Estudio Nacional de la Discapacidad 2015

Ministerio de Desarrollo Social
Gobierno de Chile

METODOLOGÍA DE DISEÑO MUESTRAL DE LA ENCUESTA DEL SEGUNDO ESTUDIO NACIONAL DE LA DISCAPACIDAD, 2015

Departamento de Estudios, Servicio Nacional de la Discapacidad (Senadis)
División Observatorio Social, Subsecretaría de Evaluación Social
Ministerio de Desarrollo Social

26 de Enero de 2016

PRESENTACION

El Segundo Estudio Nacional de la Discapacidad da cumplimiento a un compromiso de Gobierno de la Presidenta Michelle Bachelet, de entregar un diagnóstico detallado de cómo viven los chilenos y chilenas en situación de discapacidad y, a partir de este diagnóstico, avanzar en el diseño de políticas públicas pertinentes y oportunas en materia de inclusión social para las personas en situación de discapacidad del país.

El objetivo general de este estudio es determinar la prevalencia y caracterizar la discapacidad a nivel nacional, identificando las principales brechas de acceso de las personas en situación de discapacidad en Chile y a partir de ello evaluar los resultados de la aplicación de la normativa nacional e internacional y de las políticas, planes y programas existentes en la materia.

Para esto, incorpora un nuevo enfoque para la medición y caracterización de las personas en situación de discapacidad, que toma en cuenta:

- el modelo teórico y conceptual de la Clasificación Internacional de Funcionamiento, de la Discapacidad y de la Salud (CIF) (2001);
- el conjunto de ámbitos relevantes establecidos por la Convención de Derechos de las Personas con Discapacidad (2006 y ratificada en 2008).

Mediante un convenio de transferencia[1], el Ministerio de Desarrollo Social (MDS), a través de su Observatorio Social, y bajo la asesoría técnica permanente del Servicio Nacional de la Discapacidad (Senadis), encomendó al Instituto Nacional de Estadísticas (INE) el diseño muestral, el levantamiento de la prueba de campo, el levantamiento de la encuesta del II Estudio Nacional de la Discapacidad en Chile, y la construcción de la correspondiente base de datos. El presente documento de Diseño Muestral de la encuesta del II Estudio Nacional de la Discapacidad, 2015 (en adelante Endisc II), se basa en el Informe de Diseño Muestral (versión preliminar, de enero 2016) elaborado por el INE, en el marco del citado convenio.

Se informa la estrategia de muestreo desarrollada por el INE conforme a los objetivos del estudio. Mediante esta estrategia, se elaboraron y desarrollaron los términos específicos del diseño muestral, y se analizó el método óptimo de selección de las unidades muestrales, considerando tanto los objetivos del estudio, como el marco presupuestario y el plazos disponible. En particular, se informa cómo se buscó el tamaño de muestra mínimo óptimo a partir de los requerimientos técnicos expresados por el MDS, y cómo se distribuyeron luego las unidades muestrales en las distintas áreas geográficas requeridas, en forma proporcional al tamaño calculado.

Un aspecto a destacar es que la encuesta de Endisc II considera un diseño muestral basado en un método bifásico con niveles de representación estadístico y/o márgenes de error considerados razonables a nivel regional, nacional y por áreas geográficas urbana y rural. Entre las ventajas de este tipo de diseño se cuentan sus menores costos de tiempo y monetarios que la alternativa de una muestra independiente multi-etápica, pues el primer diseño ofrece un marco más actualizado de direcciones de viviendas particulares, mientras el segundo requiere enumerar las unidades de primera etapa (manzanas y secciones), previo a la selección de las viviendas particulares donde realizar entrevistas. Sin embargo, en el diseño bifásico, la selección aleatoria de nuevas unidades, introduce mayor variabilidad, aumentando sus factores de expansión.

[1] Aprobado mediante Decreto Supremo N° 39, fechado 12 de diciembre 2014, del Ministerio de Desarrollo Social.

CONTENIDO

I. SIMULACIONES DE TAMAÑO MUESTRAL

El Ministerio de Desarrollo Social encomendó al INE la misión de desarrollar la estrategia de muestreo de la encuesta del II Estudio Nacional de la Discapacidad, la cual consiste en elaborar y desarrollar los términos específicos del diseño muestral, analizando el método óptimo de selección de las unidades muestrales, acorde a los objetivos, presupuesto y plazos disponibles para el estudio.

De esta forma, el objetivo fue encontrar un tamaño de muestra mínimo óptimo en base a los requerimientos del Ministerio de Desarrollo Social (Observatorio Social, con asesoría técnica permanente de Senadis), para luego distribuirlo por las distintas áreas geográficas requeridas, en forma proporcional al tamaño calculado.

En una primera etapa de la investigación, se analizaron diferentes escenarios. El primero, consideró como variable de interés la estimación de la proporción de personas que presentaron al menos una condición de discapacidad de la encuesta Casen 2011. Este escenario permitió calcular y ajustar los primeros parámetros para dar una visión general de un tamaño muestral óptimo. Posteriormente, atendido que el cuestionario de Casen 2011 no permitía medir adecuadamente el fenómeno de la discapacidad nacional, conforme los estándares actuales, se descartó usar sus estimaciones y niveles de prevalencia para determinar el diseño muestral óptimo para el presente estudio.

Alternativamente, se analizó un segundo escenario, basado las estimaciones provenientes del I Estudio Nacional de la Discapacidad, 2004 (en adelante Endisc I), que era hasta entonces el único estudio de medición exclusiva, oficial, de la discapacidad a nivel nacional. Así, la variable de interés utilizada para esta segunda simulación de tamaño muestral, es la proporción de "personas en condición de discapacidad" estimada usando la base de datos de Endisc I.

De esta forma, las simulaciones realizadas en definitiva para el presente estudio se basan en Endisc I y consideran una estrategia de muestreo bifásico, que para efectos de selección de la muestra, y para evitar los costos en tiempo y presupuestarios asociados a enumeración de la muestra, utilizan una submuestra del marco de viviendas con encuestas a hogares logradas en la encuesta de Caracterización Socioeconómica Nacional (Casen) 2013. Con ello, se establece que la población objetivo son todas las personas residentes en las viviendas particulares de dicho marco muestral.

El presente documento se estructura como sigue. En la Sección 1, Objetivos, se informan los objetivos generales y específicos de las simulaciones de tamaño muestral para el II Estudio Nacional de la Discapacidad. En la Sección 2, Tamaño muestral con muestras complejas, se detallan consideraciones estadístico-conceptuales generales para el cálculo de los distintos estimadores empleados en las simulaciones de tamaño muestral de muestras complejas. En la Sección 3, Presentación de simulaciones, se describen los procedimientos de cálculo y criterios estadísticos utilizados para la obtención del escenario de tamaño muestral implementado.

1. Objetivos

Encontrar un tamaño muestral que reúna un conjunto de características que puedan representar de manera adecuada a la población chilena respecto de sus componentes sociodemográficos en materia de discapacidad, de modo que sea estadísticamente representativa a nivel regional, nacional y por áreas geográficas urbana y rural.

Atendidas las decisiones adoptadas respecto de los cuestionarios aplicados a nivel hogar, adulto y niño/a (este último respondido por adulto responsable), estos niveles de representación aplican para la muestra de la población adulta (de 18 años o más), en tanto para la población infantil (de 2 a 17 años), que es entrevistada sólo en hogares con niños/as, los niveles de representación son nacional y urbano/rural (dado su tamaño muestral más reducido).

Objetivo de la investigación para el diseño muestral de la encuesta: Estudiar y calcular un tamaño de muestra mínimo óptimo en base a niveles de estimación, errores de muestreo razonables, variables de diseño y cobertura geográfica, plazos y presupuesto disponible.

Niveles de estimación: Regional, Nacional, Nacional urbano/rural (población adulta); Nacional, Nacional urbano/rural (población infantil).

Variable de diseño: Proporción de personas en condición de discapacidad.

Parámetros a estimar: Tasa de discapacidad (proporción de personas en situación de discapacidad).

Márgenes de error razonables a nivel regional: Error absoluto menor al 5 puntos porcentuales, y error relativo menor al 30%-

Márgen de error razonable a nivel nacional: Error absoluto 0,4 puntos porcentuales, y error relativo menor al 3,5%.

Marco muestral para selección y levantamiento: Listado de viviendas con encuestas a hogares logradas en el levantamiento de Casen 2013.

Otras consideraciones: Los cálculos muestrales propuestos por el INE cumplen los requerimientos de acuerdo a los criterios estadísticos establecidos por el Ministerio de Desarrollo Social. Sin embargo, dado el marco presupuestario disponible, el INE propuso disminuir la cobertura geográfica, obteniendo un total de 135 comunas a levantar, en contraste de las 324 comunas incluidas en Casen 2013. De esta forma, el tamaño muestral escogido considera aquellas comunas que al acumular los tamaños totales de viviendas comunales de forma descendente dentro de la región, alcanzan como mínimo el 80% acumulado del total de viviendas a nivel regional. Con esta distribución se asegura una amplia cobertura comunal para cada región.

Adicionalmente, se hicieron ajustes de redistribución de la muestra a nivel regional y comunal para afinar criterios de proporcionalidad, y conjuntamente, para la representatividad nacional urbano-rural, se asegura una cobertura geográfica mínima del total de viviendas a nivel nacional del 80% a nivel nacional urbano y 51% a nivel nacional rural.

2. Tamaño muestral con muestras complejas

Los factores para un adecuado cumplimiento de los objetivos de una encuesta, en términos de diseño muestral, deben incluir: la variable de interés, el estimador asociado a esa variable, los niveles de estimación, los errores de muestreo razonables y disponer de una fuente de información para obtener las estimaciones de interés (censo o encuestas anteriores).

A partir de ello, se realiza un proceso iterativo donde se van ajustando los parámetros hasta obtener un tamaño muestral razonable. Este trabajo considera distintas estrategias de muestreo, escenarios a probar, como también variables de diseño y parámetros a estimar. Para simplificar la metodología se describen los elementos necesarios para el cálculo del tamaño muestral, desde sus generalidades hasta los casos particulares realizados. Posteriormente se describe y presentan los resultados obtenidos para cada uno de los escenarios estudiados.

2.1. Consideraciones Generales

En la actualidad los diseños muestrales complejos asociados a encuestas de hogares tienen múltiples características, una de ellas es que la información disponible del marco muestral está sujeta a la configuración geográfica del país, hecho relevante dado que la población objetivo se encuentra en una ubicación geográfica establecida. Para cada ubicación geográfica se pueden observar comportamientos distintos en los fenómenos de estudio y la configuración de las unidades de primera y segunda etapa incorporan mayor complejidad al diseño muestral. Esto impacta directamente en la construcción de los factores de expansión para los distintos dominios de estudio[2], ya sea en su estimación y la de sus correspondientes errores de muestreo, por lo tanto los parámetros iniciales deben considerar los elementos antes mencionados. Para representar estas características se requiere un tamaño muestral que reúna la mayor parte de esta realidad.

Se desea estimar un parámetro desconocido (en este estudio, la tasa de discapacidad) que representa cierta población objetivo, de modo que esta estimación se acerque lo más posible al parámetro poblacional. La distancia entre el parámetro poblacional y su estimación[3] puede ser usada como una medida que cuantifica cuán lejos o cerca estamos del verdadero valor. Esto se puede expresar en términos probabilísticos y tomar los márgenes de error de la estimación, como el error absoluto y relativo con cierto nivel de confianza. En esta sección se presentan algunos elementos que inician el cálculo del tamaño muestral, utilizados para encontrar un tamaño mínimo necesario para algún dominio de estudio.

3. Simulaciones

A continuación se presenta la última alternativa de tamaño muestral estudiada. Esta simulación recogió la experiencia de las versiones simuladas previamente y los consensos alcanzados, bajo una mirada técnica y presupuestaria, con el Ministerio de Desarrollo Social, a través del Observatorio Social y el Servicio Nacional de la Discapacidad.

Para esta simulación la variable de interés utilizada para el diseño muestral es la tasa de discapacidad estimada utilizando la base de datos de Endisc I (P_i). Además, considera el criterio de

[2] Se llamarán dominios de estudio a los niveles de desagregación, generalmente geográficos, para los cuales se requieren estimaciones confiables y precisas.

[3] Esta distancia o diferencia absoluta es denominada error de estimación o error de muestreo o margen de error.

cobertura total regional de 80%. La propuesta resultante de tamaños muestrales objetivo a nivel nacional, regional, y según zona se presentan en la Tabla 1. Se observa que esta simulación de tamaño muestral considera una muestra objetivo (M2) que contiene un total de 12.196 viviendas (urbano y rural), distribuida en 135 comunas del país, y con niveles de representatividad estadística Regional, Nacional y Nacional urbano-rural.

En particular este escenario persigue lograr representatividad a nivel nacional-regional, usando la variable de interés tasa de discapacidad, estableciendo errores absolutos y relativos inferiores o iguales, a nivel regional, al 5 puntos porcentuales y 30%, respectivamente, y para un error absoluto de 0,4 puntos porcentuales a nivel nacional y un error relativo a nivel nacional no supere el 3,5% , y resguardando que la distribución de la muestra por las distintas comunas requeridas, en forma proporcional a lo establecido en el marco muestral de viviendas con encuestas a hogares logradas en Casen 2013, para así salvaguardar que exista un porcentaje apropiado de viviendas de reemplazo, para aquellos casos en que se requiera reemplazar viviendas dentro de la comuna.

Cabe señalar que la estrategia de levantamiento de la muestra, sigue la modalidad de reemplazo de viviendas, por lo tanto, los cálculos que se presentan no considera la estimación de sobremuestreo. Además, la distribución de la muestra que se presenta en la Tabla N° 2 es una estimación aproximada realizada por INE, que podía variar levemente desde 2% a 5% al momento de generar la selección definitiva para el levantamiento, pues era necesario realizar un ajuste por tamaño mínimo sobre todo a nivel rural.

Tabla 1: Simulación tamaño muestral propuesto a nivel regional y por áreas urbano/rural

Región	$P_{i(2004)}$	Tamaño muestral 2004	Total viviendas marco muestral	Error absoluto propuesto	Error relativo propuesto	Muestra objetivo propuesta
Nivel País Total	**0,129**	**13.767**	**4.441.666**	**0,4**	**3,4**	**12.196**
Nivel País Urbano	**0,125**	**11.435**	**3.919.961**	**0,4**	**3,7**	**10.194**
Nivel País Rural	**0,156**	**2.332**	**521.705**	**1,3**	**8,0**	**2.002**
I de Tarapacá	0,124	228	73.430	2,6	21,1	241
II de Antofagasta	0,113	588	134.676	1,9	16,5	488
III de Atacama	0,134	595	71.377	3,7	27,3	291
IV de Coquimbo	0,133	705	186.606	1,9	14,5	546
V de Valparaíso	0,085	1.216	531.339	1,0	11,3	1.443
VI de O'Higgins	0,149	670	191.955	2,2	14,9	575
VII del Maule	0,177	791	257.056	1,9	11,0	718
VIII del Biobío	0,151	1.731	526.671	1,2	8,1	1.639
IX de la Araucanía	0,176	749	241.585	1,8	10,4	692
X de Los Lagos	0,130	689	190.036	2,2	16,9	518
XI de Aysén	0,134	546	26.510	2,6	19,4	276
XII de Magallanes	0,068	418	45.198	2,0	28,7	256
XIII Metropolitana	0,116	4.182	1.814.485	0,7	6,2	3.789
XIV de Los Ríos	0,175	298	94.621	2,9	16,4	415
XV de Arica y Parinacota	0,175	361	56.121	2,5	14,4	309

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Tabla 2: Distribución de la muestra objetivo por área urbano/rural, según región y comuna

Región	Comuna	Viviendas urbanas	Viviendas rurales	Total viviendas
Total nacional		**10.194**	**2.002**	**12.196**
I de Tarapacá	1101 Iquique	147	11	158
	1107 Alto Hospicio	83	-	83
	Total regional	**230**	**11**	**241**
II de Antofagasta	2101 Antofagasta	357	-	357
	2201 Calama	115	16	131
	Total regional	**472**	**16**	**488**
III de Atacama	3101 Copiapó	184	14	198
	3201 Chañaral	11	10	21
	3301 Vallenar	59	13	72
	Total regional	**254**	**37**	**291**
IV de Coquimbo	4101 La Serena	162	21	183
	4102 Coquimbo	138	25	163
	4201 Illapel	20	25	45
	4203 Los Vilos	30	14	44
	4301 Ovalle	65	46	111
	Total regional	**415**	**131**	**546**
V de Valparaíso	5101 Valparaíso	269	14	283
	5103 Concón	24	-	24
	5107 Quintero	18	15	33
	5109 Viña del Mar	270	-	270
	5301 Los Andes	72	11	83
	5501 Quillota	65	15	80
	5502 Calera	44	12	56
	5601 San Antonio	79	13	92
	5603 Cartagena	14	11	25
	5604 El Quisco	11	-	11
	5701 San Felipe	41	12	53
	5801 Quilpué	160	14	174
	5802 Limache	39	90	129
	5804 Villa Alemana	116	14	130
	Total regional	**1.222**	**221**	**1.443**
VI de O'Higgins	6101 Rancagua	181	7	188
	6104 Coltauco	8	8	16
	6106 Graneros	32	7	39
	6107 Las Cabras	8	13	21
	6108 Machalí	19	7	26
	6111 Olivar	8	7	15
	6112 Peumo	12	6	18
	6113 Pichidegua	7	7	14
	6115 Rengo	35	14	49
	6116 Requínoa	8	13	21
	6117 San Vicente	14	32	46
	6301 San Fernando	40	22	62
	6303 Chimbarongo	11	14	25
	6310 Santa Cruz	16	19	35
	Total regional	**399**	**176**	**575**

Continúa ►

Tabla 2: Distribución de la muestra objetivo por área urbano/rural, según región y comuna

Región	Comuna	Viviendas urbanas	Viviendas rurales	Total viviendas
VII del Maule	7101 Talca	124	13	137
	7102 Constitución	29	12	41
	7109 San Clemente	12	33	45
	7201 Cauquenes	20	17	37
	7301 Curicó	208	27	235
	7308 Teno	8	24	32
	7401 Linares	46	25	71
	7403 Longaví	6	24	30
	7404 Parral	16	17	33
	7406 San Javier	16	23	39
	7407 Villa Alegre	8	10	18
	Total regional	**493**	**225**	**718**
VIII del Biobío	8101 Concepción	169	39	208
	8102 Coronel	99	24	123
	8103 Chiguayante	105	-	105
	8105 Hualqui	11	15	26
	8106 Lota	56	-	56
	8107 Penco	41	13	54
	8108 San Pedro de la Paz	64	-	64
	8110 Talcahuano	142	7	149
	8111 Tomé	41	15	56
	8112 Hualpén	88	-	88
	8202 Arauco	35	14	49
	8203 Cañete	21	23	44
	8205 Curanilahue	42	19	61
	8301 Los Ángeles	118	110	228
	8304 Laja	11	13	24
	8305 Mulchén	24	13	37
	8306 Nacimiento	22	9	31
	8401 Chillán	150	34	184
	8406 Chillán Viejo	18	34	52
	Total regional	**1.257**	**382**	**1.639**

Continúa ►

Tabla 2: Distribución de la muestra objetivo por área urbano/rural, según región y comuna

Región	Comuna	Viviendas urbanas	Viviendas rurales	Total viviendas
IX de la Araucanía	9101 Temuco	194	26	220
	9102 Carahue	8	18	26
	9103 Cunco	3	13	16
	9105 Freire	5	28	33
	9108 Lautaro	17	12	29
	9111 Nueva Imperial	10	18	28
	9112 Padre Las Casas	38	34	72
	9114 Pitrufquén	14	18	32
	9115 Pucón	11	17	28
	9120 Villarrica	33	22	55
	9201 Angol	43	10	53
	9202 Collipulli	8	9	17
	9203 Curacautín	13	8	21
	9210 Traiguén	16	9	25
	9211 Victoria	21	16	37
	Total regional	**434**	**258**	**692**
X de Los Lagos	10101 Puerto Montt	108	43	151
	10102 Calbuco	8	33	41
	10105 Frutillar	9	9	18
	10106 Los Muermos	8	12	20
	10107 Llanquihue	8	10	18
	10201 Castro	22	18	40
	10202 Ancud	19	19	38
	10208 Quellón	17	15	32
	10301 Osorno	124	18	142
	10305 Río Negro	7	11	18
	Total regional	**330**	**188**	**518**
XI de Aysén	11101 Coyhaique	142	57	199
	11201 Aysén	67	10	77
	Total regional	**209**	**67**	**276**
XII de Magallanes	12101 Punta Arenas	228	28	256
	Total regional	**228**	**28**	**256**

Continúa ►

Tabla 2: Distribución de la muestra objetivo por área urbano/rural, según región y comuna

Región	Comuna	Viviendas urbanas	Viviendas rurales	Total viviendas
XIII Metropolitana	13101 Santiago	184	-	184
	13103 Cerro Navia	101	-	101
	13104 Conchalí	96	-	96
	13105 El Bosque	104	-	104
	13106 Estación Central	107	-	107
	13110 La Florida	231	-	231
	13111 La Granja	109	-	109
	13112 La Pintana	125	-	125
	13113 La Reina	51	-	51
	13114 Las Condes	188	-	188
	13118 Macul	86	-	86
	13119 Maipú	393	-	393
	13120 Ñuñoa	131	-	131
	13121 Pedro Aguirre Cerda	77	-	77
	13122 Peñalolén	169	-	169
	13123 Providencia	164	-	164
	13124 Pudahuel	148	-	148
	13125 Quilicura	120	-	120
	13126 Quinta Normal	77	-	77
	13127 Recoleta	107	-	107
	13128 Renca	102	-	102
	13130 San Miguel	48	-	48
	13132 Vitacura	69	-	69
	13201 Puente Alto	359	-	359
	13301 Colina	49	2	51
	13401 San Bernardo	169	6	175
	13402 Buin	36	4	40
	13501 Melipilla	50	17	67
	13601 Talagante	45	10	55
	13605 Peñaflor	50	5	55
	Total regional	**3.745**	**44**	**3.789**
XIV de Los Ríos	14101 Valdivia	175	46	221
	14104 Los Lagos	9	17	26
	14107 Paillaco	12	13	25
	14108 Panguipulli	13	31	44
	14201 La Unión	26	21	47
	14204 Río Bueno	24	28	52
	Total regional	**259**	**156**	**415**
XV de Arica y Parinacota	15101 Arica	247	62	309
	Total regional	**247**	**62**	**309**

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

II. DISEÑO MUESTRAL

1. Diseño bifásico de la muestra

Un muestreo de dos fases (bifásico), corresponde a un método de recolección de ciertos ítems de información durante una muestra inicial, o muestra de primera fase, y luego otros ítems son recolectados en una segunda fase a partir de una submuestra de la muestra inicial.

Generalmente, en un contexto práctico, las encuestas de primera fase corresponden a encuestas ya existentes que permiten identificar a subpoblaciones de interés de poca prevalencia para las cuales no existen marcos muestrales disponibles. Además, estos métodos permiten ahorros en los costos de recolección y procesamiento de datos, ya que resulta más económico recolectar datos desde una muestra grande de primera fase y luego utilizar esos datos para seleccionar la muestra de segunda fase.

En ese sentido, para la encuesta Endisc II, se optó por un método bifásico en vez del levantamiento de una muestra independiente bietápica, principalmente por el ahorro presupuestario en los ítems relacionados a la enumeración de la muestra. De esta forma, en vista de la magnitud de viviendas que contiene la encuesta Casen 2013, se decidió utilizar esta encuesta a modo de marco muestral, asumiendo el desfase temporal existente entre ambas encuestas. Por otro lado, como un efecto adicional del método bifásico, pudiese incrementar la variabilidad de los factores de expansión como consecuencia de la selección de una submuestra, en vez de una muestra completa.

En definitiva, en la primera fase se utiliza, como marco muestral, el listado de viviendas con encuestas logradas en el levantamiento de la encuesta Casen 2013, para luego, como segunda fase, seleccionar desde dicho marco la muestra de la encuesta de Endisc II.

Esta primera fase (encuesta Casen 2013) permite detectar las unidades de muestreo que pertenecen a las subpoblaciones de interés para la encuesta de segunda fase (encuesta Endisc II), de esta forma, será posible identificar los estratos que permitan optimizar la estrategia de muestreo, haciendo que las unidades de interés puedan ser seleccionadas con probabilidades de selección que varían según los estratos definidos, que en este caso son las comunas del país.

La muestra para la encuesta Endisc II está diseñada para lograr representatividad estadística a nivel nacional, nacional urbano/rural, y regional (en población adulta). Sin embargo, por razones de costos, se optó por disminuir la cobertura geográfica a un total de 135 comunas del país (de las 324 comunas incluidas en la muestra de Casen 2013). Más específicamente, la muestra de la encuesta Endisc II incluye las comunas que al acumular los tamaños totales de viviendas comunales de forma descendente dentro de la región, permiten alcanzar como mínimo el 80% del total regional de viviendas. Con esta distribución se asegura una amplia cobertura comunal para cada región.

Además, se hicieron ajustes de re-distribución de la muestra a nivel regional-comunal para afinar criterios de proporcionalidad y, para contar con la representatividad Nacional Urbano-Rural, se asegura una cobertura geográfica mínima del total de viviendas a nivel nacional del 80% a nivel Nacional-Urbano y 51% a nivel Nacional-Rural.

De esta forma, se determinó como tamaño muestral objetivo un total 12.196 viviendas. Para compensar las pérdidas asociadas a la no respuesta, la estrategia operativa de levantamiento de la

muestra considera una modalidad de reemplazo de viviendas (a diferencia de Casen, no se ocupa sobre-dimensionamiento de la muestra o sobremuestreo). En la Tabla 3 se detalla el número de viviendas de reemplazos asignadas y disponibles para cada región[4].

Tabla 3: Distribución del total de viviendas disponibles para reemplazo

Región	Muestra objetivo	Cuota de reemplazos disponibles
Total nacional	**12.196**	**14.870**
I de Tarapacá	241	554
II de Antofagasta	488	704
III de Atacama	291	483
IV de Coquimbo	546	846
V de Valparaíso	1.443	1.524
VI de O'Higgins	575	882
VII del Maule	718	903
VIII del Biobío	1.639	2.058
IX de la Araucanía	692	1.171
X de Los Lagos	518	766
XI de Aysén	276	532
XII de Magallanes	256	476
XIII Metropolitana	3.789	2.764
XIV de Los Ríos	415	654
XV de Arica y Parinacota	309	553

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

2. Población objetivo

En vista que el diseño muestral de la encuesta Endisc II se basa en un diseño muestral bifásico, se realiza un muestreo estratificado por el cruce de la división político-administrativa y el área urbano-rural del país, y posteriormente se seleccionan viviendas del listado de unidades muestrales obtenidas de la encuesta Casen 2013. Cada selección al interior de cada grupo se realiza en forma sistemática, realizando previamente un ordenamiento geográfico jerárquico.

En ese sentido, la población objetivo son los hogares que residen en viviendas particulares ocupadas del país. Y el propósito de la encuesta de Endisc II es recolectar datos de los distintos hogares insertos en las viviendas seleccionadas de la muestra, de tal forma, que permitan representar de manera adecuada a la población chilena respecto de sus componentes sociodemográficos en materia de discapacidad. De este modo, los resultados de la encuesta permiten extrapolar inferencias con niveles de representatividad estadística y/o márgenes de error razonables a nivel regional, nacional y por áreas geográficas urbana y rural.

[4] Cabe señalar que la cuota de reemplazos disponible que aparece en la Tabla 1, no conlleva a que todas esas viviendas serán visitadas, sólo es una asignación sobreestimada de apoyo en terreno para optimizar tiempos y salvaguardar que la muestra objetivo sea mayormente lograda, de acuerdo a los objetivos técnicos y operativos propuestos. De esta forma, a modo de aclaración, durante todo el levantamiento de la Encuesta Endisc II, puede que se utilice solamente entre un 10% o 30% de la cuota reemplazos, haciendo hincapié que lo que se busca lograr son las 12.196 viviendas asignadas (o lo más cercano a ello). A diferencia del sobremuestreo, que se establece comúnmente una cuota de sobremuestreo fija entre un 30% o 50% para intentar acercarse al logro de la muestra objetivo (de acuerdo a la evidencia de estudios anteriores de tasas de no respuesta). Con ello se establece un margen de sobremuestra que quizás no fue necesario incorporar, lo cual implicó un gasto extra innecesario. En cierta forma, con la estrategia de reemplazos, se busca cubrir un margen mínimo de "sobremuestra". Sin embargo, se asume el riesgo de que en pos de alcanzar el logro de la muestra objetivo, el uso de los reemplazos sobrepase el porcentaje de la cuota de "sobremuestreo", y en ese sentido, se perdería eficiencia en términos de costos (al contrastarlo con la estrategia de sobremuestreo), pero aun así, existe mayor probabilidad de logro, por el dinamismo que proporciona en terreno la cuota de reemplazos, propiamente tal.

3. Características del marco muestral

Como se enunció anteriormente, la población objetivo de la encuesta Endisc II son los hogares que residen en viviendas particulares ocupadas del país. Siguiendo las directrices del diseño muestral bifásico de la encuesta, se elaboró un listado de las viviendas con entrevistas logradas durante la aplicación de la encuesta Casen 2013, cuyo levantamiento se inició en noviembre de 2013 y finalizó el 2 de febrero de 2014. El listado de viviendas fue utilizado como marco muestral para la selección de la muestra a entrevistar en la encuesta Endisc II. En lo que sigue, se describen las características del marco muestral de la encuesta Casen 2013 (primera fase) a partir del cual se seleccionó la muestra de la encuesta Endisc II (segunda fase).

3.1. El marco muestral en Casen 2013

Un marco muestral se define como la lista o los procedimientos que permiten identificar a todos los elementos de una población objetivo (Groves et al. 2004, pág. 68).

La población objetivo de la encuesta Casen son los hogares que residen en viviendas particulares ocupadas. El INE mantiene un marco de áreas geográficas que sirve de base para la selección de viviendas, requeridas para las muestras de las encuestas de hogares, como Casen.

Un marco muestral de área contiene las unidades geográficas de un país organizadas de forma jerárquica. En Chile, esta ordenación se denomina división político-administrativa y las unidades corresponden, en orden descendiente, a región, provincia y comuna. Al interior de cada comuna se conforma la división censal que da origen a las áreas geográficas denominadas urbana y rural.

Estas áreas se encuentran definidas por la actividad económica preponderante y dan origen a las siguientes entidades:

- **Ciudad (CD):** Es un gran centro urbano conformado por uno o un conjunto de centros urbanos adyacentes con 40.000 o más habitantes.

- **Resto de Área Urbana (RAU**): Conformado por un conjunto de centros urbanos que totalizan menos de 40.000 y más de 2.000 habitantes. Esta clasificación se da cuando en una comuna existe una ciudad (CD) y entonces todos los centros urbanos restantes, si es que existen, se denominan resto de área urbana (RAU).

- **Urbano (U):** Es el centro urbano con menos de 40.000 y más de 2.000 habitantes. Esta clasificación se da cuando en la comuna no existe una ciudad (CD), por lo que cada uno de sus centros urbanos se denominan simplemente como urbanos (U).

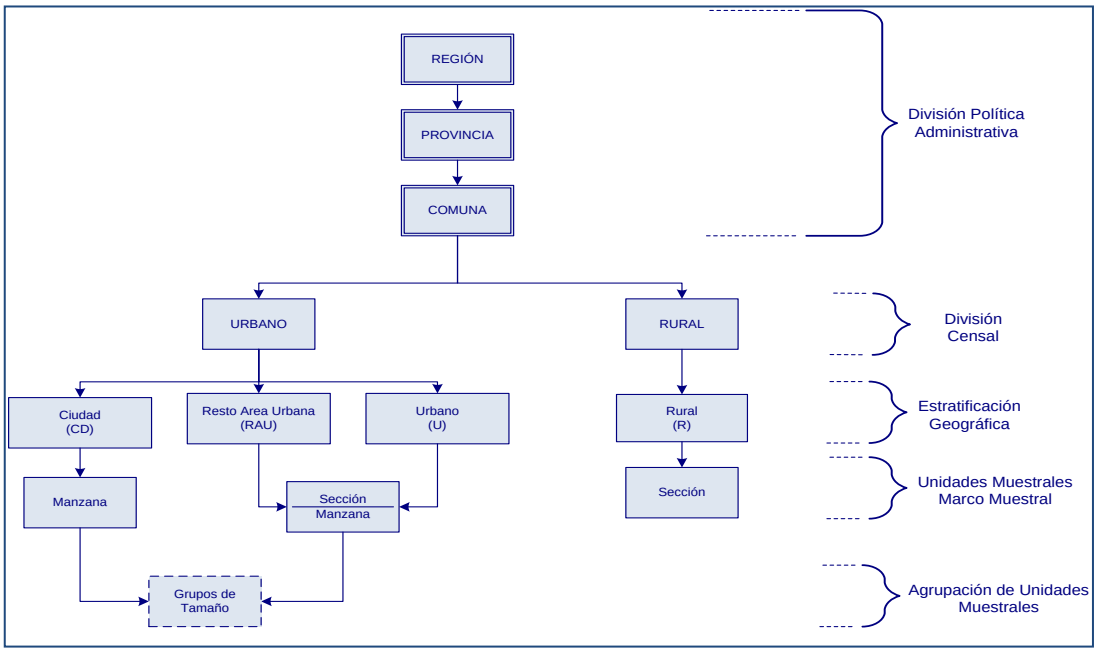
- **Rural (R):** Conformado por el conjunto de entidades clasificadas como rurales de acuerdo a un tamaño poblacional menor a 1.000 habitantes o entre 1.001 y 2.000 habitantes con predominio de población económicamente activa (PEA) dedicada a actividades primarias.[5]

Para efectos del Censo 2002, se realizaron sub-divisiones posteriores denominadas manzanas censales (en las áreas urbanas) y secciones de empadronamiento censal (en las áreas rurales).

[5] Se entiende por Actividad Primaria a toda aquella actividad relacionada con la extracción de recursos naturales. (agricultura, caza, pesca, minería, etc.).

Estas son las unidades geográficas más pequeñas y corresponden a las unidades primarias de muestreo más comúnmente utilizadas en las encuestas de hogares diseñadas por el INE. La Figura 1 ilustra las unidades geográficas descritas.

Figura 1: Estratificación e identificación de unidades primarias de muestreo



Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Para cumplir con los objetivos de investigación, el marco de muestreo "ideal" debe ser completo, preciso y actualizado. A mediados del periodo intercensal, investigaciones realizadas por el INE indican que el marco muestral en uso (MS2002) presentaba deficiencias en estos aspectos, por lo tanto a partir de 2008 el INE desarrolló un nuevo marco de muestreo, definiciones para la conformación de unidades de muestreo y procedimientos para la selección de unidades de muestreo en las áreas urbanas del país.

Bajo el nuevo marco de muestreo urbano, se mantiene la división político-administrativa, pero se cambia la conformación de las unidades primarias de muestreo. En el nuevo marco, las manzanas censales (predios urbanos delimitados por calles) reemplazan a las secciones de empadronamiento censal en las áreas urbanas, lo que permite actualizar más rápidamente las unidades muestrales que el marco antiguo por medio del plan municipal de edificaciones, logrando captar nuevos desarrollos urbanos tanto en áreas urbanas como en áreas previamente catalogadas como agrícolas.

El INE comenzó a seleccionar las muestras urbanas utilizando el nuevo marco de muestreo a partir del año 2008. Las primeras encuestas a nivel nacional, cuyas muestras urbanas fueron seleccionadas a partir del nuevo marco, fueron la Nueva Encuesta del Empleo (NENE) en 2009, la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC) en 2008. En las zonas rurales, sin embargo, el INE ha seguido utilizando el marco de secciones (MS2002) para la selección de muestras.

La selección de la muestra de la encuesta Casen 2013 se realizó sobre el marco muestral que el INE mantiene vigente al año 2013, el cual comprende el uso de dos marcos de muestreo que son mutuamente excluyentes:

- **En el área urbana y rural:** el marco de muestreo corresponde al generado a partir del Censo de Población y Vivienda del año 2002. Las unidades que componen este marco de muestreo se denominan "secciones" y corresponden, en el área rural, a aglomeraciones de viviendas particulares conformadas a partir de una o más entidades pobladas, enmarcadas generalmente dentro de un distrito censal. En el área urbana, corresponden a aglomeraciones de viviendas particulares conformadas a partir de una o más manzanas según los rangos de viviendas asignados para la sección. Las secciones no sobrepasan los límites del distrito. En adelante, denominaremos este marco muestral como Marco de Secciones (MS2002).

- **En el área Ciudad**: en el área urbana, que no está incluida en el MS2002, el marco de muestreo es aquel generado a partir de cartografía digital, actualizado al segundo semestre del año 2008. La información en el marco se actualiza con información de los registros administrativos asociados a nuevas construcciones otorgados por los municipios. Las unidades que componen este marco de muestreo se denominan "manzanas" y corresponden a delimitaciones geográficas fijas. En adelante, se denomina a este marco como Marco de Manzanas (MM2008).

Para aquellas áreas denominadas Resto de Área Urbana (RAU) y Urbano (U) sus unidades pueden pertenecer al MS2002 o al MM2008. De un total de 219 áreas de este tipo, 89 son extraídas desde el MS2002 , mientras que las restantes 130 se extraen desde el MM2008 . Cabe señalar que las 89 áreas urbanas seleccionadas a partir del MS2002, corresponden a áreas de pequeña densidad poblacional, con no más de 40.000 habitantes, y las unidades muestrales, al igual que en las áreas rurales, son secciones.

Para fines de selección muestral, se han conglomerado las unidades primarias de muestreo que conforman el MM2008 del INE para sus encuestas de hogares; el objetivo, es poder generar grupos homogéneos de manzanas, respecto de una cierta característica y, de este modo, establecer en cada grupo un número fijo de unidades secundarias de muestreo para encuestar. Para la consecución de tal propósito, se realizó un análisis de conglomerados (*clusters*), que utilizó el número de viviendas particulares, al interior de cada manzana, como variable de conglomeración. Como resultado de esta segmentación se pudo asociar las manzanas del MM2008 dentro de 5 grandes conglomerados de tamaño. Una vez definidos, y de acuerdo a una estrategia de selección de una fracción de muestreo aproximada del 25% del total de viviendas particulares de la manzana, los 5 conglomerados de tamaño se subdividieron, dando origen a 30 grupos de tamaño. Desde esta última clasificación surge la variable de conglomeración que se

añade al marco muestral, previo a la selección de cualquier muestra, y que permite escoger las unidades primarias, pero considerando grupos de manzanas homogéneas en cuanto a tamaño.

Más de la mitad de las manzanas en el marco tiene entre 8 y 44 viviendas, mientras que el grupo de tamaño más grande corresponde a manzanas con 155 o más viviendas, las cuales representan el 1,9% del total de las manzanas en el marco (ver Tabla 4). En la Tabla 5 se puede apreciar la muestra seleccionada de manzanas y sus respectivas viviendas para la muestra Casen 2013.

Tabla 4: Número de manzanas y viviendas en el MM2008, según grupo de tamaño

Grupo de tamaño	Rango de viviendas	Total de Manzanas	% de manzanas	Total de Viviendas	% de viviendas
Total		**133.360**	**100**	**4.000.762**	**100**
Grupo 0	1 a 7	13.894	10,4	53.578	1,3
Grupo 1 a 4	8 a 23	63.319	47,5	994.825	24,9
Grupo 5 a 9	24 a 44	39.267	29,4	1.216.764	30,4
Grupo 10 a 19	45 a 81	10.334	7,7	591.573	14,8
Grupo 20 a 28	82 a 154	3.990	3	439.327	11
Grupo 29 y 30	155 y más	2.556	1,9	704.695	17,6

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Tabla 5: Número de manzanas y viviendas seleccionadas para la muestra Casen 2013 en el MM2008, según grupo de tamaño

Grupo de tamaño	Rango de viviendas	Total de Manzanas	% de manzanas	Total de Viviendas	% de viviendas
Total		**7.418**	**100**	**72.663**	**100**
Grupo 0	1 a 7	0	0	0	0
Grupo 1 a 4	8 a 23	962	13	3.461	4,8
Grupo 5 a 9	24 a 44	3.972	53,5	26.217	36,1
Grupo 10 a 19	45 a 81	1.691	22,8	21.229	29,2
Grupo 20 a 28	82 a 154	286	3,9	6.722	9,3
Grupo 29 y 30	155 y más	507	6,8	15.034	20,7

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Tabla 6: Número de manzanas y viviendas en el MM2008, según grupo de tamaño y nivel socioeconómico (*)

Grupo de Tamaño	Manzanas				Viviendas			
	1 Bajo	2 Medio	3 Alto	Total	1 Bajo	2 Medio	3 Alto	Total
Total	**40.008**	**80.021**	**13.331**	**133.360**	**944.559**	**2.473.829**	**582.374**	**4.000.762**
Grupo 0	6.735	5.952	1.207	13.894	24.550	24.022	5.006	53.578
Grupo 1 a 4	19.734	37.356	6.229	63.319	305.740	593.773	95.312	994.825
Grupo 5 a 9	10.016	25.963	3.288	39.267	307.621	806.566	102.577	1.216.764
Grupo 10 a 19	2.417	6.825	1.092	10.334	138.263	388.047	65.263	591.573
Grupo 20 a 28	773	2.450	767	3.990	82.595	270.750	85.982	439.327
Grupo 29 y 30	333	1.475	748	2.556	85.790	390.671	228.234	704.695

(*)La clasificación socioeconómica fue derivada en 3 categorías según método Princals aplicado a la clasificación de los hogares del Censo 2002. Alto, el 10% de las manzanas con mejor nivel socioeconómico; Medio, el 60% de las manzanas.

3.2. Cobertura del marco muestral

Las unidades de muestreo tienen cuatro características fundamentales para el diseño muestral: (1) cubren, usualmente, la totalidad del territorio del país; (2) tienen sus límites bien definidos; (3) existen estimaciones poblacionales para las unidades; y (4) existen mapas para las unidades (Turner, 2003).

La cobertura es una propiedad estadística asociada al marco muestral que se utiliza para la selección de la muestra. La falta de cobertura denota la falla al incluir ciertos elementos (o unidades completas) de la población a encuestar a partir del marco muestral que se ha definido (Kish 1965, pág. 528). Estas fallas no son planeadas por el investigador (ej. fallas en el proceso de conteo e identificación de las viviendas previo a la selección).

Es importante distinguir la falta de cobertura (fallas no intencionadas), de las exclusiones que realiza el investigador en forma intencionada. En ese sentido, al tratarse de un muestreo bifásico, la encuesta Endisc II comparte las mismas propiedades de cobertura de la selección de viviendas de la encuesta Casen 2013. De esta manera, se pueden identificar tres tipos de exclusiones "intencionadas" en el proceso de selección de la muestra de la encuesta Casen 2013:

- Las 22 áreas geográficas que han sido catalogadas por INE como áreas de difícil acceso (ADA);
- Las manzanas y secciones incluidas en otras muestras seleccionadas por el INE para el periodo de recolección de datos de la encuesta Casen 2013;
- Las manzanas con menos de 8 viviendas.

Las áreas de difícil acceso corresponden a zonas geográficas no incluidas en el marco muestral del INE. Estas áreas no están presentes en las muestras de ninguna de las encuestas de hogares seleccionadas por el INE. En total, corresponden al 0,4% de la población de viviendas y al 0,4% de la población de personas del país.

Las 22 áreas de difícil acceso excluidas de la muestra de la Encuesta Casen 2013 son informadas en Tabla 7. Se presenta el total de viviendas según la información del Censo de Población y Vivienda del año 2002, y una aproximación del total de personas, estimadas mediante las proyecciones de población con fecha Noviembre de 2013.

Tabla 7: Áreas de difícil acceso definidas por el INE

Región	Provincia	Comuna	Total viviendas (*)	Total personas (**)
Arica y Parinacota	Parinacota	General Lagos	260	974
Tarapacá	Tamarugal	Colchane	461	1390
Antofagasta	El Loa	Ollagüe	66	159
Valparaíso	Valparaíso	Juan Fernández	206	905
	Isla de Pascua	Isla de Pascua	1.136	5.120
Los Lagos	Llanquihue	Cochamó	1.345	3.833
	Palena	Chaitén	1.830	6.164
		Futaleufú	606	1686
		Hualaihué	2.249	7.952
		Palena	558	1480
Aysén del General Carlos Ibáñez del Campo	Coihaique	Lago Verde	338	756
	Aysén	Guaitecas	383	1741
	Capitán Prat	O'Higgins	154	653
		Tortel	145	502
Magallanes y La Antártica Chilena	Magallanes	Laguna Blanca	116	278
		Río Verde	86	205
		San Gregorio	212	293
	Antártica Chilena	Cabo de Hornos (Ex - Navarino)	520	2160
		Antártica	10	21
	Tierra el Fuego	Primavera	228	340
		Timaukel	82	418
	Última Esperanza	Torres del Paine	114	534

(*) Viviendas particulares ocupadas según Censo 2002.
(**) Según proyecciones de población al 30 de noviembre de 2013.
Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

El marco muestral del INE, utilizado como base para la encuesta Casen 2013, cubre sólo a la población que reside en viviendas particulares ocupadas y, por lo tanto, no cubre a la población que reside en viviendas colectivas (ej. hogares de ancianos, hogares de niños, cárceles) ni tampoco a la población que reside en la calle.

Por otra parte, entre las viviendas particulares ocupadas identificadas durante el proceso de elaboración del marco muestral para la encuesta Endisc II, se establecieron criterios adicionales de exclusión intencionada. Por razones presupuestarias, se optó por disminuir la cobertura geográfica, obteniendo un total de 135 comunas donde aplicar la encuesta (ver Tabla 8), en contraste de las 324 comunas incluidas en muestra Casen 2013.

De esta forma, se estimó un tamaño muestral objetivo total de 12.196 viviendas a entrevistar distribuidas en 135 comunas, alcanzando gran presencia en áreas urbanas y rurales, lo que permitió lograr un diseño muestral con representatividad estadística y/o márgenes error razonables a nivel nacional, nacional urbano/rural, y regional.

Tabla 8: Cobertura geográfica de la encuesta Endisc II

Tarapacá	7201 Cauquenes	10201 Castro
1101 Iquique	7301 Curicó	10202 Ancud
1107 Alto Hospicio	7308 Teno	10208 Quellón
Antofagasta	7401 Linares	10301 Osorno
2101 Antofagasta	7403 Longaví	10305 Río Negro
2201 Calama	7404 Parral	**Aysén**
Atacama	7406 San Javier	11101 Coyhaique
3101 Copiapó	7407 Villa Alegre	11201 Aysén
3201 Chañaral	**Bío-Bío**	**Magallanes**
3301 Vallenar	8101 Concepción	12101 Punta Arenas
Coquimbo	8102 Coronel	**Santiago**
4101 La Serena	8103 Chiguayante	13101 Santiago
4102 Coquimbo	8105 Hualqui	13103 Cerro Navia
4201 Illapel	8106 Lota	13104 Conchalí
4203 Los Vilos	8107 Penco	13105 El Bosque
4301 Ovalle	8108 San Pedro de la Paz	13106 Estación Central
Valparaíso	8110 Talcahuano	13110 La Florida
5101 Valparaíso	8111 Tomé	13111 La Granja
5103 Concón	8112 Hualpén	13112 La Pintana
5107 Quintero	8202 Arauco	13113 La Reina
5109 Viña del Mar	8203 Cañete	13114 Las Condes
5301 Los Andes	8205 Curanilahue	13118 Macul
5501 Quillota	8301 Los Ángeles	13119 Maipú
5502 Calera	8304 Laja	13120 Ñuñoa
5601 San Antonio	8305 Mulchén	13121 Pedro Aguirre Cerda
5603 Cartagena	8306 Nacimiento	13122 Peñalolén
5604 El Quisco	8401 Chillán	13123 Providencia
5701 San Felipe	8406 Chillán Viejo	13124 Pudahuel
5801 Quilpué	**Araucanía**	13125 Quilicura
5802 Limache	9101 Temuco	13126 Quinta Normal
5804 Villa Alemana	9102 Carahue	13127 Recoleta
O´Higgins	9103 Cunco	13128 Renca
6101 Rancagua	9105 Freire	13130 San Miguel
6104 Coltauco	9108 Lautaro	13132 Vitacura
6106 Graneros	9111 Nueva Imperial	13201 Puente Alto
6107 Las Cabras	9112 Padre Las Casas	13301 Colina
6108 Machalí	9114 Pitrufquén	13401 San Bernardo
6111 Olivar	9115 Pucón	13402 Buin
6112 Peumo	9120 Villarrica	13501 Melipilla
6113 Pichidegua	9201 Angol	13601 Talagante
6115 Rengo	9202 Collipulli	13605 Peñaflor
6116 Requínoa	9203 Curacautín	**Los Ríos**
6117 San Vicente	9210 Traiguén	14101 Valdivia
6301 San Fernando	9211 Victoria	14104 Los Lagos
6303 Chimbarongo	**Los Lagos**	14107 Paillaco
6310 Santa Cruz	10101 Puerto Montt	14108 Panguipulli
Maule	10102 Calbuco	14201 La Unión
7101 Talca	10105 Frutillar	14204 Río Bueno
7102 Constitución	10106 Los Muermos	**Arica y Parinacota**
7109 San Clemente	10107 Llanquihue	15101 Arica

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

3.3. Estratificación del marco muestral

En el diseño de una muestra, la estratificación corresponde al proceso de agrupar a los elementos de una población en grupos homogéneos previo a la selección de la muestra. Su propósito es mejorar la precisión estadística de los estimadores agrupando las unidades del marco en clases homogéneas en su interior y que difieran de las características del resto. Los estratos deben ser mutuamente excluyentes: cada elemento en la población debe ser asignado a un solo estrato. Además, los estratos deben ser exhaustivos colectivamente, es decir, ningún elemento de la población puede quedar excluido.

La encuesta Casen ha definido tradicionalmente los estratos de selección de la muestra a partir del criterio de proximidad geográfica. Hasta 1996 los estratos se conforman a partir de comunas y grupos de comunas, según área urbano/rural. A partir de 1998 los estratos se constituyen a partir de las comunas, según área urbano/rural. Para Casen 2013 se conformaron 584 estratos, a partir de la interposición de la división política-administrativa (a nivel de comunas) y la división censal (urbano-rural). De éstos, 312 estratos se encuentran en zonas urbanas y 273 en zonas rurales.

El diseño muestral de la encuesta Endisc II comparte las mismas propiedades de los estratos de selección de la encuesta Casen 2013, con el énfasis especial de exclusión intencionada en la cobertura geográfica, que sólo considera 135 comunas del país. En ese sentido, Endisc II considera 224 estratos, de los cuales, 128 estratos se encuentran en zonas urbanas y 96 en zonas rurales.

4. Selección de la muestra

Una vez determinado el listado y/o marco muestral de viviendas con encuestas a hogares logradas en la encuesta Casen 2013, de acuerdo a los criterios de exclusión señalados en párrafos anteriores, como parte del método bifásico, se procede a la segunda fase, en la cual se seleccionan, de manera sistemática con igual probabilidad, el número de viviendas al interior de cada uno de los estratos de selección establecidos (comuna y área urbano-rural).

La selección de viviendas fue implementada en SPSS, bajo el módulo de análisis Muestras Complejas, específicamente en el procedimiento seleccionar una muestra. Bajo este procedimiento se realizó la selección sistemática de viviendas donde se utiliza una semilla fija, a fin de poder replicar la selección.

Sean $\mathrm{M_i}$ el número de viviendas que posee el estrato y $\mathrm{m_i}$ el número de viviendas a seleccionar. Para la selección de $\mathrm{m_i}$ viviendas el software sigue los siguientes pasos:

Paso 1: En primera instancia se ordenan geográficamente todos los elementos (viviendas) según la división política administrativa, del área de procedencia:

- **Área urbana:** región, comuna, distrito, zona, manzana y número de orden de vivienda.
- **Área rural:** región, comuna, distrito, código cartográfico, número de orden de vivienda.

Paso 2: Se determina un número de viviendas a encuestar por manzana o sección, a través de una distribución proporcional al tamaño en cuanto al total de viviendas elegibles y cuyo(s) hogar(es) respondió(eron) Casen 2013.

Paso 3: Al interior de cada manzana o sección se calcula el período $\mathrm{k = M_i / m_i}$. Notar que K puede ser un número real, no entero (puede tener decimales).

Paso 4: Luego se calcula el arranque A o primera selección como un número aleatorio entre 1 y el período k. Para la selección de este número se define una semilla fija que en esta selección fue 999999999.

Paso 5: Posteriormente se va sumando sucesivamente el período k al arranque A para obtener distintos valores los que van generando las sucesivas selecciones: $A, A + K, A + 2K, A + 3K, \ldots, A + (m_i - 1)K$.

La primera vivienda seleccionada es A y es un número entero, la segunda es el redondeo de $A + K$, la tercera es el redondeo de $A + 2K$, y así sucesivamente hasta la m_i selección dada por el redondeo de $A + (m_i - 1)K$.

4.1. Selección del informante Kish

Como requisito de Endisc II, existe una última etapa de la selección muestral, que corresponde a la selección aleatoria del informante dentro del hogar que responderá la entrevista. En ese sentido, ese proceso de selección está basado en el método Kish, el cual asigna la misma probabilidad de ser elegido como informante de la encuesta a todos aquellos miembros de la vivienda que cumplen determinadas características.

Para el caso de la encuesta Endisc II, el procedimiento de selección de informante Kish se realiza primero entre todos los miembros del hogar entre 2 y 17 años, y luego entre todos los miembros del hogar mayores de 18 años.

Como aspecto clave en la encuesta Endisc II, el encuestador al llegar a la vivienda, debe anotar en el Registro de Personas del Hogar (RPH) a todas las personas que componen el hogar. El ordenamiento de los miembros de la vivienda es por hogar, y dentro del hogar, es relacionado al parentesco existente entre el jefe de hogar con cada uno de los individuos.

Luego de este ordenamiento, se debe distinguir las dos poblaciones objetivo de la encuesta que habitan la vivienda, que serían: las personas de 18 años o más, y los niños, niñas y adolescentes (entre 2 y 17 años). Cabe señalar que en la encuesta Endisc II no se consideran casos especiales que deban ser excluidos de la numeración (salvo el personal de servicio puertas adentro). Si existe algún miembro del hogar que tenga dificultades para responder (por enfermedad, problemas de consumo de alcohol o drogas, etc.), esas personas deben recibir de igual modo una numeración Kish y, de ser seleccionado, el encuestador debe solicitar que sea el cuidador o la persona responsable del miembro del hogar seleccionado quien responda la encuesta por su representado. En el caso del cuestionario infantil, éste siempre debe ser respondido por el adulto responsable del niño, niña o adolescente seleccionado.

En términos generales la metodología Kish genera cruces utilizando una tabla especialmente diseñada entre una letra que es asignada a cada vivienda en el directorio, y la cantidad de personas que se enumeran de 1 a n según los criterios pre-establecidos (ver Tabla 9).

Tabla 9: Tabla de Kish utilizada para la selección de personas

Tabla aleatoria	Si el número de personas en la vivienda es (*)											
	1	2	3	4	5	6	7	8	9	10	11	12 o más
	Selecciónese a la persona con el número:											
A	1	1	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	1	1	1	1	1	2
C	1	1	1	1	1	1	1	1	2	2	2	2
D	1	1	1	1	1	2	2	2	2	2	2	3
E	1	1	1	1	2	2	2	2	3	3	3	3
F	1	1	1	2	2	2	2	3	3	3	4	4
G	1	1	1	2	2	2	3	3	3	4	4	5
H	1	1	2	2	2	3	3	3	4	4	5	5
I	1	1	2	2	3	3	3	4	4	5	5	6
J	1	1	2	2	3	3	4	4	5	5	6	6
K	1	2	2	3	3	4	4	5	5	6	6	7
L	1	2	2	3	3	4	5	5	6	6	7	7
M	1	2	2	3	4	4	5	6	6	7	7	8
N	1	2	3	3	4	5	5	6	7	7	8	8
O	1	2	3	3	4	5	6	6	7	8	8	9
P	1	2	3	4	4	5	6	7	7	8	9	10
Q	1	2	3	4	5	5	6	7	8	9	10	10
R	1	2	3	4	5	6	7	8	8	9	10	11
S	1	2	3	4	5	6	7	8	9	10	11	11
T	1	2	3	4	5	6	7	8	9	10	11	12

(*) El proceso de selección del informante Kish se realizó en forma independiente, utilizando la misma tabla para niños y adultos.

III. DISEÑO FACTORES DE EXPANSIÓN

El diseño de la encuesta Endisc II, es un diseño complejo donde sus unidades provienen de haber participado previamente en Casen 2013. Por lo tanto, la probabilidad de que un informante participe en Endisc II, está condicionada a que la vivienda haya sido seleccionada y entrevistada en Casen 2013. Por lo tanto, un insumo fundamental para determinar el factor de expansión de las unidades de muestreo y análisis, es el factor de expansión de Casen 2013, específicamente el ponderador ajustado a no respuesta, que corresponde al inverso de la probabilidades de selección de las viviendas ajustado por no respuesta, el cual podría ser interpretado como la probabilidad de que una vivienda haya sido seleccionada y haya respondido la encuesta Casen 2013. Junto a esto, para el cálculo de los factores de Endisc II se requieren las probabilidades condicionales de selección de las viviendas y personas, además de la realización de ciertos ajustes para compensar la pérdida de respuesta, así como también el cambio de estado de elegibilidad de las unidades muestrales.

Debido a que la encuesta posee un diseño bifásico, y que los factores de expansión están sujetos a alta variabilidad, lo que genera mayor varianza en las estimaciones y por consiguiente menor confiabilidad estadística para los análisis que se planea realizar, en la encuesta Endisc II se realizó truncamiento de los valores extremos. Esto consiste en determinar un punto máximo o umbral definido como ^{n} veces el factor de expansión promedio, donde $^{n = 4,5,\ldots,10}$. Luego, el peso de todos aquellos factores que exceden el umbral es distribuido sobre el resto de los factores.

Posteriormente, se evalúa cuál de los puntos es aquél que minimiza el error cuadrático medio de alguna de variable de interés.

A continuación, se describen y detallan los procedimientos utilizados para el cálculo de los factores de expansión.

1. Ponderación de vivienda

El ponderador de vivienda está constituido por dos componentes: (1) Ponderación de selección de las viviendas, (2) Ajuste por omisión de unidades en determinadas comunas.

Posteriormente, se presenta la metodología de cálculo de las probabilidades de selección de las viviendas y también el método de ajuste de omisión de unidades en determinadas comunas.

1.1. Ponderador de selección de vivienda

Las personas, adultos y niños, seleccionados en Endisc II son individuos que componen hogares, tales que al menos un integrante de estos haya participado en Casen 2013. Por lo tanto, la probabilidad de que una persona haya sido seleccionada en Endisc II es condicional a que la vivienda en que reside haya sido seleccionada y posteriormente participara en Casen 2013. Descrito lo anterior, se puede decir que el listado de viviendas cuyo(s) hogar(es) respondió(eron) Casen 2013 es el marco de muestreo de Endisc II, y por tanto de dicho listado se seleccionó un set de viviendas.

Una propuesta preliminar de diseño muestral consideró un tamaño de 27.066 viviendas con sobremuestreo, total sobre el cual se realizó la selección. Sin embargo, por razones presupuestarias y de tiempo, se determinó reducir el tamaño muestral y, a causa de esto, realizar una segunda selección, desde la muestra de viviendas seleccionadas originalmente para Endisc II, clasificando dichas unidades en muestra objetivo y muestra de reemplazo. Por otro lado, se decidió establecer una estrategia de muestreo con sustitución y no con sobremuestreo como estaba planeado originalmente, lo que significa que en lugar de visitar todas las unidades muestrales seleccionadas hasta alcanzar el tamaño muestral "objetivo", se enviaron al trabajo de campo las 12.196 viviendas (seleccionadas en la segunda instancia como muestra objetivo), y en la medida que una vivienda no era lograda (ya sea por no contacto, rechazo u otra razón), de acuerdo a un protocolo, se enviaron viviendas de reemplazo.

De acuerdo a lo anterior, en Endisc II la probabilidad de selección de una vivienda está compuesta por tres componentes:

i) **Probabilidades de selección y entrevista de las viviendas de la encuesta Casen 2013:**
Estas pueden ser estimadas como el inverso del factor de expansión ajustado por no respuesta asociado a cada vivienda seleccionada en Casen 2013 ($P_{hi}^{NR}(j)$).

$$P_{hi}^{NR}(j) = \frac{1}{w_{hij}^{NR}} \quad (1)$$

Siendo,

h Índice del estrato de muestreo, o área (urbana o rural) de una comuna.

i Índice de la unidad primaria de muestreo correspondiente al conglomerado, pudiendo ser manzana o sección.

j Índice de la vivienda seleccionada.

w_{hij}^{NR} Factor de expansión ajustado por no respuesta proveniente de Casen 2013, asociado a la vivienda j, del conglomerado i en el estrato h.

ii) Probabilidad de selección condicional de la vivienda en Endisc II:

Corresponde a la probabilidad de selección de las viviendas obtenidas desde el listado de viviendas cuyos hogares respondieron Casen 2013. Según la metodología de selección, esto debería ser calculado como el cuociente entre el total de viviendas seleccionadas originalmente para Endisc II (ν_{hi}), sobre el total de viviendas seleccionadas y cuyos hogares respondieron Casen 2013 (V_{hi}). Como el procedimiento se realizó de forma independiente en cada conglomerado (manzana y sección), el cálculo de las probabilidades debe efectuarse al interior de cada uno de ellos y, por tanto, todas las viviendas al interior de una misma manzana tienen igual probabilidad de selección y debe ser calculado tal como se señala a continuación:

$$P_{hi}\left(j|\,\theta_{hi}\right) = \frac{\nu_{hi}}{V_{hi}} \qquad (2)$$

Donde,

j Es el índice de la vivienda seleccionada.

θ_{hi} Es el conjunto de viviendas seleccionadas y que respondió Casen 2013, pertenecientes al conglomerado i del estrato de muestreo h.

$P_{hi}\left(j|\,\theta_{hi}\right)$ Es la probabilidad de selección de la vivienda j condicional a que pertenece al conjunto de viviendas seleccionadas y cuyos hogares respondieron Casen 2013, en el conglomerado i del estrato de muestreo h.

ν_{hi} Es el número de viviendas seleccionadas inicialmente en Endisc II, en el conglomerado i del estrato h.

V_{hi} Es el número de viviendas seleccionadas y cuyos hogares respondieron Casen 2013, en el conglomerado i del estrato h.

iii) Corrección de la probabilidad condicional de selección de la vivienda.

Para el caso de Endisc II, se debe considerar un paso adicional: la probabilidad que la vivienda seleccionada en la muestra original sea asignada como muestra objetivo o, por descarte, sea parte de la muestra de reemplazo.

Una vivienda posee probabilidad de selección diferente si ésta fue asignada como muestra objetivo o muestra de reemplazo. Así la probabilidad condicional que una vivienda pertenezca a la muestra objetivo está dada por la siguiente expresión:

$$P_{hi}\left(j \in \Omega_{hi}^{o} \,|\, j \in \Omega_{hi}\right) = \frac{\nu_{hi}^{o}}{\nu_{hi}} \qquad (3A)$$

Mientras que, la probabilidad condicional que una vivienda pertenezca a la muestra de reemplazo se puede calcular por la siguiente expresión:

$$P_{hi}\left(j \in \Omega_{hi}^{R} \mid j \in \Omega_{hi}\right) = \frac{v_{hi}^{R}}{v_{hi}} \tag{3B}$$

Donde,

Ω_{hi} Es el conjunto inicial (original) de viviendas seleccionadas en el conglomerado i del estrato h (Casen 2013), que en total corresponden a 27.066 unidades.

Ω_{hi}^{O} Es el conjunto de viviendas seleccionadas como muestra objetivo en el conglomerado i del estrato h, que en total corresponden a 12.196 unidades.

Ω_{hi}^{R} Es el conjunto de viviendas considerada como muestra de reemplazo en el conglomerado i del estrato h, que en total corresponden a 14.870 unidades.

$P_{hi}\left(j \in \Omega_{hi}^{O} \mid j \in \Omega_{hi}\right)$ Es la probabilidad de seleccionar una vivienda como muestra objetivo condicional a que fue seleccionada como muestra original.

v_{hi}^{O} Es el número de viviendas asignadas como muestra objetivo en el conglomerado i del estrato h.

v_{hi}^{R} Es el número de viviendas asignadas como muestra de reemplazo, en el conglomerado i del estrato h.

$$v_{hi}^{R} = v_{hi} - v_{hi}^{O} \tag{4}$$

En consecuencia, la probabilidad de que una vivienda haya sido seleccionada en Endisc II es distinta si es muestra objetivo o de reemplazo, y se calcula a partir del producto de las tres probabilidades condicionales definidas en las expresiones (1), (2) y (3A) y (1), (2) y (3B), respectivamente. Realizando algunas reducciones se llega a la siguiente expresión:

$$P_{hij}(j) = \begin{cases} \frac{1}{w_{hij}^{NR}} \cdot \frac{v_{hi}^{O}}{V_{hi}}, & vivienda\ j\ fue\ asignada\ a\ la\ muestra\ objetivo\ \ O \\ \frac{1}{w_{hij}^{NR}} \cdot \frac{v_{hi}^{R}}{V_{hi}}, & vivienda\ \ j\ fue\ asignada\ a\ la\ muestra\ reemplazo\ R \end{cases} \tag{5}$$

Luego, el ponderador de vivienda fue calculado como el inverso de la probabilidad antes descrita.

$$w_{hij} = \frac{1}{P_{hi}(j)} = \begin{cases} w_{hij}^{NR} \cdot \frac{V_{hi}}{v_{hi}^{O}}, & vivienda\ j\ fue\ asignada\ a\ la\ muestra\ objetivo\ \ O \\ w_{hij}^{NR} \cdot \frac{V_{hi}}{v_{hi}^{R}}, & vivienda\ \ j\ fue\ asignada\ a\ la\ muestra\ reemplazo\ R \end{cases} \tag{6}$$

Donde,

w_{hij} Es el ponderador de selección de vivienda de la unidad j, perteneciente al conglomerado i del estrato muestral h.

$P_{hi}(j)$ Es la probabilidad de seleccionar una vivienda en la muestra de Endisc II en el conglomerado i del estrato muestral h.

1.2. Ajuste por omisión de unidades en ciertas comunas

En Endisc II, previo a la selección, se distribuyeron las unidades muestrales en cada región y área (urbano y rural) entre las comunas de Casen 2013 que acumularan el 80% de la población. Así, por ejemplo, en Casen 2013, en el área urbana de la región de Arica y Parinacota se seleccionaron unidades de las comunas de Putre y Arica, sin embargo, en Endisc II solo se seleccionaron unidades de la comuna de Arica, ya que concentra más del 90% de las unidades.

En este contexto, en Endisc II existen comunas que no están representadas en la muestra, es decir, existe una falta de cobertura de aproximadamente un 20% de las viviendas de la población.

Una forma de corregir esta falta de cobertura es realizar un ajuste, que consiste en ponderar el inverso de la probabilidad de selección de las viviendas con la razón calculada como el cuociente entre, (1) la suma del factor de expansión ajustado por no respuesta de Casen 2013 (sobre todas las unidades que respondieron Casen 2013, en la región r y área a), (2) la suma del inverso de la probabilidad de selección de las viviendas de Endisc II. A continuación se resume este procedimiento a partir de la siguiente fórmula,

$$w'_{hij} = w_{hij} \cdot \ddot{R}_{ra} \tag{7}$$

Donde,

r Es el índice de región, tal que $r = 1, \dots, 15$.

a Es el índice de área, donde $a = urbano\ o\ rural$.

w_{hij} Es el ponderador de selección de vivienda vinculado a la vivienda j del conglomerado i del estrato de muestreo h.

$\ddot{R}_{ra}$ Es la razón de la región r y área a con la cual se pondera el factor de selección de vivienda para realizar el ajuste por omisión de unidades. Se obtiene a partir de la siguiente expresión:

$$\hat{\ddot{R}}_{ra} = \frac{\Delta_{ra}}{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in\Theta_{hi}} w_{hij}} \tag{8}$$

Siendo,

Δ_{ra} Es el número de viviendas estimadas desde la base de Casen 2013, con el factor de no respuesta en la región r, y área a.

Θ_{hi} Es el listado de viviendas seleccionadas para Endisc II en el conglomerado i del estrato h.

Θ_h Es el listado de conglomerados seleccionados en el estrato de muestreo h.

Θ_{ra} Es el listado de estratos muestrales pertenecientes a la región r y área a.

2. Ponderador de elegibilidad

La población objetivo de la encuesta son las personas que residen en forma habitual en viviendas particulares ocupadas. Sin embargo, es posible que las viviendas cambien el tipo de uso o estado.

Por esta razón, es importante ajustar las probabilidades de selección de las viviendas seleccionadas para incorporar el hecho de que una proporción de las viviendas seleccionadas en la muestra no es elegible y para otras viviendas simplemente se desconoce su elegibilidad.

Al término del trabajo de campo, todas las viviendas seleccionadas inicialmente por el INE terminan siendo clasificadas en tres grandes grupos:

1. Elegibles: las edificaciones en que el encuestador pudo determinar que se trataban de viviendas particulares ocupadas (incluye tanto entrevistas como no entrevistas).

2. No elegibles: las edificaciones identificadas como negocios, viviendas colectivas, viviendas deshabitadas, viviendas de veraneo, viviendas destruidas, etc.

3. De elegibilidad desconocida: las edificaciones en que no se pudo determinar su estado. Este es el caso, por ejemplo, de unidades que nunca fueron enviadas a terreno, viviendas a las cuales no se pudo llegar o encontrar, y otros casos similares.

El ponderador de selección de viviendas tiene valores válidos para las viviendas elegibles, no elegibles y de elegibilidad desconocida. En lo que sigue del proceso sólo se dejarán valores válidos para las viviendas elegibles.

A continuación se describen los ajustes por elegibilidad desconocida y no elegibilidad.

2.1. Ajuste por elegibilidad desconocida

A partir de los cambios en el diseño muestral propuesto preliminarmente, un gran número de unidades fueron clasificadas como viviendas de reemplazo, lo que conllevó a una sobreestimación del número de unidades que se necesitaban para compensar la no respuesta. Lo anterior resultó en que muchas de las viviendas de “reemplazo” no fueran enviadas a terreno por lo cual fueron clasificadas como unidades con elegibilidad desconocida. Si bien, en un diseño bifásico se espera un pequeño (casi nulo) número de unidades no elegibles, a medida que mayor es el tiempo transcurrido entre el levantamiento de una y otra fase, este número puede incrementarse.

El trabajo de campo de la encuesta Casen 2013 fue realizado, principalmente, en los meses de noviembre y diciembre de 2013. Mientras que el trabajo de campo de Endisc II se realizó entre 30 de junio y 4 de septiembre de 2015, es decir, transcurrió aproximadamente un año y siete meses entre un levantamiento y otro, razón por la cual podría esperarse que el número de unidades no elegibles sea distinta de cero, más aun cuando fenómenos naturales pueden llevar a cambiar el estado de una vivienda[6]

[6] Los aluviones, terremotos, tsunamis y otros fenómenos no naturales podrían haber aumentado este número. Algunos fenómenos ocurridos en 2014 y 2015 fueron: El terremoto en Iquique en abril de 2014; Alud en las regiones de Antofagasta y Atacama en marzo de 2015; erupción de volcán Calbuco en abril de 2015; entre otros siniestros como el incendio de Valparaíso en 2014.

En este contexto, se realizó un ajuste por elegibilidad desconocida, que consiste en distribuir el peso de las unidades clasificadas de elegibilidad desconocida, de forma proporcional, entre aquellas unidades no elegibles y elegibles.

En total son 10.528 las unidades clasificadas de elegibilidad desconocida, de las cuales 10.422 son unidades que fueron asignadas como muestra de reemplazo y que no fueron enviadas a terreno.

Inicialmente, el ajuste fue realizado a nivel de comuna y área, sin embargo, debido a que bajo esta desagregación existen tamaños muestrales pequeños, donde la muestra objetivo es menor que la muestra asignada como muestra de reemplazos, se decide realizar el ajuste según región y área.

Luego, el factor de expansión de vivienda ajustado por elegibilidad desconocida se puede calcular mediante la siguiente expresión:

$$w''_{hij} = w'_{hij} \cdot \ddot{R}'_{ra} \tag{9}$$

Donde,

w'_{hij} Es el Ponderador de selección ajustado por omisión de unidades en la muestra, definido en el punto (7), para la vivienda j del conglomerado i del estrato de muestreo h.

w''_{hij} Es el ponderador ajustado por elegibilidad desconocida de la vivienda j perteneciente al conglomerado i en el estrato de muestreo h.

$\ddot{R}'_{ra}$ Es la razón utilizada en el ajuste, en la región r y área a.

$$\hat{R}'_{ra} = \frac{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in\Theta_{hi}} w'_{hij}}{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in(\Omega^{E}_{hi}\cup\Omega^{NE}_{hi})} w'_{hij}} \tag{10}$$

Siendo,

Ω^{E}_{hi} Es el listado de viviendas clasificadas como elegibles en el conglomerado i del estrato de muestreo h.

Ω^{NE}_{hi} Es el listado de viviendas clasificadas como no elegibles en el conglomerado i del estrato de muestreo h.

En esta etapa w''_{hij} poseen registros válidos aquellas viviendas donde el código de disposición final de casos corresponde a viviendas elegibles y no elegibles.

2.2. Ajuste por no elegibilidad

Originalmente, la muestra contemplaba 27.066 viviendas con sobremuestreo. Posteriormente, de ese total 12.196 fueron asignadas como muestra objetivo y 14.870 como muestra de reemplazo. De estas 10.528 resultaron de “elegibilidad desconocida” una vez finalizado el trabajo de campo. Entre las 16.538 viviendas con elegibilidad conocida, 14.568 correspondían a viviendas “elegibles” y 1.970 a viviendas “no elegibles”.

Debido a que el objetivo analítico de la encuesta es producir inferencias hacia la población que reside en viviendas particulares (elegibles), a partir de esta etapa no se consideran para fines analíticos aquellas viviendas que no conforman la población objetivo de la encuesta (viviendas no elegibles, tales como oficinas de empresas, viviendas abandonadas, viviendas de veraneo, viviendas demolidas, etc.).

Lo anterior se implementa asignando un valor blanco ("missing"), en el ponderador de selección de viviendas corregido por elegibilidad, a las viviendas con clasificación "no elegible".

3. Ponderador de no respuesta

Cómo en todas las encuesta existe problemas de localización de las unidades muestrales así como también de rechazos de parte de los entrevistados, lo que se conoce como falta de respuesta. Para subsanar posibles problemas de sesgo debido a estos se realiza un ajuste a los factores de expansión.

En este contexto, existen metodologías que sugieren como tratar este tema, para el caso de la falta de respuesta en la Endisc II se corrigió distribuyendo el peso de aquellas viviendas elegibles que no respondieron entre aquellas viviendas que sí respondieron. Para ello se ponderó el factor de expansión ajustado por elegibilidad, detallado en la fórmula (9) con una razón obtenida a partir del cociente entre la estimación del total de viviendas elegibles y la estimación del total de viviendas elegibles que responde (ambos totales obtenidos con el ponderador de elegibilidad), por cada región área. A continuación se presenta la fórmula mediante la cual se puede obtener el ponderador ajustado por falta de respuesta.

$$w'''_{hij} = w''_{hij} \cdot \ddot{R}''_{ra} \tag{11}$$

Donde,

w'''_{hij} Es el ponderador de no respuesta de la vivienda j en el conglomerado i del estrato de muestreo h.

w''_{hij} Es el ponderador de elegibilidad de la vivienda j en el conglomerado i del estrato de muestreo h.

$\ddot{R}''_{ra}$ Es la razón de la región r y área a mediante la cual se ajusta el ponderador de elegibilidad a causa de la no respuesta. A continuación se presenta su fórmula de cálculo:

$$\hat{R}'_{ra} = \frac{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in\Omega^E_{hi}} w'_{hij}}{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in\Omega^R_{hi}} w'_{hij}} \tag{12}$$

Siendo,

Ω^E_{hi} Es el listado de viviendas clasificadas como elegibles en el conglomerado i del estrato de muestreo h.

Ω^R_{hi} Es el listado de viviendas que responde en el conglomerado i del estrato de muestreo h.

A partir de esta etapa solo las viviendas que responden poseen un factor de expansión asociado, es decir las viviendas elegibles que no responden son descartadas de la base de datos.

4. Ponderador de hogar

Como no existen estimaciones del total de hogares a partir de fuentes externas, similar a las proyecciones de población, para poder estimar el total de hogares -y trabajar con información asociada a éstos- se construye un factor de hogar utilizando las proyecciones de población, sin hacer distinción de sexo y tramo de edad.

Así, el factor de hogar se construye, utilizando la base de datos de personas, ponderando el factor de vivienda, ajustado por falta de respuesta, por una razón, construida a partir del cuociente entre el total de personas determinadas según las proyecciones de población, y el total de personas estimadas con el ponderador de viviendas ajustado por falta de respuesta[7]. Ambos, el numerador y denominador se obtienen para cada región. A continuación se presenta la fórmula implementada en el cálculo del ponderador de hogar.

$$w_{hijl} = w'''_{hij} \cdot \ddot{R}_r \tag{13}$$

Donde,

t Es el índice de persona.

l Es el índice de hogar.

w_{hijl} Es el ponderador de hogar, correspondiente al hogar l de la vivienda j en el conglomerado i del estrato de muestreo h.

$\hat{R}_r$ Es la razón para la región r por la cual se debe ponderar el factor de no respuesta, detallado en la fórmula (11), para obtener el número de hogares que representa en la población un hogar entrevistado en la encuesta de Endisc II.

$$\hat{R}_r = \frac{PP_r}{\sum_{h\in\Theta_r}\sum_{i\in\Theta_h}\sum_{j\in\Omega^R_{hi}}\sum_{l\in\Omega_{hij}}\sum_{t\in\Theta_{hijl}} w'''_{hij}} \tag{14}$$

Siendo,

PP_r Las proyecciones de población total estimadas con fecha 31 de julio de 2015, para la región r.

Θ_r Es el listado de estratos muestrales pertenecientes a la región r.

Ω^R_{hij} Es el listado de hogares de la vivienda j que responde, en el conglomerado i del estrato de muestreo h.

[7] En la base de datos de personas, el ponderador de vivienda ajustado por falta de respuesta se repite tantas veces como personas se registran en cada vivienda.

Θ_{hijl} Es el listado de personas elegibles que responde en el hogar l de la vivienda j, perteneciente al conglomerado i del estrato muestral h.

Luego, si algún usuario de la base de datos de la encuesta desea realizar algún análisis a nivel de hogar, deberá seleccionar un miembro de cada hogar (podría ser el jefe de hogar, pues éste se encuentra en cada hogar) y posteriormente realizar los análisis referidos a los hogares. **Se sugiere que los análisis a nivel hogar, se realicen como máximo con desagregación geográfica regional.**

5. Ponderador de personas

Todo lo realizado anteriormente corresponde a selección de viviendas y ajustes relacionados con dichas unidades muestrales. Sin embargo, el protocolo de la encuesta Endisc II señalaba que al interior de las viviendas se debían registrar todos los hogares y en éstos todos sus miembros. En cada hogar registrado, se seleccionó una persona de 18 o más años, y en caso de identificar niños se seleccionó uno entre 2 y 17 años por hogar.

Por ello, a continuación se detalla el procedimiento para realizar el cálculo del factor de expansión de personas, diferenciado según si es un/a niño/a o adolescente, o una/a adulto/a.

5.1. Ponderador de selección de personas

El factor de personas está asociado al informante kish del hogar y además está diferenciado entre adultos/as y niños/as y adolescentes (en adelante, factor de adultos y factor de niños, respectivamente).

Probabilidad de selección de niños/as y adolescentes al interior del hogar: La probabilidad de selección de los niños está dada por el cuociente entre el número de niños/as y adolescentes seleccionados, y el total de niños/as y adolescentes de 2 a 17 años registrados en cada hogar. Por lo tanto, la probabilidad de selección de un niño en el hogar l está dada por:

$$P_{hijl}(t) = \frac{1}{q_{hijl}} \eta_{hijlt} \tag{15}$$

Donde,

t Es el índice de persona.

η_{hijlt} Es la indicatriz de niño.

$$\eta_{hijlt} = \begin{cases} 1 & Si\ persona\ t\ tiene\ entre\ 2\ y\ 17\ años \\ 0 & en\ otro\ caso \end{cases}$$

$P_{hijl}(\eta)$ Es la probabilidad de selección del niño η que pertenece al hogar l de la vivienda j en el conglomerado i del estrato de muestreo h.

q_{hijl} Es el número de niños identificados en el hogar l de la vivienda j en el conglomerado i del estrato h.

$$q_{hijl} = \sum_{t \in \Theta_{hijl}} \eta_{hijl}$$

Θ_{hijl} Es el listado de personas elegibles que responde en el hogar l de la vivienda j, perteneciente al conglomerado i del estrato muestral h.

Probabilidad de selección de adultos/as al interior del hogar: La probabilidad de selección de los adultos está dada por el cuociente entre el número de adultos/as seleccionados y el total de adultos/as registrados y elegibles[8] en cada hogar. Por lo tanto el hogar l tendrá probabilidad de selección de adultos/as dada por:

$$P_{hijl}(t) = \frac{1}{\mathrm{A}_{hijl}} \alpha_{hijlt} \tag{16}$$

Donde,

α_{hijl} Es la indicatriz de adulto del hogar l.

$$\alpha_{hijlt} = \begin{cases} 1 & Si\ \ persona\ tiene\ 18\ o\ más\ años \\ 0 & en\ otro\ caso \end{cases}$$

A_{hijl} Es el número de adultos/as identificados en el hogar l, en la vivienda j del conglomerado i en el estrato h.

$$\mathrm{A}_{hijl} = \sum\nolimits_{t \in \Theta_{hijl}} \alpha_{hijl}$$

Luego, el ponderador de personas se calcula como:

$$w_{hijlt} = w'''_{hij} \cdot \frac{1}{P_{hijl}(t)} \tag{17}$$

$$w_{hijlt} = \begin{cases} w'''_{hij} \cdot q_{hijl}, & si\ t\ es\ niño \\ w'''_{hij} \cdot \mathrm{A}_{hijl}, & si\ t\ es\ adulto \end{cases} \tag{18}$$

5.2. Suavizamiento del ponderador de selección de personas

Debido a las etapas de selección, los ponderadores iniciales a medida que se van ajustando aumentan significativamente su variabilidad. Para controlar esta variablidad, en la encuesta Endisc II se realizó un suavizamiento en los factores, para lo que se requirió como insumo, una variable de interés a medir, que permitiera identificar el umbral de suavizamiento a través de la minimización del error cuadrático medio.

[8] Se consideran elegibles todas aquellas personas de 18 y más años que componen el hogar de la vivienda seleccionada, excluyendo el personal de servicio doméstico puertas adentro.

En este contexto, la variable de interés empleada fue la tasa de discapacidad calculado por macrozona, área y tramo de edad (adulto/a, o niño/a o adolescente). Para generar el indicador señalado se utilizó la base datos a nivel de informante Kish, calculándose de la siguiente forma:

a) Se construyó una variable auxiliar (proxy del indicador que se constituiría más tarde la tasa de discapacidad según Endisc II), la cual asignaba un valor 1 si el informante Kish adulto/a tenía un valor 4 o 5 en todas o al menos una de las variables: d15, d16, d4, d35, d34, d10, d20, d40, d25, d26 y d19.
b) Se construyó una variable auxiliar la cual asignaba un valor 1 si el informante Kish niño/a o adolescente tenía un valor 4 ó 5 en todas o al menos una de las variables: n3, n1, n19, n20, n21, n11, n27, n10, n12, n13, n14 y n5.
c) Se estimó la tasa de discapacidad y su varianza usando la variable generada. Además, como el diseño de la encuesta Endisc II es complejo, se indicó al *software* que usara los pseudo estratos y pseudos conglomerados creados para la muestra y cuyo procedimiento se explica en detalle en el apartado de estimación de la varianza de este documento.

Una vez obtenida la estimación, se siguieron los siguientes pasos:

i. Se inspeccionó la existencia de valores extremos en la distribución del ponderador;
ii. Se determinaron puntos de corte a partir de los cuales realizar el suavizamiento;
iii. Se suavizaron los valores extremos identificados;
iv. Se estimó el error cuadrático medio (ECM) para los distintos puntos de corte;
v. Se eligió la opción de corte que minimizaba el ECM.

Considerando lo anterior, se analizaron 7 puntos de cortes distintos definidos como sigue[9]:

$$\beta_{gak} = k \cdot \overline{F}_{gak} \ , \ \ k \ = \ 4,5,6,7,8,9,10, \qquad (19)$$

Siendo,

$$\overline{F}_{gak} = \frac{\sum_{h\in\Theta_{ga}}\sum_{i\in\Theta_h}\sum_{j\in\Omega_{hi}^R}\sum_{l\in\Omega_{hij}}\sum_{t\in\Phi_{hijl}} w_{hijlt}}{n_{ga}}$$

Donde,

Θ_{ga} Es el conjunto de estratos pertenecientes a la macrozona g y área a.

Φ_{hijl} Es el conjunto de personas que conforman el hogar l de la vivienda j en el conglomerado i del estrato h.

n_{ga} Es el número de personas que en la base de datos posee asignado un factor de expansión en la macrozona g del área a.

Por otro lado, para realizar el suavizamiento se procedió a truncar aquellos ponderadores identificados como valores extremos de la siguiente forma:

[9] Este método de suavizamiento se realizó de forma independiente para niños/as y adolescentes, y adultos/as.

$$w_{hijlt}{}^{s} = \begin{cases} w_{hijlt} & si \;\; w_{hijlt} \leq \beta_{gak} \\ \beta_{gak} & si \;\; w_{hijlt} > \beta_{gak} \end{cases} \quad (20)$$

Si se suman todos los valores $w_{hijlt}{}^{s}$, se obtiene un total de unidades estimadas inferior que al sumar los ponderadores base, por lo tanto se debe distribuir la diferencia faltante en el resto de los ponderadores que no fueron truncados. Los pesos fueron distribuidos al interior de cada grupo gak de la siguiente forma:

$$w_{hijlt}{}^{s*} = \begin{cases} w_{hijlt} \cdot \dfrac{\left(\sum_{t \in gak} w_{hijlt} - \sum_{(t \in gak \cap w_{hijlt} > \beta_{gak})} \beta_{gak}\right)}{\sum_{gak \cap (w_{hijlt} \leq \beta_{gak})} w_{hijlt}} & , \; si \; w_{hijlt} \leq \beta_{gak} \\ \beta_{gak} & , \quad si \; w_{hijlt} > \beta_{gak} \end{cases} \quad (21)$$

Donde $w_{hijlp}{}^{s*}$es el factor suavizado.

Esto es, aquellos ponderadores identificados como valores extremos fueron truncados al valor máximo establecido ($\beta_{gak} = k * \overline{F}_{gak}$), mientras que el peso "sobrante" de los ponderadores truncados fue distribuido sobre el resto de los ponderadores.

Luego, para determinar el punto de corte donde se realizó finalmente el suavizamiento, se calculó un estadígrafo que diera cuenta del sesgo y de la variabilidad. Para esto se obtuvo el ECM asociado a la variable de interés. Como en esta encuesta se pretende caracterizar la proporción de discapacitados por dos grandes grupo de edad (adultos/as, y niños/as y adolescentes), se calculó dicho indicador por macrozona y área, y su desviación típica. De esta forma, el sesgo y el ECM se calculan como:

$$sesgo(\ddot{P}_{gak}) = P_{ga} - \ddot{P}_{gak} \quad (22)$$

$$ECM(\ddot{P}_{gak}) = Sesgo^2(\ddot{P}_{gak}) + Var(\ddot{P}_{gak}) \quad (23)$$

Siendo,

P_{ga} La proporción de discapacitados en la macrozona g área a obtenido con el factor de expansión sin truncar w_{hijlt}.

$\ddot{P}_{gak}$ La proporción de discapacitados en la macrozona g área a con el factor suavizado en el punto k.

En la Tabla 10 se muestran los resultados obtenidos, donde el valor mínimo de la mediana del ECM para cada grupo (adulto/a, o niño/a o adolescente) se alcanza cuando el ponderador es truncado a 7 veces la media, por lo que finalmente es el criterio utilizado.

En la Tabla 11 se observan las estadísticas descriptivas del ponderador original y como quedan luego del suavizamiento. Los únicos grupos que no sufrieron cambios fueron en la macrozona Metropolitana área rural tanto adultos/as como niños/as y adolescentes. Por el contrario, la macrozona norte área urbana tanto adultos/as como niño/as y adolescentes presentaron una

fuerte reducción en sus valores máximos que pasaron de 27.793 a 4.876 y de 11.117 a 3.010. Para el resto de los grupos los valores máximos disminuyeron entre un 30% y un 50% aproximadamente. El número de observaciones y el promedio se mantienen igual, y con la redistribución de diferencia entre factor inicial y el truncado hace que se cambien ligeramente los valores de los cuartiles.

Posteriormente, utilizando como insumo el ponderador base suavizado, se realiza la calibración al stock poblacional, según se detalla en la siguiente sección.

Tabla 10: Estimación del error cuadrático medio por macrozona y área, en adultos/as y en niños/as y adolescentes, según punto de suavizamiento

Grupo			Factor Original			4F			5F			6F		
	Macrozona	Área	P_g	Error estándar	Varianza	$P_{g\,(4)}$	Varianza [Pg (4)]	ECM [Pg (4)]	$P_{g\,(5)}$	Varianza [Pg (5)]	ECM [Pg (5)]	$P_{g\,(6)}$	Varianza [Pg (6)]	ECM [Pg (6)]
Adultos/as	Norte	Rural	0,1731	0,0260	0,0007	0,1801	0,0006	0,0006	0,1776	0,0007	0,0007	0,1765	0,0007	0,0007
	Norte	Urbano	0,2296	0,0160	0,0003	0,2364	0,0002	0,0003	0,2348	0,0002	0,0002	0,2339	0,0002	0,0002
	Centro	Rural	0,2979	0,0223	0,0005	0,2933	0,0004	0,0004	0,2946	0,0004	0,0004	0,2954	0,0004	0,0004
	Centro	Urbano	0,2728	0,0113	0,0001	0,2720	0,0001	0,0001	0,2725	0,0001	0,0001	0,2724	0,0001	0,0001
	Sur	Rural	0,2630	0,0255	0,0006	0,2658	0,0007	0,0007	0,2645	0,0007	0,0007	0,2641	0,0007	0,0007
	Sur	Urbano	0,2523	0,0168	0,0003	0,2617	0,0003	0,0003	0,2605	0,0003	0,0003	0,2590	0,0003	0,0003
	Metropolitana	Rural	0,2848	0,0835	0,0070	0,2848	0,0070	0,0070	0,2848	0,0070	0,0070	0,2848	0,0070	0,0070
	Metropolitana	Urbano	0,2877	0,0132	0,0002	0,2969	0,0001	0,0002	0,2971	0,0002	0,0002	0,2966	0,0002	0,0002
Niños/as y adolescentes	Norte	Rural	0,0062	0,0039	0,0000	0,0070	0,0000	0,0000	0,0067	0,0000	0,0000	0,0065	0,0000	0,0000
	Norte	Urbano	0,0669	0,0097	0,0001	0,0706	0,0001	0,0001	0,0696	0,0001	0,0001	0,0691	0,0001	0,0001
	Centro	Rural	0,1444	0,0390	0,0015	0,1138	0,0004	0,0014	0,1187	0,0005	0,0012	0,1225	0,0006	0,0011
	Centro	Urbano	0,1092	0,0129	0,0002	0,1038	0,0001	0,0001	0,1047	0,0001	0,0001	0,1055	0,0001	0,0001
	Sur	Rural	0,0628	0,0124	0,0002	0,0633	0,0002	0,0002	0,0630	0,0002	0,0002	0,0628	0,0002	0,0002
	Sur	Urbano	0,1120	0,0148	0,0002	0,1181	0,0002	0,0003	0,1166	0,0002	0,0002	0,1157	0,0002	0,0002
	Metropolitana	Rural	0,1423	0,1302	0,0170	0,1423	0,0170	0,0170	0,1423	0,0170	0,0170	0,1423	0,0170	0,0170
	Metropolitana	Urbano	0,1237	0,0122	0,0001	0,1243	0,0001	0,0001	0,1242	0,0001	0,0001	0,1246	0,0001	0,0001

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Continúa ►

Tabla 10: Estimación del Error cuadrático medio por macrozona y área, en adultos/as y en niños/as y adolescentes, según punto de suavizamiento

Grupo g			7F			8F			9F			10F			
	Macrozona	Área	$P_{g\,(7)}$	Varianza [Pg (7)]	ECM [Pg (7)]	$P_{g\,(8)}$	Varianza [Pg (8)]	ECM [Pg (8)]	$P_{g\,(9)}$	Varianza [Pg (9)]	ECM [Pg (9)]	$P_{g\,(10)}$	Varianza [Pg (10)]	ECM [Pg (10)]	Mediana
Adultos/as	Norte	Rural	0,1742	0,0007	0,0007	0,1733	0,0007	0,0007	0,1731	0,0007	0,0007	0,1731	0,0007	0,0007	0,0007
	Norte	Urbano	0,2339	0,0002	0,0002	0,2341	0,0002	0,0002	0,2343	0,0002	0,0002	0,2342	0,0002	0,0003	0,0002
	Centro	Rural	0,2958	0,0004	0,0005	0,2964	0,0005	0,0005	0,2971	0,0005	0,0005	0,2978	0,0005	0,0005	0,0005
	Centro	Urbano	0,2722	0,0001	0,0001	0,2723	0,0001	0,0001	0,2724	0,0001	0,0001	0,2727	0,0001	0,0001	0,0001
	Sur	Rural	0,2637	0,0007	0,0007	0,2633	0,0007	0,0007	0,2630	0,0006	0,0006	0,2630	0,0006	0,0006	0,0007
	Sur	Urbano	0,2577	0,0003	0,0003	0,2566	0,0003	0,0003	0,2557	0,0003	0,0003	0,2548	0,0003	0,0003	0,0003
	Metropolitana	Rural	0,2848	0,0070	0,0070	0,2848	0,0070	0,0070	0,2848	0,0070	0,0070	0,2848	0,0070	0,0070	0,0070
	Metropolitana	Urbano	0,2958	0,0002	0,0002	0,2949	0,0002	0,0002	0,2942	0,0002	0,0002	0,2935	0,0002	0,0002	0,0002
Niños/as y adolescentes	Norte	Rural	0,0063	0,0000	0,0000	0,0063	0,0000	0,0000	0,0062	0,0000	0,0000	0,0062	0,0000	0,0000	0,0000
	Norte	Urbano	0,0687	0,0001	0,0001	0,0685	0,0001	0,0001	0,0683	0,0001	0,0001	0,0682	0,0001	0,0001	0,0001
	Centro	Rural	0,1266	0,0007	0,0010	0,1286	0,0008	0,0010	0,1307	0,0008	0,0010	0,1327	0,0009	0,0011	0,0011
	Centro	Urbano	0,1062	0,0001	0,0001	0,1070	0,0001	0,0001	0,1080	0,0001	0,0001	0,1091	0,0002	0,0002	0,0001
	Sur	Rural	0,0628	0,0002	0,0002	0,0628	0,0002	0,0002	0,0628	0,0002	0,0002	0,0628	0,0002	0,0002	0,0002
	Sur	Urbano	0,1149	0,0002	0,0002	0,1144	0,0002	0,0002	0,1138	0,0002	0,0002	0,1133	0,0002	0,0002	0,0002
	Metropolitana	Rural	0,1423	0,0170	0,0170	0,1423	0,0170	0,0170	0,1423	0,0170	0,0170	0,1423	0,0170	0,0170	0,0170
	Metropolitana	Urbano	0,1250	0,0001	0,0001	0,1250	0,0001	0,0001	0,1251	0,0001	0,0001	0,1250	0,0001	0,0001	0,0001

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Tabla 11. . Estadísticas descriptivas del ponderador base y ponderador suavizado

	Grupo		Observaciones		Media		Mínimo		Máximo		MODA		Cuartil 1		Cuartil 2		Cuartil 3	
	Macrozona	Área	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T	Fact. O	Fact. T
Adultos	Norte	Rural	251	251	587	587	19	19	4.809	4.108	72	72	118	118	374	377	781	786
	Norte	Urbano	1.639	1.639	697	697	47	48	27.793	4.876	542	556	315	323	516	529	846	867
	Centro	Rural	1.000	1.000	676	676	11	11	6.841	4.729	83	83	213	214	507	509	898	901
	Centro	Urbano	3.378	3.378	839	839	53	53	9.769	5.880	555	5.870	382	384	625	628	1.016	1.022
	Sur	Rural	689	689	554	554	12	12	4.866	3.879	176	177	246	247	452	453	723	725
	Sur	Urbano	1.462	1.462	674	674	8	8	10.496	4.718	616	4.718	239	244	463	473	843	861
	Metropolitana	Rural	45	45	2.986	2.986	552	552	10.822	10.822	1.656	1.656	1.472	1.472	2.524	2.524	3.625	3.625
	Metropolitana	Urbano	3.801	3.801	1.095	1.095	91	96	31.620	7.927	1.035	7.663	436	456	725	757	1.256	1.313
Niños	Norte	Rural	433	433	453	453	11	12	7.107	3.168	214	219	170	173	298	304	571	583
	Norte	Urbano	1.489	1.489	544	544	65	66	8.831	3.808	135	3.808	246	250	378	385	637	648
	Centro	Rural	105	105	465	465	19	19	4.292	3.281	28	29	125	128	220	225	408	417
	Centro	Urbano	849	849	430	430	43	44	11.117	3.010	192	197	194	199	308	316	507	520
	Sur	Rural	286	286	383	383	22	22	2.186	2.186	313	313	191	191	291	291	495	495
	Sur	Urbano	635	635	449	449	8	8	7.459	3.145	238	245	177	182	318	327	569	584
	Metropolitana	Rural	21	21	1.940	1.940	552	552	5.879	5.879	841	841	927	927	1.519	1.519	1.931	1.931
	Metropolitana	Urbano	1.697	1.697	653	653	53	54	16.491	4.653	156	160	282	289	450	460	737	754

Nota: Fact. O = factor inicial; Fact. T = factor truncado.

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

5.3. Ponderador de calibración

Este ponderador corresponde al ponderador de personas pero calibrado a algún stock poblacional. En el caso de la encuesta de Endisc II, este factor se ajusta a los totales poblacionales obtenidos con fecha 31 de julio de 2015, por sexo y dos grandes grupos de edad (2 – 17 años, y 18 y más años). Respecto a la desagregación geográfica, se realiza por regiones, ya que no es recomendable realizar mayores desagregaciones de los grupos.

El INE genera la estimación del total de personas, para la población residente y no residente en viviendas particulares, a partir de los modelos demográficos tradicionales. Sin embargo, para fines analíticos, sólo interesan las personas residentes en viviendas particulares, por ello se excluyen de la estimación todas aquellas personas residentes en hogares colectivos como hoteles, hospitales, cárceles, etc.

Por esta razón, en la encuesta de Endisc II, al igual que en la encuesta Casen, las proyecciones poblacionales son ajustadas de acuerdo a la proporción de personas residentes en viviendas particulares estimada a partir del Censo de Población y Vivienda del año 2002. Este ajuste puede ser resumido en dos pasos:

i. A partir del Censo 2002, se estima la proporción de personas que reside en viviendas particulares, obtenida como el cuociente entre el total de personas que reside en viviendas particulares y el total poblacional, según sexo y tramo de edad (0-1 ; 2-17; 18 y más años).
ii. Luego, se ponderan las proyecciones de población según la proporción obtenida en el paso (1), y se redondean.

En la Tabla 12 se presentan las proyecciones de población estimadas al 31 de julio de 2015, mientras que en la Tabla 13 se encuentra el total de personas residentes en viviendas particulares estimadas en base a las proyecciones de población y el Censo 2002. Se observa que aproximadamente 389.000 personas residen fuera de viviendas particulares, lo que representa a un 2,2% de la población.

Luego, el ponderador se construye a partir de ajuste realizado con la razón entre las proyecciones de población y el total de personas estimado con el ponderador de selección de personas, diferenciado según sexo y tramo de edad, por región. Lo anterior puede ser resumido y explicitado en la siguiente fórmula:

$$w'_{hijlt} = w_{hijlt} \cdot \ddot{R}'''_{raf} \qquad (24)$$

Donde,

f Es el índice que diferencia las personas según sexo y dos grandes tramos de edad (2- 17 años y 18 y más años).

$\ddot{R}'''ra_f$ Es la razón con la cual se calibra el factor de expansión de personas para alcanzar los totales determinados por las proyecciones del INE. Esta razón se obtiene a partir de la siguiente expresión:

$$\widehat{R}'''_f = \frac{PP_{raf}}{\sum_{h\in\Theta_{ra}}\sum_{i\in\Theta_h}\sum_{j\in\Omega_{hi}^R}\sum_{l\in\Omega_{hij}}\sum_{t\in\Theta_{hijl}} w_{hijlt}} \quad (25)$$

Siendo,

PP_{raf} Son las proyecciones de población por sexo y tramo de edad f, estimadas con fecha 31 de julio de 2015, obtenidas por región y área.

Tabla 12: Total poblacional, según proyecciones de población estimadas al 31 de julio de 2015

Región	Hombres 0 - 1 año	Mujeres 0 - 1 año	Hombres 2 - 17 años	Mujeres 2 - 17 años	Hombres 18 y más años	Mujeres 18 y más años	Total
Total País	**256.402**	**246.883**	**2.046.037**	**1.973.254**	**6.542.308**	**6.811.705**	**17.876.589**
I de Tarapacá	5.742	5.521	43.686	41.807	133.493	120.835	351.084
II de Antofagasta	10.127	9.740	75.864	72.770	230.287	209.095	607.883
II de Atacama	4.638	4.476	36.135	34.773	107.768	103.073	290.863
IV de Coquimbo	11.228	10.812	90.860	87.323	279.466	290.931	770.620
V de Valparaíso	24.962	24.009	202.520	193.478	684.566	722.545	1.852.080
VI de O'Higgins	12.527	12.067	106.042	102.424	348.221	344.723	926.004
VII del Maule	14.202	13.672	120.436	115.560	386.129	398.076	1.048.075
VIII del Biobío	28.682	27.628	238.953	230.230	766.126	808.478	2.100.097
IX de la Araucanía	14.754	14.250	121.964	117.092	363.497	379.395	1.010.952
X de Los Lagos	13.130	12.673	106.364	102.058	332.702	321.958	888.885
XI de Aysén	1.765	1.684	13.995	13.439	42.009	37.149	110.041
XII de Magallanes y La Antártica	2.195	2.091	17.963	16.819	64.572	57.570	161.210
XIII Metropolitana	104.588	100.695	805.961	782.494	2.601.934	2.802.495	7.198.167
XIV de Los Ríos	5.320	5.123	44.029	42.268	141.795	146.210	384.745
XV de Arica y Parinacota	2.542	2.442	21.265	20.719	59.743	69.172	175.883

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

Tabla 13: Total de personas residentes en viviendas particulares, según proyecciones de población ajustadas con información del Censo 2002

Región	Hombres 0 - 1 año	Mujeres 0 - 1 año	Hombres 2 - 17 años	Mujeres 2 - 17 años	Hombres 18 y más años	Mujeres 18 y más años	Total
Total País	**252.473**	**243.701**	**2.016.798**	**1.946.362**	**6.291.093**	**6.737.106**	**17.487.533**
I de Tarapacá	5.629	5.448	43.227	41.485	115.873	118.848	330.510
II de Antofagasta	9.848	9.580	75.143	72.295	199.629	205.539	572.034
II de Atacama	4.554	4.428	35.735	34.403	99.597	101.928	280.645
IV de Coquimbo	11.100	10.712	88.886	85.190	269.417	288.232	753.537
V de Valparaíso	24.594	23.668	200.733	191.977	661.036	714.399	1.816.407
VI de O'Higgins	12.407	11.959	105.048	101.333	338.154	342.354	911.255
VII del Maule	14.000	13.532	118.844	113.757	375.468	394.852	1.030.453
VIII del Biobío	28.298	27.310	234.942	226.540	742.638	800.722	2.060.450
IX de la Araucanía	14.481	14.045	116.767	111.609	352.907	374.626	984.435
X de Los Lagos	12.886	12.443	102.256	98.606	317.479	317.326	860.996
XI de Aysén	1.734	1.655	13.377	12.973	36.669	36.375	102.783
XII de Magallanes y La Antártica	2.136	2.049	17.639	16.702	55.091	56.610	150.227
XIII Metropolitana	103.123	99.428	801.846	779.021	2.539.097	2.773.637	7.096.152
XIV de Los Ríos	5.227	5.048	41.508	39.985	135.298	144.062	371.128
XV de Arica y Parinacota	2.456	2.396	20.847	20.486	52.740	67.596	166.521

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

IV. ESTIMACIÓN DE VARIANZA COMPLEJA

El diseño muestral de la encuesta de Endisc II, es un diseño bifásico complejo, donde sus unidades provienen de haber participado previamente en Casen 2013, las cuales estaban clasificadas en estratos y conglomerados de acuerdo a su ubicación geográfica. Sin embargo, cuando se analizó como quedaron compuestos éstos para la muestra seleccionada de la encuesta Endisc II, se observó que existían estratos de muestreo que poseían solo un conglomerado y que el número de unidades seleccionadas y que responde en cada conglomerado era desigual y muy variable. Por tal razón, se agruparon tanto estratos como conglomerados a fin de que los nuevos pseudo-estratos y pseudo-conglomerados constituidos, garanticen la estimación de varianzas en cada uno, y de ésta forma no sean subestimados los errores, a fin de minimizar los problemas señalados anteriormente, y siguiendo las recomendaciones internacionales.

La Tabla 14 informa, para la muestra de la encuesta de Endisc II, cuántos estratos habrían sido de acuerdo a Casen 2013 y, en contraste, cuántos son una vez que se han agregado para cumplir con lo antes expuesto.

Tabla 14. . Número de estratos de muestreo, Número de VarStrat Casen y VarStrat en encuesta de Endisc II, según región.

Región	Total estratos Muestreo Endisc II	Total VarStrat Casen	Total VarStrat Endisc II
Total País	**237**	**188**	**143**
I de Tarapacá	3	3	3
II de Antofagasta	3	3	3
III de Atacama	6	5	3
IV de Coquimbo	10	8	8
V de Valparaíso	25	18	15
VI de O´Higgins	28	19	9
VII de Maule	22	18	13
VIII de Biobío	34	26	22
IX de La Araucanía	30	23	12
X de Los Lagos	20	13	10
XI de Aysén	4	4	3
XII de Magallanes y La Antártica	2	2	2
XIII Metropolitana	36	32	30
XIV de Los Ríos	12	12	8
XV de Arica y Parinacota	2	2	2

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

A continuación se detallan los procedimientos y criterios utilizados en la creación de dichas variables.

1. Creación de pseudo estratos

Los estratos ficticios o pseudo-estratos fueron construidos con el objetivo de corregir los problemas generados por la existencia de estratos con solo un conglomerado (estratos unitarios), esto es, la subestimación de la varianza de cualquier variable de interés.

Los pseudo-estratos son construidos a través de la agrupación de dos o más estratos originales, de acuerdo a un patrón u ordenamiento jerárquico de variables geográficas o de tamaño, de modo que estos contengan al menos dos conglomerados, los que a su vez deberán contener al menos 9 unidades que responden en su interior.

La composición de los pseudo-estratos en la encuesta Endisc II es informada en Tabla 15. Se observa que la mayoría (51%) de los estratos están compuestos por dos o tres conglomerados. El tamaño máximo de los pseudo-estratos es de 23 conglomerados por estrato.

Tabla 15: Frecuencia del tamaño de conglomerados por pseudo-estrato

Total de conglomerados por pseudo-estratos	Frecuencia absoluta	Frecuencia relativa
2	40	28,2%
3	31	21,8%
4	9	6,3%
5	14	9,9%
6	8	5,6%
7	7	4,9%
8	3	2,1%
9	5	3,5%
10	9	6,3%
11	2	1,4%
12	3	2,1%
13	1	0,7%
14	3	2,1%
15	1	0,7%
16	1	0,7%
18	1	0,7%
20	1	0,7%
22	1	0,7%
23	2	1,4%

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

2. Creación de pseudo conglomerados

Los conglomerados ficticios o pseudo-conglomerados fueron construidos con el objetivo de reducir los problemas generados a causa de la diversidad de tamaños de los conglomerados (número de unidades que participa en ellos), pues a mayor variabilidad en el tamaño de los conglomerados, la varianza de los estimadores tiende a incrementarse y volverse más inestable.

Los pseudo-conglomerados fueron creados a partir de un ordenamiento jerárquico, según comuna y total de unidades que responde, al interior de cada pseudo-estrato. Luego, se unieron los conglomerados a fin de que estos en conjunto reunieran, de ser posible, 9 unidades aproximadamente. La composición de los pseudos-conglomerdos en la encuesta de Endisc II se informa en la Tabla 16. Se observa que el tamaño de los conglomerados es de mínimo 8 y máximo 34 viviendas, con la excepción de dos conglomerados uno de 3 viviendas y otro de 5 viviendas, que no pudieron agregarse, para respetar que cada estrato debía tener al menos dos conglomerados para garantizar la estimación de la varianza correspondiente (esto ocurrió en las regiones de

Tarapacá y Antofagasta, específicamente en la zona rural). El 60% los conglomerados tiene tamaños de 10 a 17 viviendas, y solo un 7% más de 22.

Tabla 16: Frecuencia del número de viviendas por pseudo-conglomerado

Total de viviendas por pseudo-conglomerado	Frecuencia absoluta	Frecuencia relativa
3	1	0,1%
5	1	0,1%
8	2	0,3%
9	23	2,9%
10	42	5,4%
11	50	6,4%
12	70	8,9%
13	84	10,7%
14	68	8,7%
15	71	9,0%
16	79	10,1%
17	60	7,6%
18	64	8,2%
19	34	4,3%
20	41	5,2%
21	26	3,3%
22	18	2,3%
23	14	1,8%
24	11	1,4%
25	8	1,0%
26	5	0,6%
27	1	0,1%
28	2	0,3%
29	7	0,9%
30	2	0,3%
34	1	0,1%

Fuente: INE, Informe preliminar Diseño muestral Endisc II, Enero 2016.

3. Estimación de variables y varianzas

Una vez creados los pseudo-estratos y pseudos-conglomerados, se calculó la variable de interés para el estudio de Endisc II, conforme a la metodología de medición de la discapacidad del estudio, usando como ponderador el factor de expansión generado.

Dicha estimación se realizará hasta el nivel regional, en el caso de la población adultos/as (de 18 ó más años de edad). Próximamente, se realizará también a nivel nacional para la población de niños/as y adolescentes (2 a 17 años) (cuyo tamaño muestral es menor al de la población adulta entrevistada en esta encuesta, puesto que en cada hogar seleccionado se debió entrevistar un adulto/a, y en hogares con niños/as o adolescentes también debía seleccionarse y entrevistarse un niño/a o adolescente).

4. Resultados a nivel zona (urbana y rural) nacional y nacional.

La Tabla 17 presenta las estimaciones, error estándar, error absoluto, intervalo de confianza a nivel nacional y por zona del porcentaje de personas adultas (de 18 años o más) por situación y grado de discapacidad.

En la encuesta Endisc II, los márgenes de error de estas estimaciones fueron determinados utilizando los parámetros Varstrat_N, Varunit_N y Factor_Persona (factor de expansión de personas), considerando la utilización de pseudo-estratos y pseudo-conglomerados. Estos cálculos se realizaron, paralelamente, usando los programas SPSS v.22 y Stata v.13.

Tabla 17: Estimaciones, error estándar, error absoluto, intervalo de confianza a nivel nacional y por zona del porcentaje de personas adultas (de 18 años ó más) por situación y grado de discapacidad, encuesta Endisc II, 2015.

Zona	Personas sin Situación de Discapacidad (PsSD)					Personas en Situación de Discapacidad (PeSD)[1]				
	Estimación (%)	Error estándar (%)	Error Absoluto (puntos porcentuales)	Intervalo de confianza (95%)		Estimación (%)	Error estándar (%)	Error Absoluto (puntos porcentuales)	Intervalo de confianza (95%)	
				Inferior (%)	Superior (%)				Inferior (%)	Superior (%)
URBANO	80,1	0,6	1,2	78,9	81,3	19,9	0,6%	1,2	18,7	21,1
RURAL	79,1	1,8	3,6	75,3	82,5	20,9	1,8%	3,6	17,5	24,7
País	80,0	0,6	1,1	78,8	81,1	20,0	0,6%	1,1	18,9	21,2

Zona	Personas en Situación de Discapacidad Leve a Moderada[2]					Personas en Situación de Discapacidad Severa[3]				
	Estimación (%)	Error estándar (%)	Error Absoluto (puntos porcentuales)	95% de intervalo de confianza		Estimación (%)	Error estándar (%)	Error Absoluto (puntos porcentuales)	95% de intervalo de confianza	
				Inferior (%)	Superior (%)				Inferior (%)	Superior (%)
URBANO	11,5	0,5	0,9	10,6	12,5	8,4	0,4	0,8	7,6	9,1
RURAL	12,9	1,8	3,6	9,7	16,9	8,0	0,8	1,6	6,6	9,8
Pais	11,7	0,5	0,9	10,8	12,6	8,3	0,4	0,7	7,6	9,0

[1] Corresponde a Personas con dificultades severas de capacidad (según índice basado en la capacidad). Incluye Personas en Situación de Discapacidad Leve a Moderada y Personas en Situación de Discapacidad Severa (graduadas según índice basado en desempeño).

[2] Corresponde a Personas con dificultades severas de capacidad y problemas leves a moderados de desempeño.

[3] Corresponde a Personas con dificultades severas de capacidad y problemas severos de desempeño.

Nota: Estimaciones realizadas por Instituto Nacional de Estadísticas y Ministerio de Desarrollo Social.

Fuente: Ministerio de Desarrollo Social, Encuesta del Segundo Estudio Nacional de la Discapacidad, 2015

5. Programas Computacionales

Se presentan a continuación las sintaxis realizadas en programas SPSS y STATA (utilizando muestras complejas) de la estimación, error estándar, intervalo de confianza al 95% de confianza a nivel nacional y por zona del porcentaje de personas adultas (de 18 años ó más) por situación y grado de discapacidad, a partir de los datos de la encuesta de Endisc II, 2015.

5.1 Sintaxis en SPSS

*** Diseño Plan de Muestras Complejas**

```
CSPLAN ANALYSIS
  /PLAN FILE='C:\Base_Datos2015_EndiscII\Plan_VarUnit_Varstrat_EndiscII.csaplan'
  /PLANVARS ANALYSISWEIGHT=Factor_Persona
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=Varstrat_N CLUSTER=Varunit_N
  /ESTIMATOR TYPE=WR.
```

Estimación de distribución de personas según situación y grado de discapacidad (personas en situación de discapacidad leve a moderada, personas en situación de discapacidad severa, y personas sin situación de discapacidad), por zona

```
* Frecuencias de muestras complejas.
CSTABULATE
  /PLAN FILE='C:\Base_Datos2015_EndiscII\Plan_VarUnit_Varstrat_EndiscII.csaplan'
  /TABLES VARIABLES=Discapacidad
  /SUBPOP TABLE=zona DISPLAY=LAYERED
  /CELLS TABLEPCT
  /STATISTICS SE CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Estimación de distribución de personas según situación de discapacidad (personas en situación de discapacidad y personas sin situación de discapacidad), por zona

```
* Frecuencias de muestras complejas.
CSTABULATE
  /PLAN FILE='C:\Base_Datos2015_EndiscII\Plan_VarUnit_Varstrat_EndiscII.csaplan'
  /TABLES VARIABLES=Discapacidad_dummy
  /SUBPOP TABLE=zona DISPLAY=LAYERED
  /CELLS TABLEPCT
  /STATISTICS SE CIN(95)
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

5.2 Sintaxis en STATA

Estimación de distribución de personas según situación y grado de discapacidad (personas en situación de discapacidad leve a moderada, personas en situación de discapacidad severa, y personas sin situación de discapacidad), por zona

```
svyset Varunit_N [w=Factor_Persona], strata(Varstrat_N)
svy: prop Discapacidad, level(95)
svy: prop Discapacidad, level(95) over(zona)
```

Estimación de distribución de personas según situación de discapacidad (personas en situación de discapacidad y personas sin situación de discapacidad), por zona

svyset Varunit_N [w=Factor_Persona], strata(Varstrat_N)
svy: prop Discapacidad_dummy, level(95)
svy: prop Discapacidad_dummy, level(95) over(zona)

V. BIBLIOGRAFÍA

- INE Chile. (2006). *Encuesta Nacional de Empleo*. Diseño muestral.
- Hansen, M; Hurwitz, W; Madow, W.(1953). Sample *Survey Method and Theory*. Volume II-Theory. John Wiley & Sons, Inc. New York.
- Cochran, W. (1977). *Sampling Techniques*. Third Edition. John Wiley & Sons, Inc. New York.
- Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, Inc. New York.
- Groves, R. (1989, 2004). *Survey Errors and Survey Cost*. John Wiley & Sons, Inc. Hoboken. New Jersey.
- Levy, P; Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*. Third Edition. Wiley & Sons, Inc. Canada.
- Lehtonen, R; Pahkinen, E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. John Wiley & Sons, Ltd. England.
- Rao, P. (2000). *Sampling Methodologies with Applications*. Chapman & Hall/CRC.US.
- Naciones Unidas. (2009). *Encuesta de hogares en los países de desarrollo y transición.* Departamento de Asuntos Económicos. División Estadística. New York.