



FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Pontificia Universidad Católica de Chile

Facultad de Matemáticas

EYP2417 - Muestreo

Proyecto de Muestreo

Encuesta CASEN 2022

Métodos y Diseño Muestral

Grupo 4

Alexander Pinto

Esteban Román

Julián Vargas

Francisca Sepúlveda

- 1 Introducción
- 2 Diseño Muestral
- 3 Plan de Análisis
- 4 Metodología Detallada



Encuesta CASEN

La Encuesta de Caracterización Socioeconómica Nacional (CASEN) tiene como objetivo medir las condiciones de vida de los hogares y la población en el territorio chileno.

- ✓ Permite estimar **indicadores de pobreza**, desigualdad e inclusión social a nivel nacional
- ✓ Representa a la población que reside en **viviendas particulares ocupadas** en todo el territorio nacional
- ✓ Excluye las denominadas **áreas especiales** (zonas de acceso restringido, alto costo o condiciones climáticas adversas)



Áreas Especiales Excluidas

CASEN 2022 excluye **11 comunas completas** y UPM específicas en 4 comunas adicionales:

Comunas excluidas:

- Ollagüe, Juan Fernández
- Isla de Pascua, Coquimbo
- Chaitén, Futaleufú
- Hualaihué, Palena
- Guaitecas, O'Higgins
- Antártica Chilena

UPM específicas:

- General Lagos (4 UPM)
- Colchane (5 UPM)
- Tortel (1 UPM)
- Cabo de Hornos (1 UPM)

Fuente: Diseño Muestral CASEN 2022; Nota Técnica N°3



Tipo de Diseño Muestral

CASEN 2022: Diseño Estratificado Bietápico

Diseño probabilístico, estratificado y bietápico

Etapa 1: Selección UPM

- PPT sistemática sin reemplazo de conglomerados (UPM)
- Variable de tamaño: **número de viviendas** por UPM
- 12.545 UPM seleccionadas
- 764 estratos (comuna × área × NSE)

Etapa 2: Selección viviendas

- MAS dentro de cada UPM
- Total: 106.856 viviendas
- Con sobremuestreo

Estratificación

Estratos definidos por: **Geografía** (335 comunas) × **Área** (urbano/rural) × **NSE**

Tamaño de Muestra CASEN 2022

Nivel	Tamaño Objetivo	Error Absoluto	Error Relativo	Tamaño con Sobremuestreo
País	71.028	0,4 %	3,3 %	106.856
Urbano	56.905	0,5 %	4,6 %	87.252
Rural	14.123	1,3 %	9,2 %	19.604

Fuente: *Manual Metodológico CASEN 2022*, p. 35



Marco Muestral y Niveles de Inferencia

Marco: MMV 2020

Base: Censo 2017 actualizado a 2020 con verificación local (335 comunas, 12.545 UPM)

Niveles de Inferencia:

- ✓ Nacional
- ✓ Nacional urbano/rural
- ✓ Regional (16 regiones)
- ✗ NO comunal

Exclusiones:

- 10 comunas especiales (Isla de Pascua, Juan Fernández, etc.)
- UPM específicas (11 UPM)
- Viviendas no elegibles

Fuente: Diseño Muestral CASEN 2022; Nota Técnica N°3



Objetivo 1: Brecha Salarial de Género

Objetivo

Cuantificar y explicar las diferencias salariales por género controlando por factores socioeconómicos, laborales y educativos

Hipótesis de Investigación

H1: Existe una brecha salarial significativa entre hombres y mujeres ($\mu_{\text{hombre}} > \mu_{\text{mujer}}$), incluso controlando por educación, edad, ocupación y composición del hogar

Variables principales:

- **sexo:** Sexo de la persona
- **ytrabajocorh:** Ingreso del trabajo principal corregido
- **esc:** Años de educación formal

Variables de control:

- **edad:** Edad de la persona
- **oficio4_08:** Ocupación
- **tot_per_h:** Total personas en el hogar



Estrategia de Contrastación

Estimación en tres etapas con creciente control de confusores

1. Análisis bivariado: Test de diferencia de medias

- `svytest(ytrabajocorh ~ sexo)`
- Reporta brecha bruta sin controles

2. Modelo ajustado (controles socioeconómicos):

- `svyglm(ytrabajocorh ~ sexo + esc + edad)`
- Aísla efecto directo del género

3. Modelo completo (controles laborales):

- Incluye `oficio4_08` y `tot_per_h`
- Evalúa mediación por segregación ocupacional

Decisión: Rechazar H_0 si $p < 0.05$ en modelo completo con coeficiente $\beta_{sexo} < 0$



Objetivo 2: Distribución de la Pobreza

Objetivo

Cuantificar disparidades territorial-educativas en la prevalencia de pobreza y sus determinantes estructurales

Hipótesis de Investigación

H2a: La tasa de pobreza en zona rural es significativamente mayor que en zona urbana ($p_{\text{rural}} > p_{\text{urbano}}$)

H2b: La educación reduce la probabilidad de pobreza, con efecto más pronunciado en zonas urbanas

Variables principales:

- **pobreza:** Condición de pobreza
- **ytotcorh:** Ingreso total corregido
- **zona:** Rural/Urbana

Variables de control:

- **esc:** Años de educación
- **región:** Región del país
- **edad:** Edad de la persona
- **expr:** Factor de expansión



Estrategia de Contrastación Secuencial

Tres análisis complementarios para triangular evidencia

1. H2a - Test bivariado:

- `svychisq(~ pobreza + zona)` - Prueba de independencia
- `svyttest(ytotcorh ~ zona)` - Diferencia de ingresos

2. H2b - Modelo logístico principal:

- `svyglm(pobreza ~ zona + esc + edad, family=quasibinomial())`
- Comparar OR de educación entre zonas

3. H2b - Análisis de mediación:

- Estimar efecto indirecto de zona vía educación
- Test de Sobel para significancia del efecto mediador



Estimadores

Estimador de Horvitz-Thompson (Media ponderada)

$$\hat{Y} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$$

donde $w_i = \text{expr}_i$ es el factor de expansión de CASEN 2022

Aplicación en CASEN

- Corrige probabilidades desiguales de selección (PPT en 1^a etapa)
- Factor w_i incluye ajustes de no-respuesta y calibración
- Se usa para medias, totales y proporciones poblacionales

Ref: Diseño Muestral CASEN 2022, ecs. (43)-(44)



Varianzas y Errores Estándar

Varianzas: EVCU/WR oficial (sin FPC, $f_h = 0$)

$$V(\hat{Y}) = \sum_h \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2$$

donde $z_{hi} = w_{hi}(y_{hi} - \hat{Y})$ es el residuo ponderado en estrato h

Implementación

VARIABLES DE DISEÑO: **varstrat** (estratos) y **varunit** (UPM). Linealización de Taylor para totales/razones.

Nota: Aproximación WR

Aunque la 1^a etapa es **sin reemplazo** (PPT sistemática), se usa fórmula **WR sin FPC** como aproximación conservadora estándar en encuestas complejas (fracciones de muestreo pequeñas \Rightarrow efecto despreciable).

Ref: Diseño Muestral CASEN 2022, ecs. (45)-(47)



Software Principal (R):

- `survey / srvyr`: Diseño complejo, svyglm()
- `mice / mitools`: Imputación múltiple
- `sandwich`: Errores robustos

Soporte (Python):

- `pandas, numpy`: Procesamiento
- Solo para QA descriptivo

Ponderadores:

- Variable: `expr` (CASEN 2022)
- Corrige probabilidades desiguales de selección
- Aplicado en todos los estimadores

Inferencia

Toda inferencia por diseño se realiza en R con `survey`



Regresiones con Diseño Complejo

Método: `survey::svyglm()` con Linealización de Taylor

Declarar diseño (`svydesign`) antes de estimar. Respeta estratos, UPM y pesos.

Modelos GLM:

- Lineal: `svyglm(y ~ x)`
- Logística: `family=quasibinomial()`
- Poisson: `family=quasipoisson()`

Errores Estándar:

- **Estándar:** EE por diseño (Taylor)
- **Sensibilidad:** HC/cluster documentada

Inferencia:

- $\alpha = 0,05$ | IC 95 % por diseño



Manejo de Datos Faltantes

Política Oficial CASEN

CASEN no imputa excepto variables de ingreso (metodología Mideplan oficial).

Análisis Principal:

- `subset()` sobre diseño (preserva pesos)
- Reportar % missingness por variable
- Justificar exclusión si >5 %

Sensibilidad (si aplica):

- IM solo si % missing sustantivo
- Justificar supuesto MAR explícitamente

Flujo IM + Diseño:

1. `mice::mice(data, m=20)`
2. `svydesign` por imputación
3. Ajustar modelo en cada una
4. `MIcombine()` (Reglas de Rubin)

Nota

Documentar mecanismo de missing y comparar con casos completos



¿Preguntas?

Gracias por su atención

Alexander Pinto | Esteban Román
Julián Vargas | Francisca Sepúlveda

Pontificia Universidad Católica de Chile