

ANEXO: METODOLOGÍA MUESTRAL

A continuación se presentará la metodología muestral a ser utilizada en el estudio.

Universo de Estudio y Cobertura geográfica

Se definirá como integrantes del Universo de Estudio a todas las personas de 15 años y más de edad, que habiten en hogares particulares ubicados en todas las comunas de la Región Metropolitana, incluyendo a todos los distritos de la Región, tanto urbanos como rurales.

Marco Muestral

Para efectos de extracción de la muestra, se trabajará con el material digitalizado en medios electromagnéticos y material cartográfico obtenido del último Censo de Población y Viviendas 2002, del Instituto Nacional de Estadísticas.

La información obtenida del Censo que se aprovechará para los efectos del presente estudio, incluye información sobre población y viviendas a nivel de manzanas, para todas las áreas urbanas del país, lo que permitirá la extracción de las muestras mediante una metodología muestral estrictamente probabilística, combinando técnicas de estratificación, conglomeración y extracción de conglomerados (generalmente manzanas) en base a probabilidades proporcionales al tamaño.

Diseño muestral replicado

Conviene detenerse a analizar el esquema de selección probabilística utilizado, el así llamado **submuestreo interpenetrante de Mahalanobis**, el que luego fue modificado y actualizado por W.E. Deming, diseño que es también conocido como “**muestreo replicado**”.

El texto de W. Edwards Deming, *Sample Design in Business Research*¹, está dedicado en su totalidad a este tipo de muestreo probabilístico específico. Más aún, el destaca que desde que conoció de esta técnica, aprendiéndola del destacado estadístico hindú Mahalanobis, no ha usado otro método en su vida profesional².

Este método se caracteriza por replicar un determinado diseño muestral, cualquiera sea su complejidad, en un número determinado de muestras probabilísticas de idéntico diseño, de menor tamaño, y cada una de ellas igualmente representativa del Universo de Estudio, las que luego son combinadas para obtener las inferencias.

¹ Editorial John Wiley & Sons, Inc., 1960.

² Ver nota al pie de página, en el texto ya citado, págs. 186-187.

Ventajas comparativas del diseño Replicado en relación a los diseños tradicionales

La muestra obtenida es científicamente válida, de la cual se pueden obtener no solamente las estimaciones deseadas, sino también estimaciones de los márgenes de error muestral a que están sometidas las estimaciones de los parámetros, a los niveles de confianza que el investigador desee, tomando en consideración todas las complejidades del diseño muestral utilizado.

Una ventaja adicional de este método, es que se cumple con el precepto de jamás efectuar reemplazos³ en una muestra probabilística, evitando por lo tanto, perder las características de una muestra probabilística totalmente válida.

Una nueva ventaja de este método es la de que permite calcular los márgenes de error muestral con algoritmos de fácil elaboración y de alta confiabilidad, en lugar del uso indiscriminado y erróneo de las fórmulas correspondientes al “muestreo aleatorio simple”, que se suele utilizar, erróneamente, a pesar de no ser este el diseño realmente utilizado en el estudio.

Una ventaja adicional del submuestreo interpenetrante es la de poder salir a terreno en forma modular, submuestra tras submuestra, hasta obtener el tamaño muestral deseado. Incluso, si por causas de fuerza mayor se debiera interrumpir el trabajo en terreno, la muestra ya obtenida sería representativa del Universo y podría ser fácilmente procesada e inflactada al Universo, a pesar de su menor tamaño.

Otra ventaja estriba en el hecho de que el número de intentos por conglomerado se mantiene prácticamente homogéneo a través de toda la muestra, vale decir se mantiene casi constante el tamaño de los conglomerados finales, lo que impide un número excesivo de entrevistas por conglomerado, lo que generalmente aumenta el margen de error muestral, especialmente para variables que posean un coeficiente de correlación intraclase elevado.

Otra ventaja de este diseño, es el de proporcionarnos una estimación del coeficiente de correlación intraclase para cualquiera de las variables bajo estudio. Esta estimación nos servirá luego para estimar con mayor propiedad el tamaño muestral necesario para obtener una precisión deseada, en futuros estudios.

Esta estimación del coeficiente de correlación intraclase, nos permite también calcular el valor del “design effect”, termino acuñado por Leslie Kish, lo que puede transformarse en una gran ayuda, casi podríamos decir imprescindible, para la realización de análisis estadísticos futuros. Por último, e igualmente muy importante, debemos mencionar el hecho que el diseño a implementar es dinámico, vale decir sus estimaciones –valores inflactados al universo- no precisan de datos estadísticos secundarios respecto a la población (estimaciones poblacionales), sino que, automáticamente, detectan el crecimiento de la población, u otro movimiento poblacional, desde la fecha del último censo hasta la fecha de la encuesta.

³ Texto de W. E. Deming, pág. 24.

Diseño muestral

Tal como se ha destacado anteriormente, el diseño muestral a ser implementado en este estudio es estrictamente probabilístico, único método conocido que permite efectuar inferencias al Universo, e interpretar los resultados estadísticamente, calculando además, *a priori*, la precisión a obtener, y, *a posteriori*, la precisión realmente obtenida. Además, permite efectuar los tests estadísticos necesarios para determinar si las diferencias encontradas entre los diversos parámetros estimados, son estadísticamente significativas. Por lo tanto, cualquier intento de selección en base a cuotas queda totalmente excluido en cualquiera de las etapas de selección.

El diseño contempla la estratificación a lo largo de la región, estratificando el marco muestral con un criterio geográfico.

Se contempla un total de 36 estratos geográficos o Zonas Finas, y la extracción de un total de 18 sub-muestras interpenetrantes, cada una de ellas representativa del Universo de Estudio. Esto proporcionará un total de 648 conglomerados (manzanas), con un total inicial esperado de al menos 3888 hogares a seleccionar en la muestra, incluyendo en este total las submuestras de reserva que se podrían requerir para suplir las no-respuestas provenientes de diversas causas, entre ellas rechazos y no habidos (hogares y/o personas seleccionadas).

La muestra, tal como se dijo, será estratificada geográficamente, con afijación proporcional al tamaño de los estratos, multi-etápica, de áreas (manzanas, hogares y finalmente una persona perteneciente al Universo de Estudio), con selección PPS a nivel de las Unidades Muestrales de Primera etapa, y probabilidades recíprocas al tamaño estimado de los conglomerados en la segunda etapa, obteniéndose finalmente una muestra autoponderada a nivel de viviendas.

Este diseño, como todo diseño probabilístico estricto, no contempla reemplazos de ninguna especie en ninguna de las etapas de selección de la muestra, evitándose con ello los peligros de sesgo de las muestras por esta causa.

Cabe hacer notar que la afijación estratificada proporcional minimiza los márgenes de error muestral para estimaciones de parámetros a nivel total del Universo.

Reemplazos

Este diseño, como todo diseño probabilístico estricto, no contempla reemplazos de ninguna especie en ninguna de las etapas de selección de la muestra, evitándose con ello los peligros de sesgo por esta causa. Para recuperar los casos de no-respuestas se utilizarán submuestras no utilizadas aun en el transcurso del trabajo en terreno, lo que facilitará la recuperación de la mayor parte de la precisión perdida debido a las “no-respuestas” (al permitir recuperar el tamaño muestral original deseado), evitando el peligro de sesgos en las estimaciones, que la mayoría de los diseños muestrales que contemplan reemplazos, introducen.

Ejemplos del peligro de introducción de sesgos es el que los reemplazantes son personas que están en el hogar cuando el seleccionado probabilístico no lo está, u hogares en la muestra original en los que no hay nadie, tal vez por ser un hogar de pocas personas,

siendo reemplazados por hogares en que si hay alguien en esa misma hora, etc. Lo más probable en estos casos es la introducción de sesgos en los resultados.

Cabe señalar que el método de recuperación de casos de no-respuesta a utilizar en este estudio, tampoco asegura una representatividad del 100% del Universo de Estudio original, sino que sólo del sub-universo de **colaboradores (unidades muestrales originales que responden)** de la Cobertura Completa Idéntica, definida a continuación:

“COBERTURA COMPLETA IDÉNTICA”:

El concepto de **Cobertura Completa Idéntica** es fundamental para comprender el significado del muestreo probabilístico.

La Cobertura Completa Idéntica consiste en el resultado que se **habría obtenido** del examen del **total** de las unidades muestrales del Marco Muestral (conglomerados, viviendas, personas), examen efectuado por los mismos encuestadores que efectuarían el trabajo en la **“encuesta por muestreo”**, utilizando las mismas definiciones, cuestionarios y procedimientos, ejerciendo el mismo cuidado que el ejercitado en la encuesta muestral, y utilizando el mismo personal que trabajaría en la codificación y digitación de la encuesta muestral, y en el mismo período.

En el caso de cualquier estudio, la Cobertura Completa Idéntica puede considerarse dividida en 2 subconjuntos: un subconjunto que sí aporta respuestas, y otro subconjunto que no aporta respuestas. Es al primer conjunto solamente que el estudio puede aspirar a representar, y no al Universo de Estudio completo, objetivo inicial del estudio.

Selección de las personas a entrevistar

La selección de las personas a entrevistar se obtendrá a través de un proceso multi-etápico, a saber:

1era. etapa: Selección de manzanas con PPS (Probabilities Proportional to Size)

2.a etapa : Selección de hogares dentro de cada manzana seleccionada, con partida aleatoria y paso sistemático utilizando el empadronamiento previo de la manzana y el tamaño adjudicado a dicha manzana (número de viviendas particulares), en el Censo del 2002. Este número de viviendas se traducirá en una etapa intermedia del diseño muestral, en unidades muestrales de 6 viviendas cada una (por ejemplo, a una manzana de 18 viviendas se le adjudicará un tamaño de 3 U.M. (Unidades Muestrales). El paso sistemático que se aplicará en cada conglomerado corresponde al tamaño adjudicado a cada manzana previo a la primera etapa de selección, basándonos en la información del Censo del 2002 del INE, medido en términos de UM.

3a etapa: Selección aleatoria de una persona al interior de cada hogar de la muestra, de entre las personas pertenecientes por definición al Universo de Estudio que habitan en cada hogar. Esta selección se efectuará utilizando una versión ampliada y modificada por G. Davidovics, de la tabla de Kish.

Estimaciones de parámetros

Para evitar sesgos matemáticos, un diseño probabilístico estricto utiliza estimadores debidamente ponderados por los recíprocos de sus probabilidades de selección, al contrario de estudios no probabilísticos o pseudo-probabilísticos en los que simplemente no se pondera, presentándose las tablas con los resultados como si la muestra fuera autoponderada, o, cuando sí se pondera, los ponderadores provienen de estadísticas secundarias no actualizadas y generalmente no correspondientes a las características geográficas (ubicación) o socio-demográficas, de cada respondente.

En este estudio, las estimaciones de valores absolutos, inflactados al Universo de Estudio, serán producto de la inflatación por los recíprocos de las probabilidades finales (overall sampling fractions) de selección de cada persona entrevistada. Además se utilizarán factores de corrección de las tablas aleatorias las que corregirán cualquier desviación de la distribución **empírica** en el uso de las tablas, en relación con la distribución **teórica** de las tablas (debido a tasas diferenciales de “no-logro” para las distintas tablas aleatorias (14 tablas)). Además, finalmente, se utilizará la técnica de post-estratificación respecto a la variable “tramos etarios” utilizando los datos de la distribución etaria del último Censo de Población y Vivienda. Ésta será la única variable que se considera conveniente para la utilización de la post-estratificación.

Probabilidad de selección en primera etapa de selección:

$$P_{1j} = \frac{X_{i,j}}{\sum X_{i,j}} = \frac{X_{i,j}}{X_j}$$

Donde:

$X_{i,j}$ = Tamaño del conglomerado “i” en la Zona Fina “j”, medido en términos de unidades muestrales (definidos como segmentos no compactos de 6 viviendas cada una).

X_j = Total de unidades muestrales en la Zona Fina “j” (estrato secundario) de la Región.

Probabilidad de selección en segunda etapa de selección:

$$P_{2j} = \frac{1}{X_{i,j}}$$

El valor 1 corresponde a una unidad muestral.

Fracción muestral con ambas etapas combinadas

$$P_j = P_{1j} * P_{2j} = \frac{X_{i,j}}{X_j} * \frac{1}{X_{i,j}} = \frac{1}{X_j}$$

Puesto que todas las Zonas Finas tienen igual tamaño, vale decir **X j es una constante**, tenemos que el inflator (recíproco de la Fracción muestral) hasta llegar al nivel de viviendas, es una **constante**.

En consecuencia, utilizando este inflator, la muestra a nivel de viviendas, será autoponderada. **Para llegar al “inflator muestral final” a nivel de personas, se deberá multiplicar cada entrevista lograda, además del inflator recién explicitado, por el recíproco de la probabilidad de selección de una persona de cada hogar.** En otras palabras, se deberá multiplicar por el número de personas de cada hogar (personas que pertenecen, por definición, al Universo de Estudio).

Obviamente, como estos valores difieren de un hogar a otro, la muestra final de personas no será autoponderada, y, en consecuencia, los resultados de la encuesta antes de ser tabulados, deberán ser inflactados para llevarlos al nivel del Universo, para sólo después poder obtener los resultados deseados. Esto en adición a otros factores de corrección que se explicarán a continuación.

Factores correctores adicionales

En cada conglomerado: Éste es un corrector consistente en la relación matemática del número de hogares seleccionados en el conglomerado, con personas que habiten en ellos (respondientes + no-respondientes), dividido por el número de viviendas respondientes.

En cada Zona Fina: Éste es un corrector consistente en la relación matemática del número de conglomerados seleccionados en cada Zona Fina, dividido por el número de conglomerados respondientes (es decir, en los que se logró al menos una vivienda con respuesta (excepto en los casos en los que en la muestra seleccionada en dicho conglomerado no exista ni una vivienda en la que habiten personas pertenecientes al Universo de Estudio).

Post-Estratificación: Éste corrector ya se explicó en un párrafo anterior.

Es necesario añadir aquí que estas correcciones se aplicarán sólo en caso que se acepte el supuesto de que la muestra resultante después de considerar los casos de no-respuesta, se pueda considerar como un subconjunto aleatorio obtenido de la muestra original.

Tamaños muestrales y precisión de la información

El tamaño muestral total propuesto es de 2000 entrevistas finales, con lo que se podrían lograr los márgenes de error muestral que se presentan a continuación, para variables dicotómicas y/o politómicas.

Suponiendo un coeficiente de correlación intraclase promedio de 0.15 para un conjunto de variables, y un número de entrevistas promedio por manzana de 3.85 el “efecto del diseño” será de:

$$deff = [1 + (\bar{n} - 1) * \rho]$$

$$(1 + (3.85 - 1) * 0.15) = \mathbf{1.4275}$$

La fórmula a utilizar para calcular el error standard es la siguiente:

$$\sqrt{\sum_h^L W_h^2 * \frac{p_h * q_h}{n_h} * [1 + (\bar{n} - 1) * \rho]} = \sqrt{\sum_h^L W_h^2 * \frac{p_h * q_h}{n_h} * deff}$$

Donde:

- L = Numero de estratos (o Zonas Finas).
- W_h = Proporción del total del Universo, en el estrato “h”.
- n_h = Tamaño muestral en el estrato “h”.
- n = Promedio de entrevistas efectivas por manzana (u otro conglomerado).
- ρ = Coeficiente de correlación intraclase.
- p_h = Proporción cuyo error standard se desea estimar (proporción que responde en una determinada categoría a una determinada pregunta)
- q_h = 1 – p_h.

Reemplazando los símbolos de la fórmula por los valores que corresponden en este estudio, tenemos:

- L = 36
- W_h = N_h / N = Z/36Z = 1/36 (pues en el diseño replicado las zonas finas Z son de igual tamaño)
- n_h = 53.9
- n = 3.85
- ρ = 0.15
- p_h = 0.50 (para trabajar con varianza máxima y así estimar error standard máximo)

Con estos valores, la fórmula precedente nos proporciona una estimación del error standard de 0.01356.

Con este valor, tenemos que, para un nivel de confianza del 95 %, el error muestral máximo sería de:

$$\} (1.96 * 0.01356) = \} 0.02658 \approx \} 2.66 \%$$

Si se deseara un mayor nivel de confianza, digamos del 99,73%, “z” es 3 y por lo tanto el error muestral máximo sería de:

$$\} (3 * 0.01356) = \} 0.04068 = \} 4.1 \%$$

Recordemos que estos valores se calcularon para varianza máxima, por lo cual, para cualquier valor de “P” distinto a 0.50 (vale decir distinto a 50%), o valores de “p_h” distintos

a 0.50 en los distintos estratos – aunque promediando 0.50 para el gran total- los márgenes de error muestral serán menores.

Obviamente, los márgenes de error muestral para subconjuntos de la muestra, serán función del tamaño muestral respectivo y mayores que el error muestral para el total del universo.

A continuación, y para mayor información, se presentan los márgenes de error muestral para el **total del Universo de Estudio**, en puntos porcentuales, a un nivel de confianza del 95%, vale decir existiendo una probabilidad de 95% que el valor real del porcentaje que se está tratando de medir, se encuentre al interior del intervalo de confianza formado por la estimación de punto más/menos el error muestral correspondiente.

**Porcentaje estimado por la muestra (estimación de punto)
Margen de error muestral al nivel de confianza del 95%**

5%	} 1.2
10%	} 1.6
15%	} 1.9
20%	} 2.1
30%	} 2.4
40%	} 2.6
50%	} 2.7
60%	} 2.6
70%	} 2.4
80%	} 2.1
85%	} 1.9
90%	} 1.6
95%	} 1.2

Ejemplo del uso de la tabla precedente

1.- En el caso hipotético de obtener a una determinada pregunta un 70% para una de las alternativas de respuesta a dicha pregunta, interesa conocer su intervalo de confianza.

2.- Se resta y se suma, simultáneamente, el margen de error muestral a esta estimación, y se obtienen los límites inferior y superior del intervalo de confianza. En este caso, estos límites serían 67.6% para el límite inferior ($70 - 2.4$) y 72.4% para el límite superior ($70 + 2.4$).

3.- Esto significa que existe una probabilidad del 95% (pues los márgenes están calculados para dicha probabilidad) de que el valor verdadero, que es el que se obtendría si el estudio se hubiese aplicado a todo el Universo de Estudio en lugar de sólo a una muestra extraída de él, estaría ubicado en algún lugar al interior del intervalo 67.6% -- 72.4%.

Para Dominios de Estudio distintos al total del Universo de Estudio, se pueden obtener los márgenes de error muestral multiplicando los márgenes de error muestral presentados en la tabla precedente por el factor,

$$\sqrt{\frac{200}{n_1}}$$

Donde n_1 es el tamaño muestral del Dominio de Estudio respecto al cual se desea conocer los márgenes de error muestral.

Consideraciones adicionales a la precisión de la información

Como fuera señalado en otros acápite, el tamaño muestral final estimado ha sido tomando principalmente en consideración, el margen de precisión y nivel de confianza deseados, para las más importantes variables bajo estudio.

A este respecto cabe agregar que en una investigación por muestreo, lo que se debe minimizar es el “error total” y no solamente el “error muestral”. El “error total” lo definiremos como la raíz cuadrada del “error cuadrático medio” el cual está compuesto por la suma de la varianza del estimador y el cuadrado del sesgo:

$$E.T. = \sqrt{\sigma_p^2 + b^2}$$

donde la varianza del estimador no es otra cosa que el cuadrado del error standard de “p”, y “b” simboliza el sesgo.

La precisión de la información entregada por la encuesta, se refería en los acápite anteriores solamente al error muestral, pero debemos estar conscientes también de la existencia de errores no muestrales (aleatorios y no aleatorios). Los primeros, es decir los aleatorios, ya están considerados en el error muestral (aunque también se deberían minimizar), de modo que ahora debemos concentrarnos en aquellos errores que no se cancelan, es decir en los sesgos.

En los sesgos no-matemáticos se incurre no solamente en los estudios muestrales, sino también en los censales, y son independientes del tamaño muestral, no disminuyendo porque el tamaño muestral aumente.

Por lo tanto, es muy importante invertir todo el esfuerzo posible en minimizar esta fuente de error (buena selección de encuestadores, buen entrenamiento, evitar sesgos en las preguntas, tanto en el cuestionario como en la forma de efectuar las preguntas, buen diseño del cuestionario, pre-tests, etc.)

A continuación detallaremos algunas de las medidas que se tomarán para minimizar los errores no muestrales, sean estos aleatorios o consistentes (sesgos).

(a) Se utilizarán cuestionarios estructurados, los que se pre-testearan en 30 entrevistas previo a la salida definitiva a terreno, para evitar errores estructurales en su diseño o redacción.

(b) Las entrevistas serán personales (*face to face*), en los hogares de las personas entrevistadas, y llevadas a cabo por un equipo de encuestadores entrenados. **El informante en cada caso, será la misma persona seleccionada aleatoriamente.**

(c) Se realizarán hasta 4 intentos para realizar cada entrevista – intento inicial y hasta 3 visitas en caso necesario, hasta ubicar a la persona seleccionada en el hogar seleccionado.

(d) Después de 4 intentos infructuosos en un hogar habitado, se le considerará como “no habido”.

(e) Estos casos no serán reemplazados, para mantener intacto el carácter probabilístico de la muestra.

(f) Los hogares deshabitados (que fueron seleccionados aleatoriamente), no serán reemplazados, por la misma razón anterior.

(g) Se controlarán, mediante visitas efectuadas por supervisores, un porcentaje determinado (de hasta 30%) del trabajo de cada encuestador(a). También se efectuará una supervisión telefónica exhaustiva del 100% de los hogares con teléfono.

(h) El 100% de los cuestionarios se someterá a un “editing” manual para verificar la consistencia interna de las respuestas, y, en caso de dudas, se tomarán las medidas correctivas que correspondan.

(i) Para evitar sesgos matemáticos, un diseño probabilístico estricto utiliza estimadores debidamente ponderados por los recíprocos de sus probabilidades de selección, al contrario de estudios no probabilísticos o pseudo-probabilísticos en los que simplemente no se pondera, presentándose las tablas con los resultados como si la muestra fuera autoponderada, o, cuando si se pondera, los ponderadores provienen de estadísticas secundarias no actualizadas y generalmente no correspondientes a las características geográficas (ubicación) o sociodemográficas, de cada respondente.

En este estudio, las estimaciones de valores absolutos, inflactados al Universo de Estudio, serán permitidos puesto que en la base de datos que se entregará al cliente, figurará para cada entrevistado(a) el factor de expansión correspondiente, el que corresponderá al producto de las diversas fracciones de muestreo en cada etapa.