Muestra representativa de la base del Registro Social de Hogares (RSH)

Contexto

Conforme a lo establecido en el Art. 6 del Decreto Supremo N°22 de 2015, en el marco del Registro Social de Hogares (RSH), la **Subsecretaría de Evaluación Social (SES)** deberá "d) ejecutar los procesos de análisis de la calidad de los datos y del comportamiento de la información contenida en el Registro Social de Hogares", "g) Administrar el proceso de tratamiento de datos personales de conformidad con lo establecido en el Decreto Supremo N°160 de 2007" y "h) Determinar por medio de uno o más actos administrativos las reglas técnicas de utilización de la información y productos del Sistema".

En su calidad de administradora del Registro o base de datos, esta Subsecretaría debe dar aplicación a los principios generales contenidos en la Ley N°19.628, sobre protección de la vida privada. En particular, esta minuta describe cómo se busca mantener un equilibrio entre obtener una muestra representativa del país del Registro Social de Hogares (RSH) y proteger la identidad de las personas incluidas en el Registro.

Para lograr una muestra representativa, se realiza un muestreo del RSH que logre representatividad a nivel regional y por tramo de la Calificación Socioeconómica (CSE). En cuanto a la protección de la identidad, se implementan medidas para garantizar la indeterminación de las personas en las unidades de análisis (comunas, unidades vecinales, hogares, etc.), evitando así la re-identificación de personas a partir de la muestra anonimizada con un identificador único correlativo a nivel de persona y hogar.

Metodología de construcción de la muestra

El diseño de la muestra tiene como objetivo, garantizar la representatividad y confidencialidad de los datos contenidos en el Registro Social de Hogares (RSH), por esto, se optó por una metodología de muestreo aleatorio estratificado. Esta metodología asegura una representación adecuada a nivel de región y tramo de la Calificación Socioeconómica (CSE) del país. Por lo que cada estrato representa una combinación única de una región específica y un tramo CSE, permitiendo así una segmentación detallada y precisa de la población.

El objetivo fue lograr una representatividad del 1% de los hogares a nivel nacional presentes en el RSH de la segunda quincena de junio de 2024. Para determinar el tamaño adecuado de la muestra para cada estrato, se utilizó la fórmula de tamaño de muestra ajustada para poblaciones finitas, asegurando así que cada estrato esté adecuadamente representado.

$$n = \frac{NZ^2pq}{E^2(N-1) + Z^2pq}$$

donde:

- N es el tamaño de la población (número de hogares en el estrato).
- Z es el valor z correspondiente al nivel de confianza deseado (1,96 para un 95% de confianza).
- p es la proporción esperada (0,5 para máxima variabilidad).
- q es $1 - p$.
- E es el margen de error deseado.
- n es el 1% de los datos, que corresponde a 90.467 hogares aproximadamente.

Una vez determinado el tamaño de la muestra total, se procedió a distribuir esta muestra entre los distintos estratos (combinación entre región y tramo de CSE) en proporción al tamaño de cada estrato en la población total. Para ello, se utilizó la función de muestreo aleatorio, asegurando que la selección fuera independiente y sin reemplazo. Esto garantiza que cada hogar dentro de un estrato tenga la misma probabilidad de ser seleccionado. Después de seleccionar los hogares, se generó la muestra final que consiste en las personas que pertenecen a dichos hogares. Luego, utilizando simulaciones de Monte Carlo y pruebas estadísticas que describirán más abajo, se coteja y corrobora que se mantuvo la representatividad de la muestra a nivel individual.

Limitaciones y recomendaciones de uso

Aunque la metodología de muestreo aleatorio estratificado utilizada para construir la muestra del Registro Social de Hogares (RSH) tiene numerosas ventajas, es importante reconocer sus limitaciones, para hacer un uso adecuado de la información contenida en la muestra:

- Representatividad condicional: Si bien la muestra se diseñó para ser representativa a nivel nacional y en términos de región y tramo CSE, puede no ser igualmente representativa para subgrupos más pequeños o variables no consideradas en la estratificación original. Esto podría afectar la generalización de los resultados a niveles más específicos.

- Errores de muestreo**:** Como en cualquier proceso de muestreo, existe una probabilidad de error de muestreo, especialmente en estratos con un número muy pequeño de hogares. Esto puede llevar a estimaciones sesgadas o imprecisas para esos estratos.

- Actualización de los datos: El RSH es una base de datos que cambia con el tiempo. La muestra seleccionada ha sido diseñada para ser representativa del RSH en el momento específico de la extracción de datos. Cualquier cambio en la composición

del RSH posterior a la selección de la muestra no estará reflejado en los análisis realizados con dicha muestra.

- Inferencia estadística: La capacidad de realizar inferencia estadística a partir de la muestra no es algo que se pueda asegurar, por lo tanto, es responsabilidad de cada investigadora o investigador realizar las consultas y pruebas que estime conveniente para tales fines.

Estadísticas descriptivas relevantes: t-student y chi-cuadrado

- Se realizó una prueba de t-student sobre la variable 'ESCOLARIDAD' con el objetivo de comparar su promedio con el de la población. Con un p-valor de 0,71 al 95% de confianza, por lo que se concluyó que no existe evidencia suficiente para decir que el promedio muestral y la media poblacional del RSH son distintos.

- Se realizó una prueba de chi-cuadrado para las variables categóricas, con el fin de determinar si la distribución de las frecuencias observadas es igual a la distribución de la población del RSH.

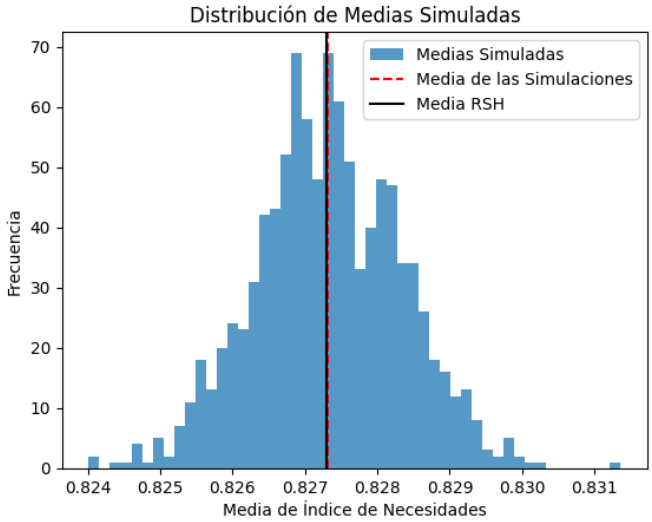
    - Para las variables de región, tramo CSE, rango etario, sexo, parentesco, nacionalidad, trabajo, discapacidad y dependencia, medios de alto valor, tipo educación, asistencia, situación de sitio, situación de vivienda, distribución de agua, sistema de baño, dormitorios ocupados y principal ocupante, el resultado de la prueba arrojó un p-valor mayor a 0,05 al 95% de confianza, por lo que se concluyó que no existe evidencia suficiente para decir que la distribución muestral es distinta a la poblacional del RSH.

    - Para las variables de pertenencia a pueblos originarios, tipo de vivienda, fuente de agua y zona de grupo familiar, el p-valor fue menor o igual a 0,05 al 95% de confianza, por lo que se concluyó que existe evidencia suficiente para decir que la distribución muestral no es igual a la distribución poblacional del RSH.

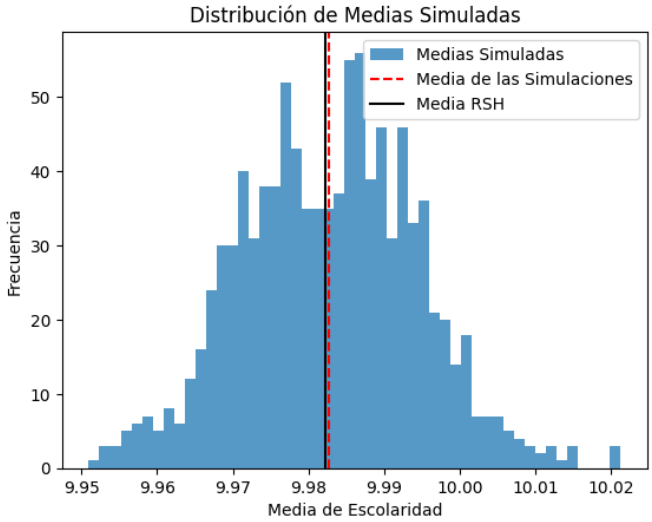
Estadísticas descriptivas relevantes: simulaciones de Montecarlo.

- Simulaciones de Monte Carlo: A partir de múltiples muestras aleatorias, se puede analizar si las distribuciones de las medias o proporciones simuladas se comparan con las de la población. Se generaron 1.000 muestras aleatorias de la población y se comparó su distribución con la del RSH y se evaluó si se encontraba dentro de los intervalos de confianza generados por las simulaciones. Dentro de las variables analizadas se encuentran:

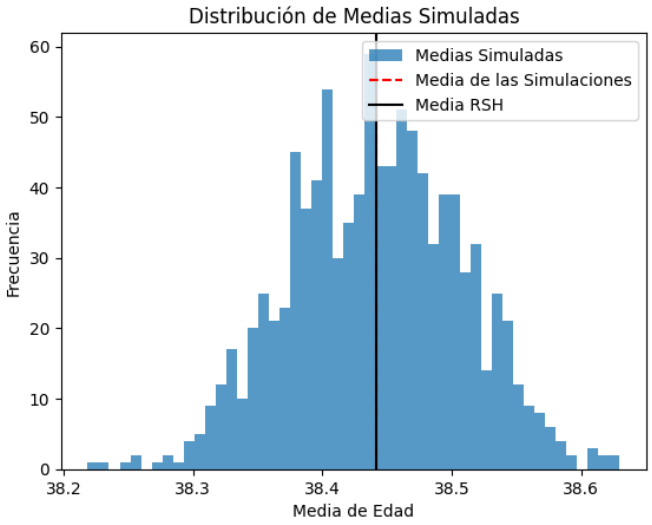
- Discapacidad – dependencia – necesidades educativas especiales (NEE): Donde la media de las medias tiene un valor de 0.82731 y la media del RSH de 0.82729. El intervalo de confianza al 95% se encuentra entre [0.82536 – 0.82928].



- Escolaridad: Donde la media de las medias tiene un valor de 9.98273 y la media del RSH de 9.98223. El intervalo de confianza al 95% se encuentra entre [9.96003 – 10.00521].



- Edad: Donde la media de las medias tiene un valor de 38.441224 y la media del RSH de 38.441994. El intervalo de confianza al 95% se encuentra entre [38.31644 - 38.56485].



Cabe mencionar que también se analizaron variables externas a la muestra, pero presentes en el RSH para asegurar que ésta se encuentre correctamente construida.

Descripción de la estructura, diccionario de códigos y campos

La base de datos del Registro Social de Hogares contiene información proporcionada por un representante de la familia, así como también, información obtenida de bases administrativas oficiales de las que dispone el Ministerio. Las variables contenidas caracterizan tanto a las personas como a los hogares de acuerdo con su tramo socioeconómico. En la muestra, cada registro corresponde a una persona que pertenece a uno de los hogares seleccionados según la metodología muestral explicada previamente.

La estructura de la base de datos es:

- Variables de caracterización de las personas: identificador ficticio, rango etario, sexo, parentesco en relación con la jefatura del hogar, nacionalidad, pertenencia a un pueblo originario, discapacidad, dependencia o necesidades educativas especiales, educación y trabajo.
- Variables de caracterización del hogar: identificador ficticio del hogar, tramo de calificación socioeconómica y posesión de medios de alto o muy alto valor.
- Variables de caracterización de la vivienda: condiciones de la vivienda, condiciones del sitio, tenencia de la propiedad y disponibilidad de servicios básicos.
- Variables territoriales: región y zona a la que pertenece el hogar.

Con el objetivo de la protección de los datos y la imposibilidad de re-indentificar a las personas, se tomó como medida adicional la agrupación de las posibles respuestas en algunas preguntas realizadas a través del formulario RSH (variables). En general, la aplicación se realiza a aquellas alternativas que tienen menos frecuencia en el total de la población.

Para comprender mejor el tratamiento realizado a los datos se presenta el siguiente caso. En el formulario original del Registro Social de Hogares la pregunta respecto del parentesco respecto de la jefatura de hogar presenta 13 alternativas de respuestas, que se detallan a continuación:

1. Jefe(a) de Hogar
2. Cónyuge o pareja
3. Hijo(a) de ambos
4. Hijo(a) sólo del jefe(a) de hogar
5. Hijo(a) sólo del cónyuge o pareja
6. Padre o madre
7. Suegro o Suegra
8. Yerno o nuera
9. Nieto(a)
10. Hermano(a)
11. Cuñado(a)
12. Otro familiar
13. No familiar

Sin embargo, se consideró apropiado realizar nuevas agrupaciones. Un caso, reagrupa las opciones de 3 a 5, que corresponden a los hijos que integran una familia. Mientras, que otro caso, las opciones desde la 6 a la 13, que son las alternativas con menor frecuencia queden en una sola categoría.

Las nuevas agrupaciones quedan de la siguiente forma:

1. Jefe de hogar
2. Cónyuge o pareja
3. Hijos/as de ambos o de sólo uno
4. Otros familiares directos o indirecto (Padres / Hermanos / Nietos / Suegros / Yerno- nuera / Cuñados/ Otros)
5. Sin información

Para esta publicación, las variables ajustadas corresponden a:

- Rango etario
- Parentesco
- Variables de vivienda (Tipo de vivienda, situación de sitio, situación de vivienda, fuente de agua, distribución de agua, sistema de baño, dormitorios ocupados y principal ocupante)

El detalle de los campos y opciones de respuesta se encuentra en el Anexo1.

Anexo 1: Diccionario de datos “Base muestral de RSH”

Glosa variable	Variable	Tipo	Opciones de respuesta	Glosa de respuesta
Periodo de referencia del RSH	PERIODO_RSH	cadena	AAAAMMDD	Indica año, mes, dia
Identificador de persona ficticio	RUNF	cadena		
Identificador de grupo familiar ficticio	FOLIOF	cadena		
Rango etario	RANGO_ETARIO	categoria	0-3	
			4-5	
			6-13	
			14-17	
			18-24	
			25-44	
			45-64	
			65+	
Sexo	PER_SEXO_ID	categoria	1	Masculino
			2	Femenino
Parentesco en relación con el jefe(a) de hogar.	PARENTESCOID	categoria	0	Sin información
			1	Jefe(a) de Hogar
			2	Cónyuge o pareja
			3	Hijo(a) de ambos o de uno sólo
			4	Otros familiares directos o indirecto (Padres / Hermanos / Nietos / Suegros / Yerno- nuera / Cuñados / Otros)
Nacionalidad	NACIONALIDAD_ID	categoria	0	Sin información
			1	Chileno
			2	Extranjero
Pertenencia a un pueblo originario	FL_PPO	categoria	0	No pertenece a un pueblo originario
			1	Pertenece a un pueblo originario
Situación laboral	TRABAJO	categoria	0	No trabaja
			1	Trabaja
Presencia de discapacidad o dependencia o necesidades educativas especiales	DISC_DEP_NEE	categoria	0	No presenta discapacidad, ni dependencia ni necesidades educativas especiales
			1	Presenta algun nivel de : discapacidad, dependencia o necesidades educativas especiales
Dispone de medios de alto valor o muy alto valor	MEDIOS_ACTIVOS	categoria	0	No tiene medios de alto valor o muy alto valor
			1	Presenta al menos un medio de alto valor o de muy alto valor
Años de escolaridad	ESCOLARIDAD	número		años de escolaridad, rango (0 a 21)
Niveles de educación	TIPO_EDUC_CURSO	categoria	Sin escolaridad	
			Básica completa	
			Básica incompleta	
			Media completa	
			Media incompleta	
			Superior completa	
			Superior incompleta	
Asistencia escolar en educación básica o media	ASISTENCIA	categoria	0	No ha asistido a algun establecimiento escolar (niveles básico y media)
			1	Ha asistido a algun establecimiento escolar (niveles básico y media)

Glosa variable	Variable	Tipo	Opciones de respuesta	Glosa de respuesta
Tipo de vivienda en la que reside	TIPOVIVIENDA	categoria	1	Casa
			2	Departamento
			3	Pieza dentro de vivienda
			4	Hospedería / Vivienda colectiva / Residencial, pensión / Mejora, mediagua / Rancho, ruca o choza/Vivienda de desecho /Caleta o punto de calle / Sin Información
Situación de sitio respecto de la propiedad	SITUACIONSITIO	categoria	1	Propio pagado o pagándose
			2	Arrendado
			3	Cedido, uso gratuito
			4	Usufructo (solo uso y goce) / Ocupación irregular / Poseedor irregular / Sin Información
Situación de la vivienda respecto de la propiedad	SITUACIONVIVIENDA	categoria	1	Propia pagada o pagándose
			2	Arrendada
			3	Cedida uso gratuito
			4	Usufructo (solo uso y goce) / Ocupación irregular / Poseedor irregular / Sin Información
Fuente de agua que abastece a la vivienda	FUENTEAGUA	categoria	1	Red pública con medidor propio
			2	Red pública con medidor compartido
			3	Red pública sin medidor / Pozo o noria /Río, vertiente o estero / Otra fuente (no potable) / Sin información
Distribución de agua al interior de la vivienda	DISTRIBUCIONAGUA	categoria	0	No tiene sistema, la acarrea / Sin Información
			1	Con llave dentro de la vivienda
			2	Con llave dentro del sitio, pero fuera de la vivienda
Sistema de baño que posee la vivienda	SISTEMABANO	categoria	0	No tiene / Sin información
			1	WC Conectado a alcantarillado
			2	WC Conectado a fosa séptica
			3	Cajón, letrina o WC sobre pozo negro, acequia, canal u otro sistema
Número de dormitorios ocupados	DORMITORIOSOCUPADOS	categoria	0	Sin información
			1	1 dormitorio
			2	2 dormitorios
			3	3 dormitorios
			4	4 o más dormitorios
Principal ocupante de la vivienda o sitio	PRINCIPALOCUPANTE	categoria	1	Principal ocupante
			2	No tiene o sin información
Código de región	C_REGION	categoria	1	Región de Tarapacá
			2	Región de Antofagasta
			3	Región de Atacama
			4	Región de Coquimbo
			5	Región de Valparaíso
			6	Región del Libertador General Bernardo O'Higgins
			7	Región del Maule
			8	Región del Biobío
			9	Región de La Araucanía
			10	Región de Los Lagos
			11	Región de Aysén del General Carlos Ibáñez del Campo
			12	Región de Magallanes y la Antártica Chilena
			13	Región Metropolitana de Santiago
			14	Región de Los Ríos
			15	Región de Arica y Parinacota
			16	Región del Ñuble
Tipo de zona	GRUPFAMI_C_ZONA	categoria	1	Urbano
			2	No urbano
Tramo de calificación socieconómica	TRAMO_CSE	categoria	40	Hogar calificados en el tramo 0-40 de mayor vulnerabilidad
			50	Hogar calificados en el tramo 41-50 de mayor vulnerabilidad
			60	Hogar calificados en el tramo 51-60 de mayor vulnerabilidad
			70	Hogar calificados en el tramo 61-70 de mayor vulnerabilidad
			80	Hogar calificados en el tramo 71-80 de menor vulnerabilidad
			90	Hogar calificados en el tramo 81-90 de menor vulnerabilidad
			100	Hogar calificados en el tramo 91-100 de menor vulnerabilidad