo intencional o de juicio

La característica principal de este tipo de muestreo es que tanto el tamaño de muestra como la selección de los elementos que la integran están sujetos al juicio del investigador, del cual se requiere suficiente conocimiento y experiencia sobre el tema. La validez de los resultados en este caso depende del nivel de conocimiento sobre el fenómeno en estudio y de evidencias estadísticas que muestren su utilidad para conocer aspectos de comportamiento.

Las modalidades del muestreo no probabilístico son aplicadas principalmente en los medios académicos y en investigaciones privadas, pero son de baja utilidad para las oficinas nacionales de estadística por lo que, en general, se opta por no utilizarlas con la excepción del muestreo intencional o de juicio, que ha resultado útil para conocer con fidelidad, tendencias y comportamientos de determinados indicadores objetivo en la generación de estadísticas económicas.

Tal es el caso cuando es posible establecer que en determinadas unidades de una población objeto de estudio se concentre un alto porcentaje del valor del Indicador Objetivo y se procede a incluirlas como parte de la muestra, hasta alcanzar el nivel de cobertura pretendido o previamente acordado para el Indicador Objetivo en cuestión.

Dicho nivel depende de los intereses y del conocimiento que se tenga, por parte del área responsable, acerca del sector de actividad económica objeto de investigación. Si bien la definición de la muestra no es representativa de la población objeto de estudio, se tiene la certeza de que su comportamiento en el tiempo caracteriza tanto al Indicador Objetivo como a algunas otras variables altamente relacionadas con ella, en tanto que el tamaño de muestra resulta muy inferior al correspondiente a un diseño probabilístico.

C. Encuestas mixtas

Las encuestas probabilísticas requieren que las probabilidades de inclusión sean mayores a cero para todas las unidades en la población objeto de estudio. Existen métodos de muestreo que emplean probabilidades de selección mayores a cero para una parte de la población objeto de estudio mientras que para la parte restante, sus probabilidades de inclusión son iguales a cero. Dichos métodos toman una posición intermedia entre muestreo probabilístico y la selección no probabilística con probabilidades de inclusión desconocidas.

Muestreo *cutoff*

El muestreo *cutoff* es una combinación de selección probabilística y no probabilística de elementos en la muestra. En esta modalidad hay una inclusión de una parte de la población objeto de estudio para ser considerada en la muestra, usando un criterio determinístico. Este procedimiento, que produce estimaciones sesgadas, puede justificarse con alguno de los siguientes argumentos: (1) el costo de la construcción y mantenimiento de un marco de muestreo confiable para la población objeto de estudio es muy alto, y sólo se obtendrá una ganancia pequeña en precisión, (2) el sesgo causado por la exclusión de algunas unidades se considera insignificante. Este procedimiento se utiliza cuando la distribución de los valores poblacionales de los indicadores objetivo es altamente asimétrica y no existe un marco de muestreo confiable para las unidades pequeñas (Särndal et al. 1992, p. 531). Un ejemplo de estas poblaciones se puede encontrar en encuestas de unidades económicas; una proporción considerable de la población objeto de estudio puede consistir en unidades económicas de pequeño tamaño cuya contribución al total del Indicador Objetivo (ingresos, número de empleados, entre otros) es modesta. En el otro extremo, tales poblaciones contienen algunas unidades económicas gigantes cuya inclusión en la muestra es virtualmente obligatoria para no obtener un error grande en la estimación de un total.

El método de muestreo *cutoff* produce estimadores sesgados, por lo que se utiliza como error de medición al error cuadrático medio (la suma de la varianza y sesgo al cuadrado, Benedetti et al, 2010).

Las siguientes 5 versiones del muestreo *cutoff* son las de mayor uso en la práctica:

Cutoff 1. Se asigna una probabilidad de inclusión igual a 1 para cualquier unidad económica con una medida de tamaño (ingresos, número de empleados, entre otros) superior a un límite definido y probabilidad de inclusión en la muestra igual a cero para las unidades debajo de tal límite. No se hace ninguna estimación para las unidades no incluidas en la muestra. Con este esquema se minimizan los costos y es adecuado cuando el Indicador Objetivo es muy asimétrico. Se asume que considerar como cero a las unidades económicas excluidas no causa un impacto significativo en el sesgo de los estimadores.

Cutoff 2. En este método se procede igual que en *cutoff* 1, excepto que la estimación se hace para las unidades que no se incluyeron en la muestra. Con esta técnica se requiere de información auxiliar para todas las unidades económicas. Se puede emplear la medida de tamaño para estimaciones modelo-asistidas como el estimador de razón.

Cutoff 3. Se establece un límite para incluir a las unidades económicas en la muestra, no obstante, algunas unidades debajo del límite también se incluyen en la muestra. Esta estratificación se conoce como 'toma todo' (*take all*) y 'toma algunos' (*take some*). Por ejemplo, una muestra aleatoria simple estratificada en donde para un estrato todas las unidades son seleccionadas. Para este método se pueden usar estimaciones basadas en diseño o modelos asistidos.

Cutoff 4. Ahora dos límites definen 3 estratos de la muestra: 'toma todo', 'toma algunos' y 'toma ninguno' (*take none*).

Cutoff 5. Las unidades económicas se ordenan de forma descendente de acuerdo con una medida de tamaño (ingresos, número de empleados, entre otros). Los datos se captan a partir de la unidad económica de mayor tamaño hasta alcanzar un límite establecido, posiblemente arbitrario. Las estimaciones son similares que en *cutoff* 2.

Los límites de inclusión pueden ser arbitrarios, sin embargo, es recomendable usar los algoritmos existentes para determinar los mencionados puntos de corte y la asignación de los tamaños de muestra en cada estrato. Este objetivo se logra usando información auxiliar con una relación alta con el Indicador Objetivo.

2.2. Criterios para elegir la modalidad de muestreo

Decidir sobre el tipo de muestreo depende, en buena medida, de la existencia o posibilidad de integrar un marco de muestreo actualizado, del cual se debe seleccionar la muestra, ya que, si ello es factible, es preferible aplicar el muestreo probabilístico. La imposibilidad de disponer, actualizar o integrar un marco, puede obligar a un muestreo determinístico (no probabilístico), u optar por el muestreo sistemático. Para elegir la modalidad de muestreo más conveniente se debe tomar en cuenta:

- Si se dispone de un marco de muestreo actualizado y las unidades del marco cuentan con la misma probabilidad de selección, es factible utilizar un muestreo aleatorio simple sin reemplazo (MAS). Aunque es el más sencillo no suele emplearse directamente para llevar a cabo selecciones en poblaciones grandes. Su mayor relevancia la debe al hecho de ser un procedimiento básico como integrante del muestreo complejo.

- Si adicionalmente el marco de muestreo ya está estratificado o se dispone de información auxiliar para hacer una estratificación, el muestreo estratificado mejora las estimaciones del MAS (estimadores del tipo

promedio o total); bajo esta situación, es preferible aplicar un muestreo sistemático con arranque aleatorio de manera independiente en cada estrato, el cual simplifica la extracción de la muestra.

- Si es necesario evitar el problema de dispersión geográfica de las unidades de observación, es útil el muestreo de conglomerados a partir de la delimitación de áreas con características similares. Los conglomerados facilitan la utilización del muestreo con probabilidad proporcional al tamaño (PPT), ya que el número de elementos en un conglomerado representa una medida natural del tamaño del conglomerado.

En la figura 2.3 se describen las características principales, ventajas y desventajas de las modalidades de muestreo presentadas en la sección 2.1.

Figura 2.3. Ventajas y desventajas de los distintos tipos de muestreo probabilístico y mixto

Modalidad de muestreo	Característica para seleccionar una muestra	Ventajas	Desventajas
Aleatorio simple (MAS)	Se selecciona una muestra de tamaño *n* de una población objeto de estudio de *N* unidades; cada elemento tiene una probabilidad de inclusión igual y conocida de *n*/*N*.	• Sencillo y de fácil comprensión. • Cálculo rápido de medias y varianzas. • Se basa en la teoría estadística y, por lo tanto, existen paquetes informáticos para analizar los datos.	• Requiere de antemano un listado completo de toda la población objeto de estudio. • Cuando se trabaja con muestras pequeñas es posible que no represente a la población objeto de estudio adecuadamente. • Tiene un costo muy alto, por lo que se vuelve poco práctico.
Sistemático	Se considera un caso particular del muestreo por conglomerados unietápico. • Conseguir un listado de los *N* elementos de la población objeto de estudio. • Determinar el tamaño muestral *n*. • Definir un intervalo *k*=*N*/*n*. • Elegir un número aleatorio, *r*, entre 1 y *k* (*r*=arranque aleatorio). • Seleccionar los elementos de la lista.	• Fácil de aplicar. • No siempre es necesario tener un listado de toda la población objeto de estudio. • Cuando la población objeto de estudio está ordenada siguiendo una tendencia conocida, asegura una cobertura de unidades de todos los tipos.	• Si la constante de muestreo está asociada con el fenómeno de interés, las estimaciones obtenidas a partir de la muestra pueden contener sesgo de selección. • Es el método con menor entropía.
Estratificado	En ocasiones será conveniente estratificar la muestra según ciertos indicadores objetivo. Para ello se debe conocer la composición estratificada de la población objeto de estudio al hacer un muestreo. Ya calculado el tamaño muestral apropiado, éste se reparte de manera proporcional entre los distintos estratos definidos en la población objeto de estudio.	• Tiende a asegurar que la muestra represente adecuadamente a la población objeto de estudio en función de variables seleccionadas. • Se obtienen estimaciones más precisas. • Su objetivo es conseguir una muestra lo más semejante	Se debe conocer la distribución en la población objeto de estudio de las variables utilizadas para la estratificación.

Modalidad de muestreo	Característica para seleccionar una muestra	Ventajas	Desventajas
		posible a la población objeto de estudio en lo que a las variables estratificadoras se refiere.	
Por conglomerados	Se realizan varias etapas de muestreo sucesivas (multietápico). La necesidad de listados de las unidades de una etapa se limita a aquellas unidades de muestreo seleccionadas en la etapa anterior.	• Es muy eficiente cuando la población objeto de estudio es muy grande y dispersa. • No es preciso tener un listado de toda la población objeto de estudio, sólo de las unidades primarias de muestreo.	• El error estándar es mayor que en el muestreo aleatorio simple o estratificado. • El cálculo del error estándar es complejo.
Longitudinal	La selección de la muestra es probabilística y las mismas unidades muestrales se incluyen en varios periodos; cuando se usa un panel rotante, una proporción de las unidades se excluyen de la muestra y se remplazan por otras unidades.	• Reduce la varianza muestral de las estimaciones muestrales ($\hat{t_1} - \hat{t_2}$ donde $\hat{t_1}$ y $\hat{t_2}$ son estimadores del total en los tiempos 1 y 2, respectivamente). • Se puede usar para obtener información del comportamiento de las unidades de observación a través del tiempo. • Los costos se pueden reducir (operativo de campo, capacitación del personal, etcétera).	• Las estimaciones y el tratamiento de la no respuesta son más complejos. • Es más complicado mantener la representatividad a través del tiempo debido a los cambios que ocurren en la población objeto de estudio, tales como el ingreso de nuevas unidades y las salidas de otras. • Se requiere que el presupuesto para el mantenimiento de la muestra se garantice para toda la duración de la encuesta.
Proporcional al tamaño	La selección de una muestra PPT se puede obtener mediante dos esquemas: (i) de tamaño fijo sin reemplazo y (ii) con reemplazo. En ambos casos las probabilidades de selección son distintas y están determinadas por una variable auxiliar correlacionada con los indicadores objetivo.	• Puede mejorar la eficiencia estadística del diseño muestral al usar información auxiliar, en comparación con el MAS y muestreo estratificado.	• Se requiere que el marco muestral se mantenga actualizado para todas las unidades para que pueda usar la medida de tamaño. • La estimación de la varianza muestral es más complicada. • Si la medida de tamaño no está correlacionada con los indicadores de interés, este diseño muestral puede ser menos estadísticamente eficiente que un MAS.
Multifase	Si no es costoso medir las variables auxiliares relacionadas con los indicadores objetivo, se puede obtener una muestra grande para	• Puede mejorar la precisión de las estimaciones en comparación con el	• Es más lenta la obtención de resultados que en una encuesta de una fase. • Puede ser más costoso

Modalidad de muestreo	Característica para seleccionar una muestra	Ventajas	Desventajas
	recolectar la información auxiliar. Esta información se considera para el diseño de una segunda muestra mucho menor que la primera; en esta segunda muestra se miden los indicadores objetivo.	MAS. • Se puede utilizar para obtener información auxiliar que no se encuentra en el marco muestral (por ejemplo, información de estratificación para el muestreo de la segunda fase). • Se puede utilizar cuando el costo de recopilación de algunas de las variables de la encuesta es costoso.	que una encuesta de una fase, ya que requiere entrevistar más de una vez a una unidad muestreada. • Si la población objeto de estudio es móvil o si las características de interés cambian con frecuencia, las demoras entre fases pueden causar problemas. • Sus fórmulas para el cálculo de estimaciones y varianza muestral pueden ser bastante complejas.
Muestreo con marcos "ab initio" (de flujo-entrada/salida)	A menudo se utiliza un marco de referencia "ab initio" y un muestreo sistemático, o un muestreo de conglomerados en dos etapas con muestreo sistemático dentro de los conglomerados muestreados. En general, existen encuestas donde la distribución espacial y temporal se debe considerar para su medición, para este tipo de encuestas se propone un diseño de muestreo en varias etapas dependiendo del estudio a realizar.	• El marco para la etapa final se puede crear mientras está en el campo.	• Puede resultar difícil relacionar la población objeto de estudio con una población conocida. Esto se debe a que las encuestas de flujo miden a los visitantes, en lugar de a las personas. • Debido a que es difícil administrar las operaciones de campo debido a los flujos variables en la población, se recomienda que las entrevistas sean breves. • Por lo general, produce tasas de respuesta bajas.
Cutoff	Se seleccionan para la muestra solo las unidades de la población objeto de estudio arriba de un límite o entre límites previamente definidos. El límite se especifica en términos de una medida de tamaño conocida.	• Puede usarse cuando el Indicador Objetivo tiene una distribución muy asimétrica. • No es necesario el mantenimiento del marco de muestreo para las unidades con medidas de tamaño pequeñas.	• No produce estimadores insesgados. • La varianza muestral es cero por definición en el estrato *take all* • Se tiene que estimar la varianza y sesgo como medidas del error muestral.
		•	•

Las figuras 2.4, 2.5, 2.6 ,2.7 y 2.8 presentan ejemplos de esquemas de diseños muestrales probabilísticos.

Figura 2.4. Muestreo estratificado

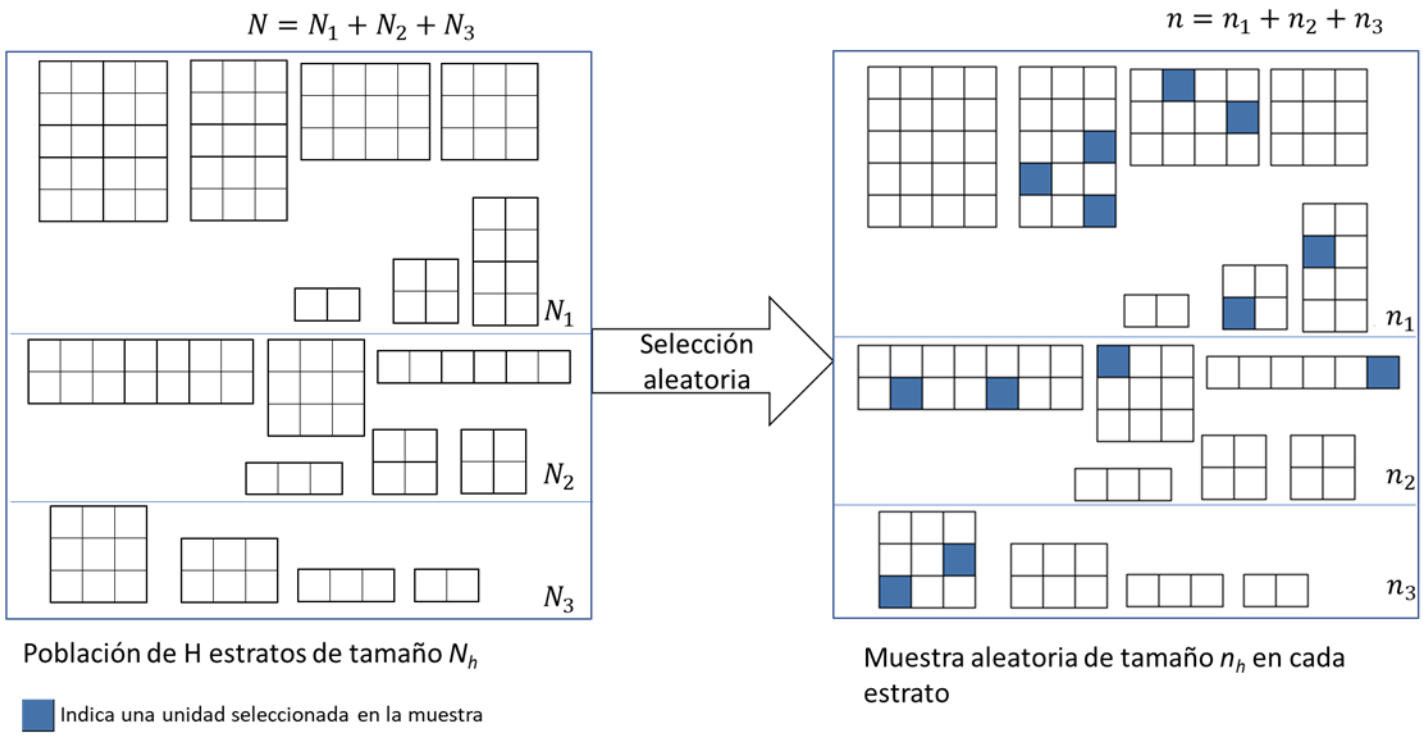


Figura 2.5. Muestreo unietápico y por conglomerados

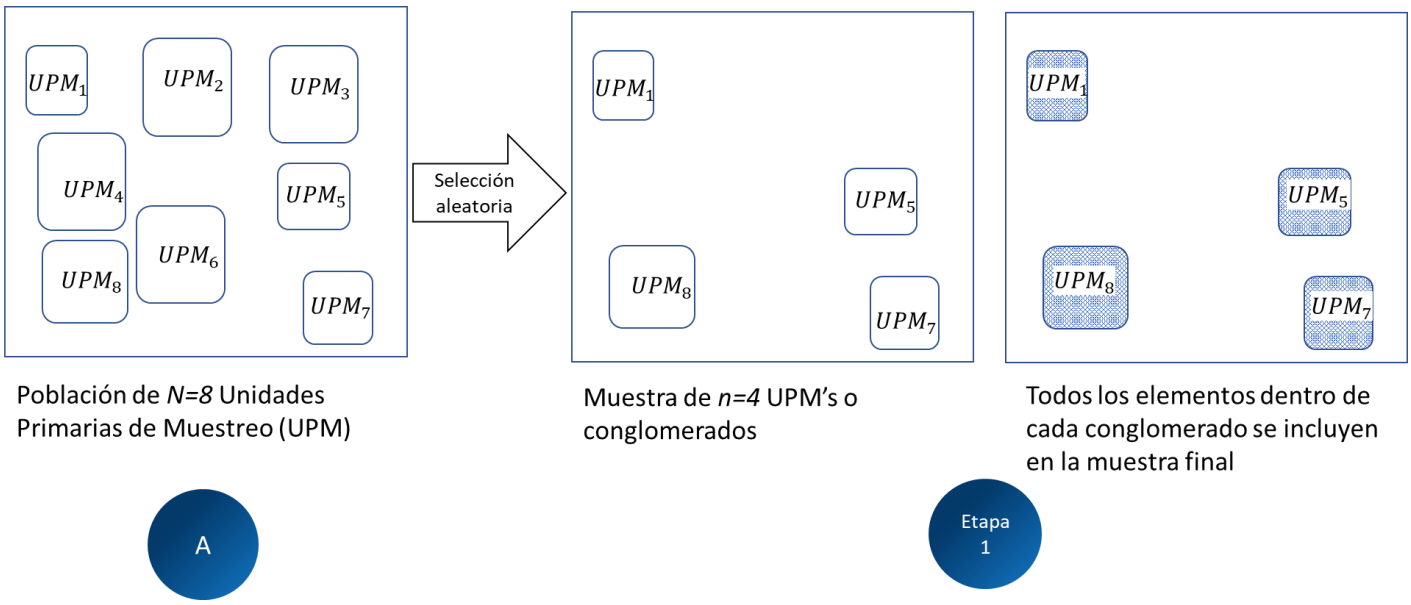
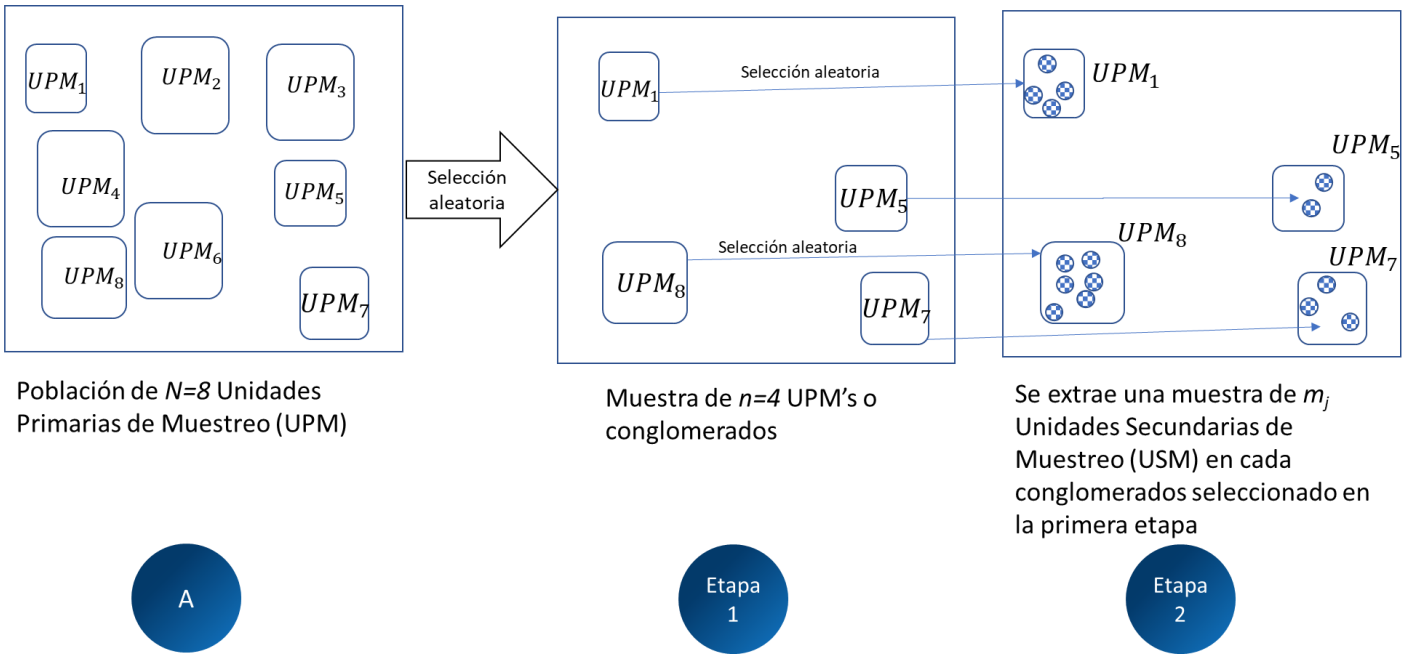


Figura 2.6. Muestreo bietápico y por conglomerados



En las figuras 2.7 y 2.8 se presentan los esquemas de muestreo de la Encuesta Nacional de Ocupación y Empleo (ENOE) y la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE), cuyo procedimiento de muestreo incluye estratificación de conglomerados y la selección de la unidad de observación se obtiene después dos o más etapas.

Figura 2.7. ENOE 2019: bietápico, estratificado y por conglomerados

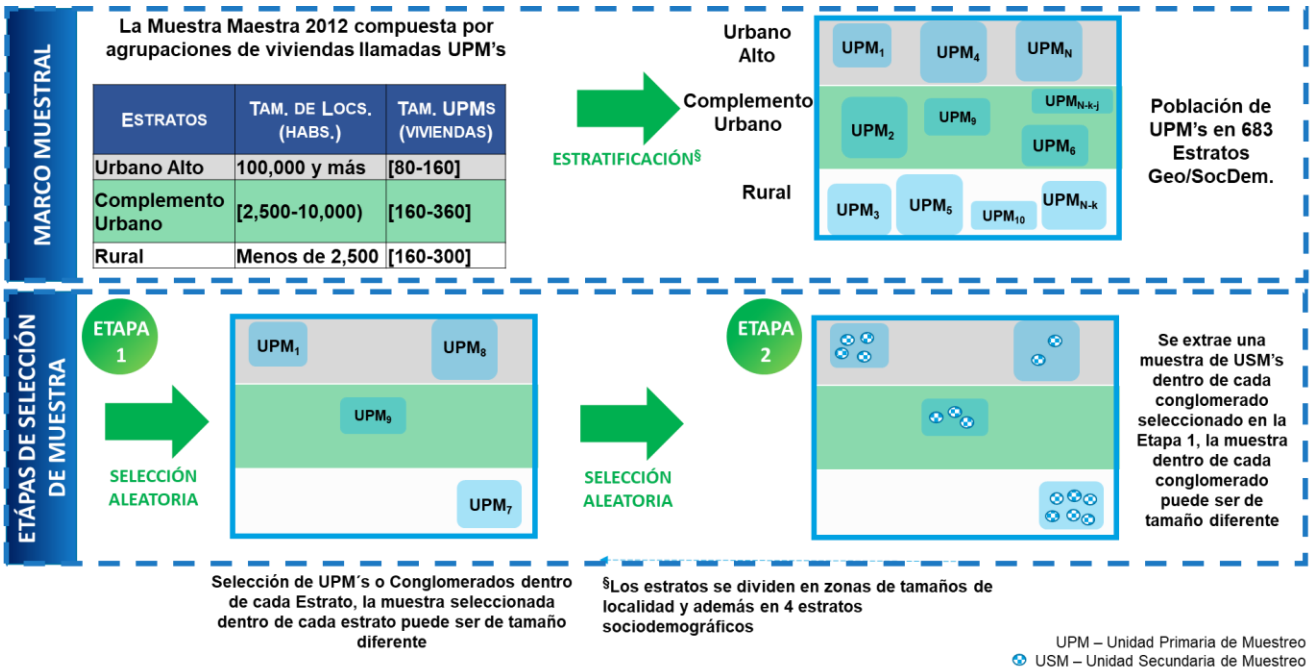
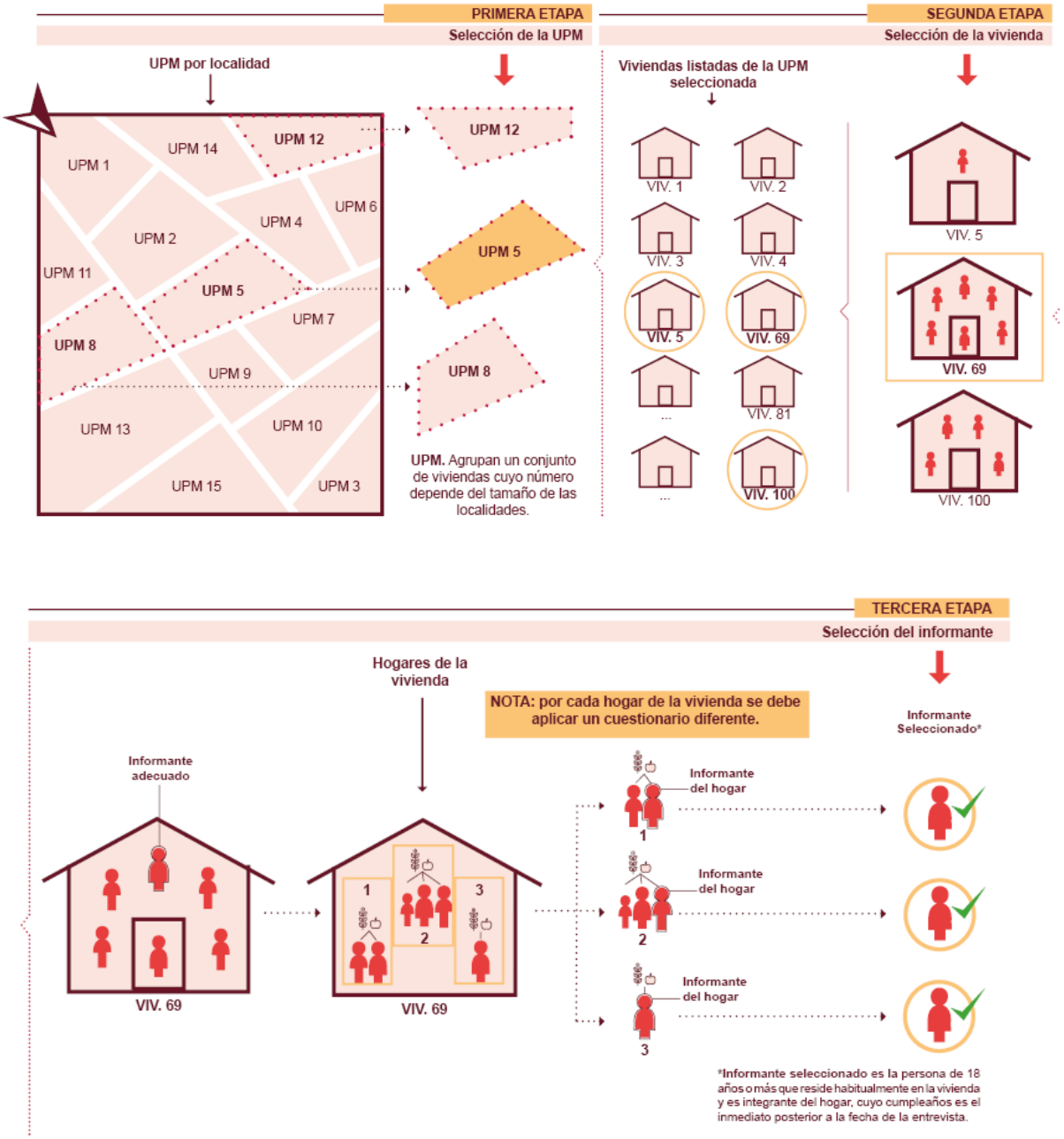


Figura 2.8. ENVIPE: trietápico, estratificado y por conglomerados



ACTIVIDAD 3: DETERMINACIÓN DE LA MUESTRA PARA MUESTREO PROBABILÍSTICO

La elección de un esquema de muestreo debe ser el resultado de varios principios. Tillé y Wilhem (2017) mencionan tres principios para seleccionar una muestra:

1. Principio de aleatorización: consiste en seleccionar con un diseño de muestreo de entropía máxima o cercana a esta.
2. Principio de sobrerrepresentación: la idea es sobremuestrear a las unidades con mayor incertidumbre debido a que la muestra debe recolectar tanta información de la población objeto de estudio como sea posible, lo que implica la tendencia a diseños muestrales con probabilidades de selección distintas.
3. Principio de restricción: consiste en seleccionar solo a las muestras con un conjunto de características deseables, es decir, utilizar un proceso de verificación (muestreo balanceado) para evitar muestras con dominios, estratos o categorías vacías.

Elegido el esquema de muestreo, ahora es necesario determinar aspectos referentes a la muestra, en particular su tamaño y su procedimiento de selección.

3.1. Cálculo del tamaño de la muestra

El tamaño de la muestra es el número de unidades de observación que deben estar incluidas en la muestra.

En el tipo de muestreo probabilístico, contrario a la creencia de que el tamaño de muestra debe guardar cierta proporción con el tamaño de la población objeto de estudio, los aspectos que se involucran en el cálculo están relacionados con la característica o fenómeno a estudiar, el nivel de precisión y confianza que se desea lograr, el esquema de muestreo que se sigue para obtener la información, así como el dominio de estudio o área a la que se desea brindar la información.

- Cuando una característica se presenta con frecuencia en la población objeto de estudio, el tamaño de muestra es menor que el requerido para una característica extraña o poco común, pues en este último caso se necesita entrevistar a una gran parte de la población objeto de estudio para obtener algunos casos que presenten la característica o fenómeno de interés.

- La variabilidad de la característica a estudiar también se involucra, pues se requiere un tamaño de muestra mayor para indicadores que toman un número infinito de valores; por ejemplo, ingreso por trabajo (que va desde 0 por trabajador familiar sin pago, hasta lo que gana el director de alguna compañía), diferente a lo que se requiere para indicadores que toman valores más acotados como estatura al nacer.

- El nivel de precisión está relacionado tanto con el error permitido (la distancia entre la estimación y el valor "real") como con la confianza con que se va a ofrecer este resultado; por ejemplo, si se desea obtener resultados con 95% de confianza y un error máximo de 8%, significa que de 100 muestras sólo 5 tendrían un error mayor al 8 por ciento.

- En el caso de las distintas subdivisiones territoriales (o dominios de estudio) indicadas por el desglose geográfico, se debe hacer un cálculo por separado para cada una de ellas; el tamaño de muestra final será la suma de los tamaños de muestra de cada subdivisión.

Es común que en encuestas de propósitos múltiples se desee brindar información de distintos indicadores objetivo, de esta manera es necesario calcular un tamaño de muestra para cada uno y elegir el tamaño de muestra mayor, ya que éste cubrirá las especificaciones de precisión de todos los indicadores.

Sin embargo, este tamaño de muestra puede resultar muy costoso, entonces se debe tomar la decisión de no incluir algunos indicadores en la encuesta, o bien, admitir para ellos un error de estimación más alto que para el resto.

Considerando todo lo anterior se debe elegir la expresión matemática que permita calcular el tamaño de muestra.

Es necesario destacar que la expresión definida considerando los aspectos anteriores, se ve afectada por la modalidad de muestreo elegida y el número de etapas realizadas para obtener la información, es decir por el efecto del diseño (DEFF), la tasa de no respuesta (TNR), y el coeficiente de variación (CV), por lo cual la expresión matemática final debe incluir cada uno de estos términos.

Por otra parte, el grado de complejidad de la ecuación también depende de los valores particulares que adquieran algunos de los aspectos arriba mencionados, así, por ejemplo, se puede tener una ecuación muy simple cuando: el error es absoluto, el tipo de parámetro es una proporción, la distribución del Indicador Objetivo es normal y el esquema por aplicar es MAS.

Este es un ejemplo del cálculo de un tamaño de muestra para una proporción, considerando el efecto de diseño (DEFF[7]), la tasa de no respuesta, además de los elementos esenciales de precisión y confianza:

$$n = \frac{z^2 qDEFF}{r^2 p(1 - tnr)}$$

Donde:

n = tamaño de muestra.

z = valor del cuantil de una distribución normal estándar para un nivel de confianza fijado.

p = estimación de la proporción de interés de una muestra piloto o de encuestas anteriores.

r = precisión relativa (error relativo fijo).

q = 1– *p*

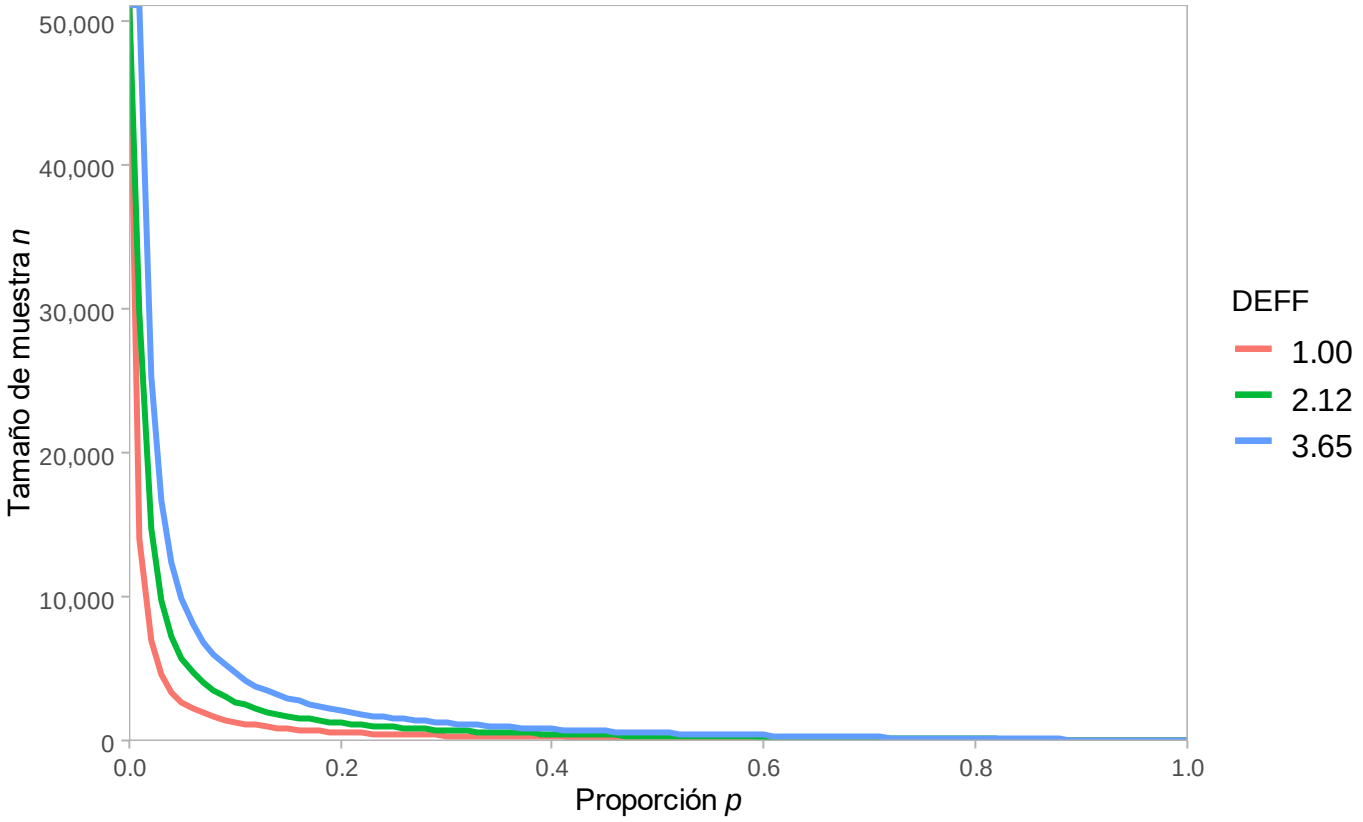
tnr = tasa de no respuesta esperada.

DEFF = Efecto de diseño.

En la figura 3.1 se presenta el comportamiento del tamaño de muestra *n* con los valores fijos de confianza del 90%, precisión de 15% y tasa de no respuesta del 15%. Los valores de la proporción de interés *p* varían entre 0.01 y 1, y se consideraron tres valores de DEFF: 1.00, 2.16 y 3.65. La figura 3.1 muestra que entre más rara es una característica en la población objeto de estudio, es decir una proporción *p* baja en la población objeto de estudio, el tamaño de muestra *n* se incrementa notoriamente. El DEFF, que involucra la varianza de la estimación del diseño empleado, es un elemento que influye para incrementar el tamaño de muestra, ya que en caso de que el DEFF sea igual a la unidad, la varianza de la estimación del diseño sería igual a un MAS con el mismo tamaño de muestra. Cuando el DEFF es mayor a uno, para lograr la misma precisión del estimador bajo un diseño muestral de MAS, el diseño tendrá un tamaño de muestra que se incrementará por un factor de la misma magnitud del DEFF; no obstante, cuando una característica tiene una prevalencia alta, valores de *p* cercanos a uno, la contribución del DEFF al incremento del tamaño de la muestra se aminorará. Por último, la tasa de no respuesta, *tnr*, incrementa a *n* por un factor de 1/(1-tnr).

[7] El cociente de la varianza en la estimación del diseño utilizado, entre la varianza obtenida, considerando un muestreo aleatorio simple para un mismo tamaño de muestra. Indica una medida de la pérdida o ganancia de precisión por usar un diseño más complejo en lugar de una muestra aleatoria simple.

Figura 3.1. Tamaño de muestra con precisión de 15%, tasa de no repuesta 15% y confiabilidad de 90%



Adicionalmente, se debe mencionar que, en el muestreo estratificado, con frecuencia los resultados se requieren para ciertos estratos de la población objeto de estudio y los límites de error deseados se establecen para cada uno de ellos; en este caso, se debe calcular por separado el tamaño en cada grupo y el tamaño de muestra final será la suma de las establecidas para cada estrato. Esta misma situación puede aplicarse cuando se requiere hacer estimaciones por separado para dominios estudio.

Ejemplo del cálculo del tamaño de muestra para la ENPOL 2016

El cálculo del tamaño de muestra se llevó a cabo por separado en cada entidad y estrato (centro penitenciario). La muestra se obtiene a partir de un diseño muestral aleatorio simple para estimar una proporción, por lo que el efecto de diseño DEFF=1, con un nivel de confianza del 90%, un error relativo del 13% y una tasa de no respuesta del 15%. El tamaño de muestra se obtiene con la siguiente ecuación:

$$n_{eh} = \frac{\frac{z^2 q_{eh}}{r^2 p_{eh}}}{1 + \frac{z^2 q_{eh}}{N_{eh} r^2 p_{eh}}} \times \frac{1}{1 - TNR_{eh}}$$

Donde:

n_{eh}= tamaño de muestra en el h-ésimo estrato, en la e-ésima entidad.

N_{eh}= total de población privada de la libertad del marco en el h-ésimo estrato, en la e-ésima entidad.

p_{eh}= proporción del h-ésimo estrato, de la e-ésima entidad.

q_{eh}=1-p_{eh}.

r=error relativo máximo aceptable.

z=valor asentado en las tablas estadísticas de la distribución normal estándar que garantiza obtener las estimaciones con una confianza prefijada.

DEFF=efecto de diseño definido como el cociente de la varianza en la estimación del diseño utilizado, entre la varianza obtenida considerando un muestreo aleatorio simple para un mismo tamaño de muestra.

TNR_{eh}=tasa de no respuesta máxima esperada en el h-ésimo estrato, en la e-ésima entidad.

Ejemplo del cálculo del tamaño de muestras para la ENOE 2019

El tamaño de la muestra para cada uno de los dominios[8] de estudio se calculó utilizando la Tasa de Desempleo Abierto (TDA), considerado Indicador Objetivo principal de la encuesta, y la que requiere el tamaño de muestra mayor, lo que garantiza que las estimaciones del resto de los indicadores objetivo queden cubiertas con este tamaño. La expresión empleada para el cálculo es la siguiente:

$$n = \frac{z^2 qDEFF}{r^2 p(1-tnr)TNP * PHV}$$

Donde:

n=tamaño de la muestra.

p=estimación de la proporción de interés: la tasa de desempleo abierto.

q=1-p.

r=error relativo máximo aceptable.

z=valor asentado en las tablas estadísticas de la distribución normal estándar que garantiza obtener las estimaciones con una confianza prefijada.

DEFF=efecto de diseño definido como el cociente de la varianza en la estimación del diseño utilizado, entre la varianza obtenida considerando un muestreo aleatorio simple para un mismo tamaño de muestra.

tnr=tasa de no respuesta máxima esperada.

TNP=tasa neta de participación.

PHV=promedio de habitantes de 15 años y más de edad por vivienda.

[8] La ENOE se compone de tres grandes dominios de estudio, los que conforma el agregado nacional, el primer dominio está conformado por las 39 ciudades autorrepresentadas, el segundo dominio es el complemento urbano de alta densidad y el tercero conformado por el rural.

Figura 3.2. Tamaños de muestra calculados para la ENOE 2019

Indicador	Nacional	Entidad	Ciudades Auto representadas
Confianza	90%	90%	90%
DEFF	4.00	1.68	1.15
TNP	0.40%	60%	64%
PHV	2.7	2.6	2.47
r	4.11%	5.00%	15%
TDA (p)	3.38%	≥ 5%	≥ 5.4%
tnr	15%	15%	15%
Tamaño de muestra	132,065	2,900	1,800
Tamaño de muestra ajustado	132,280	2,900	2,100

A nivel de ciudad auto representada los tamaños de muestra varían debido al tamaño de su población objeto de estudio y a su actividad económica. El tamaño mínimo que se requiere a nivel ciudad auto representada para dar resultados estadísticamente confiables es de 1 800 viviendas. La mayoría de las ciudades tienen un tamaño de 2 100 viviendas y solo las ciudades de Guadalajara, León, Monterrey, Puebla y Torreón tienen una muestra de 3 000 viviendas. Para el caso del Área Metropolitana de la Ciudad de México el tamaño de muestra es de 5 100 viviendas.

Valliant et. al (2018) presenta una excelente revisión para la estimación de tamaños de muestra para diseños muestrales unietápicos y multietápicos.

Las ecuaciones que se presentaron como ejemplos para el cálculo de tamaños de muestra representan las cantidades mínimas de unidades para garantizar que los criterios de errores relativos y confiabilidad planeados se lograrán.

3.2 Distribución de la muestra en los estratos

Por otra parte, aun cuando los resultados no se requieran a ese nivel en particular, se debe tener cuidado de asignar muestra a todos los estratos (L) contemplados en el marco de muestreo. En caso de que el tamaño de muestra (n) no lo permita, se deben hacer los ajustes (uniones de estratos) necesarios para no dejar a ninguno sin representación.

Al respecto, existen varias formas de distribuir o afijar la muestra dentro de los diferentes estratos (n_h); la elección de alguna de estas opciones está en función del conocimiento y comportamiento de las varianzas (S_h) y de los costos (C_h) de captar un cuestionario en cada uno de los estratos.

Distribución de igual número de cada estrato

Esta opción es la más sencilla de aplicar y asume que los estratos presentan varianzas, costos y tamaños iguales (N_h). Si esto no es así, redunda en estimaciones más pobres que las tres siguientes alternativas; sin embargo, puede ser útil cuando se pretende obtener resultados con precisiones semejantes en los diferentes estratos. Para obtener la distribución se aplica la expresión:

$$n_h = \frac{n}{L}$$

Distribución proporcional

Se usa si los estratos presentan varianzas iguales, costos también iguales y sus tamaños son distintos. La distribución se obtiene con:

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

Distribución de Neyman

Se aplica cuando los costos son iguales en todos los estratos, las varianzas son distintas y los tamaños de los estratos también son distintos. En este caso, la expresión a emplear es:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}\ n$$

Distribución óptima

Se emplea cuando se tienen costos muy diferentes por estrato, las varianzas son distintas y los tamaños de los estratos también son diferentes. La distribución se genera con:

$$n_h = \frac{\frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}\ n$$

En la distribución óptima también es posible minimizar un costo total de la muestra sujeto a una varianza fija o un coeficiente de variación fijo (Valliant, et al., 2018).

De donde para las fórmulas anteriores se debe de entender:

N = Tamaño de la población objeto de estudio

L = número de estratos

N_h = Tamaño de la población objeto de estudio en el h-ésimo estrato

n= tamaño de la muestra

n_h = tamaño de muestra en el h-ésimo estrato

S_h = varianza en el h-ésimo estrato

C_h= costo en el h-ésimo estrato

3.3. Selección de la muestra

La selección de la muestra se refiere a los procedimientos empleados para identificar a las unidades de observación que integrarán la muestra. La selección puede realizarse con o sin reemplazo. En la primera situación se permite que una observación pueda estar en la muestra más de una vez, mientras que en la segunda los elementos ya seleccionados lo hacen en forma única.

En este caso, el algoritmo empleado debe garantizar que la selección sea aleatoria y que la probabilidad de selección de las unidades de muestreo sea la que establece el esquema de muestreo elegido, para lo cual debe tenerse información clara y precisa del número de elementos que integran cada grupo de unidades del marco de muestreo sujeto de selección independiente, o en su defecto, la magnitud de la variable que se utiliza como referente para la selección. Es recomendable que la selección se haga de manera automatizada a fin de evitar al máximo los errores de conteo y el sesgo de información.

Es prudente mencionar que existen varios procedimientos de selección para un mismo esquema de muestreo; por ejemplo, en el muestreo aleatorio sin reemplazo (MAS) las probabilidades de selección en la muestra para cada elemento *k*, en una población objeto de estudio de tamaño *N,* son *n/N*, *k=1,…,N*. Los siguientes son dos ejemplos de selección de muestras MAS suponiendo la disponibilidad de un listado de elementos en la muestra:

Esquema A

Se generan ε_1, ε_2,…, números pseudo aleatorios independientes de una distribución uniforme en el intervalo $(0,1)$. Solamente si ocurre que $\varepsilon < \mathrm{n/N}$, entonces el elemento k=1 se selecciona. Para los subsecuentes elementos k=2, 3, …, n_k, es el número de elementos seleccionados de los primeros k – 1 elementos en la lista. Solamente si ocurre que

$$\varepsilon_k < \frac{n - n_k}{N - k + 1}$$

entonces el elemento *k* es seleccionado. El procedimiento se detiene cuando $n_k = n$.

Figura 3.3. Ejemplo del esquema A, población objeto de estudio N=15 y muestra de tamaño n=5

k	ε_k	n_k	$\frac{n-n_k}{N-k+1}$	¿ $\varepsilon_k < \frac{n-n_k}{N-k+1}$?
1	0.55	0	5/15=0.33	No
2	0.18	0	5/14=0.36	Sí
3	0.23	1	4/13=0.31	Sí
4	0.20	2	3/12=0.23	Sí
5	0.16	3	2/11=0.18	Sí
⋮	⋮	⋮	⋮	⋮

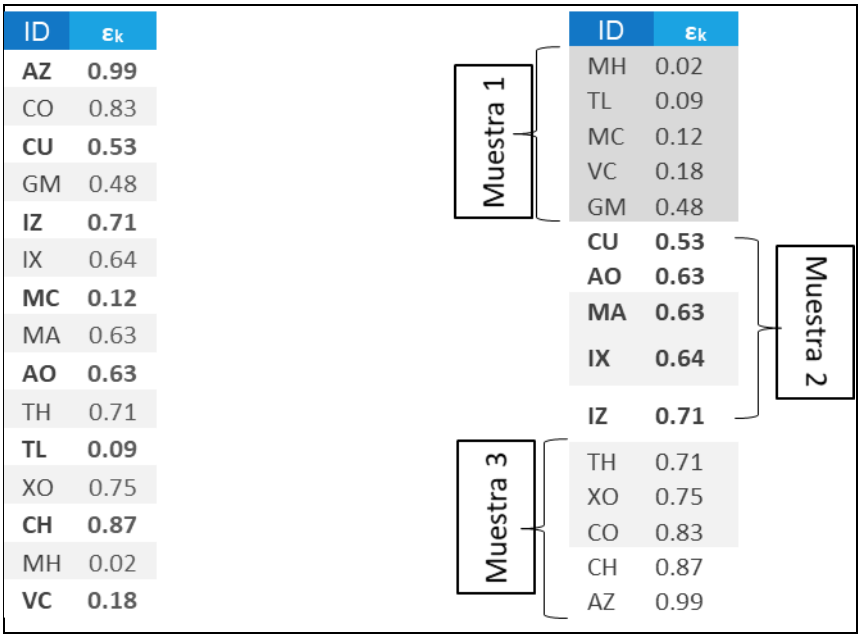
Esquema B

Este procedimiento es recomendable para poblaciones de viviendas o unidades económicas que se requiere que no sean seleccionadas frecuentemente. A este tipo de muestras sin traslape se les denominan negativamente coordinadas; las muestras con traslape máximo se conocen como positivamente coordinadas. Se generan *N* números pseudo aleatorios, con distribución uniforme en el intervalo $(0,1)$, $\varepsilon_1, \dots, \varepsilon_k, \dots, \varepsilon_N$, donde ε_k está asignado al elemento *k*. Estos números se ordenan de manera ascendente

$$\varepsilon_{(k_1)} < \varepsilon_{(k_2)} < \cdots < \varepsilon_{(k_N)}$$

Está notación indica que el i –ésimo valor más pequeño de los *N* valores ε está ligado al elemento $k_i; i = 1, \dots, N$. La primera muestra está conformada por los primeros *n* valores ε correspondientes al conjunto $\{k_1, \dots, k_n\}$, la segunda muestra corresponde a los elementos ligados con los siguientes *n* valores ε asociados con el conjunto $\{k_{n+1}, \dots, k_{2n}\}$ que no se traslapa con la primera muestra.

Figura 3.4. Ejemplo del esquema B: población objeto de estudio de tamaño N=15 y muestra de tamaño n=5



Esquema de muestreo proporcional al tamaño

El muestreo con probabilidades proporcionales al tamaño es un método de selección de muestra cuando la probabilidad de selección de una unidad es directamente proporcional a una medida de tamaño positiva, que es disponible para todas las unidades en el marco muestral y además aproximadamente proporcional al Indicador Objetivo. Ejemplos de muestreo con probabilidad proporcional al tamaño son: sin reemplazo, con reemplazo y sistemático.

Para seleccionar una muestra con probabilidad proporcional al tamaño se pueden usar varios algoritmos; se presenta como ejemplo el esquema de Poisson. Las probabilidades de inclusión se definen como $\pi_k = n \times x_k / t_x$, donde t_x es el total poblacional de la medida de tamaño x. Se generan $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ números pseudo aleatorios independientes de una distribución uniforme en el intervalo $(0,1)$. Solamente si ocurre que $\varepsilon_k < \pi_k$, entonces el elemento *k* se selecciona. Este procedimiento se repite para todos los elementos en la población objeto de estudio. En este esquema el tamaño de muestra no es fijo. Existen otros esquemas donde el tamaño de muestra es fijo, por ejemplo, el muestreo PPT sistemático, en el cual se calculan sumas parciales de la lista ordenada de x , definidas como $S_k = \sum_{j=1}^{k} \pi_j$. Se seleccionan las n unidades k que satisfacen $S_{k-1} < \varepsilon + i - 1 \leq S_k$ para $i = 1, \ldots, n$ y ε es un número pseudo aleatorio de una distribución uniforme en el intervalo $(0,1)$.

Si se emplea un muestreo estratificado o por conglomerados, la selección debe realizarse de manera independiente en cada estrato o conglomerado y se debe verificar que el número de unidades de muestreo seleccionadas por grupo coincida con el tamaño de muestra total.

Un inventario de 46 métodos para la selección de muestras aleatorias puede ser consultado en Tillé (2006).

ACTIVIDAD 4: DEFINICIÓN DE ESTIMACIONES PARA MUESTREO PROBABILÍSTICO

Dado el esquema de muestreo y las características de la muestra, se definen las bases técnicas para el cálculo de las estimaciones, lo cual implica considerar los ponderadores y las medidas de precisión.

4.1 Cálculo de ponderadores

El ponderador o factor de expansión es un concepto relacionado con la probabilidad de selección y se interpreta como la cantidad de unidades en la población objeto de estudio que representa una unidad en la muestra, llámese personas, viviendas, unidades económicas, entre otras. Dicho ponderador permite dar conclusiones sobre la población objeto de estudio.

El hecho de que a partir de una muestra se infieran estimaciones sobre indicadores objetivo para la población objeto de estudio, implica la expansión de la muestra de acuerdo con los llamados ponderadores, que técnicamente se definen como el inverso de la probabilidad de selección. Esto se aplica en dos actividades básicas: el cálculo de los ponderadores y el análisis de los ajustes que se deban considerar en estos; por ejemplo, ajuste por no respuesta, ajuste por estimación del tamaño de la población objeto de estudio, otro tipo de ajuste por característica particular del ponderador y de la encuesta.

En una primera etapa, dada su definición, los ponderadores se calculan con la información del marco de muestreo, dado que solo se requiere conocer la probabilidad de selección de cada unidad de observación acorde con el esquema de muestreo elegido. A este concepto se le denominará ponderador de diseño, ya que es el resultado del inverso de la probabilidad de selección de la muestra inicial.

Los diseños muestrales sistemático y MAS sin reemplazo producen muestras autoponderadas[9]. El muestreo estratificado con distribución proporcional de la muestra también es otro ejemplo de diseño autoponderado; su ponderador es

$$\frac{N_h}{n_h}$$

Usando el hecho de que la distribución es proporcional se verifica que el ponderador es igual para cada elemento en la población objeto de estudio

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

Cuando un diseño de la muestra no usa distribución proporcional de la muestra, debido a que se aplican ajustes por no respuesta y/o calibración (ver secciones 4.2 y 4.2.1) diferentes por cada estrato a los ponderadores, el diseño deja de ser autoponderado.

4.2 Ajuste de los ponderadores

Esta actividad se realiza hasta que culmina la Fase de Captación, sin embargo, en el subproceso del Diseño de la Muestra se definen los ajustes que se realizarán a los ponderadores. Uno de los ajustes implica evaluar el nivel de la no respuesta, procediéndose a los ajustes necesarios considerando lo siguiente:

[9] Cada elemento en la población objeto de estudio tiene la misma probabilidad de ser incluida en la muestra.

- Que las unidades con respuesta tendrán pesos mayores a los planeados en la estimación, para compensar a los valores que se perdieron debido a la no respuesta (ajuste por no respuesta).

- Que en cada dominio de estudio se obtenga la misma población a la determinada por una estimación de la población objeto de estudio confiable y referida a la misma fecha del levantamiento de la encuesta (ajuste por estimación de la población objeto de estudio).

El ajuste por no respuesta y por estimación de la población objeto de estudio pueden incluirse en un marco teórico más general propuesto por Deville & Särndal (1992) que denominaron calibración. Supóngase un diseño muestral aleatorio simple con los ponderadores de diseño $d_k = n/N$ y se cuenta con un conjunto de información auxiliar $\boldsymbol{x}$ relacionada con la población objeto de estudio, el estimador de un Indicador Objetivo es el total $\hat{t}$ calculado como

$$\hat{t} = \sum_{k=1}^{n} d_k \, y_k$$

Los ponderadores calibrados reemplazan este estimador por

$$\hat{t}_{\mathrm{W}} = \sum_{k=1}^{n} w_k \, y_k$$

el ponderador calibrado w_k se define en la siguiente ecuación

$$w_k = g_k \times d_k$$

donde g_k es el peso de corrección. Dos condiciones se tienen que cumplir en la calibración de los ponderadores:

1. Los ponderadores calibrados g_k tienen que ser muy cercanos a 1.

2. La distribución muestral calibrada de la información auxiliar debe coincidir con la distribución poblacional

$$\bar{x}_W = \frac{1}{N} \sum_{k=1}^{n} w_k \, x_k = \overline{\boldsymbol{X}}$$

La primera condición garantiza que los estimadores obtenidos serán insesgados o aproximadamente insesgados y la segunda condición garantiza que la muestra ponderada por los ponderadores calibrados w_k es representativa de la información auxiliar usada.

La función de distancia entre g_k y 1, denotada como $D(g_k, 1)$ se utiliza para minimizar

$$\sum_{k=1}^{n} D(g_k \, , 1)$$

sujeta a la condición de calibración 2. Este problema de optimización se puede resolver utilizando el método de multiplicadores de Lagrange. La función de distancia $D(g_k, 1)$ no es única, por ejemplo:

a) $D(g_k, 1) = (g_k - 1)^2$ define una calibración lineal (regresión lineal)

b) $D(g_k, 1) = g_k \, log(g_k) - g_k + 1$ define una calibración multiplicativa (*raking*)

En la figura 4.1 se presenta un ejemplo de calibración multiplicativa, se tienen los datos de la muestra y los totales marginales para la población objeto de estudio. La calibración se hace por edad y grupo de edad. En el primer paso los pesos de calibración $w_k^{(1)}$=total grupo edad poblacional/total grupo edad muestra paso inicial; en el

segundo paso los pesos son $w_k^{(2)} = w_k^{(1)} \times$ total grupo sexo poblacional/total grupo edad muestra paso 1. Por ejemplo, el peso de calibración de un adulto hombre en el paso 2 es 1.035×11.45. El procedimiento concluye cuando totales por edad y sexo de las estimaciones ajustadas por calibración son iguales a los totales poblacionales o las diferencias son muy pequeñas.

Figura 4.1. Ejemplo calibración multiplicativa

Inicio

Muestra			
	Hombres	**Mujeres**	**Total**
Joven	23	15	38
Adulto	16	17	33
Adulto M	13	16	29
Total	52	48	100

Población objeto de estudio			
	Hombres	**Mujeres**	**Total**
Joven	-	-	435
Adulto	-	-	296
Adulto M	-	-	269
Total	511	489	1000

Paso 1. Edad (ajuste por totales de renglón)

Pobl/Muestra	w_k	Muestra			
			Hombres	**Mujeres**	**Total**
435/38	11.45	**Joven**	263.29	171.71	435
296/33	8.97	**Adulto**	143.52	152.48	296
269/29	9.28	**Adulto M**	120.59	148.41	269
		Total	527.39	472.61	1000

Paso 2. Sexo (ajuste por totales de columna)

	Muestra		
	Hombres	**Mujeres**	**Total**
Joven	255.11	177.67	432.77
Adulto	139.05	157.77	296.83
Adulto M	116.84	153.56	270.40
Total	511	489	1000
w_k	0.969	1.035	
Pobl/Muestra	511/527.39	489/472.61	

Los ponderadores calibrados w_k obtenidos después de aplicar cualquier método de calibración pueden tener las siguientes desventajas:

i) Ponderadores extremadamente grandes pueden generar estimaciones altamente inestables.

ii) La calibración lineal puede producir ponderadores negativos.

No obstante, otros métodos de calibración permiten mantener los ponderadores calibrados dentro de límites preestablecidos y obtener inferencias válidas. En el caso de encuestas con diseños complejos como el muestreo por conglomerados, se obtiene información para hacer inferencias para dos poblaciones: la población de viviendas y la población de personas habitantes de las viviendas. Los métodos de calibración pueden aplicarse en ambos casos lo cual resulta en dos conjuntos de ponderadores calibrados asignadas a cada registro, lo cual complica el análisis de la información. Bethlehem, *et al.* (cap. 8, 2011) describen un procedimiento para obtener ponderadores calibrados únicos tanto para las viviendas como para personas en la muestra.

Por ejemplo, la Encuesta telefónica de Ocupación y Empleo utilizó una calibración multiplicativa, *raking ratio*, para asegurar que en los dominios de interés se obtuvieran los totales poblacionales determinados por la ENOE del primer trimestre de 2020.

Posterior a estos posibles ajustes y antes de declarar como liberados los ponderadores "definitivos", debe verificarse que en cada dominio para el que se pretende obtener estimaciones, la expansión obtenida para los indicadores objetivo sea congruente con el total obtenido con el marco.

Los ponderadores definitivos son aplicados durante la Fase de Procesamiento para la explotación de resultados.

4.3 La no respuesta en encuestas probabilísticas

La no respuesta ocurre cuando un informante seleccionado en la encuesta no proporciona la información solicitada. Existen dos tipos de no respuesta:

- Por unidad o total: un elemento seleccionado no proporciona ninguna información.
- Por variable: el informante seleccionado responde algunas preguntas, pero no todas.

La no respuesta es un problema debido a que el tamaño de la muestra se reduce, lo cual afecta la precisión planeada de los estimadores de la encuesta, y además se genera un sesgo por la no respuesta selectiva en algunos subgrupos de la población objeto de estudio o subconjuntos de preguntas en el cuestionario de la encuesta. De forma general, la no respuesta por unidad o total y por ítem son tratados con métodos de ajuste de los ponderadores muestrales y métodos de imputación, respectivamente (Bethlehem et al., cap. 12 y 14, 2011). La imputación múltiple (Schafer, 1997) puede emplearse para el tratamiento de la no respuesta por unidad; los métodos de ajuste de los ponderadores muestrales son aplicables para tratar la no respuesta por ítem cuando se requiere estimar totales, medias o proporciones de un Indicador Objetivo.

Una medida de la calidad de la información es la tasa de no respuesta. Este indicador de tasa de no respuesta para un ciclo del programa se calcula a nivel muestral conforme a los códigos de no respuesta obtenidos en campo; así como ponderado, de acuerdo con el inverso de la probabilidad de selección de las unidades en muestra:

a. La tasa de no respuesta sin ponderar, usando los conteos simples de la muestra, ofrecen una descripción útil del éxito operativo de la encuesta, es decir, se puede monitorear el éxito de la encuesta para obtener respuestas en la encuesta

b. La tasa de no respuesta ponderada se calcula como el cociente de la suma de los pesos de las unidades que responden y la suma de los pesos de todas las unidades en la muestra. Las tasas de no respuesta ponderadas ofrecen una mejor descripción del éxito de la encuesta con respecto a la población objeto de estudio. Son útiles cuando se trata de encuestas con probabilidades distintas de inclusión; en el caso de muestreo aleatorio simple las tasas de no respuesta ponderadas y sin ponderar son iguales.

El CoAC aprobó los indicadores de calidad[10] relacionados con la no respuesta:

1. Tasa de no respuesta antes de imputación a nivel unidad de observación (TNR_AI).
2. Tasa de no respuesta después de imputación a nivel unidad de observación (TNR_DI).
3. Tasa de imputación a nivel unidad de observación (TI).

En el caso de muestreo probabilístico o de un muestreo no probabilístico basado en el tamaño de alguna variable de diseño también se deberá calcular la versión ponderada de estos indicadores.

[10] https://extranet.inegi.org.mx/calidad/indicadores-de-calidad-y-evaluaciones/

4.4 Cálculo de las estimaciones y precisiones estadísticas

Liberada la base de datos definitiva y disponiéndose de los ponderadores, una actividad más que se considera en la Fase de Análisis de la Producción y en el Diseño de la Muestra, se refiere al cálculo de estimaciones y las precisiones estadísticas correspondientes.

4.4.1 Cálculo de las estimaciones

Disponiéndose de la información captada, los ponderadores y la expresión matemática para cada estimador, el cálculo de las estimaciones se realiza por sustitución. Así, para estimar el total de un Indicador Objetivo, se suman los productos generados a partir de multiplicar el valor obtenido para determinada variable de cada unidad de observación por el ponderador correspondiente. De manera similar y respetando la expresión del estimador, se construyen las estimaciones para promedios, proporciones y razones.

Idealmente, el estimador elegido debe satisfacer las siguientes características:

- Ser insesgado; en promedio las estimaciones deben ser iguales al valor poblacional.
- Consistente; cuando se incrementa el tamaño de la muestra la estimación se acerca al valor poblacional.
- Eficiente; las variaciones de los resultados de las posibles muestras deben ser pequeñas (precisión).
- Ser fácil de obtener y calcular (una combinación lineal de valores observados).

Sin embargo, las propiedades de esta lista no pueden obtenerse para la mayoría de los estimadores. En la realidad, el tamaño de muestra *n* es finito y podría considerarse "grande". El sesgo de un estimador es la diferencia entre el valor esperado, promedio de la distribución muestral, de un estimador y el valor verdadero de un estadístico poblacional. Formalmente, si θ es el estadístico poblacional y $\hat{\theta}$ es el estimador de dicho estadístico, el sesgo se define como

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Si el $Sesgo(\hat{\theta}) = 0$ entonces $\hat{\theta}$ es un estimador insesgado del valor poblacional del estadístico θ. Si el estimador es asintóticamente (teóricamente cuando *n* se va incrementando infinitamente) insesgado, entonces puede considerarse como aproximadamente insesgado cuando *n* es suficientemente grande. De esta manera la consistencia del estimador se cumple, y entonces la distribución muestral del estimador puede considerarse que se encuentra alrededor del valor poblacional (Särndal et al. 1992, cap. 5).

Existen dos situaciones en las cuales no es posible usar estimadores exactamente insesgados: (i) para varios parámetros es difícil encontrar un estimador insesgado, (ii) un estimador con sesgo puede tener un error cuadrático medio menor (derivado de tener una varianza más pequeña) que cualquier estimador insesgado.

Para varios estimadores se cuenta con fórmulas específicas para el cálculo de varianza de los estimadores. Frecuentemente se requiere estimar otras cantidades que no son funciones de totales a partir de datos de una encuesta para las cuales no se cuenta con fórmulas exactas de varianza. Por ejemplo, para un estimador de razón $\hat{R} = \frac{\widehat{\theta_y}}{\widehat{\theta_x}}$ la varianza del estimador $\hat{R}$ no es igual al cociente de las varianzas de los estimadores $\widehat{\theta_y}$ y $\widehat{\theta_x}$. Para este ejemplo se emplea una técnica de linearización para obtener un estimador aproximado de la varianza de $\hat{R}$. Además de técnicas de linearizacion para obtener estimaciones de varianzas, también se utilizan técnicas de remuestreo como *jackknife* y *bootstrap*. Una descripción de las técnicas para estimación de varianzas de estimadores para encuestas se encuentra en Wolter (2007).

4.4.2 Cálculo de las precisiones estadísticas

Como parte de la evaluación de la calidad de la información captada por la encuesta, en particular para conocer si se cumplieron las expectativas de confiabilidad de los estimadores, se calculan las precisiones de estos. Esta actividad consiste en obtener para todos los indicadores objetivo:

- El error estándar es una medida de la dispersión esperada de las estimaciones muestrales alrededor del parámetro poblacional; el cual se obtiene a partir del cálculo de la estimación de la varianza $Var(\hat{\theta})$ para el estimador $\hat{\theta}$ del esquema de muestreo empleado, de esta manera, el error estándar $EE(\hat{\theta})$ se define como $EE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$. Un error estándar grande sugiere menos confianza en el estimador muestral.

- El factor de corrección por población finita (fpc) se usa para ajustar la varianza estimada del estimador $\hat{\theta}$, debido a que se obtiene con datos de la muestra, es decir, la población finita no es completamente observada o medida. La variabilidad del estimador se debe a los datos que no están en la muestra. Si el fpc se ignora, la consecuencia es la sobreestimación del error estándar de $\hat{\theta}$. La varianza estimada debe ajustarse hacia la baja a medida que el tamaño de la muestra n se incrementa, ya que los datos observados en la muestra son completamente conocidos y la contribución a la varianza de $\hat{\theta}$ proviene de los $N - n$ casos que no se incluyen en la muestra. Para el muestreo aleatorio sin reemplazo fpc= $\frac{N-n}{N} = 1 - f$ donde $f = \frac{n}{N}$ se denomina la fracción de muestreo, la cantidad $1 - f$ se aplica a la varianza estimada de $\hat{\theta}$. Dicha varianza se reduce a cero cuando $n = N$ y se acerca al valor completo cuando el tamaño de muestra n se reduce. Esto se reduce a lo siguiente

$$fpc = \begin{cases} 0, & n \longrightarrow N \\ 1, & n \longrightarrow 0 \end{cases}$$

  El fpc se puede ignorar si la fracción de muestreo f no es mayor al 5% o incluso puede ser tan alto como 10% (Cochran, 1977). Cuando una muestra se diseña con estratificación y los fpc apropiados se aplican en cada estrato, uno o más de los estratos tendrán fracciones de muestreo f altas lo que generará una reducción considerable de los errores estándares de las estimaciones, por ejemplo, en encuestas donde hay estratos en los que todas sus unidades son seleccionadas en la muestra, la contribución a los errores estándares de las estimaciones será nula. En muestreos multiétapicos habrá fracciones de muestreo f distintas en cada etapa que se verán involucradas en la estimación de la varianza de $\hat{\theta}$. El efecto de los fpc en la estimación de la varianza depende de cómo varían los indicadores objetivo analizadas dentro de las unidades primarias de muestreo (UPM) o entre ellas. No obstante, la fracción de muestreo en la primera etapa, la muestra de UPM, puede ser ignorada si es lo suficientemente pequeña. Si esta condición se cumple entonces las variaciones de la medias o totales de las UPM incorporará automáticamente cualquier fpc aplicable al submuestreo dentro de las UPM.

  Como consecuencia, si la fracción de muestreo de la primera etapa es pequeña, entonces la contribución a la varianza $\hat{\theta}$ del submuestreo en las siguientes etapas no será considerable; está es una ventaja del método del "ultimo conglomerado". El único caso donde se tiene que poner atención especial en los diseños biétapicos sucede cuando las fracciones de muestreo de las UPM y de las unidades en la segunda etapa son grandes. El efecto en la estimación del error estándar de $\hat{\theta}$ será mayor cuando los indicadores objetivo medidos tengan una alta variabilidad entre las UPM. En este caso la inclusión de la fpc puede hacer una gran diferencia en la estimación de la varianza de $\hat{\theta}$.

- Los intervalos de confianza se determinan de acuerdo con el nivel de confianza establecido para el cálculo del tamaño de muestra y empleando la varianza del estimador. La fórmula general de un intervalo de confianza de $100(1 - \alpha)\%$ para $\hat{\theta}$, asumiendo que $\hat{\theta}$ tiene distribución normal, es

$$\hat{\theta} \pm z_{1-\alpha/2}\left[V(\hat{\theta})\right]^{1/2}$$

Los intervalos de confianza son los valores en los que se espera esté contenido el valor del parámetro con cierta probabilidad.

- El efecto de diseño $DEFF_p(\hat{\theta})$ para el estimador $\hat{\theta}$ bajo el diseño muestral p, con un tamaño de muestra n fijo

$$DEFF_p(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

Donde

$Var_p(\hat{\theta})$ es la varianza para el estimador $\hat{\theta}$ bajo el diseño muestral p.

$Var_{MAS}(\hat{\theta})$ es la varianza para el estimador $\hat{\theta}$ bajo el diseño muestral aleatorio simple sin reemplazo.

Cuando $DEFF_p(\hat{\theta}) > 1$ indica que la varianza del estimador $\hat{\theta}$ es mayor usando el diseño p en comparación con la varianza de $\hat{\theta}$ en el muestreo aleatorio simple sin reemplazo.

En la determinación del intervalo de confianza, en casos de un diseño muestral complejo, debe considerarse el efecto del diseño, multiplicando los límites de éste por la raíz cuadrada del DEFF calculado (DEFT)

$$\hat{\theta} \pm z_{1-\alpha/2}\left[Var_{MAS}(\hat{\theta})\right]^{1/2} \times DEFT$$

El DEFT es un factor que ajusta los errores estándar debido al uso de un diseño muestral complejo. Tiene las siguientes interpretaciones:

a) DEFT=1: no tiene efecto en las estimaciones de los errores estándares el diseño muestral complejo.

b) DEFT>1: el diseño muestral complejo incrementa los errores estándares de los estimadores.

c) DEFT<1: se incrementa la eficiencia usando el diseño muestral complejo (se reducen los errores estándar).

- El coeficiente de variación $CV(\hat{\theta})$ se define como

$$CV(\hat{\theta}) = \frac{\sqrt{\left(Var(\hat{\theta})\right)}}{\hat{\theta}}$$

El coeficiente de variación es una medida relativa de su precisión; conforme sus valores son más próximos a cero, la estimación es más precisa. El coeficiente de variación no tiene medidas de unidad y por lo tanto la precisión relativa de dos o más indicadores objetivo puede compararse. Para que la interpretación del coeficiente de variación sea útil, el estimador $\hat{\theta}$ debe ser positivo, ya que si $\hat{\theta}$ es muy cercano a cero, el coeficiente de variación puede ser muy inestable.

4.4.3 Indicadores de calidad (precisión) para indicadores objetivo

El CoAC adoptó un conjunto de 4 indicadores para medir la precisión de las encuestas, los cuales son de uso externo, es decir, se hacen públicos y acompañan a los resultados del Programa de Información. Los indicadores de precisión que se calculan para encuestas con muestreo probabilístico son: el coeficiente de variación, el error estándar y el intervalo de confianza. La cobertura de la variable de diseño es un indicador de calidad en encuestas con muestreo no probabilístico. La cobertura representa el porcentaje alcanzado de la variable de diseño -que puede ser la variable en la que se basa la selección no probabilística de la muestra u otra considerada en el diseño- con respecto al marco de la muestra. La fórmula de cálculo es

$$\frac{\tilde{T}_{D,m}}{T_{D,M}} \times 100$$

Donde

$\tilde{T}_{D,m}$= Total de la variable de diseño en la muestra para el dominio D

$T_{D,M}$= Total de la variable de diseño en el marco muestral para el dominio D

El Comité acordó la homologación de los umbrales y especificaciones para la semaforización del coeficiente de variación y de la cobertura de la variable de diseño en los tabulados de resultados de las encuestas con muestreo probabilístico y no probabilístico. En la figura 4.2 se presentan los umbrales y semáforos para los reportes de precisión para encuestas con muestreo probabilístico.

Figura 4.2. Umbrales aprobados para el coeficiente de variación (%) encuestas probabilísticas

Interpretación	Semaforización	Unidad de observación	
		Viviendas/ Hogares/otras unidades diferentes a las económicas	Unidades Económicas
		DGES/DGEGSPyJ	DGEE/DGEGSPyJ
Alta	Blanco	[0, 15)	[0, 20)
Moderada	Amarillo	[15, 30)	[20, 30)
Baja	Naranja	>=30	>=30
Acuerdo CAC-007/01/2018			

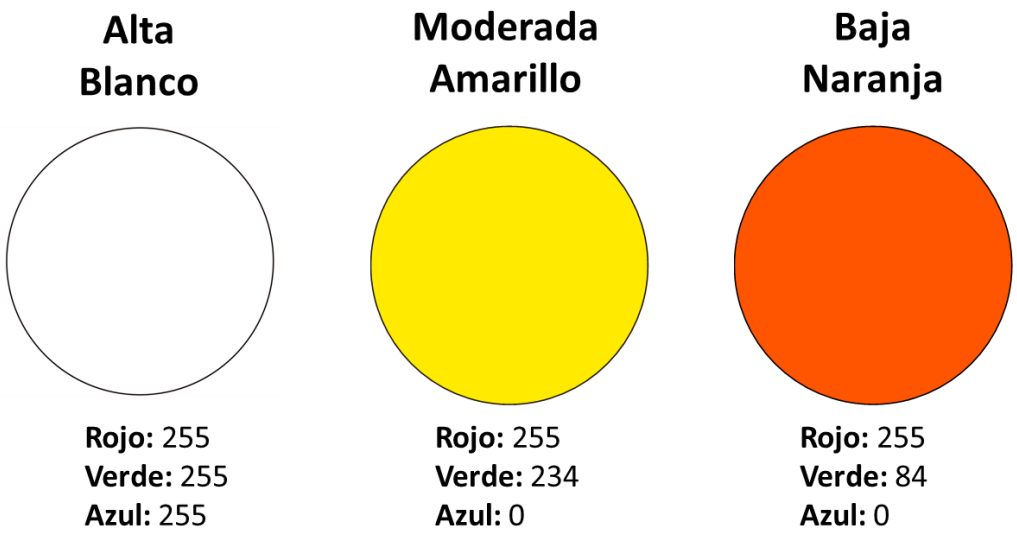
En la figura 4.3 se describen los umbrales y el semáforo de acuerdo con el nivel de cobertura de la variable de diseño para encuestas con muestreo probabilístico.

Figura 4.3. Umbrales aprobados para la cobertura de la variable de diseño (%) encuestas no probabilísticas

Interpretación	Semaforización	Unidad de observación: Unidades económicas
		DGEE
Alta	Blanco	>=80
Moderada	Amarillo	[60, 80)
Baja	Naranja	<60
Acuerdo CAC-008/01/2018		

4.4.4 Parámetros RGB para la semaforización del coeficiente de variación y de la cobertura de la variable de diseño

Figura 4.4. Semaforización del coeficiente de variación y de la cobertura de la variable de diseño



Texto explicativo para los tabulados de encuestas con muestreo probabilístico en unidades económicas

Las estimaciones que aparecen en esta figura están coloreadas, de acuerdo con su nivel de precisión, en Alta, Moderada y Baja, tomando como referencia el coeficiente de variación CV (%). Una precisión Baja requiere un uso cauteloso de la estimación en el que se analicen las causas de la alta variabilidad y se consideren otros indicadores de precisión y confiabilidad, como el intervalo de confianza.

Figura 4.5. Nivel de precisión de las estimaciones coeficiente de variación CV (%)

Alta, CV en el rango de (0,20)

Moderada, CV en el rango de [20, 30)

Baja, CV de 30% en adelante

Texto explicativo para los tabulados de encuestas con muestreo probabilístico en unidades diferentes a las económicas, incluyendo viviendas, hogares y personas

Las estimaciones que aparecen en esta figura están coloreadas, de acuerdo con su nivel de precisión, en Alta, Moderada y Baja, tomando como referencia el coeficiente de variación CV (%). Una precisión Baja requiere un uso cauteloso de la estimación en el que se analicen las causas de la alta variabilidad y se consideren otros indicadores de precisión y confiabilidad, como el intervalo de confianza.

Figura 4.6. Nivel de precisión de las estimaciones coeficiente de variación CV (%), en unidades diferentes a las económicas

Alta, CV en el rango de (0,15)

Moderada, CV en el rango de [15, 30)

Baja, CV de 30% en adelante

Texto explicativo para los tabulados en encuestas con muestreo no probabilístico en unidades económicas

Las estimaciones que aparecen en esta figura están coloreadas, de acuerdo con el nivel de cobertura de la variable de diseño (%) usada para seleccionar la muestra, en Alta, Moderada y Baja. Una cobertura Baja requiere un uso cauteloso de la información.

Figura 4.7. Nivel de cobertura de la variable de diseño (%)

Alta, cobertura en el rango de [80, 100]

Moderada, cobertura en el rango de [60, 80)

Baja, cobertura menor a 60%

ACTIVIDAD 5. IDENTIFICACIÓN DE LAS FUENTES DE ERROR TOTAL DE MUESTREO

El error total de muestreo se refiere a todas las fuentes de sesgo y varianza de estimación que pueden afectar la precisión de los datos muestrales y surgen en el diseño de la muestra, captación, procesamiento y análisis de los datos de la encuesta. Estas fuentes se pueden agrupar en dos clases básicas:

1. *Representación*. Se relaciona con la población objeto de estudio que será descrita en la encuesta. Responde a la pregunta ¿quiénes están incluidos en la muestra?
2. *Medición*. Describe los datos que se obtendrán de la unidad de observación. Responde a la pregunta ¿cuál es la temática de la encuesta?

Los costos, la carga para el informante, profesionalismo, la ética y las restricciones afectan a ambas clases de errores.

La representatividad de una muestra está afectada por las siguientes fuentes de error (Pennell, et. al 2017):

a. Error de cobertura. Las imperfecciones de un marco de muestreo causan una representatividad incompleta de la muestra para la población objeto de estudio. Esto genera una precisión deficiente en las estimaciones muestrales y sesgo de cobertura.

b. Error de muestreo. Es el error que ocurre cuando una muestra es medida en lugar de la población objeto de estudio completa. Se compone por error fijo (sesgo) y un error variable (varianza). El sesgo muestral es la falla sistemática en observar a elementos en la muestra que tiene valores distintos en el Indicador Objetivo. La varianza muestral se refiere a la variabilidad de las estimaciones por las posibles muestras obtenidas, que son afectadas por el tamaño de muestra, así como por la conglomeración, estratificación y ponderación.

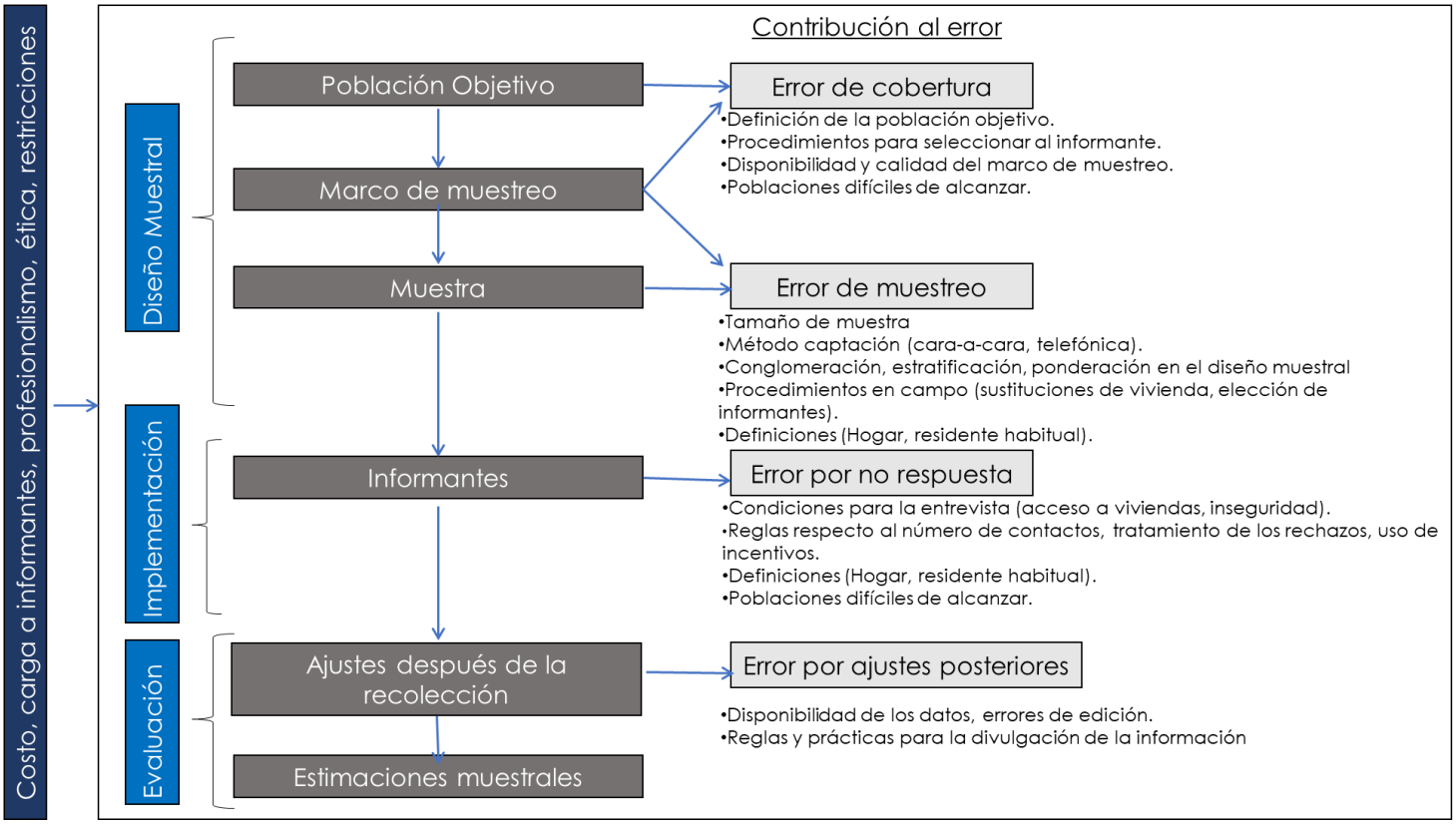
c. Error por no respuesta. La no respuesta por unidad o total sucede cuando un informante seleccionado no puede ser encontrado, se niega a ser entrevistado o existen barreras de comunicación. El sesgo por no respuesta se genera cuando la no respuesta por unidad está correlacionada con una o más indicadores objetivo.

d. Error por ajustes posteriores. Cuando se ha concluido la captación de datos de una encuesta basada en muestreo probabilístico, se llevan a cabo ajustes para considerar el sesgo de selección, errores de cobertura y errores por no respuesta. Deben notarse las diferencias entre los ajustes hechos debido al diseño de la

muestra (por ejemplo, ajustar la probabilidad de selección del informante por el tamaño del hogar) y los ajustes hechos para eliminar las diferencias entre los resultados de la estimación muestral y las estadísticas oficiales disponibles.

En la figura 5.1 se agrupan los errores por representación en cuatro categorías: de cobertura, de muestreo, por no respuesta y por ajustes posteriores.

Figura 5.1. Error total de muestreo Representación



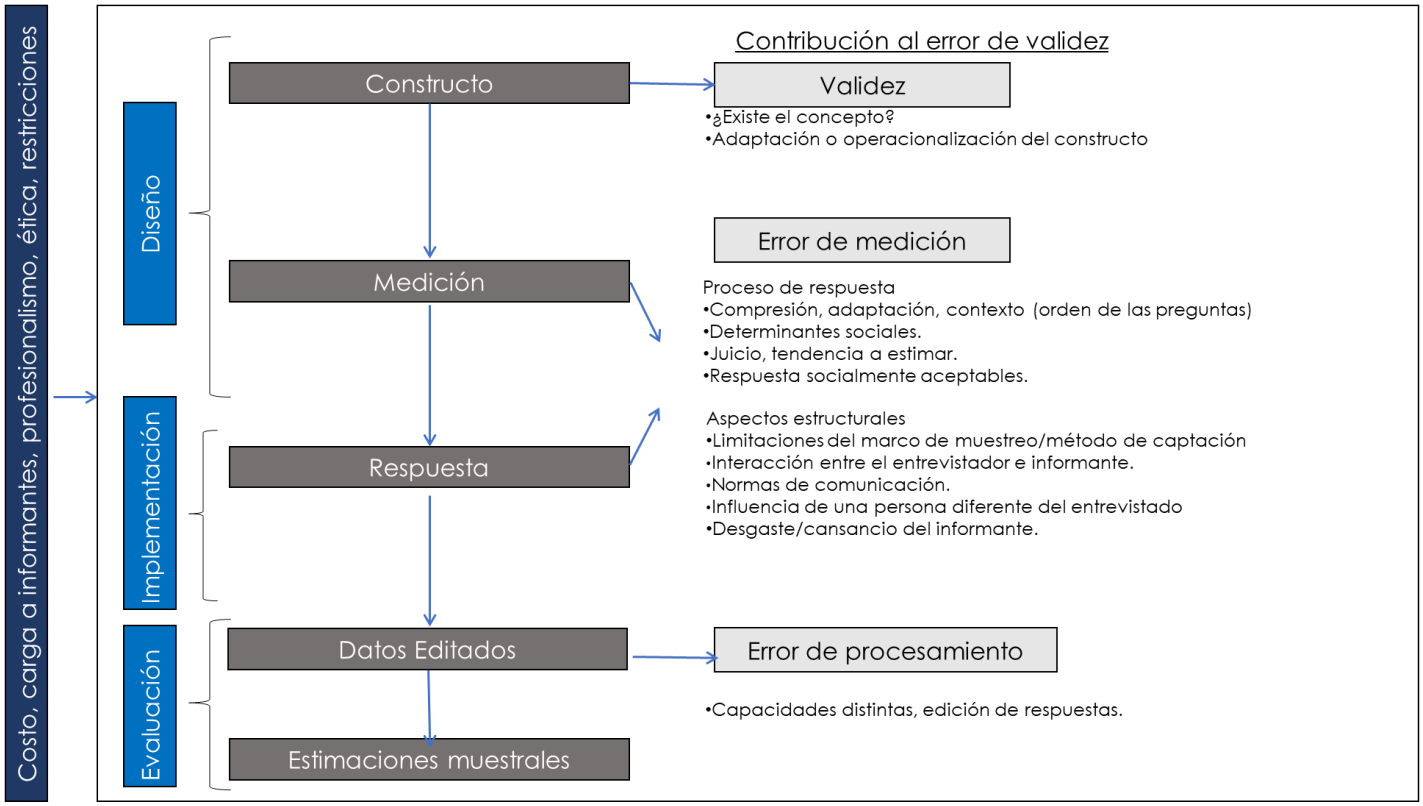
Los componentes del error de Medición, presentadas en la figura 5.2, se describen a continuación.

a. Validez. La encuesta mide correctamente el constructo planeado por la temática de la encuesta.

b. Error de medición y en el proceso de respuesta. Este tipo de errores ocurren cuando los valores verdaderos difieren del valor medido en la unidad muestreada.

c. Error de procesamiento. El procesamiento de la información de la encuesta incluye la captura, codificación y edición. El personal a cargo del procesamiento de datos puede introducir errores fijos y aleatorios.

Figura 5.2. Error total de muestreo Medición



ACTIVIDAD 6. DOCUMENTACIÓN DEL DISEÑO DE LA MUESTRA

6.1 Nota metodológica del diseño de la muestral

Finalmente, la actividad de cierre del subproceso de Diseño de la Muestra implica la elaboración de una nota metodológica del diseño de la muestra en la cual se documenten todos los aspectos importantes relacionados con el desarrollo de la propia Fase de Diseño.

Entre ellos cabe destacar, en función del caso de que se trate, los siguientes:

1. Definición de la población objeto de estudio
2. Cobertura conceptual
    - 2.1 Temporal
    - 2.2 Temática
    - 2.3 Geográfica
3. Desglose geográfico
    - 3.1 Nacional
    - 3.2 Estatal
    - 3.3 Municipal/Alcaldía
    - 3.4 Otro
4. El marco de muestreo
    - 4.1 Disponibilidad del marco muestral
    - 4.2 Fecha de actualización del marco
    - 4.3 Marco de muestreo múltiple (componentes registros administrativos, marcos de muestreo)
    - 4.4 Tamaño del marco de muestreo
    - 4.5 Descripción de las variables auxiliares contenidas del marco de muestreo
    - 4.6 Descripción de las Unidades de muestreo
        - 4.6.1 Primaria
        - 4.6.2 Secundaria
        - 4.6.3 Terciaria
        - 4.6.4 Unidad última de muestreo

5. Descripción de complejidad de la muestra

    5.1 Estratificación

    5.2 Conglomeración

    5.3 Dominios de estudio

6. El esquema de muestreo empleado

    6.1 Muestreo probabilístico

        6.1.1 Justificación para usar muestreo probabilístico

    6.2 Muestreo determinístico (no probabilístico)

        6.2.1 Justificación para usar muestreo determinístico (no probabilístico)

        6.2.2 Descripción de la metodología

    6.3 Muestreo mixto

        6.3.1 Justificación para usar muestreo mixto

        6.3.2 Descripción de la metodología (sección muestreo no probabilístico)

7. La determinación del tamaño de la muestra

    7.1 Indicador(es) objetivo principal(es) usado(s) y variables utilizadas para el cálculo de tamaños de muestra por dominios de estudio

    7.2 Nivel de precisión planeado

    7.3 Confianza planeada

    7.4 Cálculo de tamaño de muestra para dominios

    7.5 Coeficiente de Variación

    7.6 Efecto de Diseño

    7.7 Tasa de no respuesta esperada

    7.8 Número total de estratos auto representados

    7.9 Número total de estratos unidos

    7.10 Fórmula para determinar el tamaño de la muestra

    7.11 Distribución del tamaño de muestra

    7.12 Selección de la muestra

8. Métodos de captación de datos de la encuesta

    8.1 Vivienda, Unidad económica o Plaza de captación

    8.2 Entrevista directa o diferida

    8.3 Medio de captación (en papel, internet, teléfono o medios digitales)

9. Fórmulas de cálculo y ajuste de los ponderadores

    9.1 Por el nivel de no respuesta

    9.2 Ajuste por estimación de la población objeto de estudio

    9.3 Verificación del balance muestral (las congruencias del total estimado del dominio y el total obtenido en el marco muestral)

10. Fórmulas de cálculo de los estimadores empleados y errores estándares

    10.1 Nombre de los indicadores objetivo y dominios de estudio para los cuales se calculan

    10.2 Fórmulas de los estimadores

    10.3 Fórmula de estimación de la varianza y error estándar

Nota: la documentación de cálculos de las siguientes evidencias corresponde a las Fases de Análisis de la Producción y Difusión, para los cuales desde el Diseño de la Muestra se deben especificar sus fórmulas de cálculo.

11. Cálculo de las precisiones estadísticas.

    11.1 Indicadores objetivo y dominios de estudio para los cuales se calculan

    11.2 Cálculo de intervalos de confianza para los indicadores objetivo y dominios de estudio para los cuales se calculan

    11.3 Coeficiente de variación para los indicadores objetivo y dominios de estudio para los cuales se calculan

    11.4 Sólo para encuestas mixtas, para el estrato de estimación con una muestra no probabilística

        11.4.1 Nombre de las variables de diseño y cálculo del porcentaje de cobertura.

        11.4.2 Método de estimación y variables auxiliares usadas.

12. No respuesta por unidad

    12.1 Tasa de no respuesta sin ponderar

    12.2 Tasa de no respuesta ponderada

13. No respuesta por ítem

    13.1 Método de imputación

13.2 Tasa de imputación sin ponderar

13.3 Tasa de imputación ponderada

En el desarrollo de estos puntos se deben plasmar tanto las características originalmente planteadas para el Diseño de la Muestra como los ajustes realizados a las mismas durante el proceso de aplicación. Ambas situaciones deben quedar establecidas con la mayor claridad posible.

La documentación de los aspectos mencionados tiene como propósito presentar la metodología utilizada en la encuesta de una forma organizada, clara y lo más sencilla posible para facilitar su comprensión. Se pueden utilizar cuadros, gráficos e ilustraciones a fin de hacerla más atractiva.

La descripción detallada debe presentarse en el orden lógico en que se desarrolló el trabajo, considerando los problemas enfrentados durante el Diseño de la Muestra y las soluciones correspondientes que se adoptaron.

Para incrementar la utilidad de esta actividad, los expertos que participaron en este trabajo pueden aportar conclusiones y recomendaciones que se pueden incluir en un capítulo complementario.

Con ello se ofrecen elementos para realizar mejoras en el Programa actual o para encuestas futuras.

6.2 Especificación de los metadatos

Los metadatos son datos estructurados que describen las características del contenido, captura, procesamiento, calidad, condición, acceso y distribución de Información estadística o geográfica. Proporcionan a los usuarios información sobre los datos existentes, describiendo los procesos de recolección, procesamiento y evaluación que se utilizaron en su generación, así como las formas de acceder a ellos, con la finalidad de que los usuarios puedan identificar, localizar y consultar los que atiendan de mejor manera sus necesidades.

La documentación de metadatos ayuda a los usuarios con distintos niveles de especialización a:

- Encontrar los datos en los que están interesados.
- Entender qué es lo que están midiendo los datos y cómo se han creado.
- Evaluar la calidad de los datos.

El Banco Mundial (BM) y la Organización para la Cooperación y el Desarrollo Económicos (OCDE), a través del Programa Acelerado de Datos (PAD) de la Red Internacional de Encuestas de Hogares (RIEH) promovieron la Iniciativa de Documentación de Datos (DDI por sus siglas en inglés), que consiste en establecer un estándar internacional basado en archivos en formato XML para la documentación, presentación, transferencia, diseminación y preservación de datos. Este requerimiento se encuentra citado en el Art 14, fracción I, inciso c y el Art 15, fracción I, inciso e, de la NTPPIEG.

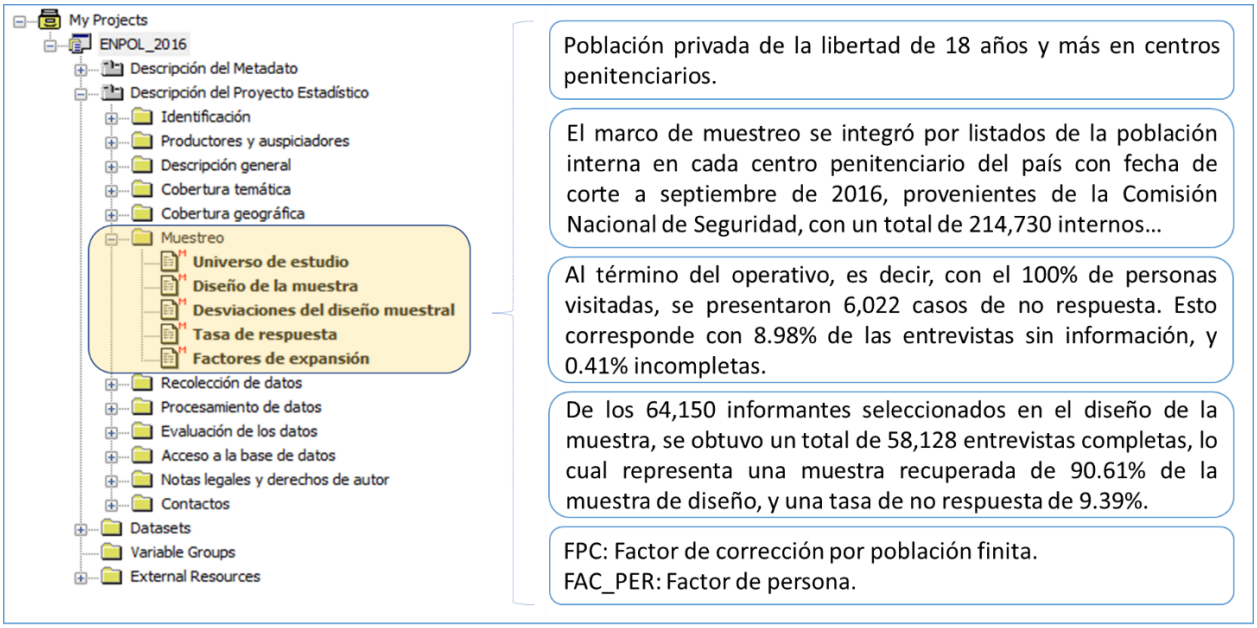
La Norma Técnica para la Elaboración de Metadatos para proyectos de generación de Información Estadística Básica y de los componentes estadísticos derivados de proyectos geográficos especifica las etiquetas, definiciones y condición de llenado de los elementos que conforman la plantilla de metadato para encuestas, descritas en la figura 6.1:

Figura 6.1. Plantilla de metadatos para los programas de información de encuestas para el Diseño de la Muestra

Etiqueta del elemento	Descripción del elemento / Contenido esperado	Condición de llenado
Universo de estudio	Descripción de la población objeto de estudio en el proyecto estadístico.	Obligatorio
Diseño de la Muestra	Información sobre el marco muestral y los procedimientos utilizados, que debe considerar los siguientes elementos: 1. Marco de la muestra. 1.1. Formación de las unidades primarias de muestreo. 1.2. Estratificación. 2. Esquema de muestreo. 3. Tamaño de la muestra. 4. Distribución del tamaño de muestra. 5. Selección de la muestra. 6. Ajustes a los factores de expansión. Deberá incluirse el nombre del campo en la base de datos que contiene las unidades primarias de muestreo, los estratos de diseño y el tamaño de la localidad.	Obligatorio
Desviaciones del diseño de la muestral	Descripción de la correspondencia entre las unidades efectivamente encuestadas y la muestra seleccionada. Se deben indicar las desviaciones importantes.	Obligatorio
Tasa de respuesta	La tasa de respuesta indica la proporción de entrevistas realizadas respecto del total de las entrevistas planeadas en el diseño de la muestra. Debe anexarse la tasa de no respuesta; en ésta se incluyen los porcentajes de rechazos de entrevista, tasa de no entrevista, así como la falta de respuesta asociada a unidades no existentes o cualquier otra condición por la que no fue posible obtener la información.	Obligatorio
Factores de expansión	En este apartado se proporcionará la lista de las variables usadas para ponderar los datos de la muestra. Si existe más de una variable ponderadora, se deberá describir la forma en que éstas difieren una de otra y cuál es el propósito de cada una de ellas.	Obligatorio

Para la documentación de los metadatos se emplea el editor Nesstar Publisher, en la figura 6.2 se presenta un ejemplo de la ENPOL.

Figura 6.2. Elementos con la Descripción del Diseño de la Muestra de la ENPOL en el editor Nesstar Publisher



GLOSARIO

Para efectos de la presente Guía se entenderá por:

Cobertura geográfica. Territorio o ámbito espacial al que se refiere la captación de datos en un proyecto estadístico.

Coeficiente de variación. Es el cociente entre la desviación estándar y la media aritmética.

Conglomerado. Es una subpoblación en la cual los elementos que la componen poseen cierta característica que les hace ser propios de cierta cualidad o atributo. Por ejemplo, lugar geográfico, grupo étnico, ideología y organización social.

Cuestionario. Tipo de instrumento de captación que presenta, bajo un orden determinado, las preguntas e indicaciones necesarias para el registro de los datos correspondientes a las unidades de observación, en un proyecto de generación de estadística básica.

Determinación de la muestra. Contempla tanto la definición del tamaño como la selección de la muestra.

Diseño Conceptual. Proceso de investigación documental y consulta a usuarios y expertos; en este se elabora el marco conceptual y los productos de información que se difundirán en sus diferentes presentaciones.

Diseño de la Muestra. Conjunto de actividades mediante las cuales se determina el método de muestreo por aplicar, el tamaño de la muestra y los procedimientos de selección, así como los elementos técnicos para la determinación de estimadores.

Distribución o afijación de la muestra. Se refiere a la distribución de la muestra entre los diferentes estratos de diseño y dominios de estudio para los que se requieren realizar estimaciones.

Documentación. Texto descriptivo utilizado para definir o describir un objeto, diseño, especificación, instrucciones o procedimiento.

Dominio de estudio. Subconjunto de la población para el cual se requiere realizar mediciones o representaciones de los conceptos de forma separada (Artículo 3, fracción XII de la NTPPIEG).

Encuesta longitudinal: Es una encuesta donde se obtienen varias observaciones en el tiempo para en un conjunto de variables de las mismas unidades en la muestra. Las encuestas de panel son encuestas longitudinales donde las observaciones durante un periodo se obtienen para una muestra fija o rotante.

Encuesta por muestreo. Método para generar información estadística mediante la captación de datos para un subconjunto de unidades seleccionadas de la población objeto de estudio.

Entrevista. Procedimiento para obtener información mediante un cuestionario en el que se presenta una serie de preguntas realizadas a un interlocutor o entrevistado.

Entropía de un diseño muestral. Es una medida del nivel de incertidumbre o de sorpresa en una muestra que será seleccionada.

Error absoluto. Es el valor positivo del error de estimación.

Error estándar. La desviación estándar de la distribución de un valor estadístico.

Error de estimación. La diferencia no conocida entre el valor estimado y el valor verdadero.

Error del marco de muestreo. Error causado por limitaciones inherentes a los insumos utilizados para producir los

datos, o por retrasos y errores en la obtención y procesamiento de los datos. La sobrecobertura, la subcobertura y los registros duplicados son los errores del marco de muestreo.

Error de muestreo. Es el error que ocurre cuando una muestra es medida en lugar de la población objeto de estudio completa. Se compone por error fijo (sesgo) y un error variable (varianza).

Error no muestral. Error en las estimaciones de la muestra que no pueden atribuirse a las fluctuaciones del muestreo. Ejemplos de errores no muestrales son la subcobertura del marco muestral, la no respuesta, los errores de medición relacionados con el cuestionario y el entrevistador y errores de procesamiento de los datos.

Error por no respuesta. Error que ocurre cuando la encuesta falla en obtener la respuesta de una, o posiblemente de todas las preguntas.

Error relativo. Es el cociente entre el error absoluto y el parámetro.

Esquema de muestreo. Es una combinación específica del tipo de muestreo, la modalidad de muestreo y el número de etapas de selección por aplicar, según las características de la población objeto de estudio y el tipo de datos a captarse.

Estadístico. Es una medida cuantitativa, derivada de un conjunto de datos de una muestra.

Estimación. Es el valor numérico de un estimador.

Estimador. Es un estadístico usado para estimar un parámetro desconocido de la población objeto de estudio.

Estimador consistente. Un estimador se dice ser consistente en el sentido de que, al incrementar el tamaño de muestra, la estimación se acerca cada vez más al parámetro poblacional.

Estimador insesgado. Es cuando el valor esperado del estimador es igual al parámetro estimado.

Estimador de razón. La estimación se forma por un cociente basado en la relación existente entre dos variables y que se miden en el mismo conjunto de elementos.

Estrato. Es una subpoblación en la cual los elementos que la componen reúnen características comunes que la hacen ser homogénea.

Estrato auto representado. Una unidad de muestreo está auto representada si es muestreada con probabilidad 1, en un muestreo con probabilidades proporcionales al tamaño, si sucede que $x_k > \sum_U x_k / n$ (n es el tamaño de muestra y x_k es la medida de tamaño) para la unidad k entonces la probabilidad de inclusión se define como $\pi_k = 1$ y la unidad k se trata como un estrato auto representado.

Factor de corrección por finitud. Se determina restando al tamaño de la población *N*, el tamaño de la muestra *n* y dividiendo este diferencial nuevamente entre el tamaño de la población, (N–n)/N; es un ajuste que se hace a la varianza cuando la población es finita.

Fase de Análisis de la Producción. Esta fase tiene por objeto asegurar que la información producida es apta para su propósito, es decir, está lista para su uso y difusión (Artículo 27 de la NTPPIEG).

Fase de Captación. Esta fase tiene por objeto captar los datos necesarios, incluyendo la obtención de Metadatos, para la generación de productos de información estadística y geográfica (Artículo 21 de la NTPPIEG).

Fase de Construcción. Esta fase tiene por objeto la construcción y prueba de la infraestructura informática, los componentes, aplicaciones y servicios de software, para crear un ambiente operacional completo que permita ejecutar la producción de información, así como la ejecución de pruebas que lo acrediten (Artículo 17 de la NTPPIEG).

Fase de Difusión. Esta fase tiene por objeto poner a disposición de los usuarios el Conjunto de Información a través del producto de información y sus diversas presentaciones y servicios (Artículo 31 de la NTPPIEG).

Fase de Diseño. Esta fase tiene por objeto diseñar los Productos de información estadística o geográfica que atenderán las Necesidades Estructuradas de Información determinadas de acuerdo con los elementos documentales recabados en la fase anterior. En esta fase se diseñarán las salidas, conceptos, metodologías, instrumentos de captación, Protocolos y Canales de intercambio; así como las estrategias generales para el desarrollo de las Fases de Construcción, Captación, Procesamiento, Análisis de la producción y Difusión, la modalidad metodológica de ejecución y otros aspectos que se consideren relevantes dentro del proceso de producción de información (Artículo 13 de la NTPPIEG).

Fase de Documentación de las Necesidades. El objetivo de esta fase es documentar las necesidades de información que sustentan al Programa de Información (Artículo 11 de la NTPPIEG).

Fase de Evaluación del Proceso. Esta fase tiene por objeto decidir si el siguiente ciclo de producción de información debe llevarse a cabo utilizando las mismas especificaciones de necesidades, diseño y construcción o si se requiere implementar alguna mejora en el mismo (Artículo 34 de la NTPPIEG).

Fase de Procesamiento. Esta fase tiene por objeto preparar los datos captados para el análisis, mediante procesos de transformación como la clasificación, codificación, geocodificación, georreferenciación, revisión, validación, edición e imputación de estos, conservando el registro de los procesos que transforman a cada dato de entrada. Además, se calculan nuevas variables, unidades, ponderadores y agregados y se preparan los archivos del Conjunto de Datos Procesados (Artículo 24 de la NTPPIEG).

Fenómeno de interés o Tema. Campo de conocimiento que se desea representar o medir (Artículo 3, fracción XII Ter de la NTPPIEG).

Indicador Objetivo. Indicador asociado a un dominio de estudio y fenómeno de interés que permite hacer mediciones directamente relacionadas con los objetivos del programa de información. El indicador deberá cumplir al menos uno de los siguientes criterios: que derive del cumplimiento de una Ley o Reglamento; que atienda un compromiso o recomendación internacional; que sea de utilidad para las políticas públicas; que forme parte de un conjunto de indicadores clave; o que contribuya a que la información sea comparable y coherente a través del tiempo y con los componentes que lo conforman (Artículo 3, fracción XIV de la NTPPIEG).

Instrumento de captación. Formato en medio impreso o electrónico, diseñado para el registro de los datos que han de obtenerse de las unidades de observación.

Intervalo de confianza. Rango o recorrido de valores numéricos dentro del cual se espera que se ubique el parámetro de estudio con un grado de confianza definido (habitualmente se emplea el 95% de confianza).

Ítem. Preguntas estandarizadas que conforman el cuestionario de una entrevista. El objetivo de un ítem es que sea presentado al informante para que produzca una respuesta, la cual puede ser abierta o limitada a un conjunto de opciones.

Marco conceptual. Estructura en la que se definen, ordenan y vinculan el fenómeno de interés, universo de análisis, dominios de estudio, variables, jerarquía de categorías e indicadores objetivo correspondientes a un Programa de Información.

Marco de muestreo o marco muestral. Listado en el cual se identifica a todos los elementos de una población objeto de estudio y que permite seleccionar una muestra de esta con fines de estimación estadística.

Metadatos. Datos estructurados que describen las características del contenido, captura, procesamiento, calidad, condición, acceso y distribución de la información estadística o geográfica para facilitar su uso y aprovechamiento (Artículo 3, fracción XXII de la NTPPIEG).

Muestra. Subconjunto de unidades seleccionadas de una población objeto de estudio, bajo condiciones preestablecidas que serán objeto de registro y captación de datos.

Muestra no probabilística. Una muestra en la cual la selección de las unidades está basada en métodos subjetivos diferentes a una selección aleatoria, por ejemplo, conveniencia, experiencia anterior o el juicio del investigador.

Muestra probabilística. Una muestra seleccionada mediante un método basado en la teoría de la probabilidad (proceso aleatorio), esto es, por medio de un método que incluye el conocimiento de las posibilidades de que alguna unidad sea seleccionada.

Muestreo aleatorio simple (MAS). Es una modalidad del muestreo probabilístico donde cada elemento de la población objeto de estudio tiene la misma probabilidad de ser seleccionado para integrar la muestra.

Muestreo aleatorio simple con reemplazo (irrestricto). Es un caso particular del muestreo aleatorio simple en donde la selección de la muestra se hace con reemplazo, es decir, puede haber muestras con elementos repetidos y cualquier elemento de la población objeto de estudio puede aparecer 0, 1, 2, …, n veces en la muestra. Genera observaciones independientes e idénticamente distribuidas, lo cual es un supuesto básico de la teoría estadística que es necesario para confirmar relaciones entre variables (véase muestreo aleatorio simple).

Muestreo aleatorio simple sin reemplazo. Es un caso particular del muestreo aleatorio simple donde la selección se hace sin reemplazo, es decir, la muestra obtenida no contendrá elementos repetidos y la selección de un elemento no tendrá influencia de la demás ya elegidas. Esta modalidad no produce observaciones independientes ni idénticamente distribuidas, pero su demanda de uso es alta, ya que sus varianzas estimadas son menores que el caso con reemplazo. Además, para solventar el inconveniente de independencia hace uso del factor de corrección por finitud.

Muestreo bietápico. La muestra se genera en dos etapas de selección.

Muestreo bola de nieve (*snowball)*. Es una modalidad del muestreo determinístico (no probabilístico) y consiste en localizar a algunos individuos que integran la población objeto de estudio, los cuales conducen a otros y éstos a su vez a otros, así hasta conseguir una muestra suficiente. Esta modalidad de muestreo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas o determinados tipos de enfermedades.

Muestreo con probabilidad proporcional al tamaño (PPT). Es una modalidad del muestreo probabilístico; puede llevarse a cabo cuando el marco de muestreo contiene información sobre variables "auxiliares" que tienen una buena relación proporcional con el Indicador Objetivo. Las unidades de observación se seleccionan con probabilidades desiguales y, a la vez, proporcionales a cierta variable auxiliar, logrando con esto que el cociente formado por el Indicador Objetivo y la auxiliar tenga mucha menor variabilidad que la de las lecturas individuales del Indicador Objetivo; este método maneja al igual que el muestreo aleatorio simple, la opción de seleccionar la muestra con o sin reemplazo, aunque el cálculo de las varianzas estimadas se obtienen suponiendo un muestreo con reemplazo, ya que el caso sin reemplazo es un problema teórico de difícil solución; es decir, se adopta una posición conservadora.

Muestreo estratificado. Es una modalidad del muestreo probabilístico y se basa en la conformación de subpoblaciones (estratos) mutuamente excluyentes, de tal manera que cada una de éstas, sea lo más homogénea posible en su interior y, a la vez, entre ellas sean lo más diferente posible. Para elaborar los estratos adecuadamente se requiere, entre otras cosas, de una selección adecuada de variables "auxiliares", de definir el número óptimo de grupos a formar y de la existencia y manejo de un sistema de cómputo adecuado; a pesar de lo anterior, en general, esta modalidad redunda en ciertos beneficios versus la opción de no conformar estratos a saber: mejores estimaciones, permite un manejo óptimo de recursos y facilita administrativamente el control en la Fase de

Captación.

Muestreo intencionado, por juicio, opinático o por conveniencia. Es una modalidad del muestreo determinístico (no probabilístico) y se basa en que los elementos que integran la muestra son resultado de la experiencia del investigador. Las principales ventajas de esta opción son la facilidad de obtener la muestra y que el costo usualmente es bajo.

Muestreo multietápico. La muestra se genera en más de dos etapas de selección.

Muestreo por conglomerados. Es una modalidad del muestreo probabilístico y consiste en la identificación de subpoblaciones considerando, para su delimitación, rasgos de proximidad territorial (se dice ser la contraparte de formar estratos); enseguida, se seleccionan sólo algunas de ellas y se capta información del total de unidades de estas. La gran ventaja de manejar esta alternativa de muestreo es que definiendo conglomerados de tamaño adecuado los entrevistadores no tienen que hacer grandes traslados para lograr más entrevistas durante la captación obteniendo con esto la muestra en un corto periodo y a bajo costo.

Muestreo por cuotas o accidental. Es una modalidad del muestreo determinístico (no probabilístico) y consiste en facilitar al entrevistador tanto el número como las características de las personas que tiene que seleccionar dejando, a su criterio la elección de estas, siempre y cuando cumplan con las características indicadas, cuando se combina con una selección por etapas generalmente se aplica en la última de ellas.

Muestreo sistemático con arranque aleatorio. Es una modalidad particular en el muestreo probabilístico, el cual divide a la población objeto de estudio en **k** subgrupos de **n** elementos y aleatoriamente toma como muestra a uno de ellos; por su forma de construcción el tamaño de muestra de los subgrupos no es constante (puede ser menor al requerido).

Muestreo unietápico. La muestra se obtiene en una sola emisión (no hay etapas de selección).

Muestreo. El proceso de seleccionar un número de casos de todos los casos en un grupo particular o población objeto de estudio.

Necesidad Estructurada de Información. Necesidad de información para la que se han definido el objetivo de la información, los conceptos a ser medidos, la población, territorio o fenómeno objeto de estudio, los dominios de estudio y la periodicidad con la que se requiere (Artículo 3, fracción XXVI de la NTPPIEG).

Nivel de confianza. Es la probabilidad de que el intervalo construido en torno a un estadístico incluya el verdadero valor del parámetro.

Número de etapas de selección. Significa que la determinación de la muestra se realiza en varias fases o etapas; dentro de cada una de ellas se aplica una selección individual; las etapas están vinculadas de tal forma que la muestra de una etapa cualesquiera es seleccionada solo en aquellas unidades que fueron extraídas en la etapa inmediata anterior.

Parámetro de interés. Cantidad numérica calculada sobre una población que resume los valores que esta toma en algún atributo (total, razón, porcentaje, índice, tasa, por mencionar algunas).

Población. El conjunto de unidades pertenecientes a un grupo de personas, empresas, establecimientos, viviendas, o cualquier otro tipo de objetos, acciones o eventos, con base en ciertas características bien definidas, incluyendo límites sobre tiempo y espacio (Artículo 3, fracción XXVIII de la NTPPIEG).

Población objeto de estudio o población objetivo. Población para la cual se requiere realizar mediciones o representaciones de los conceptos; este conjunto contiene a todos los dominios de estudio (Artículo 3, fracción XXVIII Bis de la NTPPIEG).

Ponderador. Es un concepto relacionado con la probabilidad de selección y se interpreta como la cantidad de

unidades en la población objeto de estudio que representa una unidad en la muestra, llámese personas, viviendas, áreas económicas o agrícolas, entre otras; dicho ponderador permite dar conclusiones sobre la población objeto de estudio.

Precisión. Se refiere a la correspondencia de los resultados de las mediciones obtenidas de la muestra (estimador) con respecto a los resultados que se medirían en toda la población objeto de estudio (parámetro).

Probabilidad de selección. Oportunidad quetiene cada elemento de la población objeto de estudio o universo de ser incluido en una muestra.

Proceso de Producción o Proceso. Conjunto de actividades, recursos, datos e infraestructura de información y fases que se relacionan lógicamente y se ejecutan para producir información que permita alcanzar los objetivos y metas definidos por el Programa de Información que le da origen (Artículo 3, fracción XXX de la NTPPIEG).

Programa de Información o Programa. Conjunto de actividades mediante el cual se establecen los objetivos, metas y estrategias para la ejecución de uno o más Procesos de producción para atender Necesidades Estructuradas de Información, de las cuales podrán resultar uno o más productos estadísticos y geográficos (Artículo 3, fracción XXXII de la NTPPIEG).

Representatividad de la muestra. Una muestra es representativa si los rasgos de los elementos que la integran son similares a los de toda la población objeto de estudio, es decir, si la muestra es capaz de reproducir las características de la población objeto de estudio.

Selección de la muestra. Se refiere a los procedimientos empleados para identificar las unidades de observación que integrarán la muestra. La selección puede realizarse con o sin reemplazo; en la primera situación se permite que una observación pueda estar en la muestra más de una vez, mientras que la segunda los elementos ya seleccionados lo hacen en forma única.

Sistema nacional de información estadística y geográfica o sistema (SNIEG). Conjunto de unidades organizadas a través de los subsistemas, coordinadas por el instituto y articuladas mediante la red nacional de información, con el propósito de producir y difundir la información de interés nacional.

Unidad de muestreo. Es un elemento que puede ser seleccionado del marco de muestreo con probabilidad conocida.

Unidad de observación. Elemento unitario del cual se obtienen datos con propósitos estadísticos sobre el conjunto al que pertenece.

Universo de análisis. Conjunto de elementos para los cuales se busca cuantificar y caracterizar el fenómeno de interés o tema.

Variable. Concepto que admite distintos valores para la caracterización o clasificación de elementos unitarios o conjuntos.

Varianza. Es un estimador de la dispersión de una variable aleatoria respecto de su media.

ÍNDICE DE FIGURAS

BIBLIOGRAFÍA

Benedetti, R., Bee, M., and Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4), 651

Benedetti, R., Piersimoni, F., Bee, M., and Espa, G. (Eds.). (2010). *Agricultural survey methods*. John Wiley & Sons.

Benthlehem, J., Cobben, F., and Schouten, B. (2011): Handbook of Nonresponse in Household Surveys, Nueva York: Wiley.

Chromy, J. R. *(2008). Probability proportional to size (PPS) sampling. In Encyclopedia of Survey Research Methods, P.J. Lavrakas (ed.). London: Sage.*

CEPAL (2002). Noveno Taller Regional sobre Diseño y Construcción de los Marcos de Muestreo para las Encuestas de Hogares.

Cochran, W.G. (1977). Sampling Techniques, third ed. John Wiley & Sons.

DANE. *La Calidad Estadística a Través de las Normas ISO.* Colombia. https://www.dane.gov.co/files/banco_datos/Revista/Estadisticas_al_dia_No4.pdf

DANE. *Metodología para Formulación de Planes Estadísticos.* Colombia. https://www.dane.gov.co/files/sen/planificacion/metodologia/metodologia-desarrollo-planes-estadisticos.pdf

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Elmar Wein. *The Planning of Data Editing.* Federal Statistical Office, Germany.

EUROSTAT-OECD. *Glossary Of Statistical Terms (OECD)* https://stats.oecd.org/glossary/

EUROSTAT. *Harald Sonnberger and Nick Maine. Editing and Imputation in Eurostat.* https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2000/10/sde/21.e.pdf

EUROSTAT-OECD. *Procedures and Checklists for Oecd Statistical Activities.* https://www.oecd.org/dataoecd/26/40/21687687.pdf

Felligi, I.P. Holt D. (1976). *A systematic approach to automatic edit imputation, journal of the American statistical. association.* March 1976.

Fellegi, I. (2010). Survey Methods and Practices, Statistics Canada.

Granquist, Leopold. (1997). *The new view on editing.*

Grossh Margaret E., Muñoz Juan. (1998). *Manual de diseño sobre condiciones de vida (LSMS)* documento de trabajo no. 1265.

INE. *Nuevas Tecnologías para Difundir Datos Estadísticos. España.*

INEC. (2019). Marco Maestro de Muestreo para encuestas de hogares. Instituto Nacional de Estadística y Censos, Quito-Ecuador.

INEGI. (1999). *Estrategias generales del XII Censo General de Población y Vivienda 2000.*

INEGI. (2002). *Memoria XII Censo General de Población y Vivienda 2000.*

INEGI. (2005). *Sistema de codificación automático y manual de la Encuesta Nacional de Ocupación y Empleo (ENOE).*

INEGI. (2006). *Desarrollo y documentación de software, Guía.*

INEGI. *(2010). Norma técnica para la generación de estadística básica. Noviembre de 2010. Diario Oficial de la Federación en: DOF.* http://dof.gob.mx/nota_detalle.php?codigo=5167222&fecha=12/11/2010

INEGI. *(2010). Proceso estándar para encuestas por muestreo.*

INEGI. *(2011). Diseño de la muestra en proyectos de encuesta.*

INEGI. *(2013). Captación en encuestas por muestreo.*

INEGI. (2018). *Encuesta de Viajeros Fronterizos EV, Encuestas de Viajeros Internacionales EVI. Síntesis metodológica*

INEGI. (2019). *Cómo se hace la ENOE: métodos y procedimientos.*

INEGI. (2020). *Diseño muestral de la Encuesta Nacional sobre Uso del Tiempo (ENUT) 2019.*

INEGI. (2020). *Encuesta Nacional Agropecuaria (ENA) 2019: Metodología.*

INEGI. *(2020). Norma Técnica del Proceso de Producción de Información Estadística y Geográfica para el Instituto Nacional de Estadística y Geografía. Diario Oficial de la Federación en:* https://sc.inegi.org.mx/repositorioNormateca/On_23Nov20.pdf

Kish, L *(1965). Survey Sampling.* Jonh Wiley & Sons.

Kish, L *(1975). Muestreo de Encuestas.* Ed. Trillas.

Kish, l ***Knaub, J.R., Jr.*** *(2008). Cutoff Sampling. In Encyclopedia of Survey Research Methods, P.J. Lavrakas (ed.). London: Sage.*

Lavallée, P., and Hidiroglou, M. *(1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.*

Lavrakas, P. J*. (2008). Encyclopedia of survey research methods. Thousand Oaks, Calif: SAGE Publications.*

Leiv Solheim, *How to Measure the Effect of Data Editing.* Statistics Norway.

Lohr, S., and Rao, J. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

ONU. *Diccionario de datos de definiciones oficiales.*

ONU. *Principios y recomendaciones para los censos de población y habitación.* Serie M, No. 67/rev.1.

ONU. *Recomendaciones internacionales para las estadísticas industriales.* Serie M No. 48.

Pennell, B.E., Cibelli Hibben, K.L., Lyberg, L., Mohler, P.P., and Worku, G. (2017). A total survey error perspective on surveys in multinational, multiregional, and multicultural contexts. Total Survey Error in Practice, pp.179-202.

ReStore National Centre for Research Methods. (2021, junio 18). *Finite Populations Correction.*

https://www.restore.ac.uk/PEAS/finitepop.php

Särndal, C.-E., Swensson, B., and Wretman, J. (1992*). Springer series in statistics. Model assisted survey sampling*. Springer-Verlag Publishing.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

STATCAN. *Characteristics of an Effective Statistical System* (Ivan, P. Fellegi). https://unstats.un.org/unsd/goodprac/bpform.asp?DocId=190&KeyId=25

STATCAN. (2003) *Gestión de la Calidad de los Datos en un Organismo Estadístico* (Gordon Brackstone). https://repositorio.cepal.org/handle/11362/16464.

Tillé, Y. (2006). Sampling algorithms. Springer New York.

Tillé, Y., and Haziza, D. (2010). An interesting property of the entropy of some sampling designs. *Survey Methodology*, 36, 229-231.

Tillé, Y., and Wilhelm, M. *(2017).* Probability sampling designs: principles for choice of design and balancing. *Statistical Science, 176-189.*

United Nations (2008). Designing Household Survey Samples: Practical Guidelines. https://unstats.un.org/unsd/demographic/sources/surveys/Series_F98en.pdf

Valliant, R., Dever, J.A., and Kreuter, F. (2018). Practical Tools for Designing and Weighting Survey Samples. Heidelberg, Germany: Springer

Villan, Idelfonso, Bravo, Maria Soledad. *Procedimiento de depuración de datos estadísticos.* Eustat, 1990. Seminario internacional de estadística en EUSKADI.

Wolter, K. (2007). Introduction to variance estimation. Springer Science & Business Media.