Una nueva estrategia muestral para las encuestas de empresas en Chile

Seminario para optar al título de
Ingeniero Comercial, mención Economía.

Santiago, 26 de septiembre de 2013

Participantes
Iván Gutiérrez

Profesor Guía
Michael Basch

Universidad de Chile
Facultad de Economía y Negocios

Índice general

Resumen Ejecutivo

Este artículo propone una nueva estrategia muestral para las encuestas de empresas en Chile utilizando a la VII Encuesta Nacional de Innovación como ejemplo. Para esta encuesta en particular, los resultados arrojan que no sólo existen ineficiencias en su diseño, sino que también su análisis. Los principales cambios sugeridos son: (a) reemplazar el diseño muestral actual, básicamente estratificado y sistemático, por un muestreo balanceado, (b) cesar el uso del estimador de razones (separadas) para el cálculo de totales y/o promedios y (c) cambiar la regla de auto-representación del 2 % de mayores ventas por estrato, esencialmente arbitraria, por alguna regla con sustento teórico.

Nota del Autor

Como el evaluador de esta tesis podrá notar, el presente estudio no concuerda con el título presentado originalmente. Esto se debe a que, si bien el estimador propuesto en la tesis original tenía una racionalidad teórica, las últimas simulaciones probaron que era ostensiblemente menos preciso que los estimadores ya existentes. Ciertamente la nueva tesis es mucho más corta que el promedio[1], pero confío en que la importancia y aplicabilidad del tema discutido lo compensen.

[1] De hecho, se basa en unas notas que nunca habría pensado en publicar.

1

Observaciones a la VII Encuesta de Innovación

1.1. Introducción

La Encuesta de Innovación, realizada por la Subsecretaría de Economía en conjunto con el Instituto Nacional de Estadísticas (INE), tiene por objetivo proporcionar información sobre la estructura del proceso de innovación de las empresas en Chile (insumos y resultados) y mostrar las relaciones entre dicho proceso y la estrategia de innovación de las empresas, el esfuerzo innovativo, los factores que influyen en su capacidad para innovar y el rendimiento económico de las empresas.

La Encuesta mide variables como el tipo de innovación (producto, proceso, gestión organizativa y/o marketing), grado de novedad, derechos de propiedad intelectual, las actividades innovativas, incluyendo la I+D, que realizan las empresas Chilenas, en los distintos sectores productivos y regiones del país.

El diseño del formulario y metodología de levantamiento, sigue los lineamientos generales sugeridos por la OECD y la *Community Innovation Survey* (CIS) de *Eurostat* para este tipo de encuestas, los que están plasmados en el Manual de Oslo que son aplicados en la mayoría de los países miembros. Esto con la finalidad de hacer comparables los resultados y estadísticas internacionalmente.

Sin embargo, la actual estrategia muestral de la encuesta no da cuenta de los recientes avances metodológicos en el uso de la información auxiliar en el diseño y análisis de encuestas complejas. Por ejemplo, el método de calibración de Deville & Särndal (1992) permite generar un estimador balanceado con respecto un número arbitrario de variables auxiliares, mientras que el método del cubo de Deville & Tillé (2004) posee una cualidad análoga a nivel de diseño.

Motivado por estas y otras consideraciones, este estudio analiza fríamente las fortalezas y debilidades de la actual metodología de la VII Encuesta de Innovación, así como las alternativas actualmente existentes. En el espíritu de Neyman (1934), se privilegiará la inferencia estadística basada en el diseño por sobre aquella basada en modelos específicos, aunque el enfoque alternativo se considerará cuando sea imprescindible.

El resto del estudio se organiza como sigue. La segunda sección expone en detalle los aspectos metodológicos de la VII Encuesta de Innovación. La tercera expone los errores y limitaciones de dicha metodología. La cuarta discute distintas

alternativas tanto al diseño como al análisis de la encuesta. Finalmente, la quinta presenta las conclusiones y sugiere futuras líneas de investigación.

1.2. Descripción de la actual Estrategia Muestral

Esta sección provee una breve descripción de la actual estrategia muestral de la VII Encuesta de Innovación. Este resumen no pretende en ningún caso ser exhaustivo, sino únicamente contextualizar la presente propuesta metodológica.[1]

Población Objetivo

La población objetivo se compone de las empresas naturales o jurídicas, que desarrollen su actividad dentro de los límites territoriales del país, que cuenten con declaración en el Servicio de Impuestos Internos (SII) en el año 2009 y con un nivel de ventas anuales superiores a 2.400 UF.

Unidad Estadística

La unidad estadística es la empresa, organización que tiene iniciación de actividades independiente, es decir, un RUT y contabilidad propia, y cuyo giro lo puede realizar en uno o más establecimientos, que desarrollan alguna actividad principal entre las señaladas más adelante.

Marco Muestral

El marco muestral se construye a partir del Directorio INE, año contable 2009; conformado por los registros del SII y directorios internos de levantamiento INE.

En el caso de las empresas industriales se utiliza el directorio de la Encuesta Nacional Industrial Anual (ENIA) del año 2010, con período de referencia año 2009, que contiene aquellas empresas que cuentan con establecimientos de 10 y más trabajadores, y cuyo nivel de ventas anuales son superiores a 2.400 UF.

En el caso de empresas en el área de Generación, Distribución de Energía Eléctrica, Gas y Agua, se utiliza el Directorio del Índice de Electricidad, Gas y Agua (EGA) del INE, año 2010, con periodo de referencia año 2009. Corresponde a un censo de empresas que realizan generación y distribución de electricidad, gas y agua. Las empresas Generadoras (productoras y autoproductoras) deben contar con una producción de más de 2 MWH a nivel nacional.

En Explotación de minas y canteras, se utiliza el Directorio del Índice de Minería del INE, año 2010, con período de referencia año 2009. Corresponde a un censo de empresas, de la mediana y gran minería consideradas como tales por el INE y el Servicio Nacional de Geología y Minería de Chile (SERNAGEOMIN).

Para los Otros Sectores, se emplea la información proveniente del Servicio de

[1] El lector interesado en mayores detalles, como los coeficientes de variación calculados para cada estrato, puede consultar el Informe Metodológico confeccionado por el INE (2012).

Impuestos Internos, del año 2009, considerando las empresas que declaran ventas anuales superiores a 2400 UF, que desarrollan dentro del país alguna de las siguientes actividades descritas en el Cuadro 1.1.

Cuadro 1.1: Clasificación según sector económico

Categoría	Descripción
A	Agricultura, Ganadería, Caza y Silvicultura.
B	Pesca.
C	Explotación de Minas y Canteras.
D	Manufacturas.
E	Suministro de Electricidad, Gas y Agua.
F	Construcción.
G	Comercio al por mayor y al por menor; Reparación de vehículos automotores, motocicletas, efectos personales y enseres domésticos.
H	Hoteles y Restaurantes.
I	Transporte Almacenamiento y comunicaciones.
J	Intermediación Financiera.
K	Actividades Inmobiliarias y de alquiler.
N	Actividades de servicios sociales y de salud.
O	Otras actividades de servicios comunitarios, sociales y personales del tipo servicio.

Diseño Muestral

La VII Encuesta de Innovación tiene un diseño muestral estratificado y sistemático. En primera instancia, la estratificación se realiza según región, sector económico y nivel de ventas. Los sectores económicos se clasifican de acuerdo al Cuadro 1.1, mientras que los niveles de ventas se clasifican de la siguiente manera:

Cuadro 1.2: Clasificación según ventas

Tamaño Empresa	Ventas Anuales (UF)	
	Límite Inferior	Límite Superior
Grande	100.000^a	∞^b
Mediana	25.000^a	100.000^b
Pequeña	2.400^a	25.000^b

[a] *Inclusive.* [b] *Exclusive.*

A continuación, los segmentos asociados a la industria manufacturera son reestratificados según los tipos de manufactura descritos en el Cuadro 1.3:

Cuadro 1.3: Clasificación según tipo de manufactura

Categoría	División	Descripción
D		Industria Manufacturera
	15	Elaboración de productos alimenticios y bebidas.
	20	Producción de madera y fabricación de productos de madera y de corcho, exeptos muebles, fabricación de árticulos de paja y de materiales trenzables.
	21	Fabricación de papel y productos de papel.
	24	Fabricación de sustancias y productos quimicos.
	27	Fabricación de metales comunes.
	28	Fabricación de productos elaborados de metal excepto maquinaria y equipo.
	31	Fabricación de maquinaria y aparatos eléctricos N.C.P.
	99	Resto de Industria Manufacturera.

Finalmente, cada uno de los estratos resultantes se divide en dos sub-estratos. El primero, correspondiente a los primeros dos percentiles de ventas anuales del estrato original, estará auto-representado[2]; mientras que el segundo será muestreado sistemáticamente según ventas anuales[3].

Determinación y Distribución del Tamaño Muestral

Inicialmente, el tamaño muestral (n) se determina de tal manera que, para alguna característica considerada clave en el estudio,

$$n = \text{ínf}\{m : \mathbb{P}\{|\bar{y}_N - \hat{\bar{y}}_m| \leq \epsilon\} \geq 1 - \alpha\}, \tag{1.1}$$

donde α y ϵ son valores establecidos por el diseñador[4], $\bar{y}_N$ es la media poblacional de dicha característica, y $\hat{\bar{y}}_m$ es su estimación dada una muestra de tamaño m[5]. En el caso particular de un muestreo aleatorio simple (SRS), se tiene que

$$n = \frac{z^2_{\alpha/2} S^2}{\epsilon^2 + z^2_{\alpha/2} S^2/N} \tag{1.2}$$

donde N es el tamaño poblacional, $z^2_{\alpha/2}$ es el cuantil $\alpha/2$ de una normal estándar y S^2 es la varianza poblacional de la característica considerada[6]. Ciertamente, esta solución no es válida para diseños más complejos. Sin embargo, en la medida en que el diseño efectivo genere estimadores más precisos que un SRS, (1.2) resulta

[2]Se dice que una unidad muestral está auto-representada si es muestreada con certeza. La auto-representación de unidades muestrales (especialmente primarias, o PSU) es una estrategia ampliamente utilizada en encuestas de empresas y de hogares. Vea, por ejemplo, el capítulo 4 de Heeringa (2010)

[3]En realidad, el diseño también contempla auto-representar los estratos demasiado pequeños para ser muestreados, pero este detalle es irrelevante para nuestro estudio.

[4]El parámetro ϵ se conoce como el *margen de error* de la encuesta.

[5]Por simplicidad, todos los cálculos de esta sub-sección son relativos a la sub-población no auto-representada.

[6]Esta fórmula es específica al SRS, pero es fácilmente generalizable a un SRS estratificado. Vea, por ejemplo, los capítulos 2 y 3 de Lohr (2010).

ser una estimación conservadora[7].

Por supuesto, el estimador (1.2) sólo es factible si se posee un estimador de S^2. Típicamente, este se obtiene mediante encuestas piloto. No obstante, si la variable clave sea binaria, existe una alternativa mucho más sencilla. Para poblaciones grandes y variables binarias, se tiene que $S \approx \bar{y}_N(1 - \bar{y}_N)$, valor que alcanza su máximo cuando $\bar{y}_N = 1/2$. Por lo tanto, reemplazando S^2 por $1/4$ en (1.2) se obtiene una estimación aún más conservadora que la original.[8]

Una vez establecido el tamaño muestral, este es repartido entre H los estratos no auto-representados de forma directamente proporcional a su nivel de ventas. En otras palabras, el tamaño muestral (relativo) del $h-$ésimo estrato no autorepresentado (n_h/n) estará dado por

$$n_h/n = \frac{\sum_{i=1}^{N_h} v_{ih}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} v_{ih}}, \tag{1.3}$$

donde N_h es el tamaño del $h-$ésimo estrato $(h = 1, \ldots, H)$ y v_{ih} es el nivel de ventas de su $i-$ésima empresa constituyente $(i = 1, \ldots, N_h)$.

Estimación

Suponga, por simplicidad, que se desea estimar el total de la característica y, $T_y = \sum_{h=1}^{H}\sum_{i=1}^{N_h} y_{ih}$.[9] En tal caso, el equipo investigador tras la encuesta propone dos método de estimación, dependiendo del tipo de variable que sea *"y"*. El primer método, recomendado cuando la característica de interés es cualitativa, es el estimador de Horvitz-Thompson (HT),

$$\widehat{T}_y^{\text{HT}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \pi_{ih}^{-1} y_{ih}, \tag{1.4}$$

donde $\pi_{ih} = n_h/N_h$ es el recíproco de la probabilidad de selección de la $i-$ésima empresa del $h-$ésimo estrato[10]. El segundo método, recomendado cuando la característica de interés es cuantitativa, es el estimador de razones separado (SR) que utiliza a v_{ih} como variable auxiliar,

$$\widehat{T}_y^{\text{SR}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{ih} y_{ih}, \quad \text{donde} \quad w_{ih} = \frac{\sum_{i=1}^{N_h} v_{ih}}{\sum_{i=1}^{n_h} v_{ih}} \tag{1.5}$$

[7]Resulta extraño constatar que la fórmula presentada en el Informe Metodológico coincide casi, pero no perfectamente, con esta definición. Lo más probable es que sea un error de tipeo.

[8]En realidad, el diseño también contempla un 30 % de sobre-muestro para compensar potenciales problemas de no-respuesta, pero este detalle es irrelevante para nuestro estudio.

[9]Asuma, sin pérdida de generalidad, que no existen subpoblaciones auto-representadas ni problemas de no-respuestas.

[10]Vea, por ejemplo, la sección 2.3 del capítulo de Berger & Tillé (2009).

1.3. Limitaciones de la Metodología Actual

Si bien la estrategia muestral de la VII Encuesta de Innovación es fácil de entender y de analizar (propiedades que considero tienen un gran valor), esta presenta al menos tres problemas importantes. El primero de ellos, el más obvio, es la arbitrariedad con se generan los estratos auto-representados (me refiero, específicamente, a la regla de los dos primeros percentiles).

El segundo, algo más sutil, es que ninguno de los dos estimadores propuestos posee una varianza estimada que sea diseño−insesgada. Esto se debe a que, en ausencia de una aleatorización previa, el muestreo sistemático no es más que un caso particular de un muestreo por conglomerados[11], y ya que sólo se selecciona un conglomerado por estrato, el estimador diseño-insesgado de la varianza deja de ser factible.

Note que, ya que en general el muestreo sistemático (SyS) permite obtener estimadores más precisos que un SRS[12], podríamos repetir la estrategia tras (1.2) y utilizar el estimador de la varianza apropiado para un SRS como un estimador conservador de la varianza real. Sin embargo, tal como aducen Brewer & Gregoire (2009), ¡la varianza estimada por esta vía termina siendo mayor que la que se habría obtenido utilizando un SRS desde un principio! En otras palabras, si bien la varianza teórica bajo un SyS es menor que la que se obtiene bajo un SRS, las varianzas estimadas terminan señalando todo lo contrario.

El tercero, es que el estimador SR puede tener un sesgo significativo. Si bien el estimador SyS posee en general un error cuadrático medio (MSE) menor al estimador al HT, su uso no se recomienda cuando los estratos muestrales son muy pequeños (como es precisamente el caso) pues cada una de las razones w_{ih} están sesgadas, y dichos sesgos pueden propagarse a través de los estratos[13]. Por supuesto, este problema puede ser fácilmente superado reemplazando el estimador SR por el estimador de razones combinado (CR).

$$\widehat{T}_y^{\mathrm{CR}} = \frac{\widehat{T}_y^{\mathrm{HT}}}{\widehat{T}_v^{\mathrm{HT}}} T_v, \tag{1.6}$$

Sin embargo, veremos que existen soluciones mucho más ambiciosas.

1.4. Metodologías Alternativas

Si bien no existe ninguna solución definitva a los problemas antes mencionados, sí existen una gran cantidad de soluciones por ejemplo. Por ejemplo, en relación

[11] Un hecho por lo demás bien documentado. Vea, por ejemplo, la sección 5.5 4 de Lohr (2010) y la (sub-)sección 3.2.3. de Brewer & Gregoire (2009)

[12] Puede probarse que esto ocurre siempre que el coeficiente de correlación intra-conglomerados (ICC) sea negativo. Abusando de las referencias, consulte la ecuación (5.33) del texto de Lohr (2010).

[13] Vea, por ejemplo, la sección 4.5 de Lohr (2010)

al primer problema, la regla de Glasser (1962) para determinar el punto de corte óptimo consiste en declarar como auto-representadas a todas las unidades cuya característica x (por ejemplo, las ventas) exceda $\bar{x}_N + \sqrt{NS_N^2/n}$, donde $\bar{x}_N$ y S_N^2 son la media y la varianza poblacionales de dicha característica[14].

En cuanto al segundo problema, existen varias vías de solución. En primera instancia, las varianzas de los estimadores podrían estimarse desde una perspectiva basada en un modelo razonablemente robusto a malas especificaciones[15]. Sin embargo, dado que el problema no yace en las variables sino que es inducido por el diseño muestral, los cambios al método de selección que se propondrán más adelante también son una solución satisfactoria.

Finalmente, con respecto al tercer problema, se sugiere utilizar únicamente el estimador HT, pero reemplazar el muestreo actual (que es sistemático) por el muestreo balanceado (BS) de Deville & Tillé (2004), el cual se expondremos a continuación.

El Muestro Balanceado y El Método del Cubo

Considere una *población finita* de la forma $\mathcal{F} = \{z_1, \ldots, z_N\}$. Suponga que el vector de características z_i puede ser particionado como $z_i = (y_i : x_i)$, donde x_i es un $L-$vector de variables auxiliares conocidas para el diseñador. Para un conjunto dado de probabilidades de inclusión $\pi_i = E[I(z_i \in \mathcal{S})]$, un diseño muestral $\mathcal{P}(.)$ está balanceado con respecto a x_i ssi $\widehat{T}_x^{\text{HT}} = T_x$ para toda muestra $\mathcal{S} \subseteq \mathcal{F}$ tal que $\mathcal{P}(\mathcal{S}) > 0$, o equivalentemente,

$$\widehat{T}_x^{\text{HT}} = T_x \quad a.s.[\mathcal{P}] \tag{1.7}$$

El muestreo balanceado (BS) generaliza muchos diseños consabidos. Por ejemplo, todo diseño de tamaño muestral fijo está balanceado con respecto a $\pi_i = E[I(z_i \in \mathcal{S})]$, pues

$$\widehat{T}_\pi^{\text{HT}} = \sum_{i=1}^{N} I(z_i \in \mathcal{S})\pi_i/\pi_i = \sum_{i=1}^{N} I(z_i \in \mathcal{S}) = \sum_{i=1}^{N} \pi_i = T_\pi$$

donde la tercera ecuación se debe a que el tamaño muestral, $n = \sum_{i=1}^{N} I(z_i \in \mathcal{S})$ es fijo, por lo cual $n = E[n] = \sum_{i=1}^{N} E[I(z_i \in \mathcal{S})] = \sum_{i=1}^{N} \pi_i$. En la misma línea, suponga ahora que $\mathcal{F}$ puede ser dividido en H estratos $\{\mathcal{F}_h\}_{h=1}^{H} = \{z_{1h}, \ldots, z_{N_h h}\}_{h=1}^{H}$; y que en cada uno de estos estratos se obtiene, de manera independiente, una muestra $\mathcal{S}_h$ mediante un SRS de tamaño n_h. Entonces, el diseño resultante está balanceado con a respecto a las variables $\{\delta_{ih}\}_{h=1}^{H} = \{I(z_{ih} \in \mathcal{F}_h)\}_{h=1}^{H}$ pues, para

[14]Tanto esta como otras reglas son descritos en detalle en Hidiroglou & Lavallée (2009)

[15]Por ejemplo, el estimador de la varianza de Opsomer et al. (2012) no sólo es no-paramétrico, sino que su varianza anticipada converge a la inducida por el diseño.

cada $h = 1, \ldots, H$, se tiene que

$$\widehat{T}_{\delta_h}^{\text{HT}} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} I(z_{ih} \in \mathcal{S}_h)\delta_{ih}/\pi_{ih} = \sum_{h=1}^{H} \delta_{1h} \sum_{i=1}^{N_h} I(z_{ih} \in \mathcal{S}_h)/\pi_{ih} = \sum_{h=1}^{H} \delta_{1h} N_h = T_{\delta_h}$$

Geométricamente, el muestreo balanceado puede plantearse en los siguientes términos. Defina el vector $s = (s_1, \ldots, s_N) \in \{0, 1\}^N$, donde $s_i = I(z_i \in \mathcal{S})$. Dado un vector de probabilidades de inclusión $\pi = (\pi_1, \ldots, \pi_N)$, un diseño muestral $\mathcal{P}(.)$ está balanceado con respecto a x_i si y solo si

$$s \in Q := \pi + \text{Ker}\{[\pi_1^{-1} x_1 \ldots \pi_N^{-1} x_N]\} \quad a.s.[\mathcal{P}] \tag{1.8}$$

Así pues, seleccionar un muestra mediante bajo un BS equivale a seleccionar aleatoriamente un punto de $K = \{0, 1\}^N \cap Q$ de tal forma que $E[s] = \pi$.

La complejidad del problema antes mencionado (especialmente para N grande y $L > 2$) ha suscitado una gran cantidad de algoritmos de selección. Entre estos, el más popular es el *método del cubo*[16] de Deville & Tillé (2004), el cual se basa en generar un proceso que satisfaga las restricciones impuestas por las variables auxiliares ($s \in K$) y las probabilidades de inclusión ($E[s] = \pi$) que *eventualmente* converja a un vértice del cubo $C = [0, 1]^N$. Específicamente, suponga que existe un proceso $\pi(t)$ que satisface las siguientes propiedades:

i. $E[\pi(t)] = E[\pi(t-1)] = \ldots = E[\pi(0)] = \pi$;
ii. $\sum_{i=1}^{N} \{\pi_i^{-1} x_i\} \pi_i(t) = \sum_{i=1}^{N} x_i = T_x$;
iii. Cuando el proceso $\pi(t)$ alcanza una cara de C, este no la abandona.

Entonces, en la medida en que $\pi(t)$ converja a un vértice de C, dicho límite correspondería (indirectamente) a una muestra bajo BS[17]. Por supuesto, no existe un único proceso con dichas propiedades, pero en la práctica la familia de procesos sugeridos por Deville & Tillé (2004) (Algoritmo 8.3) generan estimadores con buenas propiedades.

Note que el diseño muestral resultante será general bastante complejo, por lo que la fórmula estándar de $V(\widehat{T}_y^{\text{HT}})$ requerirá el conocimiento de las probabilidades de inclusión conjunta $\pi_{ij} = E[s_i s_j]$. Sin embargo, dado que la entropía del BS es relativamente alta, estimar $V(\widehat{T}_y^{\text{HT}})$ suponiendo un muestreo de Poisson Condicional (CPS) genera una buena aproximación[18].

[16] Podría pensarse que el nombre al algoritmo proviene del cubo $C = [0, 1]^N$. Sin embargo, esta no es un hipótesis con mucho fundamento, pues cualquier algoritmo debiese arrojar un punto en dicho conjunto. El verdadera intención del nombre del algoritmo es enfatizar la naturaleza geométrica del problema.

[17] Por supuesto, existen situaciones en las cuales no existe ningún proceso que convergenete. En ese caso, Deville & Tillé (2004) proponen un segundo algoritmo con el fin de obtener al menos una solución aproximada.

[18] Vea, por ejemplo, la sección 8.8 de Deville & Tillé (2004).

Si bien el BS fue motivado como un diseño capaz de explotar la información contenida en un vector de variables auxiliares, posee al menos dos ventajas adicionales para esta encuesta en particular. La primera, es que permite combinar el balanceo con un vector arbitrario de probabilidades de inclusión. Esto es importante pues permite, por ejemplo, definir π_i de forma directamente proporcional al volumen de ventas (lo cual es, de hecho, la práctica más común en este tipo de encuestas). La segunda, más sutil, es que permite balancear el diseño muestral no solo con respecto a *dummies* identificadoras de estratos (como expusimos anteriormente) sino también con respecto a *dummies* identificadoras de estratos traslapados[19]. Esto es especialmente relevante para la VII encuesta de innovación pues no todas las empresas tienen sus actividades en una sola región[20].

1.5. Comparación simulada entre los distintos métodos

Si bien el diseño balanceado posee varias ventajas teóricas por sobre el actual diseño, no existe ninguna manera conocida de medir su precisión relativa de manera analítica. Para suplir este vacío, en esta sección realizaremos un pequeño estudio de Monte Carlo que permita dar luces sobre el desempeño relativo de los diseños discutidos. Específicamente, se simularon $S = 500$ muestras de tamaño $n = 300$ de una población de tamaño $N = 3000$ que poseía la siguientes características:

$$
\begin{aligned}
x_{1i} &\overset{iid}{\sim} exp(1); \qquad x_{2i} \overset{iid}{\sim} exp(1); \qquad x_{3i} \overset{iid}{\sim} \mathcal{N}(2,1) \\
y_{1i} &= 2 + 2 * x_{1i} + 0{,}0x_{1i}^2 + 0x_{2i} + 0x_{3i} + \mathcal{N}(0,1); \\
y_{2i} &= 2 + 2 * x_{1i} + 0{,}3x_{1i}^2 + 0x_{2i} + 0x_{3i} + \mathcal{N}(0,1); \\
y_{3i} &= 2 + 2 * x_{1i} + 0{,}0x_{1i}^2 + 2x_{2i} + 0x_{3i} + \mathcal{N}(0,1); \\
y_{4i} &= 2 + 0 * x_{1i} + 0{,}0x_{1i}^2 + 2x_{2i} + 2x_{3i} + \mathcal{N}(0,1);
\end{aligned}
$$

durante este experimento se consideraron dos métodos de muestreo, uno sistemático (en el cual la variable x_{1i} se usó como llave) y uno balanceado en las variables x_{1i}, x_{2i}, x_{3i} y la constante. La idea de usar 4 series es medir el desempeño relativo de los métodos bajo distintos escenarios. Los primeros 2 son claramente ventajosos para el diseño sistemático (aunque son muy poco realistas), el tercero es relativamente neutro y el cuarto es abiertamente favorable al diseño balanceado.

En cada una de las 500 simulaciones, se calcularon 3 estimadores. Los primeros dos, $\widehat{T}_{1s}$ y $\widehat{T}_{2s}$, fueron los estimadores Horvitz-Thompson y de Razones para el total de la $s-$ésima muestra simulada bajo el diseño sistemático; mientras que el tercero, $\widehat{T}_{3s}$, fue el estimador Horvitz-Thompson para el total de la $s-$ésima muestra simulada bajo el diseño balanceado. Una vez que cada uno de estos 3 estimadores fue calculado 500 veces, su precisión fue aproximada mediante su

[19] Vea, por ejemplo, la (sub-)sección 8.7.2 de Deville & Tillé (2004).

[20] De hecho, el informe metodológico nunca explicita la manera en la cual trata este problema.

desviación relativa promedio:

$$\overline{\mathsf{RD}}(\widehat{T}_i) = \frac{1}{S}\sum_{s=1}^{S} 100 \times \frac{|\widehat{T}_{is} - T_i|}{T_i}, \qquad i = 1, 2, 3$$

Los resultados de este experimento se resumen en la siguiente tabla:

Cuadro 1.4: Desviación Relativa Promedio según variable y estrategia:

Estrategia Muestral	Variable			
	y_1	y_2	y_3	y_4
$\widehat{T}_1$	0.7937	1.9331	2.9381	2.6065
$\widehat{T}_2$	0.7827	1.4865	2.7390	2.4636
$\widehat{T}_3$	1.6904	2.1729	1.3067	0.7670

Los resultados de este experimento permiten afirmar que $\widehat{T}_1$ y $\widehat{T}_2$ solo superan a $\widehat{T}_3$ en las situaciones poco realistas 1 y 2, siendo claramente superados no solo en la situación 4 (la favorable al Método propuesto) sino también en la 3 (la que es relativamente neutra).

1.6. Resumen y Comentarios

Este estudio discute la fortalezas y debilidades de la actual estrategia muestral de la VII Encuesta de Innovación, así como las de sus potenciales alternativas. La principal fortaleza de la actual estrategia es su simplicidad, mientras que sus principales debilidades son (a) la arbitrariedad con que selecciona los estratos auto-representados, (b) su muestreo sistemático y (c) el potencial sesgo asociado al estimador recomendado para estimar totales y/o promedio de variables cuantitativas. Las principales fortalezas de la estrategia alternativa son: la solución aproximada que ofrece a los problemas (a) y (b), la capacidad de balancear el diseño con respecto a varias variables de manera simultánea, la posibilidad de combinar esta propiedad con un muestreo probabilístico desigual y de lidiar con estratos traslapados; mientras que su principal debilidad es el desafío que representa no sólo su análisis sino su entendimiento.

Antes de concluir, quisiera realizar dos comentarios a título personal. En primer lugar, quisiera recalcar que, si bien el estudio se enfocó en el análisis en una encuesta externa, la mayoría de las críticas a su metodología también aplican a varias encuestas en las cuales colabora el Centro de MicroDatos de la Universidad de Chile (*e.g.*, la Encuesta Longitudinal de Empresas[21]). Con esto no quiero acusar en ningún caso negligencia por parte del equipo investigador, sino la necesidad de que se abra a nuevas metodologías. En segundo lugar, quisiera expresar (una vez más) la gran falta que hace un curso de teoría del muestreo en nuestra facultad

[21] De hecho, existe un preocupante nivel de *Copy-Paste* en los Informes Metodológicos de varias de encuestas claves a nivel nacional.

(tanto a nivel de pregrado como de postgrado). En toda la carrera, si bien existen varios cursos de Estadística (y luego de Econometría), no hay un solo curso en el que se haga siquiera una mención seria al diseño y/o análisis de una encuesta compleja[22]. Sin embargo, ¿podría alguien discutir la relevancia de esta teoría en el quehacer de un economista en la actualidad? Creo sinceramente que el país no podrá volver a confiar en sus estadísticas oficiales hasta que se avance en esta materia.

[22]Aunque sí existen cursos de magíster en los cuales se exponen los conceptos elementales

A

Introducción al Muestreo de Poblaciones Finitas

A.1. Introducción

Uno de los mayores objetivos en Estadística es caracterizar poblaciones finitas de elementos *distinguibles*. Por ejemplo, se podría desear saber (*i*) el porcentaje de empresas chilenas que realizan alguna innovación, o bien (*ii*) si el actual el porcentaje de pobres en Chile ha bajado en comparación con el del año anterior. Ahora, examinar el número de innovaciones de cada empresa o el ingreso de cada persona en Chile sería excesivamente lento y costoso. Así pues, parece natural inspeccionar solo una pequeña parte de la población, una *muestra representativa*, con el fin de *estimar* las características de la población a partir de ella. Sin embargo, esta aparente *solución* no está exenta de polémicas. En particular,

- ¿Cómo *seleccionar* la muestra: determinística o aleatoriamente?
- ¿Cómo *extrapolar* las estimaciones de la muestra a la población?
- ¿Cómo medir el *sesgo* y la *precisión* de dichas estimaciones?
- ¿Cómo realizar inferencias sobre las características estudiadas?

Con el fin de responder estas y otras preguntas coherentemente desde una perspectiva estadística, adoptaremos el enfoque propuesto por Neyman (1934), quien defiende el *muestreo probabilístico* en conjunto con una inferencia *basada en el diseño*. Mostraremos que, bajo ciertas condiciones de regularidad, este enfoque permite realizar inferencias precisas incluso utilizando muestras relativamente pequeñas. Los métodos que discutiremos a continuación solo consideran el caso ideal de una población grande, estática, perfectamente identificada y libre de no−respuestas. Una teoría mucho más elaborada ha sido desarrollada en los últimos 50 años para enfrentar las violaciones a estos supuestos. Remitimos al lector al libro de Chaudhuri & Stenger (2005) para una introducción seria a estas materias, así como al manual editado por Pfeffermann & Rao (2009) para consultas específicas.

A.2. Definiciones Elementales

Considere un conjunto compuesto por N **unidades estadísticas** o UE (*v.gr.*, escuelas, empresas o personas). Cada una de estas unidades (digamos, la $i-$ésima) posee una **etiqueta identificadora** $i \in \{1, \ldots, N\}$ y un **vector de características** $y_i = (y_{1i}, \ldots, y_{Ki})' \in \mathbb{R}^K$, el cual es inicialmente desconocido para un investigador que desea estimar un **parámetro de interés** $\psi = \Psi(\{(y_i, i) : i \in \mathcal{U}\})$, tal

como

- *El Total Poblacional*, $t_y = \sum_{i \in \mathcal{U}} y_i$, o bien
- *La Media Poblacional*, $m_y = N^{-1} \sum_{i \in \mathcal{U}} y_i$.

El conjunto $\mathcal{U} = \{1, \ldots, N\}$ recibe el nombre de **población**, mientras que el conjunto $\mathcal{F} = \{(y_i, i) : i \in \mathcal{U}\}$ se conoce como **información poblacional**. Seleccionando n etiquetas (no necesariamente distintas) de esta población es posible generar un vector $s = (s_1, \ldots, s_N) \in \mathbb{N}_0^N$ cuyo $i-$ésimo elemento denota el número de veces que la etiqueta i es seleccionada. El vector s recibe el nombre de **muestra**, mientras que el conjunto $\mathcal{S} = \{(y_i s_i, s_i) : i \in \mathcal{U}\}$ se conoce como la **información muestral**. Una vez que una muestra $s \in \mathbb{N}_0^N$ es escogida, los números $n = \sum_{i \in \mathcal{U}} s_i$ y $n^* = \sum_{i \in \mathcal{U}} I(s_i > 0)$ reciben los nombres de **tamaño muestral bruto y efectivo**, respectivamente.

A.2.1. Observabilidad Completa

Tal como hemos señalado, el principal objetivo de la *teoría del muestreo* es realizar inferencias sobre un parámetro $\psi = \Psi(\mathcal{F})$ utilizando información muestral $\mathcal{S}$ que sea, en un sentido que pronto discutiremos, *representativa* de la información poblacional. Obviamente, esta estrategia solo es viable en la medida en que la información muestral sea al menos parcialmente observada. En este capítulo, asumiremos que dicha observación es **completa**:

Supuesto A.1 (Observabilidad Completa)**.** *Una vez que una muestra $s \in \mathbb{N}_0^N$ es escogida es posible, para cada $i \in \mathcal{U}$, observar y_i exactamente s_i veces, o equivalentemente, el conjunto $\mathcal{S} = \{(y_i s_i, s_i) : i \in \mathcal{U}\}$.* □

Por supuesto, esta es una situación ideal raramente observada en la práctica. De hecho, la gran mayoría de las encuestas exhiben, en mayor o menor grado, **errores de medición** y problemas de **no-respuesta**. Sin embargo, en la medida en que estos problemas sean tenues, las conclusiones derivadas de este supuesto siguen siendo razonables.

A.2.2. Muestreo Probabilístico

Si bien el mecanismo de muestreo puede ser completamente determinado por el investigador, en la teoría moderna del muestreo el interés cae casi exclusivamente en mecanismos de **muestreo probabilístico**:

Supuesto A.2 (Muestreo Probabilístico)**.** *Dada la información poblacional $\mathcal{F}$, la muestra $s \in \mathbb{N}_0^N$ es una variable aleatoria cuya medida de probabilidad $P(.|\mathcal{F})$ es conocida por el investigador.* □

La popularidad del muestreo probabilístico se debe, fundamentalmente, a dos motivos. El primero, es su imparcialidad. Mientras otros mecanismos de muestreo están sujetos a la manipulación por parte del investigador, en un muestreo probabilístico toda muestra tiene cierta probabilidad de ser escogida. El segundo, es

su robustez. Mientras otros mecanismos de muestreo requieren de supuestos paramétricos adicionales para hacer inferencias precisas, la inferencia posterior a un muestreo probabilístico es esencialmente no-paramétrica.

Note, sin embargo, que ni siquiera este mecanismo es completamente neutral o no-paramétrico. Esto se debe, entre otras causas, a que:

- No todas las muestras tienen necesariamente la misma probabilidad de ocurrencia (esta es, de hecho, la excepción más que la regla).
- Algunas preguntas de interés (en general las analíticas y en particular las causales) seguirán exigiendo la formulación de hipótesis no-testeables.

Así pues, el muestreo probabilístico no debiese ser visto como una panacea, sino como una herramienta estadística cuyos beneficios dependen, en gran medida, de la sabiduría con la cual se escoja el **diseño muestral** [*i.e.*, la *p.m.* $P(.|\mathcal{F})$]

Dependiendo de la forma particular del diseño muestral, podemos hacer las siguientes clasificaciones generales:

Definición A.3 (Diseño Informativo)**.** *Un diseño muestral es informativo si depende de las características de las unidades estadísticas.* □

Definición A.4 (Diseño sin Reemplazos)**.** *Un diseño muestral es sin reemplazos si, para todas las muestras posibles, el tamaño muestral bruto equivale al efectivo.* □

Definición A.5 (Diseño Informativo)**.** *Un diseño muestral es de tamaño fijo si todas las muestras posibles tienen el tamaño muestral.* □

A.2.3. Estadísticos Muestrales

Como en tantas otras áreas de la Estadística, la inferencia estadística en la Teoría del Muestreo sigue descansa sobre el concepto de **estadístico**:

Definición A.6 (Estadístico)**.** *Se conoce como estadístico a cualquier función conocida de la información muestral. A saber,* $\hat{\psi} = \hat{\Psi}(\mathcal{S})$. □

Ejemplos clásicos de estadísticos son:

- La *Muestra Efectiva*, $r = [I(s_i > 0)] \in \mathbb{R}^N$.
- La *Media Muestral*, $\hat{m}_y = n^{-1} \sum_{i \in \mathcal{U}_N} y_i s_i \in \mathbb{R}^K$.
- El *Total Muestral*, $\hat{t}_y = N\hat{m}_y \in \mathbb{R}^K$.

Si bien se asume (como en tantas otras ramas de la Estadística) que los estadísticos son realizaciones de variables aleatorias, en la Teoría del Muestreo esta aleatoriedad posee dos fuentes: una inducida por la información poblacional y otra inducida por el proceso de muestreo. Esta es una característica notable, pues permite redefinir varios conceptos estadísticos familiares condicionándolos a una realización particular de $\mathcal{F}$. A saber,

Definición A.7 (D-Esperanza). *Dado un estadístico* $\hat{\psi} = \hat{\Psi}(\mathcal{S})$*, se define su esperanza inducida por diseño, o* D-Esperanza*, como*

$$\mathsf{E}(\hat{\psi}|\mathcal{F}) = \sum_{s \in \mathbb{N}^N} \hat{\Psi}(\mathcal{S}) P(s|\mathcal{F}) \tag{A.1}$$

En otras palabras, la D-Esperanza *de un estadístico* $\hat{\psi}$ *corresponde a la Esperanza Matemática de* $\hat{\psi}$ *condicional a la información poblacional.* □

Definición A.8 (D-Insesgamiento). *Dados un estadístico* $\hat{\psi} = \hat{\Psi}(\mathcal{S})$ *y un parámetro* ψ*, se dice que* $\hat{\psi}$ *es un estimador* D-Insesgado *de* ψ *si* $\mathsf{E}(\hat{\psi}|\mathcal{F}) = \psi$*. En caso contrario, se dice que es* D-Sesgado *y la diferencia* $\mathsf{Bias}(\hat{\psi}|\mathcal{F}) = \mathsf{E}(\hat{\psi}|\mathcal{F}) - \psi$ *recibe el nombre de* D-Sesgo*.* □

Definición A.9 (D-Varianza). *Dado un estadístico* $\hat{\psi} = \hat{\Psi}(\mathcal{S})$*, se define su varianza inducida por diseño, o* D-Varianza*, como*

$$\mathsf{V}(\hat{\psi}|\mathcal{F}) = \sum_{s \in \mathbb{N}^N} [\hat{\Psi}(\mathcal{S}) - \mathsf{E}(\hat{\psi}|\mathcal{F})]^2 P(s|\mathcal{F}) \tag{A.2}$$

En otras palabras, la D-Varianza *de un estadístico* $\hat{\psi}$ *corresponde a la Varianza Matemática de* $\hat{\psi}$ *condicional a la información poblacional.* □

El Cuadro A.1 reune algunas D-Esperanzas y D-Varianzas de uso recurrente:

Cuadro A.1: D-Esperanzas y D-Varianzas Notables

Símbolo	Definición
μ	$\mathsf{E}[s\|\mathcal{F}]$
M	$\mathsf{E}[ss'\|\mathcal{F}]$
π	$\mathsf{E}[r\|\mathcal{F}]$
Π	$\mathsf{E}[rr'\|\mathcal{F}]$
Σ	$\mathsf{V}[s\|\mathcal{F}]$
Δ	$\mathsf{V}[r\|\mathcal{F}]$

Note que π y Π también pueden interpretarse como probabilidades. Específicamente, el $i-$ésimo elemento de $\pi \in \mathbb{R}^N$ equivale a la Probabilidad de que la $i-$ésima etiqueta aparezca al menos una vez en la muestra,

$$\pi_i = Pr(s_i > 0|\mathcal{F}) \tag{A.3}$$

mientras que el $(i,j)-$ésimo elemento de $\Pi \in \mathbb{M}_{N\times N}$ corresponde a la Probabilidad de que tanto la $i-$ésima como la $j-$ésima etiquetas aparezcan (conjuntamente) al menos una vez en la muestra,

$$\pi_{ij} = Pr(s_i > 0 \cap s_j > 0|\mathcal{F}) \tag{A.4}$$

Estas probabilidades, comúnmente conocidas como **probabilidades de inclusión de 1° y 2° orden**, jugarán un rol fundamental en secciones posteriores.

A.3. Estimación del Total y la Media Poblacional

La inferencia sobre el total y la media poblacional constituye el núcleo de la teoría del muestreo. Esto se debe, entre otras razones, a que

- La mayoría de las encuestas (tanto privadas como públicas) tienen como principal objetivo la estimación del total y/o la media poblacional de una pocas características (*v.gr.*, es bien sabido que el objetivo político de la encuesta CASEN es medir aspectos particulares la pobreza).
- Son dos de los pocos parámetros para los cuales existen estimadores D-Insesgados bien comportados.
- Si bien no todos los parámetros relevantes son un promedio o un total, muchos son una función de alguno (*v.gr.*, la razón entre dos totales, el estimador *Máximo Verosímil* y el estimador del *Método de los Momentos*).

En esta sección, presentaremos los dos estimadores elementales de estos parámetros: el de **Hansen-Hurwitz** y el de **Horvitz-Thompson**.

A.3.1. El estimador de Hansen-Hurwitz

Considere por un momento el total muestral $\hat{t}_y = (N/n)\sum_{i\in\mathcal{U}} y_i s_i$. Si bien este parece ser el estimador *natural* del total poblacional $t_y = \sum_{i\in\mathcal{U}} y_i$, resulta que ni siquiera es un estimador D-Insesgado pues, al menos en un diseño de tamaño fijo, su D-Esperanza está dada por

$$\mathsf{E}(\hat{t}_y|\mathcal{F}) = (N/n)\sum_{i\in\mathcal{U}} y_i \mathsf{E}(s_i|\mathcal{F}) = (N/n)\sum_{i\in\mathcal{U}} y_i\mu_i \tag{A.5}$$

valor que, en general, es distinto de t_y. Esto se debe a que las observaciones con mayor probabilidad de inclusión tienden a estar sobrerrepresentadas en el total muestral y viceversa (siendo μ_i la medida exacta de esta sobre- o sub-representación). **El estimador Hansen-Hurwitz** (HH) corrige este problema igualando la *representatividad* de cada observación:

Definición A.10 (Estimador Hansen-Hurwitz)**.** *El estimador Hansen-Hurwitz de un total* t_y (*cf.* Hansen & Hurwitz (1943)) *está dado por*

$$\hat{t}_y^{HH} = \sum_{i\in\mathcal{U}} y_i s_i/\mu_i \tag{A.6}$$

en la medida en que $\mu_i > 0$ *para todo* $i \in \mathcal{U}$. □

No es difícil probar que la D-Esperanza y D-Varianza de este estimador están dadas por el siguiente par de ecuaciones:

$$\mathsf{E}(\hat{t}_y^{\text{HH}}|\mathcal{F}) = t_y, \tag{A.7}$$

$$\mathsf{V}(\hat{t}_y^{\text{HH}}|\mathcal{F}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{y_i y_j \Sigma_{ij}}{\mu_i\mu_j} \tag{A.8}$$

Para ver la base del primer resultado, note que y_i está contenido en $\mathcal{F}_N$. Por lo tanto, este puede ser considerado como una constante al momento de calcular la D-Esperanza:

$$\mathsf{E}(\hat{t}_y^{\text{HH}}|\mathcal{F}) = \sum_{i\in\mathcal{U}} y_i \mathsf{E}(s_i|\mathcal{F})/\mu_i = \sum_{i\in\mathcal{U}} y_i\mu_i/\mu_i = \sum_{i\in\mathcal{U}} y_i = t_y \tag{A.9}$$

Para ver la base del segundo resultado, simplemente defina el vector $\breve{y} = (y_1/\mu_1, \ldots, y_N/\mu_N)$ y repita el argumento del párrafo anterior:

$$\mathsf{V}(\hat{t}_y^{\text{HH}}|\mathcal{F}) = \mathsf{V}(\breve{y}'s|\mathcal{F}) = \breve{y}'\mathsf{V}(s|\mathcal{F})\breve{y} = \breve{y}'\Sigma\breve{y} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{y_i y_j \Sigma_{ij}}{\mu_i\mu_j} \tag{A.10}$$

En general, la D-Varianza del estimador Hansen-Hurwitz es desconocida para el investigador. Afortunadamente, es posible estimarla de manera relativamente sencilla. Específicamente, se tiene que

$$\widehat{\mathsf{V}}(\hat{t}_y^{\text{HH}}|\mathcal{F}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{s_i s_j y_i y_j \Sigma_{ij}}{\mu_{ij}\mu_i\mu_j} \tag{A.11}$$

es un estimador D-Insesgado de $\mathsf{V}(\hat{t}_y^{\text{HH}}|\mathcal{F})$, en la medida en que $\mu_{ij} > 0$ para todo $i, j \in \mathcal{U}$. La prueba de esta aseveración es análoga a las anteriores, por lo que se deja como ejercicio.

A.3.2. El estimador de Horvitz-Thompson

Suponga por un momento el muestreo es *sin reemplazos*. En dicho caso, se tiene que $s = r$, $\mu = \pi$ y $\Sigma = \Pi$, de manera tal que el estimador Hansen-Hurwitz se reduce a

$$\hat{t}_y^{\text{HT}} = \sum_{i\in\mathcal{U}} y_i r_i/\pi_i \tag{A.12}$$

Este es el estimador de **Horvitz-Thompson** (HT). Si efectivamente el muestreo es sin reemplazos, este estimador equivaldrá al de Hansen-Hurwitz y heredará sus propiedades. Sin embargo, si el muestreo tiene reemplazos este estimador puede no ser siquiera D-Insesgado, por lo que el estimador preferido será el que posea un menor Error Cuadrático Medio inducido por el diseño.

A.3.3. Estimación de la Media Poblacional

Suponga, por simplicidad, que dada ha estimado un total poblacional t_y con el estimador de Horvitz-Thompson, y que el muestreo ha sido sin reemplazos. Entonces, un estimador obvio de la media poblacional sería

$$\hat{m}_y^{\text{HT}} = N^{-1}\sum_{i\in\mathcal{U}} y_i r_i/\pi_i \tag{A.13}$$

Por desgracia, el tamaño muestral N no siempre es conocido por el investigador. En tal caso, la solución más común es reemplazar $N = \sum_{i\in\mathcal{U}} r_i$ por su propio estimador Horvitz-Thompson, $\hat{N}^{\text{HT}} = \sum_{i\in\mathcal{U}} r_i^2/\pi_i = \sum_{i\in\mathcal{U}} r_i^2/\pi_i$. Por supuesto, este estimador ya no será D-Insesgado, pero seguirá siendo un estimador razonable en la medida en que $\hat{N}^{\text{HT}}$ sea una buena aproximación del tamaño muestral[1].

A.4. Algunos Diseños Muestrales Populares

Hasta ahora, solo hemos hecho algunas definiciones elementales y deducido algunos estimadores generales. Es tiempo de ver cómo se aplican estos conceptos a algunos diseños en particular. Por simplicidad, esta sección se enfoca en los 4 tipos de muestreo más simples y populares (al menos en Chile). Sin embargo, existen decenas (quizás cientos) de diseños muestrales teóricamente atractivos. El lector queda referido al texto de Tillé (2006) para una introducción a estos diseños.

A.4.1. Muestreo Aleatorio Simple

En un muestreo aleatorio simple, una muestra de tamaño fijo n es seleccionada de manera secuencial, siendo una etiqueta es seleccionada de manera completamente aleatoria en cada etapa hasta reunir n de ellas. Si cada etiqueta puede ser seleccionada a lo más una vez se dice que es un muestreo aleatorio simple sin reemplazo (SRSWOR). En caso contrario, se dice que un muestreo aleatorio simple con reemplazo (SRSWR).

En el caso particular del SRSWOR, las fórmulas (A.9)-(A.11) se reducen a:

$$\hat{t}_y^{\text{HT}} = \hat{t}_y \tag{A.14}$$

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \frac{N^2}{n} S^2 \left(1 - \frac{n}{N}\right) \tag{A.15}$$

$$\widehat{\mathsf{V}}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \frac{N^2}{n} s^2 \left(1 - \frac{n}{N}\right) \tag{A.16}$$

donde $S^2 = \sum_{i\in\mathcal{U}_N}(y_i - m_y)^2/(N-1)$ y $s^2 = \sum_{i\in\mathcal{U}_N} r_i(y_i - \hat{m}_y)^2/(n-1)$ representan la varianza poblacional y muestral, respectivamente. Para entender la razón del primer resultado, simplemente note que solo $\binom{1}{1}\binom{N-1}{n-1}$ de las $\binom{N}{n}$ muestras posibles contienen la etiqueta i. Por lo tanto, su probabilidad de inclusión está dada por

$$\pi_i = \binom{N}{n}^{-1}\binom{1}{1}\binom{N-1}{n-1} = \frac{n}{N}$$

y el estimador HT se reduce a $\hat{t}_y^{\text{HT}} = (N/n)\sum_{i\in\mathcal{U}} y_i r_i = \hat{t}_y$, tal como se declaraba.

Para probar la segunda ecuación, simplemente note que solo $\binom{1}{1}\binom{N-2}{n-2}$ de las $\binom{N}{n}$ muestras posibles contienen las etiquetas i j cuando $i \neq j$. Por lo tanto, sus

[1] Puede demostrarse que este estimador alternativo no es más que un caso particular del estimador de razones, vea la sección.

probabilidades de inclusión de 2^{o} orden están dadas por

$$\pi_{ij} = \begin{cases} \binom{N}{n}^{-1}\binom{1}{1}\binom{1}{1}\binom{N-2}{n-2} = \frac{n(n-1)}{N(N-1)} & \text{si } i \neq j \\ \binom{N}{n}^{-1}\binom{1}{1}\binom{N-1}{n-1} = \frac{n}{N} & \text{en caso contrario} \end{cases}$$

de forma tal que cada uno de los coeficientes $A_{ij} \equiv \mu_i^{-1}\mu_j^{-1}\Sigma_{ij}$ se reducen a

$$\begin{aligned} A_{ij} &= \pi_i^{-1}\pi_j^{-1}\Sigma_{ij} \\ &= \pi_i^{-1}\pi_j^{-1}(\pi_{ij} - \pi_i\pi_j) \\ &= \begin{cases} \frac{N}{n} - 1 & \text{si } i = j \\ \frac{N(n-1)}{n(N-1)} - 1 & \text{en caso contrario} \end{cases} \end{aligned}$$

Reemplazando estos coeficientes en (2.10) obtenemos la siguiente fórmula

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = y'Ay$$

donde, como es costumbre, A es una matriz cuyo $(i,j)-$ésimo elemento está dado por A_{ij}. Se podría pensar que A es una matriz irregular, pero tiene mucha más estructura de la que aparenta. De hecho, basta examinarla con cuidado para notar que

$$A = \frac{N(N-n)}{n(N-1)} M_{\mathbb{1}_N}$$

donde, como también es habitual, $M_{\mathbb{1}_N} \equiv I_N - \mathbb{1}_N\mathbb{1}_N'/N$. Uniendo estas últimas ecuaciones (y recordando que $S^2 = y'M_{\mathbb{1}_N}y/(N-1)$), se deduce el resultado

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \frac{N(N-n)}{n(N-1)} y'M_{\mathbb{1}_N}y = \frac{N^2}{n}S^2\left(1 - \frac{n}{N}\right)$$

La prueba del resultado (A.16) es similar a la anterior, por lo que se deja propuesta.

A.4.2. Muestreo Estratificado

Considere una población $\mathcal{U}$ dividida en H partes no traslapadas o **estratos** de tamaños $N_1, \ldots, N_H$. Es decir, $\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h$, donde $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ para todo $i \neq j$. Un diseño dice ser estratificado si en cada estrato $\mathcal{U}_h$ se selecciona una muestra aleatoria de tamaño fijo n_h de manera estadísticamente independiente.

Al igual que en el SRSWOR, el estimador HT y su varianza pueden deducirse directamente de su definición. Sin embargo, dado que las muestras seleccionadas en cada estrato son independientes entre sí, el estimador HT, su varianza y su

varianza estimada se reducen a

$$\hat{t}_y^{\text{HT}} = \sum_{h=1}^{H} \hat{t}_{yh}^{\text{HT}} \tag{A.17}$$

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \mathsf{V}(\hat{t}_{yh}^{\text{HT}}|\mathcal{F}) \tag{A.18}$$

$$\widehat{\mathsf{V}}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \widehat{\mathsf{V}}(\hat{t}_{yh}^{\text{HT}}|\mathcal{F}) \tag{A.19}$$

donde $\hat{t}_{yh}^{\text{HT}}$, $\mathsf{V}(\hat{t}_{yh}^{\text{HT}}|\mathcal{F})$ y $\widehat{\mathsf{V}}(\hat{t}_{yh}^{\text{HT}}|\mathcal{F})$ corresponden al estimador HT, la varianza teórica y la varianza estimada del total de $h-$ésimo estrato, respectivamente. La prueba del primer resultado es directa:

$$\hat{t}_y^{\text{HT}} = \sum_{i\in\mathcal{U}} \left(r_i y_i/\pi_i^{-1}\right) = \sum_{h=1}^{H} \left(\sum_{i\in\mathcal{U}_h} r_i y_i/\pi_i^{-1} \right) = \sum_{h=1}^{H} \hat{t}_{yh}^{\text{HT}}$$

mientras que la del segundo se debe fundamentalmente a que la independencia de las muestras de cada estrato se transmite a sus estimadores HT:

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \mathsf{V}(\sum_{h=1}^{H} \hat{t}_{yh}^{\text{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \mathsf{V}(\hat{t}_{yh}^{\text{HT}}|\mathcal{F})$$

En particular, para el caso de un SRSWOR estratificado, se tiene que:

$$\hat{t}_y^{\text{HT}} = \sum_{h=1}^{H} \hat{t}_{yh}$$

$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} S_h^2 \left(1 - \frac{n_h}{N_h}\right)$$

$$\widehat{\mathsf{V}}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} s_h^2 \left(1 - \frac{n_h}{N_h}\right)$$

A.4.3. Muestreo por Conglomerados

Considere nuevamente una población $\mathcal{U}$ dividida en M conjuntos o **conglomerados** de tamaños $N_1, \ldots, N_M$. Es decir, $\mathcal{U} = \bigcup_{i=1}^{M} \mathcal{U}_i$, donde $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ para todo $i \neq j$. Un diseño dice ser por conglomerados si un número fijo m de dichos conglomerados son seleccionados mediante algún tipo de muestreo probabilístico.

En el caso particular de un SRSWOR, se tiene que:

$$\hat{t}_y^{\text{HT}} = \sum_{h=1}^{H} t_{yh}$$
$$\mathsf{V}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \frac{N^2}{n} S_t^2 \left(1 - \frac{n}{N}\right)$$
$$\widehat{\mathsf{V}}(\hat{t}_y^{\text{HT}}|\mathcal{F}) = \frac{N^2}{n} s_t^2 \left(1 - \frac{n}{N}\right)$$

donde t_{yi} es el total del $i-$ésimo conglomerado (el cual, recordemos, sí es observado), S_t^2 es la Varianza teórica de los totales de dichos conglomerados y s_t^2 es su varianza estimada.

La intuición tras estos resultados es bastante simple. Claramente un SRSWOR por conglomerados equivale a un SRSWOR ordinario si definimos a los conglomerados como las unidades estadísticas. Por lo tanto, las fórmulas (2.14)−(2.16) siguen siendo aplicables, siempre y cuando se utilicen para calcular el total de una característica que sea completamente observada a nivel de conglomerado. Notando que el total de cada conglomerado satisface esta condición, se infieren los tres resultados.

Note que ahora N representa el número de conglomerados, no el de observaciones en la población. Por lo tanto, para utilizar el estimador de la varianza antes mencionado es necesario contar con al menos dos conglomerados.

A.4.4. Muestreo Sistemático

Considere una población finita de la forma $\mathcal{U} = \{1, \ldots, N\}$, donde N es un múltiplo de n, el tamaño pretendido para una muestra. Se dice que un diseño es sistemático si las n etiquetas seleccionadas son $\mathcal{U}_k = \{k+i[N/n] : i = 0, \ldots, n-1\}$, donde k es un entero entre 1 y $[N/n]$ escogido completamente al azar. A diferencia de los diseños vistos anteriormente, las propiedades del diseño sistemático depende del criterio mediante el cual se etiquetaron las unidades estadísticas. Si las etiquetas fueron dispuestas completamente al azar, el muestreo sistemático se asemeja a un SRSWOR. Sin embargo, si estas fueron dispuestas de manera intencionada, el diseño sistemático resulta ser un caso particular del diseño por conglomerados, en el cual los conglomerados son los conjuntos $\mathcal{U}_1, \ldots, \mathcal{U}_{[N/n]}$. Note, sin embargo, que solo uno de estos conglomerados es seleccionado, de manera tal que el estimador de la varianza derivado en la sección anterior no es aplicable. Existen numerosas *pseudo−soluciones* a este problema, pero ninguna es enteramente satisfactoria[2] (vea, por ejemplo, el capítulo 8 del libro de Wolter (2007))

[2]En cualquier caso, ninguna de ellas parece ser considerada en los cálculos presentados en el manual de la encuesta.

A.5. Determinación del tamaño muestral

Hasta ahora, siempre hemos considerado al tamaño muestral como un número definido de manera arbitraria. Sin embargo, en la práctica resulta vital utilizar un tamaño muestral que asegure cierta precisión por parte de los estimadores. En esta sección, explicaremos cómo las reglas más utilizadas para el SRSWOR y el SRSWOR estratificado.

A.5.1. Reglas para un SRSWOR

Suponga que desea establecer el mínimo tamaño muestral n^* necesario para que, con una probabilidad mayor o igual a $1-\alpha$, el estadístico $(\hat{m}_y^{\text{HT}} - m_y)$ tenga un margen de error menor o igual a ϵ utilizando un SRSWOR. En términos más formales:

$$n^* = \text{ínf}\{n : \mathbb{P}\{|\hat{m}_y^{\text{HT}} - m_y| \leq \epsilon\} \geq 1-\alpha\}, \tag{A.20}$$

Por supuesto, resulta imposible calcular n^* sin conocer la distribución de $|\hat{m}_y^{\text{HT}} - m_y|$. Sin embargo, en la medida en que las muestras discutidas sean suficientemente grandes, la aproximación $[\widehat{\mathsf{V}}(\hat{m}_y^{\text{HT}}|\mathcal{F})]^{-1/2}(\hat{m}_y^{\text{HT}} - m_y) \sim \mathcal{N}(0,1)$ continúa siendo razonable. Combinando esta aproximación con las ecuaciones (A.14)−(A.15) y (A.20), se deduce que

$$n^* = \frac{z_{\alpha/2}^2 S^2}{\epsilon^2 + z_{\alpha/2}^2 S^2/N} \tag{A.21}$$

que no es más que la fórmula conjeturada en el primer capítulo de esta tesis.

A.5.2. Reglas para un SRSWOR estratificado

Suponga ahora que desea establecer el mínimo tamaño muestral n^* necesario para que, con una probabilidad mayor o igual a $1-\alpha$, el estadístico $(\hat{m}_y^{\text{HT}} - m_y)$ tenga un margen de error menor o igual a ϵ utilizando un SRSWOR estratificado. Al igual que antes, el problema se formaliza como

$$n^* = \text{ínf}\{n : \mathbb{P}\{|\hat{m}_y^{\text{HT}} - m_y| \leq \epsilon\} \geq 1-\alpha\},$$

Sin embargo, el problema es ahora más complicado pues no solo es necesario establecer dicho tamaño muestral sino también cómo será repartido entre los estratos. Típicamente, este problema se resuelve en dos etapas. En la primera, los tamaños muestrales relativos n_h/n son determinados mediante alguna técnica externa (por ejemplo, puede determinarse que el tamaño de cada estrato sea proporcional a cierta variable auxiliar [*i.e.* una variable que siempre es observada]). En la segunda, por otra parte, el tamaño muestral n^* es determinado utilizando

la aproximación normal y la siguiente aproximación conservadora de $\mathsf{V}(\hat{m}_y^{\mathrm{HT}}|\mathcal{F})$:

$$\mathsf{V}(\hat{m}_y^{\mathrm{HT}}|\mathcal{F}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \leq \frac{1}{n}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{n}{n_h}\right) S_h^2 \equiv \frac{v}{n} \quad \text{(A.22)}$$

Combinando la aproximación normal con las ecuaciones (A.20) y (A.22), se deduce que el mínimo tamaño muestral necesario es $n^* = z_{\alpha/2}^2 v/\epsilon^2$.

B

Códigos Relacionados

Cone el fin de implementar el Método del Cubo en MATLAB, he desarrollado una bateria de programa reunidos en el archivo `matlabcodes.7z`. El evaluador no debería tener problemas ejecutando el programa y las simulaciones, siempre y cuando guarde todos los programas en un directorio adecuado. Ahora bien, el programa principal, `balancedsampling.m` trae un pequeño manual y algunas notas aclaratorias en su cuerpo, pero debo reconocer que el programa es demasiado complicado para que un tercero pueda verificar su validez. Para facilitar esta tarea, adjunto los pseudo-códigos de cada una de las fases del Método en Cuestión:

Algoritmo 2. Fase de Aterrizaje (Algoritmo Enumerativo)

Argumentos:

i. El (pseudo−)vector de inclusión π^* calculado en el Algoritmo 1.

ii. La matriz A calculada en el Algoritmo 1..

Resultados:

i. Un vector de inclusión: $s = (s_1, \ldots, s_N) \in \mathbb{M}_{1\times N}$.

Procedimiento:

i. Defina el siguiente conjunto:

$$\mathcal{C}(\pi^*) = \{s \in [0,1]^N : (\pi_i^* \in \{0,1\} \Rightarrow \pi_i^* = s_i)\}$$

ii. Defina la siguiente función de pérdida:

$$Cost(s) = ||P_{A'}(s - \pi^*)||^2 = (s - \pi^*)'A'(AA')^{-1}A(s - \pi^*),$$

iii. Encuentre el diseño muestral óptimo resolviendo el siguiente PPL

$$\begin{aligned} \min_{p^*(.)} \quad & \sum_{s\in\mathcal{C}(\pi^*)} Cost(s)p^*(s) \\ s.a. \quad & \sum_{s\in\mathcal{C}(\pi^*)} p^*(s) = 1; \\ & \sum_{s\in\mathcal{C}(\pi^*)} sp^*(s) = \pi^*; \\ & p^*(s) \in [0,1] \quad \forall s \in \mathcal{C}(\pi^*) \end{aligned}$$

iv. Extraiga una muestra a partir del diseño muestral hallado en *iii*.

Algoritmo 1. Fase de Despegue (Algoritmo de Chauvet & Tillé (2006))

Argumentos:

i. Un vector de probabilidades de inclusión: $\pi = (\pi_1, \ldots, \pi_N) \in \mathbb{M}_{1\times N^*}$, donde N^* es el tamaño poblacional.

ii. Una *matriz de balanceo*: $X = [x_1 \ \ldots \ x_N] \in \mathbb{M}_{p\times N^*}$, donde p es el número de variables auxiliares.

Resultados:

i. Un (pseudo−)vector de inclusión: $\pi^* = (\pi_1^*, \ldots, \pi_N^*) \in \mathbb{M}_{1\times N}$.

Inicialización:

i. Descarte todas las observaciones cuya inclusión sea trivial. Llame al número de observaciones restantes N.

ii. Inicialice las siguientes matrices, vectores y escalares:

$$A = [\pi_1^{-1}x_1 \ \ldots \ \pi_N^{-1}x_N]; \qquad s = (\pi_1, \ldots, \pi_N); \qquad r = (1, \ldots, p+1);$$
$$B = [\pi_i^{-1}x_1 \ \ldots \ \pi_{p+1}^{-1}x_{p+1}]; \qquad \psi = (\pi_1, \ldots, \pi_{p+1}); \qquad k = p+2;$$

Actualización:

WHILE $k \leq N$; DO

i. Genere un vector $u' \in \mathrm{Ker}B$.

ii. Calcule los siguientes escalares:

$$\lambda_1^* = \sup\{\lambda_1 : \psi + \lambda_1 u \in [0,1]\}; \quad \lambda_2^* = \sup\{\lambda_2 : \psi - \lambda_2 u \in [0,1]\};$$

iii. Actualice:

$$\psi = \begin{cases} \psi + \lambda_1^* u, & \text{con probabilidad } \lambda_2^*/(\lambda_1^* + \lambda_2^*) \\ \psi - \lambda_2^* u, & \text{con probabilidad } \lambda_1^*/(\lambda_1^* + \lambda_2^*) \end{cases}$$

iv. FOR $i = 1, \ldots, p+1$, DO

IF $\psi(i) \in \{0, 1\}$, DO

IF $k \leq N$, DO

$\pi^*(r(i)) = \psi(i)$;
$r(i) = k$;
$\psi(i) = \pi^*(k)$;
FOR $j = 1, \ldots, p$, DO $B(j, i) = A(j, k)$; ENDFOR;
$k = k + 1$;

ELSE, DO

FOR $j = 1, \ldots, p+1$, DO $\pi^*(r(j)) = \psi(j)$; ENDFOR;

ENDIF;

ENDIF;

ENDFOR;

ENDWHILE;

C

Bibliografía

Berger, Y. & Tillé, Y. (2009) Sampling with Unequal Probabilities. In Pfeffermann, D. & Rao, C.R. (eds) *Handbook of Statistics Vol #29A: Sample Surveys: Design, Methods and Applications*, chapter 2. Amsterdam: Elsevier.

Brewer, K. & Gregoire, T. (2009) Introduction to Survey Sampling. In Pfeffermann, D. & Rao, C.R. (eds) *Handbook of Statistics Vol #29A: Sample Surveys: Design, Methods and Applications*, chapter 1. Amsterdam: Elsevier.

Chaudhuri, A. & Stenger, S. (2009) *Survey Sampling: Theory and Methods, Second Edition (Statistics: A Series of Textbooks and Monographs)*, CRC Press.

Chauvet, G., Tillé, Y. (2006) *A fast algorithm of balanced sampling.* Journal of Computational Statistics, v.21, n.1.

Deville, J., Särndal, C. (1992) *Calibration estimators in survey sampling.* Journal of the American Statistical Association 87, 376–382.

Deville, J. & Tillé, Y. (2004) *Efficient balanced sampling: The cube method.* Biometrika, 91, 893–912.

Hansen, M. and Hurwitz, W. (1943) *On the theory of sampling from finite populations.* Annals of Mathematical Statistics, 14, 333–362.

Heeringa, S. (2010) *Applied Survey Data Analysis.* Chapman & Hall/CRC.

Hidiroglou, M. & Lavallée, P. (2009) Sampling and Estimation in Business Surveys. In Pfeffermann, D. & Rao, C.R. (eds) *Handbook of Statistics Vol #29A: Sample Surveys: Design, Methods and Applications*, chapter 17. Amsterdam: Elsevier.

Instituto Nacional de Estadísticas (2012) *Informe Metodológico Muestra Efectiva - VII Encuesta de Innovación*.

Kott, P. (2009) Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In Pfeffermann, D. & Rao, C.R. (eds) *Handbook of Statistics Vol #29B: Sample Surveys: Inference and Analysis*, chapter 25. Amsterdam: Elsevier.

Lohr, S. (2010) *Sampling: Design and Analysis.* Brooks/Cole, Boston.

Neyman, J. (1934) *On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection.* Journal of the Royal Statistical Society, 97, 558–606, 1934.

Opsomer, J., Francisco-Fernández, M. & Li, X. (2012) *Model-Based Non-parametric Variance Estimation for Systematic Sampling*. Scandinavian Journal of Statistics, 39: 528–542.

Pfeffermann, D. & Rao, C.R. (eds.) (2009) *Handbook of Statistics 29A, Volume 29: Sample Surveys: Design, Methods and Applications*. North Holland.

Tillé, Y. (2006) *Sampling Algorithms*. Springer-Verlag, New York.

Wolter, K. (2007) *Introduction to Variance Estimation*. Springer-Verlag, New York.