

DISEÑO MUESTRAL

2do documento técnico del proyecto

El presente documento representa el segundo entregable del proyecto denominado “Diseño de la evaluación de alternativas de esquemas de capacitación de la EsIAN”. Este documento aborda los aspectos técnico-estadísticos asociados a la implementación del diseño muestral que se propone utilizar para cubrir los objetivos del proyecto, mismos que ya fueron planteados en el documento de la *propuesta metodológica y de diseño de la evaluación* contenidos en el primer documento entregable del proyecto.

0. Considerandos

Son indispensables los siguientes considerandos en la lectura del presente documento:

- a) A la fecha, la base de datos provista por el Programa de Desarrollo Humano Oportunidades a utilizar como marco muestral tiene información incompleta (representa aproximadamente un 54% de lo planeado y de ese 54% se tiene un 60% de los registros con datos vacíos). Por su parte, de aquellos registros con información no se tiene pureza ni tampoco uniformidad de captura en los datos. Esto se documenta a detalle en la parte de Anexos.
- b) Un marco muestral de baja cobertura y de baja calidad imposibilita la extracción y/o generación de muestras a utilizar con el presente diseño de muestreo. En estas condiciones, representa un verdadero reto el proponer un diseño de muestreo.
- c) Por los considerandos anteriores, el presente documento aborda la propuesta de del diseño muestral de manera genérica y evita particularizar en detalles técnicos específicos. Es decir, se plantea un diseño de muestreo hipotético que considera la provisión de un marco muestral viable.
- d) Por último, es importante mencionar que todos los cálculos de tamaños de muestra realizados en este documento toman en cuenta a *la población planeada* a capacitarse de proveedores de la segunda cascada (alrededor de 20,000). Por su parte, el marco muestral provisto a la fecha contiene aproximadamente 10,500 registros (incluyendo aquellos con datos vacíos), que representan el 54% referido anteriormente.

1. Introducción

A efecto de homologar lenguaje y el orden de las ideas es necesario establecer algunas definiciones básicas y objetivos técnicos propios del ejercicio de implementación de técnicas de muestreo. Es importante, por ejemplo, definir el enfoque teórico-filosófico a seguir como parte de la solución del problema de muestreo planteado.

Aunque es poco común en literatura no especializada en muestreo, es pertinente establecer el enfoque teórico-filosófico a utilizar en el presente documento técnico del diseño muestral. Lo anterior, con la finalidad de evitar cualquier confusión que pudiera derivar en una discusión innecesaria de posiciones aparentemente contrapuestas.

1.1. Recordando el objetivo del muestreo

Como un primer paso, es necesario definir el objetivo o problema que responde el muestreo. Partiendo de una de las referencias más citadas en muestreo moderno (1), dicho objetivo o problema de muestreo puede resumirse como: inferir sobre ciertas propiedades de una población a partir de información parcial de ésta, i.e. una muestra. Existen varios tipos de muestras: probabilísticas y no probabilísticas; confinamos el presente documento a aquellas que son *probabilísticas* ya que interesa la producción de medidas de error en las estimaciones generadas utilizando teoría de la Estadística.

1.2. Enfoques de muestreo

De la misma manera en que es posible resolver un mismo problema matemático de diferentes formas o diferentes puntos de vista y obtener resultados similares, en muestreo también es posible utilizar diferentes enfoques o puntos de vista. Existen tres grandes enfoques que la literatura contemporánea de muestreo identifica (2, 3). Estos 3 enfoques de muestreo son:

- a) El enfoque *basado en modelos*, de predicción o con super-poblaciones (*model-based approach*),
- b) El enfoque *basado en diseño* o aleatorizado (*design-based approach*),
- c) El enfoque *asistido por modelos* (*model-assisted approach*).

La diferencia principal entre los enfoques (a) y (b) reside en dónde se asume el componente estocástico (1, 3, 4). Esto quiere decir en términos coloquiales, aunque filosóficamente más profundos, en dónde reside lo aleatorio de un fenómeno estudiado:

en el fenómeno mismo que es aleatorio intrínsecamente, o bien, en la ignorancia que tenemos como observadores respecto al comportamiento del fenómeno. Por último, el enfoque (c) es aquel que combina aspectos técnicos de ambos enfoques (a) y (b).

Se sabe que cada uno de los enfoques tiene sus ventajas y desventajas, y éstas deben ser consideradas al momento de decidir cuál enfoque adoptar para un problema de muestreo en particular (5). Además de esto último y para ser más específicos al proyecto, cabe mencionar que bajo los 3 enfoques hay definiciones que comparten nomenclatura y que al mismo tiempo significan cosas totalmente diferentes. Un ejemplo es la definición de *muestra aleatoria*, que en el enfoque basado en modelos es una colección de observaciones conocidas que provienen de variables aleatorias independientes idénticamente distribuidas; mientras que en el enfoque basado en diseño, *grosso modo*, es una colección de mediciones que son fijas en la población pero desconocidas para el investigador donde lo aleatorio reside en la composición de la muestra sujeta a medición de la cual no se conocen los individuos que la componen.

En concordancia con los objetivos del proyecto, planteados en el documento de la *propuesta metodológica y de diseño de la evaluación*, se propone utilizar un enfoque basado en diseño. Esto es, el enfoque (b), por tratarse de aquel enfoque que goza de mayor objetividad de entre los tres listados, ya que no depende de modelos impuestos y tampoco asume independencia entre elementos de la población. Es decir, en el enfoque (b) a utilizar, una vez acordado el procedimiento aleatorizado de selección de la muestra, se cancela cualquier discusión que evoque subjetividad.

La principal desventaja del enfoque (b), usualmente aducida en textos sobre muestreo basado en modelos (5) es que, como las estimaciones no utilizan un modelo, requieren de un mayor tamaño de muestra comparado con el que se necesitaría si se asumiera un modelo (6). El hecho de no utilizar un modelo en el presente proyecto es de hecho un elemento necesario y crucial para la validez de la evaluación de resultados pues se requiere que la evaluación sea enteramente objetiva.

1.3. Definiciones necesarias y problemática del proyecto

Bajo un enfoque basado en diseño cobra especial importancia la correcta definición de lo que será la población objetivo (la población a ser muestreada), las unidades muestrales (lo que se va a muestrear de la población objetivo) y el diseño de muestreo (cómo se va a seleccionar la muestra y la función que determina la probabilidad de obtener cierta muestra y que inducirá las probabilidades de inclusión en muestra de cada

unidad muestral en la población). Estas definiciones son relevantes a efecto de evitar inferencias equivocadas sobre la población objetivo partiendo de la muestra. En otras palabras, para evitar generalizaciones equivocadas (7).

Para lograr una correcta definición de los elementos que están en juego en un problema de muestreo, es necesario responder las siguientes preguntas: ¿Qué se va a muestrear? ¿De dónde se va a muestrear? ¿Hacia qué población se va a inferir (*expandir*)? ¿Cómo se va a muestrear? ¿Cuántas etapas tendrá el diseño de muestreo? ¿Cómo son las probabilidades de inclusión de las unidades muestrales?, y finalmente ¿qué tamaños de muestra corresponden con la tolerancia a errores de estimación?

Para responder dichas preguntas, es necesario entender bien los objetivos del proyecto. Para resumir de manera esquemática la problemática que enfrenta el proyecto de diseño de la evaluación de alternativas de esquemas de capacitación de la EsIAN, se presenta la Figura 1.

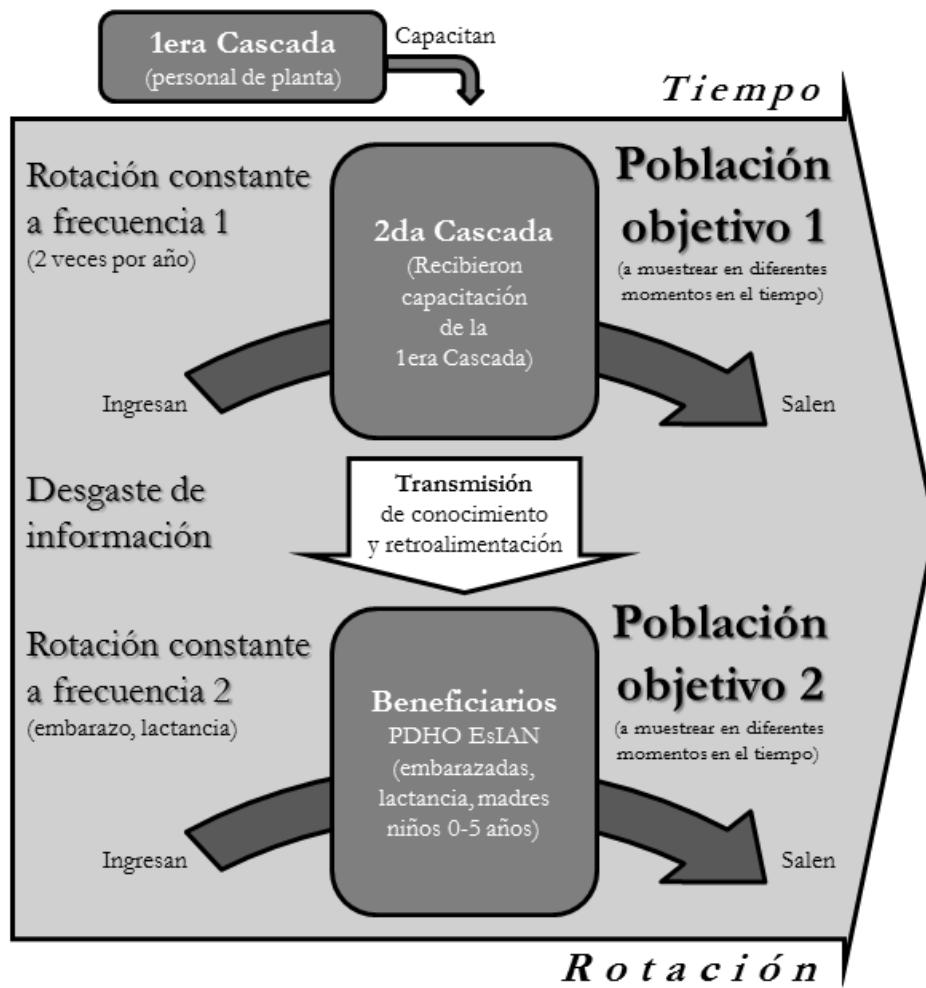
1.4. Dos poblaciones objetivo

La Figura 1 exhibe de inicio la existencia de 2 poblaciones objetivo. Por un lado, una población de personal capacitado de la segunda cascada (proveedores) y por otro lado, una población de beneficiarios de interés particular. Estas son 2 poblaciones objetivo de donde serán extraídas muestras en diferentes momentos del tiempo, de manera compaginada, para evaluar resultados de manera recurrente.

Lo anterior ya fue expresado en la *propuesta metodológica y de diseño de la evaluación* del primer documento entregable del proyecto. También la Figura 1 pone de manifiesto la constante rotación que existe en cada una de las 2 poblaciones objetivo y el desgaste que hay en la transmisión de información. Es de notar que la rotación de las 2 poblaciones objetivo ocurre a diferentes velocidades. La combinación de ambas velocidades es lo que complica aún más la problemática que enfrenta el proyecto.

Es importante señalar también, que el efecto cruzado de la rotación de ambas poblaciones sugiere muestras transversales en favor de un estudio de tipo panel. Una amplia discusión sobre las fortalezas y debilidades de un estudio tipo panel para el proyecto en cuestión se encuentra en el documento de la *propuesta metodológica y de diseño de la evaluación*.

Figura 1 Problemática de rotación, desgaste, e identificación de poblaciones objetivo.



Desventajas de un estudio tipo panel para el presente proyecto

En particular para el presente proyecto, es pertinente recordar las conocidas desventajas de un estudio tipo panel y el riesgo que tienen esas desventajas de agravarse debido a la alta rotación presente (8, 9). A continuación, se listan algunas de las desventajas intentando hacerlas particulares al proyecto:

- a) Atrición de unidades en el panel (mortalidad de panel): La alta rotación que hay en la población de proveedores y la alta rotación que hay en la población de beneficiarios pueden provocar una acelerada *muerte* de unidades contenidas en el panel. Esto complicaría la generalización de resultados a las poblaciones objetivo en cada momento que se realicen mediciones. Se tendrían que ajustar y recalibrar los pesos de unidades en el panel prácticamente de manera continua.

- b) Sobre-estimación de impacto (acumulación de conocimiento): Aquellos beneficiarios que resulten ser parte del panel tendrán una creciente acumulación de información que reciben, por ejemplo sobre lactancia, nutrición y demás. Esta memoria de conocimientos en los beneficiarios de la muestra panel puede ocasionar grandes sesgos en la estimación de conocimientos en un sentido positivo cuando en realidad tal conocimiento no fue permeado con el mismo éxito en la población en general.
- c) Sobre-estimación de impacto (incentivos a mostrar mayor conocimiento): Aquellos beneficiarios o incluso proveedores que resulten ser parte del panel tendrán incentivos a mostrar mayor conocimiento, o incluso se prepararse como si fueran a presentar un examen escolar, para las preguntas de conocimiento cuando se les haga una visita (*panel conditioning* (8)). Este sentimiento a sentirse evaluado por parte de los beneficiarios de la muestra panel puede ocasionar grandes sesgos en la estimación de conocimientos en sentido positivo cuando, como en el caso anterior, tal conocimiento mostrado no se puede generalizar a aquellos fuera del panel (falta de validez externa).

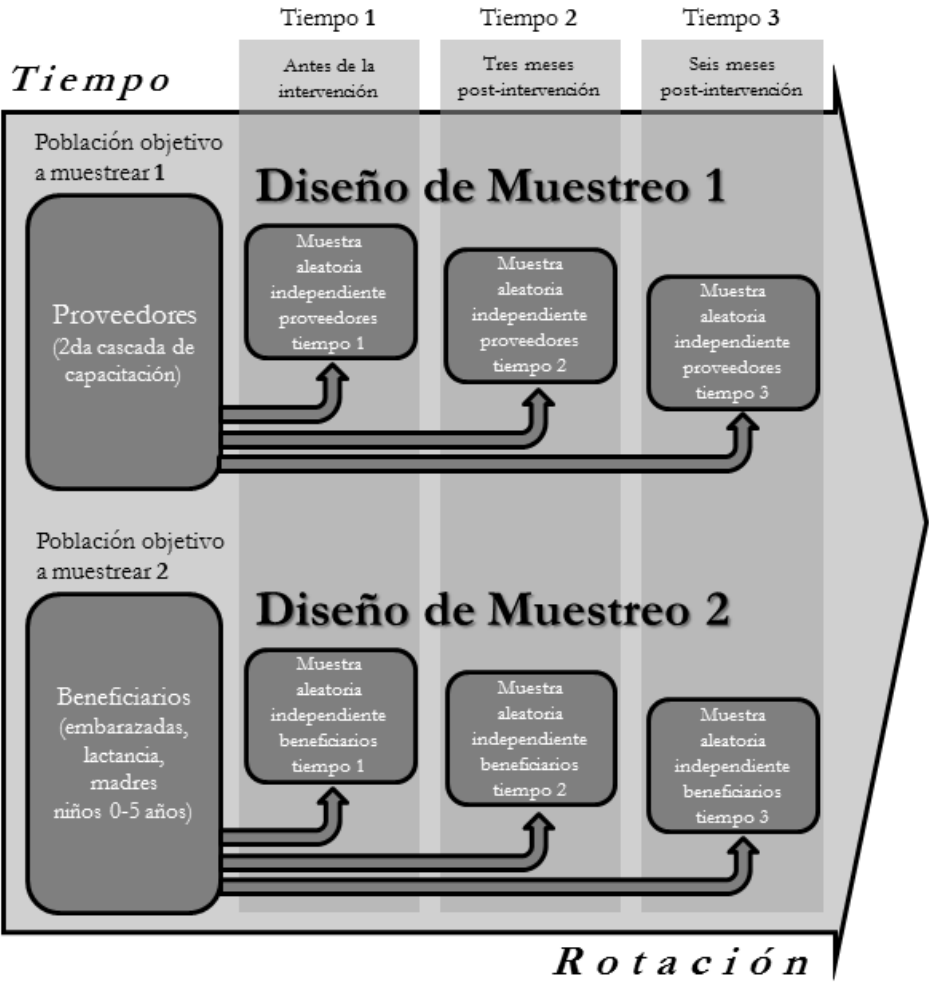
Es probable que tales desventajas, entre otras, tengan un impacto mayor debido a lo cambiante de las 2 poblaciones objetivo (tiempos muy cortos) y a que precisamente el proyecto tiene como materia la capacitación y transmisión de conocimientos o información. Las fortalezas y debilidades de un estudio tipo panel para el proyecto se discuten con mayor profundidad en el documento de la *propuesta metodológica y de diseño de la evaluación*.

2. Dos diseños muestrales

2.1. Dos poblaciones objetivo, dos diseños de muestreo

Al tratarse de dos poblaciones objetivo, una de proveedores y una de beneficiarios, se tienen que definir dos diseños de muestreo diferentes, uno para cada población objetivo; como se ilustra en la Figura 2 a continuación.

Figura 2. Descripción esquemática de poblaciones objetivo y muestras a extraer.



3. *Diseño de muestreo - Población de proveedores*

3.1. *Marco muestral*

Para muestrear en la población de proveedores, es necesario tener un marco de muestreo que liste a aquellos proveedores de servicios de salud capacitados en la segunda cascada. Dicho marco muestral se puede obtener de los registros en posesión de los capacitados de la primera cascada. Es decir, el marco muestral de la población de proveedores estará conformado por aquel listado nominal concentrado de capacitados en la segunda cascada. Tomando en cuenta que este listado no se reunió en el marco muestral provisto por el Programa de Desarrollo Humano Oportunidades, al final del presente documento se propone la alternativa para estudios futuros de utilizar al mismo personal de la institución o empresa encuestadora para capacitar y tomar mediciones antes y después de la capacitación.

La limpieza, veracidad y cobertura de tal listado nominal es crucial para el buen desarrollo de un ejercicio demoscópico (10). Aquí es importante resaltar que en este listado se debe disponer (de manera obligatoria) de información de contacto de los proveedores. De otro modo no pueden ser parte de la población a muestrear, esto es, se tendrían que descartar, pero este descarte pudiera estar altamente correlacionado con lo que se intenta medir (e.g. alcances de capacitación, resultados).

También, como este listado será utilizado en el otro diseño de muestreo asociado a la población de beneficiarios, es necesario (obligatoriamente) que se cuente con la clave de las unidades de salud a las que pertenecen los proveedores. Finalmente, cualquier tipo de estratificación en la población objetivo tendría que hacerse a partir de la información auxiliar contenida en dicho listado nominal. A continuación se lista la información que en su momento se propuso para conformar el listado nominal o marco muestral de proveedores:

NOMBRE(S) (capacitado 1era cascada)
APELLIDO PATERNO (capacitado 1era cascada)
APELLIDO MATERNO (capacitado 1era cascada)
CURP (capacitado 1era cascada)
PERFIL (capacitado 1era cascada)
CARGO (capacitado 1era cascada)

INSTITUCION (capacitado 1era cascada)
ENTIDAD (capacitado 1era cascada)
MUNICIPIO (capacitado 1era cascada)
JURISDICCION (capacitado 1era cascada)
NOMBRE (capacitado 2da cascada)
APELLIDO PATERNO (capacitado 2da cascada)
APELLIDO MATERNO (capacitado 2da cascada)
FECHA DE NACIMIENTO (capacitado 2da cascada)
CURP (capacitado 2da cascada)
SEXO (capacitado 2da cascada)
LADA TELEFONO (CASA) (capacitado 2da cascada)
TELEFONO (CASA) (capacitado 2da cascada)
LADA CELULAR (capacitado 2da cascada)
CELULAR (capacitado 2da cascada)
CORREO ELECTRÓNICO (capacitado 2da cascada)
NIVEL DE COMPUTACION (capacitado 2da cascada)
MAXIMO GRADO DE ESCOLARIDAD (capacitado 2da cascada)
OCUPACION (capacitado 2da cascada)
CARGO (capacitado 2da cascada)
ENTIDAD FEDERATIVA (capacitado 2da cascada)
MUNICIPIO (capacitado 2da cascada)
LOCALIDAD (capacitado 2da cascada)
TIPO DE LOCALIDAD DONDE LABORA (capacitado 2da cascada)
INSTITUCIÓN (capacitado 2da cascada)
CLUES (capacitado 2da cascada)
JURISDICCIÓN (capacitado 2da cascada)
ESPECIFICAR JURISDICCIÓN (capacitado 2da cascada)
DIRECCIÓN DE LA UNIDAD DE SALUD (capacitado 2da cascada)
ES PERSONAL PERMANENTE O TEMPORAL (capacitado 2da cascada)
EN CASO DE TEMPORAL, MES QUE DEJA DE LABORAR (capacitado 2da cascada)

De modo que, por ejemplo, conociendo la distribución de sexo es posible estratificar a la población para controlar posibles desviaciones en la extracción de la muestra. Como una aproximación, y con el objeto de caracterizar a la población de proveedores, en el siguiente Cuadro 1, se describe a la población **estimada** de proveedores (capacitados de la segunda cascada). Cabe mencionar que tal información fue con base en datos planeados y/o estimados por parte de los encargados de la organización de las tareas de capacitación.

En los Anexos se presenta el Cuadro A, que describe a la población objetivo de capacitados en la segunda cascada que efectivamente están en el marco muestral disponible a la fecha. Se puede observar que no se cuenta con un marco muestral

aceptable en términos de completos y pureza de los datos provistos para ser usados como marco muestral. Por lo tanto, no es posible extraer una muestra.

Cuadro 1. Población (*estimada*) de proveedores (capacitados de la segunda cascada).

Entidad Federativa	Personal a capacitar (proveedores, segunda cascada)						Total por Entidad Federativa
	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades	
	Rama Médica*		Rama Enfermería*		Rama Promotores / Comunitaria		
Chihuahua	599	156	780	299	488	30	2,352
Oaxaca	2,533	519	3,725	969	834	83	8,663
Veracruz	2,833	564	2,860	1,107	1,108	81	8,553
Total	5,965	1,239	7,365	2,375	2,430	194	19,568

* Nota: Incluye tanto personal con actividades fijas, como personal pasante, considerando que las actividades de capacitación son aplicables para todos los prestadores de servicios participantes en el Programa. De las cifras que se reportan, se han restado los asistentes al reciente evento en León, Guanajuato por considerarse que estos pertenecerían a la primera cascada.

3.2. Estratificación

En la población de proveedores, por tratarse del escenario ideal para un ejercicio de muestreo, i.e. un muestreo directo de elementos, es posible mejorar el diseño estratificando a la población de acuerdo a cierto conjunto de variables de estratificación. Tales variables necesariamente tienen que estar disponibles en el marco muestral, en este caso, el marco muestral de proveedores.

Se sabe que la estratificación tiene varios beneficios (11). De entre ellos, hay dos principales ventajas que aplican de manera directa al proyecto de evaluación de resultados de capacitación EsIAN-PDHO. Por un lado, servirá para controlar posibles desviaciones en la extracción de la muestra y por otro lado mejorará la precisión de las estimaciones producidas. Como se sabe, esta mejora en precisión dependerá de la relación (correlación) que tengan las variables de estratificación con el fenómeno que es

medido (12). Desde luego, ambas ventajas dependerán del tamaño de muestra disponible, pues la cantidad de estratos está limitada por la cantidad de muestra a distribuir en los estratos.

A la fecha de elaboración del presente documento, no es posible determinar de manera definitiva las variables de estratificación a utilizar. Lo anterior se debe a la indisponibilidad de un marco muestral útil. Aunque es aventurado determinar tal estratificación, es posible sugerir que contemple, por lo menos, 36 estratos en la población de proveedores. Dicha cantidad de estratos se obtiene de considerar las 18 subpoblaciones representadas en el Cuadro 1 y adicionando la variable sexo dentro de cada una de las subpoblaciones.

3.3. Esquema de selección de las muestras de proveedores

Por tratarse de un muestreo directo de elementos en la población de proveedores, ya que se asume que es posible disponer del listado concentrado de capacitados de la segunda cascada con la debida información de contacto, se propone seleccionar a los proveedores con probabilidades iguales dentro de un mismo estrato y sin reemplazo. La justificación para utilizar muestreo sin reemplazo se puede encontrar en Gabler (13), donde se discute la superioridad de utilizar diseños de muestreo sin reemplazo sobre aquellos con reemplazo.

También, se propone no utilizar un diseño de muestreo sistemático, ya que sufre de tener probabilidades de inclusión de segundo orden igual a cero (14). Además, el muestreo sistemático se trata de un diseño de muestreo de muy baja entropía, es decir de baja aleatorización (15, 16). Cabe mencionar que una mayor aleatorización permite una mejor generalización de las estimaciones obtenidas. Desafortunadamente en México es muy popular el uso de diseños de muestreo sistemáticos, muy probablemente porque no se requiere de equipo de cómputo.

A continuación en la Figura 3 se esquematiza la selección de las muestras para la población de proveedores.

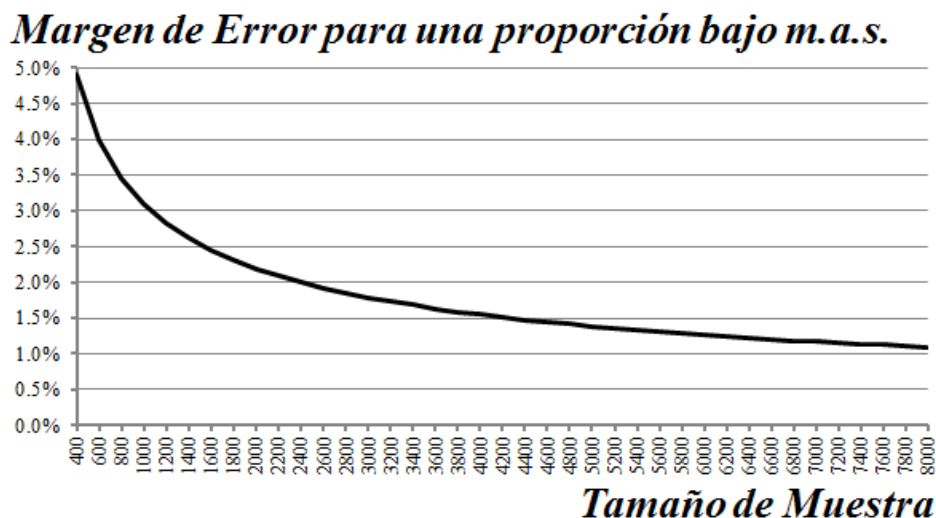
Figura 3. Esquema del diseño de muestreo 1 para la población de proveedores.



3.4. Cálculo de tamaño de muestra

A continuación, la Figura 4 exhibe los tamaños de muestra correspondientes a un muestreo aleatorio simple, suponiendo la máxima incertidumbre en las proporciones a calcular. Es pertinente mencionar que los tamaños de muestra calculados y en general el diseño de muestreo para la población de proveedores es el mismo en las mediciones pre y post capacitación. Es importante señalar que los tamaños de muestra ilustrados en el gráfico 4 son calculados para proporciones bajo un diseño de muestreo aleatorio simple. Por lo tanto, son tamaños de muestra teóricos de carácter ilustrativo y conservadores (es decir, ligeramente más grandes), ya que se propuso utilizar un diseño de muestreo aleatorio simple con estratificación. El diseño propuesto produce errores de muestreo menores al aleatorio simple. Además de tratarse de un diseño en una etapa donde sólo hay estratificación y no hay conglomeración.

Figura 4. Relación entre el margen de error y tamaño de muestra.



Se sabe que la estratificación mejora la precisión de las estimaciones y tal mejora dependerá de la relación que existe entre las variables de estratificación y las variables medidas en la encuesta para proveedores (11, 12). De cualquier forma, como ya se comentó, se espera una muy moderada ganancia en precisión con respecto a un muestreo aleatorio simple debido a que se usará una distribución de muestra en estratos con método proporcional dado que se prevé que el marco muestral tiene poca información relacionada con el fenómeno que se estudia en los proveedores. Esto último se detalla en apartados siguientes.

Como resultado del cálculo específico de tamaños de muestra a utilizar en la población de proveedores se obtuvieron los siguientes tamaños de muestra representados en el Cuadro 2 siguiente:

Cuadro 2. Tamaño de muestra de la población de proveedores.

Entidad Federativa	Tamaño de muestra						Total de muestral por Entidad Federativa
	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades	
	Rama Médica*		Rama Enfermería*		Rama Promotores / Comunitaria		
Chihuahua	234	111	257	168	215	28	1,013
Oaxaca	334	221	348	275	263	68	1,509
Veracruz	338	229	339	285	285	67	1,543
Total	906	561	944	728	763	163	4,065

Para el cálculo específico de tamaños de muestra se consideraron las siguientes restricciones y/o especificaciones:

- Se utilizaron 18 dominios de estimación. Esto es, se anticipan 18 subpoblaciones de interés para las que se generarán estimaciones. Estos dominios están representados en el Cuadro 3, a continuación:

Cuadro 3. Dominios de estimación de la población de proveedores.

Entidad Federativa	Personal de la segunda cascada de capacitación					
	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades	Secretaría de Salud	IMSS-Oportunidades
	Rama Médica*		Rama Enfermería*		Rama Promotores / Comunitaria	
Chihuahua	Dominio 1	Dominio 2	Dominio 3	Dominio 4	Dominio 5	Dominio 6

Oaxaca	Dominio 7	Dominio 8	Dominio 9	Dominio 10	Dominio 11	Dominio 12
Veracruz	Dominio 13	Dominio 14	Dominio 15	Dominio 16	Dominio 17	Dominio 18

- Para el cálculo de tamaño de muestra en cada dominio de estimación se contempla un margen de error de $\pm 5.000\%$ a un nivel de 95% de confianza en las estimaciones que correspondan a proporciones de la presencia de algún atributo.
- Lo anterior conlleva un margen de error de (a lo más) $\pm 2.323\%$ a un nivel del 95% de confianza en las estimaciones que correspondan a proporciones para el estado de Chihuahua sin desagregar por dominio. Análogamente, se tienen márgenes de error de (a lo más) ± 2.293 para Oaxaca y ± 2.259 para Veracruz, sin desagregaciones.
- Por su parte, de manera global, considerando a las 3 entidades federativas en conjunto, se tendrá un margen de error de (a lo más) ± 1.368 , a un nivel de confianza del 95% para las estimaciones de proporciones de la presencia de algún atributo.
- Es importante mencionar que los errores teóricos arriba listados dependen de la magnitud de las proporciones estimadas. Estos márgenes de error calculados no son los adecuados si estuviese en juego la estimación de proporciones muy pequeñas (menores al 5%) o proporciones muy grandes (mayores al 95%) en la cuantificación de presencia de cierto atributo.
- De acuerdo a lo descrito en incisos anteriores, se anticipa que será posible captar cambios estadísticamente significativos (a un nivel de confianza del 95%) si en mediciones consecutivas (con los mismos márgenes de error, tamaños de muestras, etc.) se observan cambios de al menos 10% en las estimaciones puntuales por dominio de estimación.
- De la misma manera, se anticipa será posible captar cambios estadísticamente significativos (a un nivel de confianza del 95%) si en mediciones consecutivas se observan cambios de al menos 4.6% aproximadamente en las estimaciones puntuales por entidad federativa, i.e. para Chihuahua, Oaxaca o Veracruz sin desagregar.
- Finalmente, de manera correspondiente, si se logran captar cambios de al menos 2.7% en las estimaciones puntuales globales (i.e. las que consideran a las 3

entidades federativas en conjunto), estos cambios podrán considerarse estadísticamente significativos (al 95% de confianza) si se preservan los tamaños de muestra propuestos y los márgenes de error arriba mencionados para las encuestas subsecuentes.

- Para el cálculo de tamaño de muestra en cada dominio se asumió un efecto de diseño de 1.00. Este supuesto es conservador y se sostiene muy fácilmente ya que se estratificará al interior de cada dominio de estimación por la variable sexo del proveedor.
- También, para el cálculo de tamaño de muestra en cada dominio se asumió máxima variabilidad de las proporciones que se medirán. Es decir, se asumirá total desconocimiento de lo que se intenta cuantificar. Este supuesto es conservador igualmente y se sostiene de manera fácil debido a que se trata de una encuesta que no sólo mide un aspecto en las unidades de observación, sino más bien un conjunto de variables de diversa índole en torno a los objetivos del proyecto.
- Para los cálculos se asumió una tasa de respuesta del 100%. Es decir, se asume que los proveedores contestarán los cuestionarios aplicados en las entrevistas y que no habrá rechazos a responder o datos faltantes. Aunque esto no se sostuviera el esquema de selección aleatoria propuesto se puede hacer mediante un ordenamiento aleatorio que permita pasar al siguiente proveedor en caso de que no conteste hasta obtener el tamaño de muestra objetivo. Esto último es una estrategia particularmente útil para no encarecer las labores de campo y cuando no se tiene experiencia previa de las tasas de respuesta.
- Los márgenes de error arriba calculados toman en cuenta el marco muestral estimado o planeado. Una vez que la población objetivo de proveedores exista, estos márgenes de error teóricos pudieran sufrir cambios. No obstante, los márgenes presentados ilustran las magnitudes de los errores muestrales que están en juego en el presente proyecto.
- Por último, se asume en todo momento que el levantamiento de información se realiza con veracidad y calidad. En otras palabras, se asume profesionalismo en las labores de recolección de la información, integridad ética por parte de los encuestadores y se asume la no existencia de conflictos de interés de la empresa, institución u organización encargada de la recolección de la información en

campo. Si este supuesto no se cumple existe el riesgo de derivar en mediciones totalmente erróneas a las que no apliquen los márgenes de error arriba mencionados.

3.5. Distribución de muestra en los estratos

La distribución de muestra en los estratos es algo a definir una vez que se tenga disponible el marco muestral de proveedores (cuando la población exista). Dependiendo de lo observado en el marco muestral, i.e. la población de proveedores, se podrían utilizar métodos de asignación proporcional al tamaño de los estratos construidos o bien, usar asignación de Neyman o alguna variante de éste en donde se asigna más muestra en aquellos estratos en donde se concentra más variabilidad de lo que se quiere medir. Se prevé que el marco muestral disponible tendrá información muy limitada y que por tanto habría que utilizarse asignación proporcional.

Sea n_h el tamaño de muestra para el estrato $h = 1, \dots, H$ con H el número de estratos que hay en la población. Y sea n el tamaño de muestra global, N_h el tamaño del estrato h -ésimo y N el tamaño de la población, i.e. el número total de proveedores, entonces se tiene que la asignación proporcional se hace siguiendo la siguiente expresión:

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

Hay que decir que, en general, la mejora en precisión con este tipo de asignación de muestra a estratos es baja; esta asignación sirve más para controlar la distribución espacial de la muestra que para aumentar la precisión en las estimaciones. Por último, hay que mencionar que no será posible utilizar otro tipo de asignación de muestra en estratos si el marco muestral disponible carece de información auxiliar. Se prevé que éste será el caso para la población de proveedores en cuestión y que sólo será posible sub-estratificar, cuando mucho, por sexo al interior de los estratos definidos por los dominios de estimación arriba representados en el Cuadro 3.

3.6. Función diseño de muestreo y probabilidades de inclusión

Sea $U = \{1, \dots, k, \dots, N\}$ la población objetivo de tamaño N , la estratificación define una partición de U en H estratos denotados $U_1, \dots, U_h, \dots, U_H$, de tamaño N_h , para $h = 1, \dots, H$, tal que $N = \sum_{h=1}^H N_h$. Sea $s_h \subset U_h$ una muestra de tamaño $n_h \leq N_h$.

Tomando en consideración el esquema de selección de la muestra probabilística descrito en apartados anteriores, i.e. muestreo directo de elementos sin reemplazo con probabilidades iguales intra-estrato (un muestreo aleatorio simple sin reemplazo estratificado), se tiene la siguiente *función diseño de muestreo* para cada estrato $p_h(\cdot)$ que describe cómo se seleccionó la muestra s_h de manera independiente para cada estrato $h = 1, \dots, H$ y determina las probabilidades $p_h(s_h)$ de seleccionar $s_h \in \mathcal{S}_h$, con \mathcal{S}_h denotando el conjunto de todas las muestras posibles para el estrato $h = 1, \dots, H$.

Se tiene que $s = s_1 \cup \dots \cup s_H$ es la muestra global de la población de proveedores y que la función diseño de muestreo global por independencia de estratos es $p(s) = p_1(s_1) \dots p_H(s_H)$. De modo que $p_h(\cdot)$ induce las *probabilidades de inclusión* de primer y segundo orden, respectivamente:

$$\pi_k = \frac{n_h}{N_h}$$

$$\pi_{kl} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}$$

Es a partir de estas probabilidades de inclusión que se generan las expresiones de los estimadores particulares al diseño de muestreo utilizado.

3.7. Estimación puntual

En particular para la población de proveedores y bajo el diseño de muestreo descrito anteriormente, la producción de estimaciones puntuales siguen las siguientes expresiones utilizando la notación definida en párrafos anteriores. Tales expresiones se obtienen a partir de los estimadores de Narain (1951) y Horvitz & Thompson (1952), sustituyendo los valores de las probabilidades de inclusión π_k y π_{kl} correspondientes (17, 18).

3.7.1. Estimación de un total

Bajo un diseño de muestreo directo de elementos aleatorio simple estratificado, un estimador de un total t de una variable de interés y de la población de proveedores es,

$$\hat{t}_y = \sum_{h=1}^H N_h \bar{y}_{s_h}$$

donde

$$\bar{y}_{s_h} = \sum_{k \in s_h} \frac{y_k}{n_h}$$

es el estimador de la media de la variable de interés y para el estrato h -ésimo.

La varianza del estimador \hat{t}_y es

$$V(\hat{t}_y) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} S_{y_{U_h}}^2$$

con $f_h = \frac{n_h}{N_h}$, la *fracción de muestreo* en el estrato h -ésimo y con

$$S_{y_{U_h}}^2 = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})^2$$

la varianza poblacional y \bar{y}_{U_h} la media poblacional de la variable de interés y de la población de proveedores dentro del estrato h -ésimo. Un estimador insesgado de esta varianza es,

$$\hat{V}(\hat{t}_y) = \sum_{h=1}^H N_h^2 \frac{1-f_h}{n_h} S_{y_{s_h}}^2$$

con $S_{y_{s_h}}^2$ la varianza muestral de la variable de interés y en el estrato h -ésimo definida:

$$S_{y_{s_h}}^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_{s_h})^2$$

3.7.2. Estimación de una media

En correspondencia con las expresiones anteriores, para la estimación de una media de una variable de interés y bajo el diseño de muestreo descrito en la población de proveedores, se utilizan las mismas expresiones de un total pero con adaptaciones para una media, es decir:

$$\hat{\bar{y}} = \sum_{h=1}^H W_h \bar{y}_{s_h}$$

donde

$$W_h = N_h/N$$

denota el peso del estrato h -ésimo. Su correspondiente estimador de varianza es

$$\hat{V}(\hat{y}) = \sum_{h=1}^H W_h^2 \frac{1-f_h}{n_h} S_{y_{s_h}}^2$$

3.7.3. *Estimación de una proporción*

La estimación de una proporción utiliza las mismas expresiones para estimar una media pero recodificando a la variable de interés y en variables indicadoras. Es decir, variables dicotómicas 0-1, en donde 1 significa que el individuo k -ésimo tiene la característica bajo estudio y 0 se asigna en cualquier otro caso.

3.7.4. *Estimación de parámetros en un modelo lineal o de funciones*

Es posible, siguiendo la misma teoría, desarrollar la estimación de parámetros en un modelo. Por ejemplo un coeficiente de regresión o por ejemplo el intercepto de un modelo.

El modelo en cuestión puede ser cualquier modelo, siempre y cuando sus parámetros a estimar se puedan expresar como funciones de totales o medias en donde cada uno de esos totales o medias se estiman con las expresiones utilizadas. Para la estimación de la varianza de funciones de totales, tal función no importa que sea no-lineal (e.g. una razón de totales). Para ello será necesario utilizar métodos de estimación de varianza por linealización o por remuestreo.

Se propone el uso de software especializado como el *Módulo Muestras Complejas* del paquete estadístico *IBM-SPSS®* o bien hacer los cálculos en *R* (R Core Team, 2014) utilizando, por ejemplo el paquete especializado en estimación de errores de muestreo *samplingVarEst* (19).

3.8. *Estimación por intervalos de confianza*

Una vez que se cuenta con las estimaciones puntuales de los parámetros de interés en la población de proveedores y su correspondiente estimación de varianza, se procede a la generación de estimaciones por intervalos de confianza. Suponiendo que es de interés la estimación para un parámetro θ que puede ser un total, una media, una proporción o una función de totales, por ejemplo, estos se calculan de la siguiente manera:

$$IC(\theta) = \hat{\theta} \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{\theta})}$$

donde $z_{\alpha/2}$ denota el cuantil de una distribución Normal que acumula una probabilidad de $\alpha/2$. Esto asumiendo una rápida convergencia a una distribución Normal (el caso común en la práctica).

De ser necesario, si se considera que hay un sesgo considerable en la estimación puntual para cierto parámetro θ cuyo estimador utilizado no es insesgado, e.g. estimación de funciones no-lineales, será necesario utilizar los siguientes intervalos de confianza de *error cuadrático medio*:

$$IC(\theta) = \hat{\theta} \pm z_{\alpha/2} \sqrt{\widehat{ECM}(\hat{\theta})}$$

donde

$$\widehat{ECM} = \hat{V}(\hat{\theta}) + [\hat{B}(\hat{\theta})]^2$$

con $\hat{B}(\hat{\theta})$ denotando un estimador insesgado del sesgo del estimador $\hat{\theta}$.

3.9. Estimación de coeficientes de variación y estimación de efectos de diseño

Como medida de calidad de las estimaciones generadas para la población de proveedores se usan los coeficientes de variación estimados. Estos se definen para $\hat{\theta}$ como:

$$cve(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}$$

Por su parte para medir la pertinencia o no del diseño de muestreo utilizado en la selección de la muestra para cada una de las variables de interés se utilizan los efectos de diseño estimados, definidos como:

$$deff(\hat{\theta}) = \frac{\hat{V}(\hat{\theta})}{\hat{V}_{MAS}(\hat{\theta})}$$

donde $\hat{V}_{MAS}(\hat{\theta})$ denota la estimación de varianza del estimador $\hat{\theta}$ bajo muestreo aleatorio simple.

Aunque poco común en la publicación de encuestas en México y en general en otros países, estas medidas son de extrema importancia para determinar qué tan bien resultó estimado lo que se está estimando en una encuesta. Adicionalmente, estas medidas sirven para que el estadístico muestrista tome decisiones al momento de repetir un estudio demoscópico.

Estas medidas son fáciles de obtener utilizando el software referido en párrafos anteriores. Para el caso de México, cabe mencionar que el Instituto Nacional de Estadística, Geografía e Informática (INEGI) menciona que las calcula pero es raro encontrar que las reporte.

4. Diseño de muestreo - población de beneficiarios

4.1. Marco muestral

En el caso de la población de beneficiarios no es posible contar de manera directa o inmediata con un listado de beneficiarios de interés particular (embarazadas, con niños 0-5 años) que hayan sido atendidos por personal capacitado de la segunda cascada. Además, debido a la rotación presente en esta población, no es posible tener actualizado el estatus de embarazo o lactancia en el padrón de beneficiarios del PDHO-EsIAN.

No obstante, es factible encontrar un listado actualizado, con las características de interés en los beneficiarios, en las unidades de salud en donde se atienden. De modo que es necesario utilizar conglomeración en el diseño de muestreo para mitigar este problema de marco muestral. Es decir, se propone un diseño de muestreo en más de una etapa.

En la primera etapa se propone como unidad primaria de muestreo a las unidades de salud o clínicas. En la segunda etapa de muestreo se seleccionarán beneficiarios con el estatus y características de interés. El esquema de selección será detallado en apartados posteriores.

Entonces, como marco muestral inicial, es decir en la primera etapa de muestreo, se utiliza un listado exhaustivo de unidades de salud o clínicas para las entidades federativas objetivo (Chihuahua, Oaxaca y Veracruz).

Como se ha comentado en párrafos anteriores es crucial la limpieza, veracidad y cobertura de tal listado de unidades de salud. También, cualquier tipo de estratificación en el marco muestral de clínicas tendrá que hacerse a partir de la información auxiliar contenida en dicho listado de unidades de salud, por ejemplo el número de beneficiarios que atiende con las características de interés (embarazadas, con niños de 0 a 2 años de edad, etc).

A continuación, se describe en el siguiente cuadro la población de unidades de salud a considerar para la primera etapa de muestreo.

Cuadro 4. Población de unidades de salud.

Entidad Federativa	Modalidad		Total por Entidad Federativa
	Rural	Urbano	
Chihuahua	349	49	398
Oaxaca	1,349	56	1,405
Veracruz	1,265	141	1,406
Total	2,963	246	3,209

Nota: Se excluyen un total de 26 unidades de salud que están fuera de operación.

Por su parte, la extracción de muestra en la segunda etapa de muestreo tendrá que correr a cargo de quien sea responsable del operativo de campo. Es decir, de quienes visiten las unidades de salud que cayeron en muestra en la primera etapa. Cabe mencionar que no es posible contar con un listado completo y actualizado de los beneficiarios de interés asociados a cada unidad de salud sino hasta haber contactado o visitado la unidad de salud. Una vez obtenido el listado de los beneficiarios, la selección se puede hacer utilizando un ordenamiento aleatorio de la lista. Así se preserva la aleatoriedad de selección y se completa el tamaño de muestra objetivo. Para tal ordenamiento basta con utilizar una tabla de números aleatorios, una calculadora simple, o un *smart-phone* que genere números aleatorios y se crea otra lista ordenada de manera aleatoria. Después se utiliza esa lista desordenada de manera aleatoria de los beneficiarios y se les busca en su domicilio o se agenda una visita si hay información de contacto. Utilizar un ordenamiento aleatorio es mejor que utilizar una muestra fija ya que se evitan los problemas de tasas de no respuesta o problemas de revisitas a un individuo caído en muestra utilizando esquemas tradicionales.

Para ejemplificarlo de manera didáctica. Suponga que se tiene un listado de 8 beneficiarios en la unidad de salud: A, B, C, D, E, F, G.

- En un esquema tradicional la muestra sería encuestar a C y F (por ejemplo), y tener que visitar a C y a F tantas veces como sea necesario hasta lograr obtener mediciones. Se tendría que considerar una tasa de no respuesta y en lugar de tener una muestra de tamaño 2 se necesitaría una muestra, digamos, de tamaño 3. Por ejemplo, C, F y G.

- En el esquema de ordenamiento aleatorio, la lista original se desordena aleatoriamente, por ejemplo: E, G, F, C, D, A, B, C. Se procede a buscar a E, si no se haya se busca a G, y así sucesivamente hasta completar 2. A esto se le llama *Permanent Random Numbering*. La muestra de tamaño 2 resultante será una muestra aleatoria simple. En campo es más fácil tener una ruta trazada que tener que estar revisitando intentando *capturar* al entrevistado.

En apartados posteriores se detallan más aspectos de la selección tanto de unidades de salud como de beneficiarios.

4.2. Estratificación

En la población de beneficiarios, por tratarse de un muestreo en 2 etapas, la estratificación puede llevarse a cabo en cada etapa. Se sugiere que gran parte de la estratificación se lleve a cabo en la primera etapa, que es en donde se concentra la mayor cantidad de variabilidad de entre las dos etapas de muestreo y sólo dejar para la segunda etapa aquella estratificación correspondiente a los dominios de estimación. Es decir, se propone una estratificación de unidades de salud de acuerdo a cierto conjunto de variables de estratificación. Por ejemplo, variables geográficas, variables construidas de niveles de densidad de familias beneficiarias y variables de distribución de capacitados en la segunda cascada (proveedores). Esto último debiera poderse obtener como resultado de compaginar los 2 marcos muestrales, el de proveedores y el de unidades de salud si se dispone de información de la unidad de salud en ambos marcos muestrales. Tales variables de estratificación o los insumos para construirlas necesariamente tienen que estar disponibles en el o los marcos muestrales.

Al igual que lo comentado para la población de proveedores, la estratificación tiene dos ventajas principales en este contexto. Por un lado, para controlar la muestra ante desviaciones y por otro lado, para mejorar las estimaciones. Esto último se logra, intuitivamente, partiendo un problema grande de estimación en varios problemas pequeños de estimación.

La mejora en precisión de las estimaciones dependerá de la relación (correlación) que tengan las variables de estratificación con el fenómeno medido. Ambas ventajas dependerán del tamaño de muestra disponible pues la cantidad de

estratos, como ya se mencionó, está limitada por la cantidad de muestra a distribuir en los estratos. En apartados siguientes se exhiben los tamaños de muestra calculados para la población de beneficiarios de interés.

Entonces, se proponen como variables de estatificación de la población de unidades de salud a la entidad federativa (Chihuahua, Oaxaca y Veracruz) y la modalidad de atención (Urbano/Rural). Adicionalmente se propone sub-estratificar a la población de beneficiarios de interés de acuerdo con los grandes dominios de interés (en la segunda etapa de muestreo), es decir por el estatus del beneficiario (mujeres embarazadas, con niños de 0-2 años, con niños de 3-5 años). De modo que se propone un total de 18 estratos que conforman los 18 dominios de estimación de donde se prevé serán desagregadas las estimaciones generadas. En apartados posteriores se especifican los tamaños de muestra de cada estrato.

Se podrían utilizar más estratos de variables geográficas, no obstante esta estratificación se piensa que no mejoraría las estimaciones y sólo encarecería de manera innecesaria la muestra pues se obtendría mayor dispersión geográfica. Notar que mayor dispersión geográfica no necesariamente mejora las estimaciones. Sólo las mejora cuando la geografía tiene un efecto importante en lo que se quiere medir. Viendo los tamaños de muestra a utilizar se podrá notar que se está concentrando el tamaño de muestra en tener un mayor número de unidades de salud y tener pocas observaciones por unidad de salud; de esta forma, se obtiene una buena dispersión en términos de unidades de salud y no en términos geográficos (notar que hay algunas jurisdicciones en Chihuahua que tienen muy pocas unidades de salud –alrededor de 5- y que reportaron tener 1 o 2 mujeres embarazadas a finales del año 2013). Entonces, obligar a tener todas las jurisdicciones en muestra mediante estratificación sería desaprovechar la oportunidad de captar elementos o insumos para el proyecto de evaluación de alternativas de esquemas de capacitación de la EsIAN.

4.3. Esquema de selección de las muestras de beneficiarios de interés

Por tratarse de un muestreo de beneficiarios de interés en dos etapas, las unidades primarias de muestreo (UMPs) serán las unidades de salud y las unidades secundarias de muestreo (USMs) serán los beneficiarios con las características de interés (mujeres embarazadas, con niños de 0-2 años, con niños de 3-5 años).

4.3.1. Selección de UPMs – unidades de salud

Se propone seleccionar a las UPMs utilizando probabilidades proporcionales al tamaño sin reemplazo, i.e. conducir la selección de UPMs con probabilidades de inclusión proporcionales a cierta variable auxiliar (comúnmente conocida como medida de tamaño, MDT).

Es importante resaltar que aunque se le llama muestreo con *probabilidades proporcionales al tamaño* este muestreo no necesariamente tiene que ver con tamaño. Es más bien tener probabilidades de inclusión proporcional a cierta variable que, idealmente, esté relacionada con lo que se quiere medir.

Habrán 3 levantamientos, se propone que en todas las ocasiones las unidades de salud sean seleccionadas con probabilidades proporcionales a la cantidad de proveedores de la 2da cascada a efecto de captar con mayor probabilidad a los beneficiarios de interés en la segunda etapa una vez que ya hay intervención, i.e. aquellos beneficiarios de interés atendidos por proveedores capacitados en la segunda cascada. De nuevo, se propone utilizar en la selección de UPMs un muestreo sin reemplazo y no sistemático por las razones técnicas explicadas en párrafos anteriores.

Es importante resaltar que si no se dispone de un marco muestral de proveedores que describa la distribución geográfica de los proveedores de 2da cascada, no será posible obtener una muestra de unidades de salud donde con alta probabilidad se capte a aquellos beneficiarios atendidos por capacitados de la 2da cascada.

En particular, se propone utilizar un muestreo de Hájek (20) de máxima entropía, también conocido como muestreo Poisson Condicional, o uno del alta entropía como el de Rao (21) y Sampford (22). Estos muestreos permiten aproximar numéricamente de manera adecuada las probabilidades de inclusión de 2do orden (14, 15, 16, 23) involucradas en la estimación de varianzas (descrita más adelante).

Respecto a la estimación de varianzas, es **crucial** una correcta y precisa estimación de varianzas en la producción de estimaciones ya que lo que se intenta medir son cambios en el tiempo. Utilizar software comercial de uso común (e.g. versiones comunes de STATA, SAS, SPSS, MINITAB y otros) para la generación de estimaciones y sus varianzas puede derivar en conclusiones erróneas debido a lo *grueso* de la estimación de varianzas, i.e. lo grueso al momento de determinar si hubo o no un cambios estadísticamente significativos. Esto cobra importancia si los cambios son muy pequeños. Adicionalmente, es en este tema donde vuelve a cobrar relevancia el no

utilizar un muestreo sistemático. La práctica común en México es utilizar un muestreo sistemático en la extracción de la muestra y luego asumir otro diseño de muestreo en la generación de estimaciones y sus varianzas (debido a que el muestreo sistemático tiene probabilidades de inclusión de segundo orden igual a 0, y por tanto, no es posible calcular varianzas).

En general, la práctica de extraer muestras de una forma y producir estimaciones suponiendo otra forma de extracción de muestras deriva en estimaciones erróneas (1, 2, 4, 7, 10, 12).

4.3.2. Selección de USMs – beneficiarios con características de interés

Para la selección de USMs se propone que, una vez que el personal de campo se encuentre en las unidades de salud seleccionadas en la primera etapa, sea solicitado al responsable de la unidad de salud un listado actualizado de aquellos beneficiarios con las características de interés (mujeres embarazadas, con niños de 0-2 años, con niños de 3-5 años). Este listado tendrá información de contacto de tales beneficiarios (teléfono, domicilio) de modo que será posible agendar o planificar visitas domiciliarias a tales beneficiarios. La cantidad de estos beneficiarios a entrevistar se detalla en apartados posteriores.

La información de registro de beneficiarios provista por las unidades de salud caídas en muestra (los totales por unidad de salud en muestra) permitirá realizar una expansión de los resultados obtenidos mediante los factores de expansión pertinentes en las fórmulas de los estimadores.

La selección de los beneficiarios de interés (mujeres embarazadas, con niños de 0-2 años, con niños de 3-5 años) se hará de manera aleatoria con ayuda de una calculadora, una tabla de números aleatorios o el esquema que el operador de campo determine más fácil pero que garantice la aleatoriedad en la selección. Cualquiera que sea el caso se asumirá en esa segunda etapa de muestreo, un muestreo aleatorio simple con igual probabilidad.

De nuevo, como se mencionó en apartados anteriores, se asume en todo momento que el levantamiento de información se realiza con veracidad y calidad. En otras palabras, se asume profesionalismo en las labores de recolección de la información, integridad ética por parte de los encuestadores y se asume la no existencia

de conflictos de interés de la empresa, institución u organización encargada de la recolección de la información en campo.

Respecto a la selección de beneficiarios, obviando la parte ética y profesionalismo en el operativo de campo, se tiene presente la dificultad de que se depende fuertemente en si las unidades de salud tienen la información de beneficiarios (de interés) de manera organizada y actualizada.

En concordancia con los tamaños de muestra propuestos (en apartados posteriores), se espera que los supuestos hechos en la segunda etapa simplifiquen el levantamiento y al mismo tiempo no afecten las estimaciones generadas. Aquí es pertinente recordar que es sabido que alrededor del 75-80% de la variabilidad en un muestreo en 2 etapas se concentra en la selección de muestra correspondiente a la primera etapa. De modo que es posible flexibilizar la parte técnica en la segunda etapa según la realidad que enfrente el equipo de levantamiento en campo.

4.4. Cálculo de tamaño de muestra

De acuerdo con el diseño de muestreo propuesto en 2 etapas, será necesario determinar el tamaño de muestra de la primera etapa, i.e. de UPMs, de acuerdo a los márgenes de error propuestos y también será necesario fijar un tamaño de muestra tope en la segunda etapa a efecto de que los responsables de la recolección de información puedan determinar sus costos en la implementación del operativo de campo correspondiente.

Una aproximación conservadora al tamaño de muestra de UPMs se logra utilizando la Figura 4, que aunque determina la relación entre márgenes de error y tamaños de muestra para un muestreo aleatorio simple, es útil para dar cuenta del tamaño de muestra necesario de UPMs a cierto nivel de margen de error máximo, i.e. el caso más pesimista y por lo tanto, conservador porque con un muestreo con probabilidades proporcionales al tamaño se obtendrán márgenes de error más pequeños. En otras palabras, bajo un muestreo con probabilidades proporcionales al tamaño se tiene mayor precisión en las estimaciones que en un muestreo aleatorio simple y entonces, determinar un tamaño de muestra bajo muestreo aleatorio simple es conservador (siempre y cuando se mantenga un tamaño de muestra relativamente pequeño en la segunda etapa de muestreo). Así, dichos márgenes de error bajo muestreo aleatorio simple se pueden esperar que no sean superados.

Al igual que en la población de proveedores, es importante mencionar que el tamaño de muestra también dependerá del nivel de desagregación que se desee en los resultados generados en la población de beneficiarios. Es decir, si se querrán arrojar estimaciones por entidad federativa, para mujeres embarazadas dentro de cierta entidad federativa o para mujeres embarazadas en general, por ejemplo. De nueva cuenta, el tamaño de muestra será especialmente sensible y tenderá a aumentar mientras más dominios de estimación existen. Todo esto se detalla a continuación.

Como resultado del cálculo específico de tamaños de muestra a utilizar en la población de unidades de salud se obtuvieron los siguientes tamaños de muestra representados en el Cuadro 5 siguiente

Cuadro 5. Tamaño de muestra de la población de unidades de salud
(Primera etapa de muestreo)

Entidad Federativa	Modalidad		Total por Entidad Federativa
	Rural	Urbano	
Chihuahua	40	30	70
Oaxaca	155	34	189
Veracruz	145	86	231
Total	340	150	490

Pensando en la viabilidad del operativo de campo para la recolección de información, es decir, pensando en costos, trazos de rutas, la organización de equipos de encuestadores (3 encuestadores y 1 supervisor por cada equipo), el rendimiento diario por encuestador dado que tiene que contactar, agendar y/o localizar a los beneficiarios de interés según los registros provistos por la unidad de salud (estimado en 12 encuestas diarias por equipo).

Se propone como tamaño de muestra en la segunda etapa, un muestreo de 12 beneficiarios de interés distribuidos en 4 mujeres embarazadas, 4 mujeres con niños 0-2 años y 4 mujeres con niños 3-5 años. Cabe mencionar que en algunos casos no será posible encuestar 12 beneficiarios de interés pues pudiera ocurrir que hay menos de esa

cantidad registrados para algunas unidades de salud en muestra. No obstante, para ser realistas, es necesario topar como máximo esta cifra.

De manera que el tamaño de muestra máximo inducido de beneficiarios quedará como se describe a continuación en el siguiente cuadro:

Cuadro 6. Tamaño de muestra máximo inducido de beneficiarios de interés

(Segunda etapa de muestreo)

Entidad Federativa	Modalidad						Total por Entidad Federativa
	Rural			Urbano			
	Embarazadas	con niños 0-2	con niños 3-5	Embarazadas	con niños 0-2	con niños 3-5	
Chihuahua	160	160	160	120	120	120	840
Oaxaca	619	619	619	137	137	137	2,268
Veracruz	581	581	581	344	344	344	2,775
Total	1,360	1,360	1,360	601	601	601	5,883

Para el cálculo específico de tamaños de muestra se consideraron las siguientes restricciones y/o especificaciones:

- Se utilizaron 6 dominios de estimación. Esto es, se anticipan 6 subpoblaciones de interés para las que se generarán estimaciones (como lo solicitó la Dirección de Enlace de Evaluación Externa para abaratar los costos del operativo de campo, en lugar de considerar 18 si se toman entidades federativas por separado). Estos dominios están representados en el Cuadro 7, a continuación:

Cuadro 7. Dominios de estimación de la población de beneficiarios

Modalidad					
Rural			Urbano		
Embarazadas	con niños 0-2	con niños 3-5	Embarazadas	con niños 0-2	con niños 3-5
Dominio 1	Dominio 2	Dominio 3	Dominio 4	Dominio 5	Dominio 6

- Con los tamaños de muestra descritos en los Cuadros 5 y 6 se espera un margen de error para **Embarazadas-Rural** de aproximadamente entre +/- 2.55% y 5.26%, (estos son el extremo optimista y el extremo pesimista, respectivamente) en las estimaciones de proporciones. Se está asumiendo un total de 16,479 embarazadas de acuerdo con datos de noviembre a diciembre de 2013 (cifra variable en el tiempo por la problemática de rotación del proyecto descrita en párrafos anteriores), y suponiendo un nivel del 95% de confianza.
- Para las proporciones estimadas del dominio **Embarazadas-Urbano** se espera un margen de error de aproximado entre +/- 1.92% y 4.99% asumiendo 2,859 embarazadas de acuerdo con el 2013 y un 95% de confianza.
- Si se está hablando de proporciones estimadas con menor desagregación entonces se espera obtener menores márgenes de error para éstas. Por ejemplo, se esperaría un margen de error para **Embarazadas** de aproximadamente entre +/- 2.098% y 4.37%, (optimista y pesimista, respectivamente). Esto suponiendo 19,338 embarazadas de acuerdo con el 2013 a un nivel de 95% de confianza.
- Es importante mencionar que los errores teóricos arriba listados dependen de la magnitud de las proporciones estimadas. Estos márgenes de error calculados no son los adecuados si estuviese en juego la estimación de proporciones muy pequeñas (menores al 5%) o proporciones muy grandes (mayores al 95%) en la cuantificación de presencia de cierto atributo.
- De acuerdo con lo descrito en incisos anteriores, se anticipa que será posible captar cambios estadísticamente significativos (a un nivel de confianza del 95%) si en mediciones consecutivas (con los mismos márgenes de error, tamaños de muestras, etc.) se observan cambios de al menos 7.81% para la desagregación

Embarazadas-Rural en estimaciones puntuales de proporciones (si se toma un promedio entre el margen de error optimista y el pesimista).

- De manera correspondiente se podría hablar de cambios estadísticamente significativos si se observan cambios de al menos: 6.91% para **Embarazadas-Urbano**, 6.468% para **Embarazadas** en general en la población conformada por los 3 estados de manera conjunta.
- Los márgenes de error calculados en incisos anteriores son ilustrativos. Hay que recordar que están calculados con base en datos de la población de beneficiarios de interés del 2013 en los meses de noviembre a diciembre. Se reitera, la población sufre una alta rotación y estas cifras son para cierta temporalidad en particular. Las mujeres embarazadas, por ejemplo, pueden distribuirse de manera asimétrica aunque para plazos más largos, e.g. un año, haya una distribución con cierta uniformidad en las áreas geográficas.
- Algo favorecedor para este ejercicio de muestreo es que se pueden esperar márgenes de error optimistas (más pequeños) si se considera el hecho de que se utilizará un muestreo con probabilidades proporcionales al tamaño según la distribución de familias (en el levantamiento pre-intervención) y según la distribución de proveedores de la segunda cascada (en levantamientos subsecuentes). De manera tal que se esperarían estimaciones más precisas de lo esperado.
- Una vez generadas las estimaciones quien realice el análisis deberá reportar los márgenes de error estimados realistas y no teóricos. Esto debiera ser posible si se hace uso de software especializado en muestreo.
- Para los cálculos de tamaño de muestra se asumió máxima variabilidad de las proporciones que se medirán. Es decir, se asumirá total desconocimiento de lo que se intenta cuantificar. De igual forma que para la población de proveedores, este supuesto es conservador y se sostiene de manera fácil debido a que se trata de una encuesta que no sólo mide un aspecto en las unidades de observación.
- Se asumió una tasa de respuesta del 100%. Es decir, se asume que los beneficiarios contestarán los cuestionarios aplicados en las entrevistas y que no habrá rechazos a responder o datos faltantes.
- Es pertinente mencionar que se trata de márgenes de error teóricos que son de carácter ilustrativo y que son muy conservadores, i.e. son ligeramente más

grandes a lo esperado ya que se utilizará un diseño de muestreo más preciso al que fue utilizado para calcular los márgenes de error. En este momento, no es posible utilizar el diseño de muestreo con probabilidades proporcionales al tamaño para calcular los márgenes de error porque se requeriría de añadir *efectos de diseño* (12) de estudios pasados (comúnmente llamados *deffs*) que midan o estudien temas similares. Con los *deffs* es posible mejorar un diseño de muestreo en el tiempo, no obstante es rara su publicación. Al respecto, hay que mencionar el mismo INEGI dice que los calcula pero no es fácil encontrar su publicación.

- Finalmente, de nueva cuenta, se asume en todo momento que el levantamiento de información se realiza con veracidad y calidad. En otras palabras, se asume profesionalismo en las labores de recolección de la información, integridad ética por parte de los encuestadores y se asume la no existencia de conflictos de interés de la empresa, institución u organización encargada de la recolección de la información en campo. Si este supuesto no se cumple existe el riesgo de derivar en mediciones totalmente erróneas a las que no apliquen los márgenes de error arriba mencionados.

A continuación se exhiben las fracciones de muestreo asociadas a los tamaños de muestra utilizados en la primera etapa de muestreo:

Cuadro 8. Fracciones de muestreo de la primera etapa de muestreo

Entidad Federativa	Modalidad		Por Entidad Federativa
	Rural	Urbano	
Chihuahua	11.5%	61.0%	17.6%
Oaxaca	11.5%	61.0%	13.4%
Veracruz	11.5%	61.0%	16.4%
Total	11.5%	61.0%	15.3%

En el Cuadro 8 es posible observar que los márgenes de error discutidos anteriormente siguen un comportamiento similar a estas fracciones de muestreo. Es decir, se espera un margen de error menor (por la existencia de una fracción de muestreo mayor) para el dominio Embarazadas-Urbano que para Embarazadas-Rural.

También, como ya se mencionó en los incisos anteriores, se observa una muy elevada fracción de muestreo en las modalidades urbano, de modo que se estarían utilizando a las UPMs como una especie de estratificación y entonces se tendría un efecto similar al de sólo estratificar USMs, lo que sería equivalente a un muestreo en una etapa estratificado. Estos comentarios son relevantes para dar cuenta de lo conservador de los cálculos de márgenes de error.

4.5. Distribución de muestra en los estratos

El marco muestral de unidades de salud disponible tiene información limitada, i.e. sólo contiene identificadores y no tiene mediciones del pasado similares a lo que es de interés medir. La distribución de muestra para la primera etapa de muestreo está determinada por el Cuadro 5.

En general, la asignación de muestra en estratos a utilizar sirve más para controlar la distribución espacial de la muestra que para aumentar la precisión de las estimaciones. La precisión en las estimaciones se recargará, para la población de beneficiarios objetivo, más en el hecho de que se está utilizando un muestreo con probabilidades proporcionales al tamaño.

4.6. Probabilidades de inclusión

En adelante, para simplificar la notación, utilizaremos expresiones matemáticas suponiendo un solo estrato o dominio, i.e. para $H = 1$. Esto es algo común en literatura de muestreo para evitar el uso excesivo de subíndices una vez que ya quedó definida la estratificación.

Entonces, considerando el esquema de selección de la muestra probabilística de unidades de salud descrito en apartados anteriores, i.e. muestreo sin reemplazo con probabilidades proporcionales al número de familias beneficiarias (en el primer levantamiento, antes de la intervención) o número de proveedores de la segunda cascada (en los siguientes levantamientos), y considerando el esquema de selección de los beneficiarios de interés descritos, se tienen las siguientes *probabilidades de inclusión* de primer orden del elemento (beneficiario de interés) k -ésimo que pertenece a la UPM i -ésima (unidad de salud):

$$\pi_k = \pi_{Ii}\pi_{k|i} = \frac{n_I x_{Ii}}{\sum_{j=1}^{N_I} x_{Ij}} \frac{n_{Ii}}{N_{Ii}}$$

donde π_{li} es la probabilidad de inclusión de primer orden de la UPM i -ésima; $\pi_{k|i}$ es la probabilidad de inclusión de la USM k -ésima dada la selección de la UPM i -ésima; n_l es el tamaño de muestra de UPMs; x_{li} es la variable auxiliar (en el primer levantamiento es el número de familias beneficiarias, en levantamientos subsecuentes este valor depende de la distribución de capacitados de la segunda cascada, i.e. proveedores); N_l es el número de UPMs en la población de UPMs; n_{li} es el número de encuestados en la segunda etapa, i.e. el número de beneficiarios de interés encuestados en la unidad de salud; finalmente, y N_{li} el número de beneficiarios de interés inscritos en la unidad de salud.

Por su parte, respecto a las probabilidades de inclusión de segundo orden se propone utilizar la aproximación de Hájek (20). Usar esta aproximación es válido debido a que se propuso el uso de un muestreo no de baja entropía. Berger (23) demuestra y da suficientes condiciones para que se puedan utilizar este tipo de aproximaciones cuando el diseño de muestreo es de alta entropía diferente al muestreo de máxima entropía. De modo que,

$$\pi_{lij} \approx \pi_{li}\pi_{lj} \left\{ 1 - \frac{(1 - \pi_{li})(1 - \pi_{lj})}{d_l} \right\}$$

con $d_l = \sum_{i=1}^{N_l} \pi_{li}(1 - \pi_{li})$ o bien si resulta demasiado cómputo intensivo es posible utilizar $\hat{d}_l = \sum_{i=1}^{n_l} (1 - \pi_{li})$ en lugar de d (15, 24). Así, tenemos que las probabilidades de segundo orden para los beneficiarios de interés son, si los elementos (beneficiarios de interés) k -ésimo y l -ésimo pertenecen a la UPM i -ésima (unidad de salud):

$$\pi_{kl} = \pi_{li}\pi_{kl|i} = \frac{n_l x_{li}}{\sum_{j=1}^{N_l} x_{lj}} \frac{n_{li}(n_{li} - 1)}{N_{li}(N_{li} - 1)}$$

Y si el elemento k -ésimo pertenece a la UPM i -ésima y el elemento l -ésimo pertenece a la UPM j -ésima:

$$\pi_{kl} = \pi_{lij}\pi_{k|i}\pi_{l|j} = \pi_{lij} \frac{n_{li}}{N_{li}} \frac{n_{lj}}{N_{lj}}$$

4.7. Estimación puntual

Para población de beneficiarios de interés y bajo el diseño de muestreo propuesto en dos etapas, las estimaciones puntuales descritas anteriormente, la producción de estimaciones puntuales siguen las expresiones siguientes obtenidas de las fórmulas

generales de los estimadores de Narain (17) y Horvitz & Thompson (18) substituyendo los valores de las probabilidades de inclusión π_k y π_{kl} correspondientes.

4.7.1. Estimación de un total

Bajo un diseño de muestreo en dos etapas, un estimador de un total t de una variable de interés y de la población de beneficiarios de interés es,

$$\hat{t}_y = \sum_{i=1}^{n_I} \frac{\hat{t}_i}{\pi_{Ii}}$$

donde \hat{t}_i es un estimador de Narain-Horvitz-Thompson del total t_i de la variable de interés y correspondiente a la UPM i -ésima con respecto a la segunda etapa.

Un estimador de la varianza del estimador \hat{t}_y , que utiliza el *principio de conglomerados últimos* es:

$$\hat{V}(\hat{t}_y) = \sum_{i \in S_I} \sum_{j \in S_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}}$$

4.7.2. Estimación de una media

Para la estimación de una media de una variable de interés y bajo el diseño de muestreo descrito en la población de beneficiarios de interés, se utiliza la expresión:

$$\hat{y} = \frac{\hat{t}_y}{N}$$

y su correspondiente estimador de varianza es

$$\hat{V}(\hat{y}) = \frac{\hat{V}(\hat{t}_y)}{N^2}$$

4.7.3. Estimación de una proporción

Como se ilustró para la población de proveedores, para estimar una proporción se utilizan las mismas expresiones matemáticas que para una media pero recodificando la variable de interés por una variable indicadora.

4.7.4. Estimación de parámetros en un modelo lineal o de funciones

Al igual que en la población de proveedores, es posible la estimación de parámetros en un modelo con parámetros expresables como funciones de totales o medias en donde cada uno de esos totales o medias se estiman con las expresiones de expansión simple.

La estimación de varianza se realiza mediante métodos de estimación de varianza por linealización o por remuestreo. Más detalles se encuentran en lo descrito para la población de proveedores.

4.8. Estimación por intervalos de confianza, coeficientes de variación y estimación de efectos de diseño

En este apartado aplican las definiciones utilizadas en la población de proveedores.

5. Comentarios sobre el uso de una sola muestra de unidades de salud

A continuación se discuten algunos aspectos sobre el uso de una sola muestra de unidades de salud para de esta misma obtener información de los proveedores.

- En conversaciones de trabajo surgió la preocupación de una muy elevada dispersión geográfica en la muestra de proveedores debido a que estos se están muestreando directamente. Existe la posibilidad de que no se pudieran contactar vía telefónica de manera directa se pudiera encarecer la obtención de información de los proveedores. Esta preocupación es pertinente.
- En tal caso se tendría que cambiar completamente la parte del diseño muestral correspondiente a la población de proveedores, es decir toda la parte que conforma la sección 3 de este documento. Esto es posible, pero sería necesario discutirlo y acordarlo.
- Como se comentó en las primeras reuniones de trabajo. Si se tienen registros veraces y completos de los proveedores de la segunda cascada con información de contacto confiable un muestreo directo de elementos, i.e. en una etapa sería el caso ideal pues se estaría utilizando un diseño de muestreo óptimo y muy económico. Por ejemplo si se asume un costo aproximado de \$100.00 pesos por cuestionario telefónico, se estaría hablando de una muestra de \$400,000.00 pesos, considerando un tamaño de muestra de 4,000. También, como se describe en apartados anteriores un diseño de muestreo en una etapa sería capaz de captar

cambios estadísticamente significativos si se tienen cambios en estimaciones puntuales de proporciones de al menos 2.7% aproximadamente.

- Como una regla de decisión se propone lo siguiente: si se prevé que habrán cambios bastante mayores a 2.7% entre cifras de estimaciones puntuales de proporciones (retomando el inciso anterior) para la población de proveedores, entonces será más fácil captar esos cambios de manera significativa con muestras menos precisas y entonces se favorece la decisión de utilizar dos etapas de muestreo, i.e. utilizar una sola muestra de unidades de salud. Por el contrario, si se prevén cambios difíciles de captar y es inviable llevar a efecto un aumento en el tamaño de muestra de unidades de salud (primera etapa en un diseño de dos etapas) entonces el tema estaría en utilizar muestreo directo en una sola etapa pero se tendría que garantizar la calidad del marco muestral de proveedores, i.e. el listado nominal de los capacitadores de la segunda cascada con información útil de contacto.
- Para concluir, esta sección, notar que en realidad lo que está en juego en esta decisión son:
 1. Cuestiones de tipo pecuniario, que dependen por un lado de quien pague el estudio cuantitativo y por otro lado de la empresa, institución u organismo que realice el trabajo de recolección de información en campo (telefónico o en vivienda).
 2. Calidad en los datos a utilizar como marco muestral, que depende de la organización de las labores de capacitación y de la disciplina que tengan los capacitadores de la primera cascada y de sus decisiones de a quienes capacitar en la segunda cascada.
 3. Cambios esperados en las mediciones y que interesan captarse con las muestras, que depende por un lado de quienes tomarán decisiones a partir de los cambios estimados de manera estadísticamente significativa y por otro lado de quien desarrolló o planteó de inicio el que se llevara a cabo la capacitación o modalidad de capacitación cuyos resultados se quieren evaluar, i.e. de quien propuso el uso del CD de capacitación como una herramienta útil y del impacto que anticipó.
- Se termina esta sección afirmando que el diseño de muestreo se debe adaptar a las necesidades del consumidor del estudio demoscópico y de lo que es viable llevar a la práctica. No obstante para lograr tal adaptación es indispensable que

el o los consumidores del estudio tengan clara la finalidad y objetivos de éste, i.e. qué se tenga claro de manera integral lo que les importa más, qué les preocupa menos y los recursos que destinarán.

6. Comentario final

Con el marco muestral de proveedores provisto a la fecha no es viable seleccionar una muestra válida a efecto de cubrir con los objetivos del proyecto. Esto debido a la incompletez y baja calidad de la base de datos a usar como marco muestral. Esto se documenta con cifras en la parte de Anexos.

Por último se tienen las siguientes recomendaciones:

- Se sugiere no llevar a cabo el levantamiento de información a partir de este marco muestral disponible. No sería posible sostener inferencias válidas.
- Recordar que no tiene sentido utilizar una muestra de unidades de salud para los objetivos del proyecto ya que no se sabe en donde están aquellos capacitados de la 2da cascada. No hay que olvidar que lo que interesa son dos cosas:
 1. la población beneficiaria atendida por capacitados de la 2da cascada
 2. la población de capacitados de la 2da cascada
- Los problemas hallados en la recolección de información sugieren dar un paso atrás en todo el proyecto. Es decir, pareciera que en lugar de hablar de *Evaluación de Resultados* habría que plantearse una *Evaluación de Procesos* de la capacitación de la EsIAN.
- Como se mencionó en párrafos anteriores, una alternativa sería dejar fuera al aparato gubernamental por completo en la capacitación y mejor utilizar a los encuestadores de la evaluación para llevar a cabo también la capacitación. Es decir, que los encuestadores hagan ambas cosas: mediciones (pre y post), y la capacitación misma. Notar que el encuestador tendrá incentivos a hacer un adecuado registro de la

información de contacto del capacitado ya que lo tendrá que volver a encuestar en el futuro.

Referencias

1. Särndal, C.-E., Swensson, B. & Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, Inc.
2. Pfeffermann, D. & Rao, C.R. (eds.) (2009). Handbook of Statistics 29A. Sample Surveys: Designs, Methods and Applications. North-Holland.
3. Brewer, K. R. W. & Gregoire, T. G. (2009). Introduction to survey sampling. Handbook of Statistics – Vol. 29A. Sample Surveys: Design, Methods and Applications. (D. Pfeffermann and C. R. Rao, eds.) Elsevier B. V. pp. 9-37.
4. Smith, T. M. F. (2001). Biometrika centenary: Sample surveys. Biometrika 88 (1): 167–243.
5. Rao, J.N.K. (2003). Small Area Estimation. John Wiley & Sons, Inc.
6. Valliant, R., Dorfman, A.H. & Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. John Wiley & Sons, Inc.
7. Lohr, S.L. (2009). Sampling: Design and Analysis. Duxbury Press.
8. Lynn, P. (ed.) (2009). Methodology of Longitudinal Surveys. John Wiley & Sons, Inc.
9. Andreß, H.-J., Golsch, K. & Schmidt, A. W. (2013). Applied Panel Data Analysis for Economic & Social Surveys. Springer-Verlag, Inc.
10. Statistics Canada. (2003). Catalogue No. 12-587-XPE, Social Survey Methods Division. Statistics Canada.
11. Kish, L. & Frankel M.R. (1974). Inference from complex samples. Journal of the Royal Statistical Society B. 36, pp. 1-37
12. Kish, L. (1965). Survey Sampling. John Wiley & Sons, Inc
13. Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. Biometrika, 71, pp. 171-175.
14. Tillé, Y. (2006). Sampling Algorithms. New York: Springer.
15. Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. Biometrika, 94, 953-964.
16. Haziza, D., Mecatti, F. & Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. Metron - International Journal of Statistics LXVI, 91-108.
17. Narain, R. D. (1951). On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics 3, 169-174.
18. Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.
19. Escobar, E. L. & Barrios, E. (2012). SamplingVarEst: Sampling Variance Estimation. R package version 0.9-9.
20. Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35, 1491-1523.

21. Rao, J. N. K. (1965). On two simple schemas of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
22. Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
23. Berger, Y. G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics* 27, 407–426.
24. Escobar, E. L. & Berger, Y. G. (2013). A jackknife variance estimator for selfweighted two-stage samples. *Statistica Sinica*, 23, 595-613.

Anexos

En esta sección se muestra el grado de completez y calidad del marco muestral de proveedores (capacitados de la segunda cascada) disponible a la fecha. Los resultados muestran que el marco muestral es insuficiente.

Como primer acercamiento, es importante notar que en la población planeada para capacitación del Cuadro 1 se contempla un total de 19,568 capacitados de la 2da cascada. El marco muestral provisto sólo tiene 10,582. Es decir, se podría decir que se tiene una cobertura de aproximadamente un 54% en el marco muestral provisto. Un marco muestral con un porcentaje de cobertura tan bajo no sería útil si lo que se quiere es detectar cambios estadísticamente significativos que se prevén serán pequeños.

El Cuadro A, desglosa la información disponible por entidad y por capacitados de cada cascada.

Cuadro A. Capacitados en el marco muestral disponible a la fecha

Entidad Federativa	Capacitados	
	1era cascada	2da cascada
Chihuahua	22	1,856
Oaxaca	75	3,388
Veracruz	66	5,338
Total	163	10,582

El siguiente Cuadro B desglosa aún más la información de acuerdo al número de capacitados de la segunda cascada por cada capacitador de la primera cascada. Se omite el nombre del capacitador por confidencialidad. Es importante mencionar que el objetivo planeado eran 100 capacitados de la segunda cascada por cada capacitador de la primera.

Cuadro B. Capacitados en 2da cascada por cada capacitador de la 1era cascada

Institución	Entidad	Municipio	Jurisdicción	Cap 2da. Cascada
IMSS OPORT.	OAXACA	SANTIAGO JUXTLAHUACA	MIXTECA	3
IMSS OPORT.	OAXACA	MIAHUATLAN DE PORFIRIO DIAZ	VALLES CENTRALES	4
IMSS OPORT.	OAXACA	SANTIAGO JUXTLAHUACA	MIXTECA	4
	OAXACA			6
IMSS OPORT.	OAXACA	SANTIAGO JAMILTEPEC	COSTA	6
IMSS OPORT.	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	7
IMSS OPORT.	OAXACA	MATIAS ROMERO AVENDAÑO	ISTMO	7
IMSS OPORT.	OAXACA	MIAHUATLAN DE PORFIRIO DIAZ	VALLES CENTRALES	8
IMSS OPORT.	OAXACA	HEROICA CIUDAD DE TLAXIACO	MIXTECA	10
IMSS OPORT.	OAXACA		MIXTECA	10
IMSS OPORT.	OAXACA	HEROICA CIUDAD DE TLAXIACO	MIXTECA	11
IMSS OPORT.	VERACRUZ	CHICONTEPEC	PÁNUCO	11
IMSS OPORT.	VERACRUZ			11
	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	12
IMSS OPORT.	OAXACA	HEROICA CIUDAD DE TLAXIACO	MIXTECA	13
IMSS OPORT.	VERACRUZ	COSCOMATEPEC	CÓRDOBA	13
IMSS OPORT.	CHIHUAHUA	HIDALGO DEL PARRAL		14
IMSS OPORT.	CHIHUAHUA	HIDALGO DEL PARRAL	PARRAL	14
IMSS OPORT.	OAXACA	HEROICA CIUDAD DE TLAXIACO	MIXTECA	14
IMSS OPORT.	OAXACA	MATIAS ROMERO AVENDAÑO	ISTMO	14
IMSS OPORT.	OAXACA	MIAHUATLAN DE PORFIRIO DIAZ	VALLES CENTRALES	15
IMSS OPORT.	OAXACA	SANTIAGO JAMILTEPEC	COSTA	17
IMSS OPORT.	OAXACA	SAN JUAN BAUTISTA TUXTEPEC	TUXTEPEC	18
IMSS OPORT.	VERACRUZ	CHICONTEPEC	PÁNUCO	18
SSA	OAXACA	VILLA DE ZAACHILA	VALLES CENTRALES	19
IMSS OPORT.	VERACRUZ	HUAYACOCOTLA	POZA RICA	19
	OAXACA			21
IMSS OPORT.	VERACRUZ	ATZALAN	MARTÍNEZ DE LA TORRE	22
IMSS OPORT.	VERACRUZ	PAPANTLA	POZA RICA	25
IMSS OPORT.	VERACRUZ	PAPANTLA	POZA RICA	26
SSA	VERACRUZ	PANUCO	PANUCO	27
IMSS OPORT.	CHIHUAHUA	GUACHOCHI	GUACHOCHI	30
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	36
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	SAN FRANCISCO LACHIGOLÓ	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
SSA	OAXACA	TLACOLULA DE MATAMOROS	SIERRA	37
IMSS OPORT.	VERACRUZ	ORIZABA	ORIZABA	37
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	40
SSA	OAXACA	SAN FELIPE USILA	TUXTEPEC	40
SSA	OAXACA	SAN PEDRO IXCATLAN	TUXTEPEC	41
SSA	OAXACA	SAN FELIPE JALAPA DE DÍAZ	TUXTEPEC	41
SSA	OAXACA	SAN MIGUEL SOYALTEPEC	TUXTEPEC	41
SSA	OAXACA	SAN PEDRO IXCATLAN	TUXTEPEC	41
SSA	CHIHUAHUA	HIDALGO DEL PARRAL	PARRAL	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	SAN JUAN LALANA	TUXTEPEC	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	SANTIAGO JOCOTEPEC	TUXTEPEC	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	ISTMO	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	AYOTZINTEPEC	TUXTEPEC	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	HUAJUAPAN DE LEÓN	MIXTECA	42
SSA	OAXACA	SAN FELIPE JALAPA DE DÍAZ	TUXTEPEC	42
SSA	OAXACA	SAN MIGUEL SOYALTEPEC	TUXTEPEC	42
SSA	VERACRUZ	TUXPAN	TUXPAN	42
SSA	VERACRUZ	TUXPAN	TUXPAN	42
SSA	VERACRUZ	TUXPAN	TUXPAN	43
SSA	VERACRUZ	TUXPAN	TUXPAN	43
SSA	VERACRUZ	TUXPAN	TUXPAN	45
SSA	VERACRUZ	TUXPAN	TUXPAN	45
SSA	VERACRUZ	TUXPAN	TUXPAN	45
IMSS OPORT.	VERACRUZ	ORIZABA	CÓRDOBA	47
SSA	CHIHUAHUA	CAMARGO	CAMARGO	50
IMSS OPORT.	VERACRUZ	ORIZABA		51
SSA	VERACRUZ	TUXPAN	TUXPAN	54
SSA	CHIHUAHUA	GOMEZ FARIAS	GOMEZ FARIAS	55
IMSS OPORT.	VERACRUZ	VERACRUZ	ORIZABA	56
	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	57
SSA	CHIHUAHUA	GUACHOCHI	GUACHOCHI	59
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	61
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	61
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	61

Institución	Entidad	Municipio	Jurisdicción	Cap 2 da. Cascada
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	61
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	62
SSA	VERACRUZ	COSAMALOAPAN DE CARPIO	COSAMALOAPAN	62
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	62
SSA	VERACRUZ	MARTÍNEZ DE LA TORRE	MARTÍNEZ DE LA TORRE	62
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	SANTO DOMINGO TEHUANTEPEC	ISTMO	63
SSA	OAXACA	CIUDAD IXTEPEC	ISTMO	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	63
SSA	CHIHUAHUA			64
SSA	CHIHUAHUA	JUÁREZ	JUÁREZ	64
SSA	CHIHUAHUA			65
SSA	OAXACA	JUCHITÁN DE ZARAGOZA	ISTMO	65
SSA	CHIHUAHUA			66
SSA	CHIHUAHUA			66
IMSS OPORT.	VERACRUZ	COSCOMATEPEC	CÓRDOBA	66
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	68
IMSS OPORT.	VERACRUZ	ORIZABA	ORIZABA	70
SSA	CHIHUAHUA	GUACHOCHI	GUACHOCHI	72
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	72
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	73
SSA	CHIHUAHUA	HIDALGO DEL PARRAL	PARRAL	78
SSA	CHIHUAHUA	CUAUHTÉMOC	CUAUHTÉMOC	80
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	81
SSA	CHIHUAHUA	BOCOYNA	CREEL	82
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	82
SSA	VERACRUZ	COSAMALOAPAN DE CARPIO	COSAMALOAPAN	85
SSA	OAXACA	SAN PEDRO MIXTEPEC -DTO. 22 -	COSTA	86
SSA	VERACRUZ	COSAMALOAPAN DE CARPIO	COSAMALOAPAN	92
SSA	VERACRUZ	BOCA DEL RÍO	VERACRUZ	96
SSA	VERACRUZ	PANUCO	PANUCO	96
SSA	VERACRUZ	VERACRUZ	VERACRUZ	97
SSA	VERACRUZ	VERACRUZ	VERACRUZ	98
SSA	VERACRUZ	VERACRUZ	VERACRUZ	99
SSA	OAXACA	OAXACA DE JUAREZ	VAILES CENTRALES	100
SSA	OAXACA	OAXACA DE JUAREZ	VAILES CENTRALES	100
SSA	OAXACA	OAXACA DE JUAREZ	VAILES CENTRALES	100
SSA	OAXACA	ZIMATLAN DE ALVAREZ	VAILES CENTRALES	100
SSA	OAXACA	OAXACA DE JUAREZ	VAILES CENTRALES	100
SSA	OAXACA	OAXACA DE JUAREZ	VAILES CENTRALES	100
SSA	OAXACA	SOLEDAD ETLA	VAILES CENTRALES	100
SSA	VERACRUZ	CÓRDOBA	CÓRDOBA	100
SSA	VERACRUZ	CÓRDOBA	CÓRDOBA	100
SSA	VERACRUZ	VERACRUZ	VERACRUZ	100
SSA	VERACRUZ	CÓRDOBA	CÓRDOBA	100
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	101
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	101
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	101
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	101
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	103
SSA	VERACRUZ	COATZACOALCOS	COATZACOALCOS	105
SSA	VERACRUZ	XALAPA	VERACRUZ	110
SSA	CHIHUAHUA	CHIHUAHUA	CHIHUAHUA	112
SSA	VERACRUZ	XALAPA	VERACRUZ	112
SSA	CHIHUAHUA	NUEVO CASAS GRANDES	NUEVO CASAS GRANDES	114
SSA	VERACRUZ	SAN ANDRÉS TUXTLA	SAN ANDRÉS TUXTLA	117
SSA	CHIHUAHUA	CUAUHTÉMOC	CUAUHTÉMOC	120
SSA	VERACRUZ	IXTACZOQUITLÁN	ORIZABA	120
SSA	VERACRUZ	ORIZABA	ORIZABA	120
SSA	VERACRUZ	NOGALES	ORIZABA	120
SSA	VERACRUZ	ORIZABA	ORIZABA	120
SSA	VERACRUZ	ORIZABA	ORIZABA	120
SSA	VERACRUZ	XALAPA	VERACRUZ	121
SSA	VERACRUZ	XALAPA	VERACRUZ	123
SSA	VERACRUZ	XALAPA	VERACRUZ	126
SSA	VERACRUZ	CÓRDOBA	CÓRDOBA	132
SSA	VERACRUZ	POZA RICA DE HIDALGO	POZA RICA	133
	OAXACA			136
SSA	VERACRUZ	POZA RICA DE HIDALGO	POZA RICA	137
SSA	VERACRUZ	XALAPA	VERACRUZ	138
SSA	CHIHUAHUA	BOCOYNA	CREEL	146
SSA	VERACRUZ	SAN ANDRÉS TUXTLA	SAN ANDRÉS TUXTLA	183
SSA	VERACRUZ	POZA RICA DE HIDALGO	POZA RICA	199
SSA	CHIHUAHUA	CHIHUAHUA	CHIHUAHUA	201
SSA	CHIHUAHUA	JUAREZ	JUAREZ	262

Cuadro C. Distribución de vacíos en el marco muestral disponible a la fecha

Detalle de vacíos (lo más relevante)	Las 3 entidades		Chihuahua		Oaxaca		Veracruz	
	Total	%	Total	%	Total	%	Total	%
Registros	10582	100.0%	1856	100.0%	3388	100.0%	5338	100.0%
Vacío: nombre (cap. 2da. cascada)	2	0.0%	0	0.0%	0	0.0%	2	0.0%
Vacío: apellido paterno (cap. 2da. cascada)	206	1.9%	0	0.0%	8	0.2%	1	0.0%
Vacío: apellido materno (cap. 2da. cascada)	274	2.6%	131	7.1%	26	0.8%	11	0.2%
Vacío: teléfono (casa) (cap. 2da. cascada)	6703	63.3%	1734	93.4%	916	27.0%	4053	75.9%
Vacío: teléfono (celular) (cap. 2da. cascada)	6147	58.1%	1343	72.4%	959	28.3%	3845	72.0%
Vacío: correo electrónico (cap. 2da. cascada)	5638	53.3%	1353	72.9%	1113	32.9%	3172	59.4%
Vacío: municipio (cap. 2da. cascada)	932	8.8%	509	27.4%	101	3.0%	322	6.0%
Vacío: localidad (cap. 2da. cascada)	1242	11.7%	775	41.8%	30	0.9%	437	8.2%
Vacío: institución (cap. 2da. cascada)	1402	13.2%	574	30.9%	2	0.1%	826	15.5%
Vacío: CLUES (cap. 2da. cascada)	2235	21.1%	940	50.6%	864	25.5%	431	8.1%
Vacío: dirección unidad de salud (cap. 2da. cascada)	4843	45.8%	1176	63.4%	764	22.6%	2903	54.4%
Vacío: teléfonos (casa y celular) (cap. 2da. cascada)	5302	50.1%						

El Cuadro C exhibe un elevado porcentaje de celdas vacías en el marco muestral disponible. En particular para la información de contacto, que es crucial para el diseño de muestreo propuesto a efecto de captar la atención de beneficiarios por capacitados de la 2da cascada. Aquí es importante recordar que este cuadro considera un total de 10,582 registros de los aproximadamente 19,568 capacitados planeados (Cuadro 1). Es decir, el 54% de los capacitados planeados. Y, de ese 54% podemos apreciar que alrededor de un 60% no hay forma de contactarlos o corroborarlos.

Adicionalmente, hay que considerar lo siguiente:

- Es importante reflexionar lo siguiente: ¿Qué características tienen aquellos capacitadores de segunda cascada sin información de contacto? ¿Por qué el capacitador de la 1era cascada no reporta información de los de la 2da cascada? ¿Qué características tienen aquellos que sí tienen forma de contactarse? Si el hecho de poder contactarlos está relacionado con lo que se quiere medir (efectividad en la capacitación y ulterior atención de beneficiarios de interés) entonces se puede incurrir en un sesgo severo a favor de la efectividad del programa de capacitación.
- De las celdas con datos, i.e. aquellas con algo de información, hay que considerar que se pueden tener errores de captura, registros equívocos o registros incompletos, e.g. teléfonos con 3 dígitos o con letras.
- En varias celdas con información se tienen duplicados o información arrastrada/repetida y no es posible determinar cuál es la información correcta.

Un marco muestral con baja calidad en los datos no es útil para muestrear. Seleccionar una muestra de tal marco puede derivar en un ejercicio demoscópico muy costoso e inválido pues será muy difícil captar a la verdadera población objetivo y las conclusiones pueden ser totalmente erróneas.