CENTRO MICRO DATOS

Departamento de Economía

Universidad de Chile

DISEÑO MUESTRAL
ENCUESTA DE OCUPACIÓN Y DESOCUPACIÓN EN EL GRAN SANTIAGO

2016

Departamento de Economía
Universidad de Chile

Diagonal Paraguay 257, torre 26,
of. 1601 · Santiago, Chile

(56-2) 2978 34 21

www.microdatos.cl

FACULTAD ECONOMÍA Y NEGOCIOS UNIVERSIDAD DE CHILE

UNIVERSIDAD DE CHILE

CENTRO MICRO DATOS
Departamento de Economía
Universidad de Chile

1 Contenido

1 Presentación

En el presente documento, el Centro de Microdatos del Departamento de Economía de la Universidad de Chile, detalla el Diseño Muestral de la Encuesta de Ocupación y Desocupación en el Gran Santiago (EOD).

El objetivo de este documento es facilitar a los usuarios información de la Encuesta, para lo cual entrega antecedentes conceptuales y metodológicos, así como su estructura operativa. En ese contexto, se incluye una breve revisión de la historia de EOD, la descripción de requerimientos y diseño de la muestra, las diferentes etapas de selección, el esquema de rotación y sus implicancias sobre el cálculo de estimadores (medias y varianzas) de las situación ocupacional.

2 Reseña Histórica

La primera EOD en el Gran Santiago fue realizada en octubre de 1956 y fue financiada por el Banco Central de Chile, ASIMET, la Fundación Rockefeller y el Gobierno de Chile.

El diseño metodológico estuvo a cargo del entonces Director del Instituto de Economía de la Universidad de Chile, el economista estadounidense Joseph Grunwald y tuvo como referente la encuesta de empleo de Estados Unidos (Current Population Survey[1]). La aplicación de la encuesta estuvo a cargo del Instituto de Economía bajo la supervisión del experto de Naciones Unidas, Roe Goodman.

En aquella oportunidad se seleccionaron 2.330 hogares, de los cuales un 98,2% fueron encuestados. Posteriormente, la EOD se repitió en junio de 1957, junio de 1958, marzo y junio de 1959. A partir de 1960, la encuesta comenzó a realizarse trimestralmente, teniendo como períodos de referencia una semana completa de los meses de marzo, junio, septiembre y diciembre de cada año[2].

En la actualidad, el Centro de Microdatos (CMD) está a cargo de la EOD, que hoy es representativa de 34 comunas del Gran Santiago y se aplica a una muestra de 3.060 hogares distribuida en ocho estratos geográficos compuestos por las comunas de acuerdo a su tamaño y proximidad geográfica[3]; el trabajo de campo se realiza cada año trimestralmente tomando como período de referencia una semana completa de los meses de marzo, junio, septiembre y diciembre.

[1] Para mayor detalle ver http://www.census.gov/programs-surveys/cps.html

[2] Para el detalle de los cambios realizados en la historia de la EOD ver anexo 2.

[3] Ver anexo 1 para detalle de las comunas que conforman los estratos.

3 Objetivos de la Encuesta

El objetivo general de la EOD es establecer un sistema continuo de información estadística sobre las características sociodemográficas y económicas de la población para los sectores público y privado, cuya unidad de observación y análisis es el hogar.

Los objetivos específicos de la encuesta son los siguientes:

- Recolectar datos acerca de las características sociodemográficas de la población y su relación con variables laborales como la "condición de ocupación".
- Recolectar y producir información sobre los niveles de ocupación y desocupación en relación con la rama de actividad, la ocupación principal y la posición en el trabajo.
- Indagar acerca de características ocupacionales, tales como trabajo desempeñado, categoría ocupacional y rama de actividad del último trabajo que tuvieron las personas que dejaron su trabajo, e identificar los motivos principales por los cuales dejaron dicho empleo.
- Indagar acerca de la heterogeneidad del mercado de trabajo, determinando características de los establecimientos (sector de propiedad, rama de actividad, etc.) en los que se insertan los ocupados.
- Indagar acerca de las condiciones de trabajo (horas, forma de pago, ingresos y prestaciones laborales) de los ocupados.
- Indagar acerca de las modalidades de empleo de la población plenamente ocupada, para diferenciarla de la población parcialmente ocupada.
- Recolectar datos que permitan estudiar con mayor profundidad el fenómeno del desempleo abierto.
- Recolectar datos acerca de los inactivos y así conocer el grado de disponibilidad para integrarse al mercado de trabajo o los motivos de su no disponibilidad para trabajar.
- Determinar la presión que ejercen sobre el mercado del trabajo los ocupados que buscan otro empleo.

4 Diseño Muestral

A continuación se detalla el diseño muestral de la EOD y el proceso de su implementación.

4.1 Población Objetivo

La población objetivo de la EOD incluye a todas las personas mayores de 14 años que residen en viviendas particulares habitadas (o habitables) ubicadas en las 32 comunas urbanas de la provincia de Santiago, además de las comunas de Puente Alto y San Bernardo.

4.2 Cobertura

La cobertura de la encuesta corresponde a las 34 comunas del Gran Santiago, lo que corresponde a un 87,6% de la población total de la Región Metropolitana.

4.3 Marco Muestral

Se usa como marco muestral el último Censo disponible del que deriva un marco muestral secundario. Éste se conforma a partir de la selección de manzanas censales en las que se actualiza el número de viviendas mediante procesos de enumeración y empadronamiento para minimizar los problemas de cobertura que pudiera tener el marco por su antigüedad.

Se excluyen del marco muestral las personas que se encuentran viviendo en instituciones o viviendas colectivas como recintos militares, cárceles, hogares de ancianos, hospitales de cuidado permanente, etc.

4.4 Diseño Muestral

Desde sus inicios el diseño muestral de la EOD ha sido un muestreo probabilístico de áreas para seleccionar a los hogares que forman parte de la muestra. En consecuencia el diseño se caracteriza por ser:

Probabilístico porque tiene una probabilidad de selección conocida distinta de cero para cada elemento del marco muestral, lo que permite realizar inferencia estadística sobre la población.

Estratificado porque las unidades primarias de muestreo son clasificadas en estratos.

Con probabilidad de selección proporcional al tamaño porque la probabilidad de selección de las manzanas censales es función del número de hogares que contienen.

Bietápico pues los elementos pertenecientes a la muestra se seleccionan en dos etapas: 1) selección de manzanas censales y 2) selección de segmentos compactos de viviendas.

4.5 Tamaño Muestral

En 1983 se realizó el último cambio al tamaño de la muestra de EOD debido a una reducción presupuestaria del proyecto, quedando en 3.060 hogares. Este tamaño de muestra es consistente con coeficientes de variación de la tasa de desempleo en el rango [0,053; 0,068]. Para esta estimación se utilizó el promedio mínimo de la tasa de desempleo y cuasivarianza poblacional de (0.065 y 0.061) y valores promedios máximos de (0.104 y 0.093).

La determinación del tamaño de la muestra requiere de la consideración de los siguientes parámetros:

- La variable más importante de la encuesta, o variable de muestreo, es la tasa de desempleo.
- Los requerimientos de precisión se expresan en términos del coeficiente de variación (CV).
- Para efectos de cálculo del tamaño de la muestra se asume que el costo de la implementación de la encuesta es una función de la cantidad de hogares entrevistados.

La formula utilizada es:

$$n = \frac{1}{\left(\frac{V_o}{S^2} + \frac{1}{N}\right)}$$

Donde: $V_o = V_{srs} * Deff$; $Deff = 1 + (m-1) * \rho$

n= número de hogares
N= número de hogares de la población objetivo
S^2= cuasivarianza poblacional de la tasa de desempleo
V_o= varianza objetivo de la tasa de desempleo con muestreo complejo
V_{srs}= varianza de la tasa de desempleo con muestreo aleatorio simple
m= promedio de hogares a entrevistar por manzana

ρ=tasa de homogeneidad de la tasa de desempleo
Deff= efecto diseño de la tasa de desempleo

4.6 Estratificación de la Muestra

Como se mencionó anteriormente, el diseño muestral de la EOD es estratificado con el fin de mejorar la precisión de los estimadores y permitir un mejor control de la distribución de la muestra.

En este sondeo la variable de estratificación es la comuna, ya que se usa como supuesto que el desempleo estaría correlacionado con el estatus socioeconómico de las personas. Luego se asume que la comuna es una buena aproximación de esta variable, por lo que se tienen buenas razones para esperar mejoras en la precisión de los estimadores producto de la estratificación de la muestra. Si estos supuestos no son válidos, entonces solo queda el nivel de precisión del diseño muestral respectivo (sin "mejoras" debido a la estratificación).

Los ocho estratos de la EOD agrupan geográficamente las comunas del Gran Santiago, lo que asegura la selección de unidades primarias de muestreo en cada uno de los estratos y cada una de las comunas del Gran Santiago.

La determinación del tamaño de la muestra es proporcional al tamaño de cada estrato en la población. Entonces, el proceso de empadronamiento y las proyecciones oficiales del INE para cada año permiten estimar la redistribución de la población objetivo y de esa manera tener una estimación más precisa.

Cuadro 1: Distribución de los hogares de la muestra de la EOD para marzo de 2015.

Estratos	Universo		Distribución Muestra EOD	
	N° de Hogares	Distribución	N° de Hogares	Distribución
1	156.573	11%	380	12%
2	172.062	12%	220	7%
3	144.570	10%	176	6%
4	248.223	18%	404	13%
5	165.203	12%	294	10%
6	160.478	12%	504	16%
7	124.698	9%	426	14%
8	222.097	16%	656	21%
Total	**1.393.904**	**100%**	**3.060**	**100%**

Fuente: Estadísticas del Censo de Población y Vivienda 2002

4.7 Unidades de Selección

Unidades Primarias de Muestreo (UPM)

Las UPM de la EOD están conformadas por manzanas identificadas en el último Censo de Población y Vivienda disponible. La selección de cada una de las UPM se realiza en forma independiente al interior de cada uno de los estratos definidos en la muestra de acuerdo a la probabilidad proporcional al tamaño (PPT) y al algoritmo de selección aleatoria sistemática.

El método PPT consiste en acumular las medidas de tamaño de las UPM (número de hogares en cada manzana) y seleccionar las UPM de acuerdo a este tamaño acumulado. Este procedimiento determina que las UPM de mayor tamaño sean seleccionadas con mayor probabilidad que las UPM con menor población.

El algoritmo de selección sistematica consiste en tomar cada k-ésima UPM a partir de una arranque aleatorio r*, donde K_h es el intervalo de selección en el estrato h y es igual a N_h/n_h (N_h=total de hogares en estrato h y n_h= número de hogares seleccionados en el estrato h).

Debido al sistema de rotación de la EOD, cada año son seleccionadas aproximadamente 296 nuevas UPM (en total, entre los ocho estratos de la muestra). La fracción de muestreo en cada estrato h de esta primera estapa de selección esta dada por:

$$f_h^1 = \frac{a_h M_{hi}}{\sum_{l=1}^{a_h} M_{hi}}$$

Donde:
a_h : número de manzanas (UPM) a seleccionar en el estrato h
M_{hi} : número de hogares en la manzana i en el estrato h (según registros del Censo)

Unidades Secundarias de Muestreo (USM)

Las USM están compuestas por bloques contiguos de aproximadamente diez viviendas particulares (habitadas permanentemente, o aptas para habitarse) que se encuentren ubicadas en las UPM seleccionadas en la muestra.

Para la selección de las USM es necesario contar con información actualizada del número de hogares existentes en las UPM seleccionadas. Dos procedimientos se realizan en forma rutinaria para estos efectos: conteo rápido de viviendas y empadronamiento.

El conteo rápido de viviendas en las manzanas seleccionadas es realizado por supervisores del CMD en forma previa al proceso formal de listado de viviendas (empadronamiento). El objetivo es determinar rápidamente las UPM que presentan crecimientos anormales respecto del último Censo. En esta circunstancia, la UPM se divide en segmentos de menor tamaño y el equipo técnico a cargo de la EOD sortea en forma aleatoria uno de los segmentos.

El empadronamiento es realizado en terreno por especialistas del CMD, y consiste en visitar las manzanas seleccionadas y realizar un listado con el número de hogares, las direcciones y el tipo de uso de las estructuras existentes, por ejemplo, habitables, en el caso de viviendas; no habitables, en locales comerciales, sitios eriazos, etc. Para actualizar el número de hogares, los empadronadores establecen contacto con las viviendas y preguntan por el número de hogares existentes en cada vivienda de la manzana seleccionada.

A partir de este listado se puede actualizar la información censal referente al número de hogares que integran la UPM seleccionada. Este procedimiento permite:

- Determinar la cantidad exacta de hogares a seleccionar en cada manzana.
- Estimar anualmente la distribución de la población en los estratos del Gran Santiago para informar el recálculo de la afijación proprocional de la muestra a los estratos.

El proceso de actualización del marco muestral es desarrollado e implementado por el equipo técnico a cargo de la EOD.

La primera etapa de selección (UPM) se realiza sobre los datos del Censo y un tamaño promedio de diez viviendas por USM. Luego, cuando no existen errores de cobertura, la probabilidad de selección de la segunda etapa en cada estrato h es:

$$f_h^2 = \frac{b}{M_{hi}}$$

Donde:
b : número original de hogares a seleccionar en el estrato h (aprox. 10)
M_{hi} : número de hogares en la manzana i en el estrato h (según registros del Censo)

Producto del proceso de empadronamiento se producen discrepancias con el registro censal producto de la obsolescencia del marco muestral. Por esa razón, para mantener fija la probabilidad de selección de las USM, se ajusta la cantidad de viviendas a seleccionar en cada manzana y de esa manera compensar el efecto de los crecimientos/decrecimientos experimentados por la UPM con respecto al Censo. Así, al

utilizar la información actualizada en el proceso de empadroamiento, la probabilidad de selección de la segunda etapa en el estrato h queda definida por:

$$f_h^{2*} = \frac{b^*}{M_{hi}^*}$$

$$b^* = \left(\frac{b}{M_{hi}}\right) * M_{hi}^*$$

Donde:
b^* : número ajustado de hogares a seleccionar en el estrato h.
M_{hi}^*: número de hogares en la manzana i en el estrato h (según registros del empadronamiento en fecha actual).
b : número original de hogares a seleccionar en el estrato h (aprox. 10).
M_{hi} : número de hogares en la manzana i en el estrato h (según registros del Censo).

El resultado de este procedimiento es que se produce una redistribución de los hogares seleccionados al interior de cada estrato. De esta manera se seleccionan menos hogares en las manzanas que han decrecido.

Una vez que se ha determinado la cantidad de hogares a entrevistar en cada manzana se procede a la selección de las USM. Se sortea un número aleatorio entre 1 y M_{hi}^* para cada UPM, que define el hogar seleccionado para la entrevista. En forma automática, se incluyen en las muestra los $b^* - 1$ hogares contiguos al recién sorteado. Luego, formalmente existe sólo una selección aleatoria que determina la inclusión de un conjunto de b^* hogares en la muestra. Esta última selección corresponde formalmente a la Unidad Secundaria de Muestreo conocida también como selección de segmentos compactos, ya que los hogares son selecciones de conglomerados (bloques) de b hogares.

El crecimiento significativo de algunas áreas geográficas en las comunas del Gran Santiago representan un desafío para el diseño y selección de la muestra. Por esa razón, se introduce una etapa adicional de selección a nivel de la USM y se establece una cota máxima al número de hogares a seleccionar en una misma manzana.

En el proceso de visitas para el conteo de las viviendas se identifican aquellas manzanas más pobladas (que originalmente no lo eran) las que son subdivididas en segmentos más pequeños y se sortea en forma aleatoria el segmento que será empadronado para formar parte de la muestra.

En general, se sigue el siguiente procedimiento:

- En manzanas residenciales donde solo hay viviendas del tipo casa y no edificios, se mantiene el criterio de subdividir en 40 viviendas.
- En manzanas compuestas por casas y edificios se trata de conservar la manzana de acuerdo a su composición original incluso mezclando casas con departamentos dependiendo del alto del mismo y las viviendas por piso que tenga el edificio.
- En manzanas compuestas solo por edificios, éstos son particionados en grupos por piso. Es decir, si un edificio tiene 88 departamentos y cuatro departamentos por piso, entonces el edificio se subdivide en dos submanzanas de 44 viviendas cada una.

Una vez que la manzana es subdividida se genera una tabla con el número de viviendas de cada submanzana, luego en la columna siguiente se calcula el acumulado y finalmente se elige un número aleatorio entre 1 y el número total de viviendas en la manzana y se ubica en la columna de acumulado para seleccionar la submanzana a empadronar, tal como muestra el ejemplo a continuación.

Cuadro 2. Ejemplo de tabla selección submanzana.

Submanzana	Número de viviendas	Acumulado
1	45	45
2	45	90
3	45	135
4	45	180
5	47	127
6	43	170
Aleatorio:		**48**

Seleccionada la submanzana se envía a empadronar, luego se digita el empadronamiento y se realiza la selección de las viviendas a encuestar.

Selección del Informante al Interior del Hogar

Una vez seleccionados los hogares a encuestar se envía la muestra a trabajo de campo. Para ello, al comenzar la entrevista el encuestador/a realiza un listado de todos los miembros del hogar, desde los infantes hasta los adultos mayores, a quienes aplica una batería de preguntas demográficas. Las preguntas del cuestionario referentes a la situación laboral e ingreso se aplican sólo a los mayores de 14 años.

En general, si las personas que conforman el hogar se encuentran presentes se les aplica la entrevista a cada uno, de lo contrario, la entrevista se realiza a un solo miembro del hogar quien proporciona su información y del resto de los miembros de su hogar. Si la

persona desconoce la información respecto de alguno de los integrantes del hogar, entonces se realizan esfuerzos especiales para tratar de entrevistar a esa persona directamente.

Esta última etapa es formalmente un "censo", ya que se entrevista a todas las personas mayores de 14 años en la USM seleccionada. Por lo tanto, la probabilidad de selección de las personas en el hogar en esta tercera etapa es igual a 1.

5 Esquema de Rotación de la Muestra

La EOD es una encuesta de panel rotativo que consiste en mantener parte del panel en forma permanente y una muestra de corte transversal completamente nueva. Este esquema permite balancear la minimización de:

- Varianza de los estimadores de cambio de trimestre a trimestre: dos cuartos de la muestra es la misma de trimestre a trimestre.
- Varianza de los estimadores de cambio de año a año: la mitad de la muestra es la misma en el mes de encuestaje de años consecutivos.
- Varianza de otros estimadores de cambio: la muestra que sale es reemplazada por una muestra que probablemente tiene las mismas características.
- Fatiga del informante: las cuatro entrevistas en las que participa cada hogar se dispersan a través de 18 meses.

La muestra de 3.060 hogares de cada aplicación de la EOD se divide en cuatro submuestas de 765 hogares aproximadamente en cada una (denominados paneles o cuartos), en donde cada una de las submuestras es independiente y representan al Gran Santiago. Esto implica que cada cuarto (submuestra) es seleccionado siguiendo el mismo diseño muestral y comparten el mismo nivel de precisión. De esta manera se espera que los hogares que se salen definitivamente de la muestra (ya fueron entrevistados en cuatro oportunidades en 18 meses) sean reemplazados por hogares que provienen de los mismos estratos geográficos y comparten características similares.

El diseño de rotación es del tipo 2-2-2, es decir, que cada vivienda seleccionada y su respectivo hogar es entrevistado en dos rondas seguidas, no se visita en las siguientes dos rondas, y luego se vuelve a entrevistar en dos rondas consecutivas. Todo en un período que abarca 18 meses en total, tal como se observa en el cuadro 3.

Para ejemplificar, se describe el ingreso del cuarto 166 a la muestra. Las aproximadamente 74 nuevas UPM seleccionadas en 2014 pasan a formar el "cuarto 166" e ingresan por primera vez a la muestra en marzo 2014. En la siguiente aplicación de la encuesta (junio 2014) los hogares del cuarto 166 son entrevistados por segunda vez. En las próximas dos encuestas (septiembre y diciembre) esos hogares no serán contactados, pero sí los entrevistarán nuevamente en marzo y junio de 2015 para participar por tercera y cuarta vez, respectivamente. Terminada la cuarta visita después de 18 meses desde la primera entrevista, los hogares del cuarto 166 dejan de formar parte de la muestra en forma definitiva.

Estructura de rotación 2-2-2 de la EOD, para cuartos seleccionados en el período 2013-2015

Submuestras (cuartos)	**2013**				**2014**				**2015**			
	M	**J**	**S**	**D**	**M**	**J**	**S**	**D**	**M**	**J**	**S**	**D**
162	1	2			3	4						
163		1	2			3	4					
164			1	2			3	4				
165				1	2			3	4			
166					1	2			3	4		
167						1	2			3	4	
168							1	2			3	4
169								1	2			3
170									1	2		
171										1	2	
172											1	2
173												1

6 Estimación de Estadísticas Laborales

La EOD es una encuesta probabilística de diseño complejo desde la cual se producen estadísticas laborales para la población de 14 años y más en el Gran Santiago. El proceso de estimación proveniente de este tipo de encuesta requiere de ajustes debido a la presencia de: (1) distintas probabilidades de selección para sub-poblaciones de interés, (2) distintas tasas de respuesta entre sub-poblaciones de interés y (3) distorsión en la distribución de variables demográficas que son informadas por controles poblacionales externos. En la literatura actual, se utiliza un enfoque que combina el uso de ponderadores para la corrección de (1), (2) y (3), junto con el uso de imputación para corregir la no respuesta en las preguntas de interés[4] (Kalton y Kasprzyk, 1986).

La metodología de la EOD[5], que se ha mantenido inalterada, establece un tratamiento distinto para estos problemas:

a) Para (1) se establece un cálculo de ponderadores de acuerdo a la probabilidad de selección de cada una de las unidades de la muestra, que en el caso de la EOD son iguales (EPSEM);

b) En cuanto a las tasas de respuesta se utiliza la metodología de *Hot deck*, que reemplaza la no respuesta por un hogar de similares características. Dado que la EOD es un panel rotativo, esta metodología se adecúa de mejor forma ya que se tiene mayor información para realizar el proceso de imputación y permite la consistencia interna de los datos, sin generar una caracterización de la muestra muy distinta a la muestra original, puesconsidera solo la comparación de los hogares que respondieron la encuesta durante el trabajo de campo.[6].

c) La distorsión en la distribución de variables demográficas se controla sólo a nivel de tamaño poblacional de acuerdo a las proyecciones de población del INE, sin distinguir poblaciones de interés, porque la representatividad de la muestra es a nivel de Gran Santiago y sus resultados son inferidos sobre esta población.

[4] La imputación es la asignación de una o más respuestas a un campo de datos que previamente no tenía respuestas o que incluía respuestas incorrectas o inverosímiles. La EOD no realiza imputación de valores para corregir la no respuesta de ningún ítem de la encuesta.

[5] Ver el Informe Metodológico de la Encuesta de Ocupación del Gran Santiago para una descripción del trato de cada uno de estos elementos en la EOD.

[6] Un ejemplo, además de la EOD es la VII EPF, ver fase C de análisis comparativo de métodos de imputación en documento publicado en http://www.ine.cl/epf/files/documentacion/metodos_de_imputacion_VII_EPF.pdf

6.1 Cálculo de Ponderadores

Por definición los ponderadores corresponden al inverso de la probabilidad de selección de los entrevistados $W_{ij} = \frac{1}{\pi_{ij}}$

Donde:

W_{ij}:es el ponderador de las personas seleccionadas en el segmento compacto j (USM) correspondiente a la manzana i (UPM)
π_{ij}: es la probabilidad de selección

En este estudio la estrategia de selección de las personas determina que todos los segmentos compactos de hogares tengan la misma probabilidad de selección (EPSEM, *equal probability selection method*). Esto implica que todas las personas entrevistadas comparten esa misma probabilidad de selección, ya que se entrevista a todos los individuos mayores de 14 años en cada hogar, lo que tiene características similares a un censo de población objetivo. De esta manera, la probabilidad de selección final en cada estrato h viene dada por la probabilidad conjunta de la primera y la segunda etapa de selección.

$$f_h = f_h^1 * f_h^{2*} = \left(\frac{a_h M_{hi}}{\sum_{\alpha=1}^{ah} M_{hi}}\right) * \left(\frac{b^*}{M_{hi}^*}\right)$$

$$f_h = \frac{a_h M_{hi}}{\sum_{\alpha=1}^{ah} M_{hi}} * \frac{1}{M_{hi}^*} * \left(\frac{b^*}{M_{hi}}\right) * M_{hi}^* = \frac{a_h b}{\sum_{\alpha=1}^{ah} M_{hi}}$$

Como se puede apreciar la última expresión es una constante, lo que implica que la probabilidad de selección de los segmentos compactos ij es la misma para todos al interior del estrato h. Por otra parte, tenemos que la fracción de muestreo f_h, correspondiente al estrato h, es la misma para todos los estratos de la medición, ya que la EOD se encuentra estratificada en forma proporcional al tamaño de los estratos, por lo tanto, $f_h = f$. La implicancia de este resultado es que todos los elementos de la muestra comparten el mismo ponderador de selección, el cual viene dado por la siguiente expresión:

$$W_{ij}^b = \frac{1}{f} = \frac{\sum_{\alpha=1}^{ah} M_{hi}}{a_h b}$$

Por ejemplo, para la encuesta de marzo 2016 el ponderador de selección de todos los hogares de la muestra es 581,2; lo que se puede interpretar como que cada hogar entrevistado en la encuesta se representa a sí mismo y a otros 580 hogares del Gran

Santiago. Los diseños muestrales con estas características se denominan autoponderados ya que todos los hogares tienen el mismo ponderador de selección base.

$$W_{ij}^{b} = \frac{1}{f} = \frac{\sum_{\alpha=1}^{ah} M_{hi}}{a_h b} = \frac{6.479.607\ personas}{11.147\ personas} = 581{,}2$$

El total de personas (M_{hi}) corresponde a la proyección del total de la población para el Gran Santiago, que se estima a partir de las estadísticas oficiales publicadas por el INE[7], mientras el total de personas en la muestra ($a_h b$) se obtiene en forma directa a partir de la encuesta.

6.1.1 Ajuste Por No Respuesta

Todas las encuestas a hogares están sujetas, en mayor o menor medida, a fallas en el proceso de contactar y lograr la cooperación de los habitantes de los hogares seleccionados para formar parte de la muestra. Esta falla se conoce como no respuesta a la unidad seleccionada, cuyos principales componentes son: (1) falla en realizar el contacto con el hogar y (2) falla en obtener la cooperación del seleccionado.

La no respuesta implica que al final del trabajo de terreno, se entrevista a un número menor a los 3.060 hogares seleccionados. El efecto de la no respuesta en los estimadores es doble: Primero reduce los niveles de precisión por reducción en el tamaño de la muestra; segundo, pone en riesgo la obtención de estimadores insesgado, debido a la "auto selección" de los hogares de la muestra. Por estos motivos, particularmente el último, es que se requiere de algún mecanismo que permita corregir el problema de no respuesta a la unidad.

A diferencia del proceso de selección de la muestra, que es definido por el investigador a través del diseño de probabilidades de selección de los elementos de la muestra, el proceso de contacto y cooperación de los hogares en la encuesta no esta bajo el control del investigador y las probabilidades de "selección" le son desconocidas. En consecuencia, se deben desarrollar modelos que expliquen las distintas probabilidades de participación de los hogares y corrijan potenciales sesgos de no respuesta.

El modelo de participación de los entrevistados más conocido es el MAR ("*missing at random*", Rubin, 1987), que asume que quienes responden son una muestra aleatoria de los seleccionados. Bajo este modelo, no tiene sentido hacer ajustes de no respuesta

[7] Para mayor detalle ver: http://www.ine.cl/canales/chile_estadistico/familias/demograficas_vitales.php

y los analistas pueden realizar estimaciones en forma directa utilizando sólo datos de los encuestados (sin realizar ningún ajuste).

El análisis de la muestra enviada a terreno y la efectivamente obtenida indican que probablemente las personas que no responden la Encuesta no son una muestra aleatoria de la muestra original. El modelo a estimar para la corrección de este tipo de no respuesta es MCAR ("*missing completely at random*", Rubin 1987) y asume que al interior de ciertas celdas de ajuste las personas efectivamente entrevistadas son una muestra aleatoria de los seleccionados originalmente en la muestra.

Las variables de agrupación a utilizar para construir las celdas de ajuste tienen que cumplir con tres características: (1) tienen que estar disponibles tanto para quienes responden como para quienes no responden la encuesta; (2) tienen que estar relacionadas con el fenómeno de participación, y (3) tienen que estar relacionadas con la variable de interés (tasa de desempleo).

En el marco muestral de la EOD la única variable que de alguna forma satisface estas características es la ubicación geográfica de las viviendas seleccionadas, por lo tanto, esta variable se utiliza para la construcción de las celdas de ajuste para la no respuesta.

El proceso de ajuste para la EOD se realiza de la siguiente forma:

1. Se forman ocho celdas de ajuste a partir de los ocho estratos geográficos utilizados para la selección de la muestra.
2. En cada celda se incluyen los hogares que respondieron y los que no respondieron a la encuesta.
3. Para cada hogar que no respondió se sortea en forma probabilística (aleatoria sistemática) un hogar que respondió.
4. Cada hogar que no respondió es reemplazado en la base de datos final por un hogar que sí respondió.

Este procedimiento de reemplazo se denomina *Hotdeck* y constituye una herramienta validada para la corrección de la no respuesta a la unidad (Kish, 1990). Aparte de la simpleza de su implementación, una de sus ventajas es que permite mantener la estructura de correlación de las variables bajo estudio, lo que es fundamental ya que no se realiza ningún otro tipo de ajuste a nivel de ponderación de los datos.

6.1.2 Ajuste de Post-Estratificación

El último ajuste de ponderadores tiene como objetivo que la distribución de ciertas características demográficas de la muestra sean idénticas a la distribución de la población objetivo, según la fuente de datos externa considerada como referencia, que generalmente es el Censo, o bien, la Encuesta Casen que es la encuesta a hogares con mayor cobertura a nivel nacional. Este ajuste también es conocido como "control de población externo", "calibración", o "post-estratificación" y se realiza no sólo con la pretensión de simular la distribución externa, sino también lograr mejoras en la eficiencia estadística y en la cobertura de la población objetivo.

En el caso de la EOD solo se aplica un ajuste poblacional asociado a la estimación de la población total del Gran Santiago, sin distinguir proyecciones para dominios de interés como sexo, tramos etarios u otra característica demográfica de la población objetivo.

6.2 Estimadores de Estadísticas Laborales

La estadística de mayor interés en la EOD es la tasa de desocupación enel Gran Santiago. A continuación se presenta su forma de cálculo y otros estimadores de interés, que pueden ser expresados en función del estimador de totales, por lo tanto, lo utilizaremos de ejemplo para explicar la estimación con muestreo estratificado.

El estimador de totales para una muestra estratificada esta dado por la siguiente expresión:

$$\widehat{t_\pi} = \sum_{sh} \widehat{t_{h\pi}} = \sum_{sh} N_h \bar{Y}_{sh}$$

Donde el total de cada estrato $t_{h\pi}$ es estimado a partir del producto entre el total de la población en cada estrato (N_h) y el promedio muestral de la característica en cada estrato (Y_{sh}). Por ejemplo, para calcular el total de personas desocupadas en el Gran Santiago se debe calcular primero el total de personas desocupadas en cada uno de los ocho estratos (h) de EOD y después sumar los ocho totales para obtener el total deseado.

Sin embargo, como la EOD tiene un diseño estratificado, entonces el estimador total se puede calcular directamente a través de la muestra completa, sin necesidad de calcular el total para cada estrato en forma independiente. Esta simplificación en el cálculo se produce exclusivamente por la afijación proporcional de la muestra en los estratos. Otro tipo de alocaciones (óptima, porporcional al total, etc.) no se benefician de esta simplificación.

$$\widehat{t_\pi} = \sum_h N_h \bar{Y}_{sh} = N\, \bar{Y}_s$$

En adelante, se presentan las fórmulas de los estimadores sin considerar la estratificación de la muestra.

6.2.1 Estimación de Totales

Para la estimación de totales se utiliza el estimador Horvitz-Thompson (1952), también conocido como estimador (∏). El estimador Horvitz-Thompson es insesgado y produce la expansión de una característica medida a partir de la muestra hacia su valor poblacional. Esto último se logra a través del cuociente entre el valor de cada variable y su respectiva probabilidad de selección.

$$\widehat{t_{y\pi}} = \sum_s \frac{Y_k}{\pi_k}$$

En la EOD todos los elementos de la muestra tienen la misma probabilidad de selección ($\pi_k = \pi$), por lo tanto, la fórmula anterior se simplifica a:

$$\widehat{t_{y\pi}} = \frac{1}{\pi}\sum_s Y_k$$

6.2.2 Estimación de Tasas y Proporciones

El estimador de tasas, también conocido como estimador de razón, es simplemente el cuociente entre dos totales de interés. Por ejemplo, la tasa de desempleo corresponde al total de personas ocupadas ($t_{y\pi}$) dividido por el total de personas en la fuerza de trabajo ($t_{x\pi}$).

$$\hat{R} = \frac{\widehat{t_{y\pi}}}{\widehat{t_{x\pi}}}$$

Esta misma expresión se utiliza para el cálculo de proporciones cuando la variable de interés (y) es binaria (0,1). Por ejemplo, la proporción de personas de 14 años o más corresponde al total de personas de 14 años o más ($t_{y\pi}$) dividido por el total de personas en la muestra ($t_{x\pi}$). La expresión simplificada que se obtiene debido al diseño autoponderado de la EOD es:

$$\hat{R} = \frac{\sum_s Y_k}{\sum_s X_k}$$

6.2.3 Estimación de Promedios

Finalmente, el estimador de promedios se calcula a través del cuociente entre el total de la variable de interés y el total de personas en la muestra. Por ejemplo, el ingreso promedio de la población corresponde al total de los ingresos declarados ($t_{y\pi}$) dividido por el total de personas en la población (N).

$$\hat{\bar{Y}}_{y\pi} = \frac{1}{N}\,\widehat{t_{y\pi}}$$

Esta última fórmula asume que el total de personas en la población (N) es un número conocido. Cuando se desconoce N o se desconfía de su precisión (por ejemplo, debido a desactualizaciones en las estimaciones censales) se utiliza un estimador del total de la población obtenido a partir de la muestra. La fórmula a utilizar en este caso es:

$$\tilde{Y}_{ys} = \frac{\widehat{t_{y\pi}}}{N} = \frac{\frac{\sum_s Y_k}{\pi_k}}{\frac{\sum_s 1}{\pi_k}}$$

La expresión simplificada que se obtiene debido al diseño autoponderado de la EOD viene dada por :

$$\tilde{Y}_{ys} = \frac{\sum_s Y_k}{n}$$

7 Estimación de Varianza

Debido a que la selección de la muestra de EOD es mediante un mecanismo aleatorio sistemático de unidades primarias de muestreo, ello complica la estimación de la varianza, porque no tenemos un estimador insesgado para la varianza con muestreo de selección sistemática, entonces no se puede medir la variabilidad muestral de los estimadores puntuales presentados anteriormente. En este caso la variable no es medible y no se puede calcular únicamente a partir de los datos de la muestra.

Existen dos enfoques para tratar este problema: Uno es utilizar modelos para la estimación de la varianza; el otro, utilizar múltiples selecciones aleatorias. La decisión sobre cual enfoque utilizar queda a criterio del investigador y de la capacidad computacional disponible. A continuación se describen ambas alternativas.

7.1 Estimación con Múltiples Réplicas

Si bien la selección de la muestra de la EOD utiliza un solo arranque aleatorio al interior de cada uno de los ocho estratos, se dispone de cuatro réplicas (cuartos) para la estimación de las estadísticas laborales en cada aplicación. Esto significa que en cada estrato se realizan efectivamente cuatro selecciones aleatorias, por lo tanto, el calculo de la varianza es posible y su fórmula cuando se dispone de K réplicas independientes es:

$$\hat{V}_{REP}(\hat{\theta}^*) = \frac{1}{K(1-K)} \sum_{k=1}^{K} (\widehat{\theta_k} - \hat{\theta}^*)^2$$

Donde:

$\widehat{\theta_k}$:es el estimador puntual calculado a partir de la replica k
$\hat{\theta}^* = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}_k$:es el estimador puntual de la muestra completa calculado como el promedio de los estimadores puntuales de las K réplicas (Sarndal, Swensson y Wretman, 1992).

En teoría, este es el estimador de la varianza que mejor refleja el diseño complejo de la EOD. Sin embargo, la estimación de la varianza puede resultar inestable debido a los pocos grados de libertad (K-1) que implica contar con sólo cuatro réplicas.

El investigador tendrá que ponderar la pérdida en precisión que implica la estimación de la varianza con sólo cuatro réplicas versus el sesgo que puede introducir la estimación de la varianza a través de los modelos que se presentan en la sección siguiente.

7.2 Modelos de Estimación de Varianza

A continuación, como una alternativa a las múltiples selecciones aleatorias, se presentan tres modelos de estimación de varianza para los resultados de la EOD: Muestreo Aleatorio Simple, Diferencias Pareadas y Técnicas de Replicación.

7.2.1 Modelo de Muestreo Aleatorio Simple

El modelo más sencillo asume que el muestreo sistemático (SY) es por lo menos tan eficiente como el muestreo aleatorio simple (SI), por lo tanto, utiliza la fórmula de estimación de la varianza SI para el cálculo de la estimación de varianza SY. Se trata de un supuesto razonable cuando, por ejemplo, la lista se encuentra ordenada en forma aleatoria al momento de realizar la selección sistemática. En cambio, si se sospecha de la existencia de un ciclo en el orden de la lista, no es un buen supuesto.

La varianza se puede estimar a partir de la siguiente expresión, donde los grados de libertad disponibles son (n-1):

$$V_{SY}(\widehat{t_\pi}) \approx V_{SI}(\widehat{t_\pi}) = N^2 \frac{(1-f)}{n} S_y^2$$

Si el muestreo sistemático es más preciso que el muestreo aleatorio simple, entonces el uso de la fórmula SI estará sobreestimando $V_{SY}(\widehat{t_\pi})$. En este caso se dice que el enfoque SI es conservador, ya que los intervalos de confianza de 95% que se estiman a partir de $V_{SI}(\widehat{t_\pi})$ incluirán el parámetro de interés a una tasa mayor que 95% en repetidas muestras (Kish, 1965).

Es necesario recordar que en la EOD la lista a partir de la cual se seleccionan las UPM se encuentra ordenada en forma geográfica, por lo tanto, puede ser discutible que este modelo de estimación sea el más adecuado para la encuesta.

7.2.2 Modelo de Diferencias Pareadas

A veces es razonable asumir que cada par sucesivo de UPM fue seleccionado en forma aleatoria a partir de una zona implícita. Por ejemplo, para la selección sistemática de 20 manzanas en un estrato con 400 manzanas: (1) se crearon diez zonas implícitas de 40 manzanas cada; (2) las manzanas se ordenaron en forma aleatoria al interior de cada zona, y (3) se seleccionaron dos manzanas en forma sistemática en cada zona implícita.

Bajo este supuesto, podemos ordenar los pares de UPM según el orden en que fueron seleccionados para la muestra y comparar la primera UPM con la segunda; la tercera UPM con la cuarta, la quinta UPM con la sexta, y hasta la UPM (n-1) con la n (Kish,

1965). La varianza se puede estimar a partir de la siguiente expresión, donde los grados de libertad disponibles son n/2:

$$V_{SY}(\widehat{t_\pi}) \approx V_{DP}(\widehat{t_\pi}) = N^2 \frac{(1-f)}{n^2} \sum_h^{n/2} (y_{ha} - y_{hb})^2$$

Debido al orden geográfico que presenta la EOD, el Modelo de Diferencias Pareadas se presenta como una alternativa bastante cercana al diseño muestral y su implementación es relativamente sencilla en términos computacionales. Se recomienda su uso por sobre el modelo del muestro aleatorio simple y como una alternativa menos compleja a otros modelos de replicación.

7.2.3 Técnicas de Replicación

Las técnicas de replicación imitan el proceso descrito en la estimación vía réplicas independientes, pero utilizando sub-muestras de la muestra completa. La técnica de Replicaciones Repetidas Jacknife (JRR) define las réplicas mediante la eliminación de una UPM a la vez. Si A es el numero de UPM en la muestra, se pueden construir (A-1) réplicas a partir de la muestra y calcular (A-1) estimadores puntuales. Finalmente se estima la varianza utilizando la fórmula presentada en la sección 8.1.

El problema de este modelo es que las (A-1) submuestras seleccionadas no son independientes unas de otras (a diferencia de las réplicas de la EOD discutidas en la sección 6), por lo tanto, se produce sesgo debido a la dependencia de las submuetras. De todas maneras, el sesgo es pequeño en encuestas grandes.

Otras técnicas de replicación utilizadas son el *Bootstrap* y las Semi-Muestras Balanceadas (*Balance Half Samples*, BHS). Para una revisión de técnicas de estimación de varianza ver Kovar, Rao y Wu (1988). Un resumen comparativo de técnicas de estimación de varianza se encuentra también en Rust (1985).

8 ANEXO 1: Estratos de la EOD

Comunas integrantes de cada estrato de la EOD desde 1957 a hoy

Estrato	Jun 1957-Dic 1997	Dic 1996- presente
1	Ñuñoa La Reina La Florida	Ñuñoa La Reina Macul Peñalolén
2	San Miguel	San Miguel La Cisterna San Joaquín La Granja San Ramón Pedro Aguirre Cerda Lo Espejo
3	La Cisterna La Granja San Bernardo	El Bosque La Pintana San Bernardo
4	Maipú Quinta Normal Pudahuel	Maipú Cerrillos Pudahuel Lo Prado Cerro Navia
5	Conchalí Renca Quilicura	Recoleta Independencia Conchalí Renca Quilicura Huechuraba
6	Providencia Santiago Oriente Las Condes	Providencia Vitacura Las Condes Lo Barnechea
7	Santiago Centro	Santiago Estación Central Quinta Normal
8	Santiago Poniente	La Florida Puente Alto

Notas: En diciembre de 1996 se inicio el proceso de traspado desde la antigua definición de los estratos (17 comunas) hacia la nueva definición (34 comunas). La encuesta de diciembre 1996 incluyó un cuarto con la nueva definición y tres cuartos con la definición antigua; en marzo de 1997 se incluyeron dos nuevos cuartos y dos antiguos; en junio de 1997, dos nuevos y dos antiguos; en septiembre de 1997 dos nuevos y dos antiguos y finalmente en diciembre 1997 se incluyeron tres nuevos y uno antiguo. A partir de marzo 1998 los cuatro cuartos utilizan la nueva definición de estratos (en base a 34 comunas).

9 ANEXO 2: Cronología de los cambios metodológicos más importantes

- **Junio 1958:** Introducción de muestras adicionales en las ciudades de Valparaíso y Viña del Mar.
- **Marzo 1959**: Introducción de muestra adicional en Gran Concepción.
- **Junio 1959**: Introducción de muestras adicionales en las ciudades de Valdivia y Los Lagos.
- **Septiembre 1960**: Introducción de muestras adicionales en las ciudades de La Serena y Antofagasta.
- **Septiembre 1960**: **Cambio de la base muestral utilizada desde 1956**: Se introducen los siguientes cambios:
    - Aumento del tamaño muestral a 3.500 hogares en el Gran Santiago.
    - División de la muestra total en cuartos, esto es, submuestras del 25% del total de los hogares a encuestar, en reemplazo de los deciles utilizados previamente.
    - Modificación de porcentaje de rotación muestral, de tal forma que, a partir de entonces, el 50% de los hogares de la muestra cambia entre encuestas sucesivas (en lugar del 20% variable previamente).
    - Incorporación de la revisión anual o bianual de los segmentos, con reemplazos trimestrales, en lugar de la base fija mantenida hasta entonces.
- **Diciembre 1960**: Introducción de muestras adicionales en las ciudades de Iquique y Coquimbo.

- **Marzo 1961**: Introducción de muestras adicionales en las ciudades de Puerto Montt y Castro.

- **1970:** El Banco Central de Chile decide financiar solamente las muestras del Gran Valparaíso, Gran Concepción y Gran Santiago.

- **Septiembre 1973:** El golpe de Estado impidió realizar el procesamiento de datos correspondiente al mes de septiembre, a pesar de que ya se había entrevistado a alrededor del 95% de la muestra. En diciembre se reanudó la aplicación y procesamiento normal de la Encuesta.

- **1974:** Reducción del tamaño muestral a 3.400 hogares, distribuidos en 296 segmentos censales, número que se mantiene inalterado hasta la actualidad.

- **1974:** Recodificación de la variable Actividad Económica en las encuestas del período 1957-1973. Este proceso, financiado por el Banco Central de Chile, estuvo a cargo de los académicos Isabel Heskia y Luis Riveros y fue supervisado por representantes de la Universidad de Chile.

- **1978:** Se revisa el diseño muestral y se propone que la distribución de segmentos pase de aleatoria (sólo eran 74 por cuarto) a ser aleatoria sistemática para permitir una mayor dispersión de la muestra.

- **1980:** Tras una solicitud del académico Arnold Harberger se agregó una pregunta sobre "deseos de trabajar" para medir la oferta potencial de mano de obra (desocupados más inactivos con deseos de trabajar).

- **Marzo 1980**: Ampliación del tamaño muestral para lograr representatividad a nivel nacional, urbano y rural, en los meses de marzo y septiembre de cada año. La encuesta tuvo representatividad nacional hasta septiembre de 1990.

- **1982**: A petición del académico Arnold Harberger se agrega a los levantamientos de ocupacion, la "Encuesta Especial a los Desocupados", que entrevista en el mes de julio a una submuestra de los desocupados identificados en la medición de junio.

- **1983:** El Banco Central de Chile disminuye su aporte financiero a la encuesta y como consecuencia el tamaño muestral se reduce en un 10%. Desde entonces el tamaño muestral de la EOD se mantiene en 3.060 hogares.

- **1998:** Se amplía de cinco a siete las categorías definidas para la variable tipo de educación, lo cual permite un conocimiento más detallado del último ciclo y modalidad de educación alcanzado por los miembros de los hogares encuestados.

- **Marzo 2001:** El Banco Central de Chile solicita agregar el cuestionario complementario "Encuesta de Percepción y Expectativas sobre la Situación Económica". El suplemento se aplica en los cuatro levantamientos anuales, a continuación del cuestionario principal.

- **Septiembre 2007:** Se realizan dos cambios metodológicos, el primero asociado al tratamiento insesgado de la no respuesta y el segundo al uso de ajustes de post-estratificación[8].

[8] Para mayor detalle ver documento asociado, el que puede descargar del siguiente link http://www.microdatos.cl/Documentos/Encuestas/Ocupacion/1/Documento_Cambio_Metodologia_Mayo_2008.pdf

FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

UNIVERSIDAD DE CHILE