





Muestreo Probabilístico

Giovany Babativa, PhD

Sobre Mi

PhD en Estadística, MSc en Big Data, MSc en Estadística. Con 15 años de experiencia, actual director de analítica en el CNC, miembro del comité de expertos en pobreza en el DANE y consultor de la División de Estadística de la CEPAL. Ex-decano de la Facultad de Estadística USTA, ex-director de operaciones en el ICFES,...

Puedes encontrarme en:

-  [Google scholar](#)
-  [GitHub. https://github.com/jgbabativam](https://github.com/jgbabativam)
-  [linkedin](#)
-  j.babativamarquez@uniandes.edu.co

MUESTREO PROBABILÍSTICO

Defina a U un universo¹ de elementos $\{U_1, \dots, U_N\}$ **finito y conocido** de antemano con una variable de interés Y que toma valores $\{y_1, \dots, y_N\}$. Sea el parámetro θ (medida del universo) una función de (y_1, \dots, y_N) de esta manera a $\theta(y_1, \dots, y_N)$ se denomina parámetro y se denota θ .

¹ En adelante se denominará universo a la población objetivo

Se define probabilidad de inclusión de primer orden del elemento k

$$\pi_k = \sum_{k \in s_i} p(s_i)$$

Sea:

$$I_k = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{en otro caso} \end{cases}$$

Entonces $\pi_k = P(I_k = 1)$

Parámetros de interés

Para un universo U de tamaño N , sea y la característica de interés entonces podríamos estar interesados en:

- **Total:** $t_y = \sum_U y_k$ (personas con cierta enfermedad)
- **Media:** $y_U = \frac{\sum_U y_k}{N} = \frac{t_y}{N}$ (dinero)
- **Proporción:** $p_U = \frac{\sum_U y_k}{N} = \frac{t_y}{N}$ para $y_k = \{1, 0\}$ (desplazados)
- **Razón:** $R = \frac{t_y}{t_z}$. Unidades del producto por establecimiento con la intención de venderlo.

Nótese que todos los parámetros pueden ser expresados como función de totales, por tanto hay un particular interés encontrar estimadores para este parámetro.

Estimador de Horvitz-Thompson (1952)

Para un universo U se desea estimar el total de una característica de interés y denotado como t_y . Por ejemplo,

Para $\theta = t_y = \sum_U y_k$ se define:

[Math Processing Error]

$\frac{1}{\pi_k}$ se denomina

Cada elemento se representa a sí mismo y a una fracción de la población.

Estimador de Horvitz-Thompson (1952)

Muestreo Aleatorio Simple

- *[Math Processing Error]*
- *[Math Processing Error]*
- *[Math Processing Error]*

Estimador de Horvitz-Thompson (1952)

Muestreo Bernoulli

- *[Math Processing Error]*
- *[Math Processing Error]*
- *[Math Processing Error]*

Estimación de la media poblacional

Muestreo Aleatorio Simple

- *[Math Processing Error]*
- *[Math Processing Error]*
- *[Math Processing Error]*

Al factor $(1 - \frac{n}{N})$ se le conoce como factor de corrección para poblaciones finitas.

Estimador de Hansen-Hurwitz (1943)

[Math Processing Error]

En donde p_{k_i} es la probabilidad de selección del elemento k . Cada elemento se representa así mismo y al resto del universo \rightarrow promedio.

Estimador de Hansen-Hurwitz (1943)

- *[Math Processing Error]*
- *[Math Processing Error]*
- *[Math Processing Error]*

ESTIMACIÓN

Con frecuencia se desean estimaciones de subpoblaciones no consideradas dentro del diseño muestral, dicha subpoblación es denominada un **dominio**, en general no sabemos cuántos elementos pertenecen a un dominio hasta obtener los resultados de la muestra. Por ejemplo, al diseñar una muestra para medir la intención de voto de los Colombianos el interés se centra en conocer el total de personas que votarán por los diferentes candidatos pero previo a los resultados no se conoce cuántos votarán por cada uno.

Estimación de un dominio: Notación

El total de un dominio es:

$$\sum_{U_d} y_k$$

con U_d la población que compone al dominio.

Sea

$$y_{dk} = \begin{cases} y_k & \text{si } k \in U_d \\ 0 & \text{en otro caso} \end{cases}$$

Estimación de un dominio: Notación

Así $t_{yd} = \sum_U y_{dk}$ será el total de la variable de interés.

Si $y_{dk} = 1$ cuando $k \in U_d$ entonces $N_d = \sum_U y_{dk}$ es el tamaño poblacional del dominio.

Luego la media del dominio es:

$$\bar{y}_{Ud} = \frac{t_{yd}}{N_d}$$

Estimación de un dominio: Ejemplo

- A nivel general el estimador de Horvitz-Thompson es:

[Math Processing Error]

- Particularmente para MAS se tiene que:

[Math Processing Error]

[Math Processing Error]

Estimación de un dominio: Aplicación

Ejemplo:

Suponga que por medio de un muestreo a hogares se desea determinar la estructura del gasto según los ingresos familiares. El cliente desea estimar un cuadro de salida así:

Total de gasto de las familias en la población según ingresos familiares.

gasto	total_2_smmlv	cve_2_smmlv	total_2_a_4_smmlv	cve_2_a_4_smmlv	total_4_smmlv	cve_4_smmlv	total_general	cve_total
Vivienda								
Alimentación								
Educación								
Transporte								
Esparcimiento								
Otros								
Total								

Estimación de un dominio: Aplicación

1. *Ejemplo Hogares - Marco Muestral.xls* seleccione una muestra $MAS(N, 10)$.
2. *Dominio Hogares.xls* estime en Excel el cuadro de salida solicitado por el cliente.
3. *Dominio Hogares.xls* asigne arbitrariamente los códigos de hogares seleccionados en el literal 1 y documente la base.
4. *Dominio Hogares.xls* suba los datos a *R*.
5. Haga un cruce de los datos con la muestra y la base con la información levantada de tal forma que los pesos muestrales queden en el archivo *Dominio Hogares.xls*.
6. Estime en **R** el cuadro de salida solicitado por el cliente.

35:00

19

Estimación de un dominio: Aplicación

Tu turno

Lohr (1999) En el censo de Agricultura de Estados Unidos que se realiza cada 5 años, se reúnen datos de todas las fincas-granjas donde se producen y venden más de \$100.000 USD en productos agrícolas al mes. Dentro de las variables que se miden está el número de granjas y los acres¹ dedicados a las granjas. El archivo **agpop.sav** contiene los datos de la población medida en el censo de $N = 3078$. El archivo **Agrsrs.sav** contiene los datos de una muestra obtenida mediante el diseño MAS(3078, 300). Estime el promedio y la cantidad de acres dedicados a la agricultura en el año 1992 en los Estados Unidos.

1 Un acre es una medida de superficie usada en agricultura equivalente a 43560 pies²

Cargue de los datos

```
1 #remotes::install_github("tidy-survey-r/srvyexploR")
2 library(pacman)
3 p_load(tidyverse, survey, srvyr, remotes, haven, srvyexploR, skimr)
4
5 censo <- "https://github.com/jgbabativam/Muestreo-I/raw/main/datos/agpop.sav"
6 encuesta <- "https://github.com/jgbabativam/Muestreo-I/raw/main/datos/Agrsrs.sav"
7
8 marco <- read_sav(censo)
9 datos <- read_sav(encuesta)
```

Explore los datos usando `skim()` y `glimpse()`

Código R: Objeto de diseño

```
1 datos$FEXP <- nrow(marco)/nrow(datos)
2
3 dsg <- datos |>
4   mutate(NI = nrow(marco)) |>
5   as_survey_design(ids = 1,
6                     fpc = NI,
7                     weights = FEXP,
8                     nest = T)
```



```
1 (estimacion <- dsg |>
2   summarise(Total = survey_total(acres92, vartype = c("se", "cv", "ci")),
3   Promedio = survey_mean(acres92, vartype = c("se", "cv", "ci"))))

# A tibble: 1 × 10
  Total Total_se Total_cv Total_low Total_upp Promedio Promedio_se Promedio_cv
  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>       <dbl>   <dbl>
1 9.17e8 5.82e7 0.0634 8.02e8 1.03e9 297897.    18898.    0.0634
# i 2 more variables: Promedio_low <dbl>, Promedio_upp <dbl>

1 ty <- sum(marco$acres92)
2 ybU <- mean(marco$acres92)
```

¿Cómo estimaría el la proporción de condados con menos de 200.000 acres de granjas?. ¿A partir de los universos del marco, cree que esta fue una buena muestra?

Estimación de un dominio: Promedio

MAS

- $y_{U_d} = \frac{t_{yd}}{N_d} = \frac{N/n \sum_U y_{dk}}{N_d}$
- *[Math Processing Error]*, con $z_{dk} = 1$ si $k \in U_d$

Puesto que la cantidad del numerador y del denominador son variables aleatorias el estimador se denomina de razón y la estimación de la varianza se obtiene por aproximación mediante el método de linealización de Taylor.

Estimación de un dominio: Promedio

Tu turno

Para cada región estime el promedio y el total de acres por condado en 1992.

05:00

25

TAMAÑO DE LA MUESTRA

Tamaño de Muestra y asignación por estrato

En general no existe una fórmula mágica que indique el tamaño de muestra apropiado sino que éste debe ser establecido de acuerdo a la estrategia muestral seleccionada.

Se debe controlar la **precisión** y la **confiabilidad**, tenga en cuenta que bajo un diseño muestral $P(\cdot)$ un intervalo de confianza de $100(1 - \alpha)\%$ para la media poblacional está dado por:

$$\left(y_s - z_{1-\alpha/2} \sqrt{V_P(y_s)}, y_s + z_{1-\alpha/2} \sqrt{V_P(y_s)} \right)$$

margen de error margen de error

Tamaño de Muestra

Recuerde que por T.C.L se tiene que $\sum_i X_i$ sigue una distribución normal y por tanto:

$$P\left(-z_{1-\alpha/2} < \frac{y_s - y_U}{\sqrt{V_P(y_s)}} < z_{1-\alpha/2}\right) = 1 - \alpha$$
$$P(|y_s - y_U| < z_{1-\alpha/2} \sqrt{V_P(y_s)}) = 1 - \alpha$$
$$P(|y_s - y_U| < e) = 1 - \alpha$$

La cantidad a minimizar es e . En encuestas donde el parámetro de interés es una proporción se sugiere $e = 0.03$ y $1 - \alpha = 0.95$.

Ejemplo: M.A.S.

En MAS:

[Math Processing Error]

Un intervalo de confianza del $100(1 - \alpha)\%$ para y_U bajo diseño MAS es:

$$\bar{y}_s \pm z_{1-\alpha/2} \sqrt{\frac{1}{n} \left(1 - \frac{n}{N}\right) S_{y_s}^2}$$

Ejemplo: M.A.S.

Se desea:

$$z_{1-\alpha/2} \sqrt{\frac{1}{n} \left(1 - \frac{n}{N}\right) S_{y_s}^2} < e$$

Al despejar n se obtiene:

$$\begin{aligned} n &> \frac{z_{1-\alpha/2}^2 S_{y_s}^2}{e^2 + \frac{z_{1-\alpha/2}^2 S_{y_s}^2}{N}} \\ &= \frac{n_0}{1 + \frac{n_0}{N}} \end{aligned}$$

$$\text{con } n_0 = \frac{z_{1-\alpha/2}^2 S_{y_s}^2}{e^2}.$$

Ejemplo M.A.S. caso Proporción

Suponga que en un municipio de $N = 1251$ habitantes se desea estimar la proporción de personas que piensan votar por un determinado candidato, para ello se aplicará un muestreo aleatorio simple y se desea obtener un margen de error máximo del 3% con una confiabilidad de 95%. Calcule el tamaño de muestra que debe utilizarse en el proyecto.

$$\begin{aligned} s_y^2 &= \frac{1}{N-1} \sum_U (y_k - y_U)^2 \\ &= \frac{1}{N-1} (\sum_U y_k^2 - N y_U^2) \\ &= \frac{N}{N-1} P_d (1 - P_d) \doteq P_d (1 - P_d) \end{aligned}$$

Graficar la función para determinar valor adecuado de P_d .

Tamaño de Muestra: M.A.S. - Proporción

$$n_0 = \frac{1.96^2 (0.5)(1 - 0.5)}{0.03^2} \approx 1067$$

Entonces el tamaño de muestra requerido es:

$$n = \frac{1067}{1 + \frac{1067}{1251}} = 576$$

Note que n_0 corresponde al tamaño de muestra al ignorar la corrección para poblaciones finitas.

Ejercicio

Tu turno

Suponga varios márgenes de error y grafique la función de tamaños de muestra para este ejercicio. Luego suponga un universo más grande y haga el mismo ejercicio ¿En términos proporcionales que puede decir del tamaño de la muestra frente al universo en ambos ejercicios?

Tamaño de Muestra: M.A.S. - Medias, Totales

En el caso de estimación de la media o de totales es necesario considerar una precisión relativa, siguiendo los mismo pasos se llega a:

$$P\left(\left|\frac{y_s - y_U}{y_U}\right| < e\right) = 1 - \alpha$$

$$P(|y_s - y_U| < e|y_U|) = 1 - \alpha$$

Así que en la fórmula del tamaño de la muestra se debe reemplazar e por $e|y_U|$.

Tamaño de Muestra: M.A.S. - Medias, Totales

$$n = \frac{z_{1-\alpha/2}^2 S_y^2}{(ey_U)^2 + \frac{z_{1-\alpha/2}^2 S_y^2}{N}}$$

$$= \frac{z_{1-\alpha/2}^2 CV^2(y)}{e^2 + \frac{z_{1-\alpha/2}^2 CV^2(y)}{N}}$$

Tamaño de Muestra: M.A.S. - Medias, Totales

Use la base y extraiga una muestra aleatoria simple de tamaño n asuma que esta fue una muestra piloto utilizada para validar las herramientas de medición pero a partir de ella también puede estimar el comportamiento de las variables de interés para hallar el $CV(y)$. Suponga que se desea un error relativo inferior al 10% determine el tamaño de la muestra que necesitaría para ejecutar el proyecto.

- Para hacer los cálculos de tamaño de muestra use una prueba piloto de la investigación. Lohr (2000) establece que este es probablemente el mejor método.
- Use información secundaria, es raro que sea la primera vez que se estudie algo relativo a su investigación. Es posible tener acceso a fuentes administrativas (Ejemplo: DANE, Ministerios, etc) que pueden tener estimaciones de la varianza de cifras relacionadas con el estudio.

- Use una distribución empírica para estimar la varianza. Por ejemplo, algunos autores sugieren que las variables económicas suelen tener distribuciones sesgadas a la derecha que pueden modelarse con distribuciones chi-cuadrado.
- En la medida que el diseño muestral se vuelve más complejo (estratificado, multietápico de conglomerados o elementos en algunas etapas) no es tan fácil llegar a fórmulas que proporcionen el tamaño de muestra adecuado, así que es necesario recurrir a métodos de Monte Carlo para hallar los tamaños de muestra basados en las fórmulas de la varianza del diseño muestral.

Tamaño de muestra en diseños complejos

$$n \geq \frac{z_{1-\alpha/2}^2 S_y^2 \text{DEFF}}{\varepsilon^2 + \frac{z_{1-\alpha/2}^2 S_y^2 \text{DEFF}}{N}}$$

$$\text{DEFF} = \frac{\hat{V}_P(\theta)}{\hat{V}_{\text{MAS}}(\theta)}$$

MUESTREO ESTRATIFICADO

Definiciones

1. $U = \{1, 2, \dots, N\}$ sea $\{U_1, U_2, \dots, U_H\}$ una partición de U .
2. $U_h, h = 1, \dots, H$ se aplica de forma **INDEPENDIENTE** el diseño muestral $P_h(\cdot)$.
3. $t_y = \sum_U y_k = \sum_{h=1}^H t_{yh}$ de tal manera que el estimador insesgado de t_y es:

[Math Processing Error]

[Math Processing Error]

1. Cada estrato U_h es de tamaño N_h , así $\sum_{h=1}^H N_h = N$.
2. Una partición de U es:
 - $N_h \neq 0, h = 1, \dots, H$.
 - $U_i \cap U_j = \emptyset, i \neq j; i, j = 1, \dots, H$
 - $\bigcup_{h=1}^H U_h = U$
3. **Estrato es diferente de dominio**, los estratos se construyen a partir del diseño de la muestra, los dominios no. Por tanto la manera de hacer estimación es diferente.

Razones para estratificar

1. Minimizar varianza. Por ejemplo, el DANE en las encuestas económicas estratifica por empresas grandes, medianas y pequeñas.
2. Necesidad de contar con información con alta precisión para algunos niveles de desagregación. Por ejemplo, en un estudio en Bogotá se desean estimaciones por localidad pero de no controlarse la muestra es posible que salgan muy pocos elementos para algunas localidades.
3. Reducción de costos. Por ejemplo, los gastos de desplazamiento pueden llegar a aumentar de manera considerable los costos de un proyecto entonces es posible estratificar por ciudades grandes, intermedias y pequeñas.

Problemas técnicos de estratificar

1. La cantidad de estratos es directamente proporcional a la var y proporcional a los costos. En general al llegar a determinado número de estratos no se gana mucho en disminuir varianza y por el contrario se invierte mucho dinero.
2. Delimitación de los estratos (Hidroglou, Dalenius, Análisis multivariante de clúster).
3. Variable de estratificación - ¿con respecto a qué variable se debe estratificar?.
4. Asignación de muestra por estratos.

GRACIAS!

Referencias

- Gutiérrez, H. A. (2016). Estrategias de muestreo: Diseño de encuestas y estimación de parámetros. Ediciones de la U.
- Lohr, S. L. (2021). Sampling: design and analysis. Chapman and Hall/CRC.
- Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. Springer Science & Business Media.
- Valliant, R., Dever, J. A., & Kreuter, F. (2013). Practical tools for designing and weighting survey samples (Vol. 1). New York: Springer.

Citación y derechos de autor

Este material ha sido creado por [Giovany Babativa-Márquez](#) y es de libre distribución bajo la licencia [Creative Commons Attribution-ShareAlike 4.0](#).

Si se copia parcial o totalmente, debe citar la fuente como:

Babativa-Márquez, J.G. *Diapositivas del curso de muestreo probabilístico*. URL: <https://jgbabativam.github.io/Muestreo-I/Semana3.html>. 2024.