Muestreo Probabilístico

Giovany Babativa, PhD

Sobre Mi

PhD en Estadística, MSc en Big Data, MSc en Estadística. C 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
- linkedin
- j.babativamarquez@uniandes.edu.co

MUESTREO PROBABILÍSTICO

Diapositivas disponibles en GitHub.

Notación

Defina a U un universo[1] de elementos $\{U_1, \ldots, U_N\}$ **finito** y **conocido** de antemano con una variable de interés Y que toma valores $\{y_1, \ldots, y_N\}$. Sea el parámetro θ (medida del universo) una función de $(y_1, \ldots, y_N)$ de esta manera a $\theta(y_1, \ldots, y_N)$ se denomina parámetro y se denota θ.

1. En adelante se denominará universo a la población objetivo

Probabilidades de Inclusión

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 inter entonces podríamos estar interesados en:

- **Total**: $t_y = \sum_U y_k$ (personas con cierta enfermedad)
- **Media**: $\bar{y}_U = \frac{\sum_U y_k}{N} = \frac{t_y}{N}$ (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.

6

Estimador de Horvitz-Thompson (1952)

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

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

$$\hat{\theta} = \hat{t}_{y,\pi} = \sum_{s_i} \frac{y_k}{\pi_k}$$

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

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

Estimador de Horvitz-Thompson (1952)

Muestreo Aleatorio Simple

- $\hat{t}_{y,\pi} = \sum_{s_i} \frac{y_k}{\pi_k} = \frac{N}{n} \sum_{s_i} y_k$
- $V_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S^2_{yU}$
- $\widehat{V}_{MAS}(\hat{t}_{y,\pi}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S^2_{ys_i}$

Estimador de Horvitz-Thompson (1952)

Muestreo Bernoulli

- $\hat{t}_{y,\pi} = \sum_{s_i} \frac{y_k}{\pi_k} = \sum_{s_i} \frac{y_k}{\pi}$
- $V_{Ber}(\hat{t}_{y,\pi}) = \left(\frac{1}{\pi} - 1\right) \sum_U y_k^2$
- $\widehat{V}_{Ber}(\hat{t}_{y,\pi}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_{s_i} y_k^2$

Estimación de la media poblacional

Muestreo Aleatorio Simple

- $\hat{\bar{y}}_U = \frac{\hat{t}_{y,\pi}}{N}$
- $V_{MAS}(\hat{\bar{y}}_U) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{yU}$
- $\widehat{V}_{MAS}(\hat{\bar{y}}_U) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{ys_i}$

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

Por una cultura de la vida, su calidad y su sentido

Estimador de Hansen-Hurwitz (1943)

$$\hat{t}_{y,MCR} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}$$

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

Estimador de Hansen-Hurwitz (1943)

- $V\left(\hat{t}_{y,MCR}\right) = \frac{1}{m}\sum_U p_k\left(\frac{y_k}{p_k} - t_y\right)^2$
- $\widehat{V}\left(\hat{t}_{y,MCR}\right) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{y,MCR}\right)^2$
- $E\left(\widehat{V}\left(\hat{t}_{y,MCR}\right)\right) = V\left(\hat{t}_{y,MCR}\right)$

Estimación de un dominio: Notación

El total de un dominio es:

$$\hat{t}_{yd} = \sum_{s} \frac{y_{kd}}{\pi_k}$$

donde

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

Tamaño de la muestra

UNIVERSIDAD EL BOSQUE
Por una cultura de la vida, su calidad y su sentido

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

$$DEFF = \frac{V_P(\hat{\theta})}{V_{MAS}(\hat{\theta})}$$

MUESTREO ESTRATIFICADO

UNIVERSIDAD EL BOSQUE

Por una cultura de la vida, su calidad y su sentido

Definiciones

1. $U=\{1,2,\ldots,N\}$ sea $\{U_1,U_2,\ldots,U_H\}$ una partición de
2. $U_h, h=1,\ldots,H$ 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:

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

$$\widehat{V}\left(\hat{t}_y\right)=\widehat{V}\left(\sum_{h=1}^H \hat{t}_{yh}\right)=\sum_{h=1}^H \widehat{V}\left(\hat{t}_{yh}\right)$$

Muestreo ESTMAS

Un caso particular es cuando se estratifica y cada uno de los estr
se aplica de manera independiente un $MAS(N_h, n_h)$. Encuesta de
Anual de servicios, manufacturera y comercio (DANE).

1. $\hat{t}_{yh} = \sum_{sh} \frac{N_h}{n_h} y_k$
2. $\hat{t}_{y} = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \sum_{sh} \frac{N_h}{n_h} y_k$
3. $\widehat{V}\left(\hat{t}_{yh}\right) = \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{ysh}^2$
4. $\widehat{V}\left(\hat{t}_{y}\right) = \sum_{h=1}^{H} \widehat{V}\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{ysh}^2$

UNIVERSIDAD EL BOSQUE

Por una cultura de la vida, su calidad y su sentido

Ejercicio

Tu turno

Use las bases **Agpop.sav** y **AgStrat.sav** que corresponde a una muestra usando un diseño ESTMAS y realice los siguientes ejercicios:

1. Seleccione una muestra de **Agpop.sav** teniendo en cuenta los mismos n_h de **AgStrat.sav**
2. Estime el promedio de acres por región con sus respectivos errores de muestreo (hágalo sacando las medidas de resumen N_h, n_h, S^2).
3. Estime en R el promedio de acres por región con sus respectivos errores de muestreo.

Ejercicio

Suponga que se planea un muestreo estratificado que permita m
la intención de voto de los colombianos en las próximas elecciones presidenciales. Se ha decidido partir la población en 3 estratos. El analista Estadístico propone hacer en la primera etapa los siguientes diseños muestrales:

- Estrato 1: Diseño de inclusión forzosa (censo).
- Estrato 2: Diseño de Probabilidad Proporcional al Tamaño (PPT), x_k es la población de 18 años o más en cada ciudad.
- Estrato 3: Muestreo Aleatorio Simple.

Determine las fórmulas necesarias para calcular el total (con su cve) de personas que votarán en las próximas elecciones. ¿Cómo estimaría la proporción de personas que están a favor del candidato A?.

Ejercicio

Se realizó un muestreo de sedes educativas mediante un diseño ESTMAS estratificando por NSE. Con el objetivo de hacer auditoria de los recursos del estado, se investigó por la cantidad de dinero (en millones de pesos) que el estado había girado a los colegios, los resultados son:

NSE	N_h	n_h	$\sum_{s_h} y_k$	s^2_{yh}
1	400	98	2361.8	5575
2	30	10	256	4064
3	61	37	9901.2	347556
4	18	6	1074	22798
5	70	39	11454.3	123578
6	120	21	697.2	9795

Estime el total de recursos girados a los colegios de la población y por NSE, calcule intervalos de confianza al 95% y muestre los cve.

Ejercicio

Se desea lanzar un nuevo producto a un target específico pero se desea saber si la preferencia depende del género para tomar una decisión sobre como enfocar la publicidad. A partir de un muestreo ESTMAS se levantaron los siguientes datos:

Ciudad	N_h	n_h	**Mujeres (%)**
Medellín	9100	636	38
Ibagué	1950	451	27
Barranquilla	5500	481	18
Bogotá	10850	611	19
Cúcuta	2100	493	36
Bucaramanga	5500	575	13
Cali	9000	588	26

Estime la proporción de mujeres de la población que preferirán el producto, calcule el cve.

Ejercicio

Para el ejercicio de los acres del censo de agricultura de Estados Unidos suponga que se hace un muestreo ESTMAS así:

Estrato	Nro. Granj	Muestra
Noroeste (NE)	220	21
Norte-Centro (NC)	1054	103
Sur (S)	1382	135
Oeste (W)	422	41
Total	3078	300

Ejercicio

1. Use R para extraer una muestra del archivo **agpop.sav** bajo el diseño propuesto.
2. Usando el archivo **agstrat.sav** estime el total de acres dedicados a la agricultura dentro de cada región.
3. Compare la estrategia de dominios vs estratos usando efecto de diseño. Concluya.

Asignación de Muestra por Estrato

Con frecuencia se requiere controlar la muestra con el fin de disminuir varianza, garantizar dispersión, entre otros. Si ya se tiene calculado un tamaño de muestra global (para obtener resultados a nivel general y no por estrato) se pueden asignar los tamaños de muestra de diferentes maneras según el objetivo que se busque. Debe tener en cuenta que si su objetivo es comparar los resultados entre los diferentes estratos debe calcular un tamaño de muestra de forma independiente para cada estrato y así garantizar una precisión determinada en cada estrato.

Afijación Proporcional

Se debe usar cuando se puede suponer que la variable de interés tiene varianzas similares en todos los estratos.

$$n_j = n * \frac{N_j S^2_{U_j}}{\sum_{h=1}^{H} N_h S^2_{U_h}} = n * \frac{N_j}{\sum_{h=1}^{H} N_h} = n * \frac{N_j}{N}$$

. . .

Muy usada a menudo pero sin probar que $S^2_{y_{U_h}} = S^2_{y_h}$ lo cual evidentemente está mal.

Afijación Proporcional

Para la muestra de **agStrat** calcule un tamaño de muestra global y asigne los tamaños de muestra por estrato de manera proporcional. Discuta el procedimiento.

Afijación X-óptima

Se debe usar cuando S_h^2 varían mucho. El objetivo de la asignación óptima consiste obtener la mayor cantidad de información al menor costo. Se debe contar con x altamente correlacionado con y o con una prueba piloto. La función de costo:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h,$$

Afijación X-óptima

con C el costo total, c_0 los costos adicionales y c_h el costo en el estrato h que varía dependiendo de h.

$$n_j = n * \left(\frac{\frac{N_j S_{x_j}}{\sqrt{c_j}}}{\sum_{h=1}^{H} \frac{N_h S_{x_h}}{\sqrt{c_h}}} \right)$$

Afijación de Neyman

Es un caso particular de la asignación óptima y se usa si $c_h = c_f$, es decir, los costos son constantes para todos los estratos.

$$n_j = n * \left(\frac{N_j S_{x_j}}{\sum_{h=1}^{H} N_h S_{x_h}} \right)$$

En general esta asignación proporciona un estimador con menor varianza que la asignación proporcional.

Afijación Proporcional al Tamaño

Por una cultura de la vida, su calidad y su sentido

Es una variación de la asignación proporcional, se usa si se cumple que el coeficiente de variación $CV_h = CV_f$ para todos los estratos:

$$n_j = n * \frac{t_{X_{U_j}}}{t_X}$$

Afijación de potencia

Sea $0 \leq \alpha \leq 1$ entonces:

$$n_j = n * \frac{t^{\alpha}_{X_{U_j}}}{\sum_{h=1}^{H} t^{\alpha}_{X_h}}$$

Analice los casos $\alpha = 1$ y $\alpha = 0$.

Delimitación de los estratos

UNIVERSIDAD EL BOSQUE
Por una cultura de la vida, su calidad y su sentido

En muestreo estratificado se busca homogeneidad dentro y heterogeneidad entre estratos. Este es un problema de clasificación que se puede resolver mediante métodos multivariados (cluster analysis). Sin embargo en muestreo se han establecido algunos métodos como:

- Caso $IF - MAS$: Método de Hidroglou (1986).
- Método de Dalenius y Hodges (1957).
- LH
- Geométrico
- Enfoque multivariante

MUESTREO DE CONGLOMERADOS Y BIETAPICO

Introducción

Hemos visto muestreos donde la población está al alcance por m
de un marco de lista. Sin embargo, en la mayoría de situaciones esto no ocurre y se debe llegar a la unidad de observación a partir de unidades de muestreo que las contengan.

Suponga que desea estimar el porcentaje de estudiantes de colegios que han sido víctimas de algún tipo de agresión, claramente usted no tiene acceso a un listado de todos los estudiantes de todos los colegios (marco de lista), pero si puede tener acceso al listado de colegios y direcciones (marco de áreas) haciendo un derecho de petición a la secretaría de educación que corresponda o al Ministerio de Educación Nacional.

Muestreo de conglomerados y muestreo bietápico

Entonces usted puede seleccionar n_I colegios y hacer a todos lo estudiantes o seleccionar n_i colegios y hacer n_j estudiantes en cada colegio.

1. La primera situación se denomina: **Muestreo de Conglomerados** y consiste en que una vez seleccionada una unidad de muestreo se aplica censo.
2. La segunda situación se denomina: **Muestreo Bietápico** y consiste en que una vez seleccionada una unidad de muestreo se selecciona otra muestra al interior manteniendo algunos principios fundamentales.

En este ejemplo el colegio es la unidad de muestreo, siendo también el conglomerado en la primera situación y para ambos casos el estudiante es la unidad de observación

Muestreo de conglomerados

Suponga que la población de elementos

$$U = \{1, \ldots, k, \ldots, N\}.$$

se divide en N_I sub-grupos poblacionales, llamados y denotados como $U_I = \{U_1, \ldots, U_{N_I}\}$.

La población de conglomerados estará dada, sin pérdida de generalidad, por

$$U_I = \{1, \ldots, N_I\}.$$

Muestreo de conglomerados

Estos definen una partición de la población en tal forma que

El número de unidades N_i en el conglomerado i-ésimo se llama tal que

$$N = \sum_{i=1}^{N_I} N_i,$$

donde N es el tamaño de la población U.

Muestreo de conglomerados

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \sum_{i=1}^{N_I} t_{yi}$$

donde $t_{yi} = \sum_{k \in U_i} y_k$ es el total del i-ésimo conglomerado.

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{1}{N} \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \frac{1}{N} \sum_{i=1}^{N_I} N_i \bar{y}_i$$

donde $\bar{y}_i = \dfrac{1}{N_i} \sum_{k \in U_i} y_k$ es la media del i-ésimo conglomerado.

Muestreo aleatorio simple de conglomerados - MASC

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I}\sum_{S_I} t_{yi}$$

$$Var_{MAC}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_{yU_I}}$$

$$\widehat{Var}_{MAC}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1-\frac{n_I}{N_I}\right)S^2_{t_{ys_I}}$$

con

$$S^2_{t_{ys_I}} = \frac{1}{n_I-1}\sum_{i\in s_I}(\hat{t}_{ysi}-\bar{\hat{t}}_{ys_I})^2$$

Ejercicio

El objetivo de una encuesta es estimar el ingreso medio en un ba
de la ciudad. Asuma que en ese barrio existen $N_I = 60$ manzanas. Se realiza un diseño de muestreo aleatorio simple de conglomerados y se seleccionan $n_I = 5$ manzanas, en las cuales se entrevistan a todos los hogares.

ID Manzana	Hogares en la manzana	Ingreso total en la manzana
AW45	120	25000
AW02	100	24000
AW31	80	19000
AW28	95	20100
AW44	80	18000

Ejercicio

- Estime el ingreso total de los hogares en el barrio. Reporte el coeficiente de variación estimado.
- Estime el número de hogares en el barrio. Reporte el coeficiente de variación estimado.
- Asumiendo que en el barrio hay $N = 2000$ hogares, estime el ingreso medio de los hogares en el barrio. Reporte el coeficiente de variación estimado.

Muestreo en varias etapas

No se tiene acceso a un marco de muestreo de elementos, entor
se debe llegar a las unidades de observación a partir de la selección de áreas que los contengan. El proceso jerárquico se realiza l veces con los siguientes pasos:

1. Construir l marcos de muestreo de unidades (conglomerados en las primeras $l-1$ etapas del diseño muestral y de elementos en la última etapa).
2. Aplicar un diseño muestral y selección de la muestras (o sub-muestras) de cada marco de muestreo.

Muestreo en varias etapas

Esquema de selección por unidades de muestreo.

$$\underbrace{\textbf{Ciudad}}_{\text{Población}U} \Rrightarrow \underbrace{\textbf{Escuelas}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Niveles}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Alumnos}}_{\text{UTM}}$$

- **UPM**: Unidad Primaria de Muestreo, es la primera subdivisión en conglomerados de la población original.
- **USM**: Unidad Secundaria de Muestreo es la sub-subdivisión de la población, es decir la subdivisión de las **UPM**.
- **UTM**: Unidad Terciaria de Muestreo corresponde a los elementos de la población objetivo, que en este caso particular son los alumnos de la ciudad.

Muestreo en varias etapas

$$\underbrace{\textbf{Ciudad}}_{\text{Población}U} \Rrightarrow \underbrace{\textbf{Sección}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Manzana}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{UTM}} \Rrightarrow \underbrace{\textbf{P}}_{\text{UCM}}$$

El principio básico de una estrategia de muestreo en varias etapas es construir estimaciones desde abajo hasta arriba.

- **Invariancia:** sugiere que la probabilidad de selección de una muestra de unidades de muestreo (conglomerados o elementos) no depende del diseño de muestreo de la anterior etapa.
- **Independencia:** interpretado como que el sub-muestreo de cualquier unidad de muestreo se lleva a cabo de manera independiente con las otras unidades de muestreo, en la misma etapa o en etapas superiores o inferiores.

UNIVERSIDAD EL BOSQUE

Por una cultura de la vida, su calidad y su sentido

Muestreo con dos etapas

Los parámetros poblacionales de interés pueden escribirse com

Total

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \sum_{i=1}^{N_I} t_{yi}$$

donde $t_{yi} = \sum_{k \in U_i} y_k$ es el total de la i-ésima unidad primaria de muestreo $i = 1, \ldots, N_I$.

Muestreo con dos etapas

Los parámetros poblacionales de interés pueden escribirse com

Media

$$\bar{y}_U = \frac{\sum_{k\in U} y_k}{N} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k\in U_i} y_k = \frac{1}{N}\sum_{i=1}^{N_I} N_i \bar{y}_i$$

donde $\bar{y}_i = \frac{1}{N_i}\sum_{k\in U_i} y_k$ es la media de la i-ésima unidad primaria de muestreo $i = 1, \ldots, N_I$.

Diseño de muestreo MAS-MAS

Este diseño de muestreo supone que la población está divida en unidades primarias de muestreo, de las cuales se selecciona una muestra s_I de n_I unidades mediante un diseño de muestreo aleatorio simple. El sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple. Es decir, para cada unidad primaria de muestreo seleccionada $i \in s_{Ih}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i.

Diseño de muestreo MAS-MAS

El algoritmo de selección es

- Separar la población en N_I unidades primarias de muestreo mediante el marco de muestreo de conglomerados. – Realizar una selección de n_I conglomerados mediante cualquiera de los métodos expuestos como el método coordinado negativo o por el método de Fan-Muller-Rezucha.
- Para cada unidad primaria seleccionada en la muestra de la primera etapa s_I, realizar una selección de $n_i \; i \in S_I$ elementos mediante cualquiera de los métodos expuestos para un MAS.

Estimación de un total en un diseño MAS-MAS

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i}{n_i} \sum_{k \in S_i} y_k$$

con estimación insesgada de la varianza dada por

$$\widehat{Var}_{MM}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I}\left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{t}_y S_I} + \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$$

donde $S^2_{t_y s_I}$ es la varianza de los totales $\hat{t}_{yi}$ $i \in s_I$ de todas y cada una de las unidades primarias de muestreo y $S^2_{y_{s_i}}$ es la varianza entre los elementos dentro de cada unidad primaria de muestreo.

Universidad El Bosque
Por una cultura de la vida, su calidad y su sentido

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