UNIVERSIDAD EL BOSQUE

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

Muestreo Probabilístico

Giovany Babativa, PhD

Diapositivas disponibles en GitHub.

Sobre Mi

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

PhD en Estadística, MSc en Big Data, MSc en Estadística. C[illegible] 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

UNIVERSIDAD EL BOSQUE

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

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

Conceptos básicos

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

Algunos parámetros de interés en un estudio por muestreo:

- **Total:** $t_y = \sum_U y_k$
- **Promedio:** $\bar{y}_U = \frac{1}{N} \sum_U y_k$
- **Razón:** $R = \frac{\sum_U y_k}{\sum_U z_k} = \frac{t_y}{t_z}$

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 $\pi_k = P(I_k = 1)$

Probabilidades de Inclusión

Se define probabilidad de inclusión de segundo orden de los elementos k y l

$$\pi_{kl} = \sum_{k,l \in s_i} p(s_i)$$

Entonces $\pi_{k,l} = P(I_k I_l = 1)$

Probabilidades de Inclusión

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

- $\pi_{kk} = \pi_k$.
- Por definición de muestra probabilística $\pi_k > 0$.
- En muestreo de elementos π_k para todo $k = 1, \ldots, N$ son conocidos de antemano.

Ejercicio

Tu turno

- Use el espacio muestral del diseño MAS para calcular la probabilidad de inclusión del elemento 1, verifique que $\pi_k = \frac{n}{N}$.
- Use el espacio muestral del diseño BER para calcular la probabilidad de inclusión del elemento 1, verifique que $\pi_k = \pi$.

10:00

Estadística y Estimador

Sea $\hat{\theta}$ una estadística o estimador entonces bajo el diseño mues
$p(\cdot)$ se define:

- Valor esperado: $E_P(\hat{\theta}) = \sum_S p(s_i)\hat{\theta}$
- Sesgo: $B(\hat{\theta}) = E_P(\hat{\theta}) - \theta$
- Varianza: $V_P(\hat{\theta}) = \sum_S p(s_i)\left[\hat{\theta} - E(\hat{\theta})\right]^2$
- Error Cuadrático Medio: $ECM(\hat{\theta}) = V(\hat{\theta}) + B^2(\hat{\theta})$

Estadística y Estimador

Sea $\hat{\theta}$ una estadística o estimador entonces bajo el diseño mues
$p(\cdot)$ se define:

- Error estándar: $\sqrt{V_P(\hat{\theta})}$
- Coeficiente de variación (error relativo): $\frac{\sqrt{V_P(\hat{\theta})}}{E_P(\hat{\theta})}$
- Coeficiente de variación estimado: $cve(\%) = 100 * \frac{\sqrt{\widehat{V}_P(\hat{\theta})}}{\hat{\theta}}$

Ejercicio

Tu turno

Calcule $E_P(\hat{\theta}) = \sum_S p(s_i)\hat{\theta}$ para:

$$\hat{\theta} = \frac{1}{n}\sum_s y_k$$

Use el espacio muestral de un MAS

10:00

Estadística I_k

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

Ejemplo:

$$\begin{aligned} E(I_k) &= \sum_S p(s_i) I_k(s_i) \\ &= \sum_{k \in s_i} p(s_i) \\ &= \pi_k \end{aligned}$$

Estadística I_k

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

Ejemplo:

$$
\begin{aligned}
V(I_k) &= \sum_S p(s_i)[I_k(s_i) - E(I_k(s_i))]^2 \\
&= \sum_S p(s_i) I_k^2(s_i) - 2\sum_S p(s_i) I_k(s_i)\pi_k + \sum_S p(s_i)\pi_k^2 \\
&= \pi_k - 2\pi_k^2 + \pi_k^2 \\
&= \pi_k(1-\pi_k)
\end{aligned}
$$

- ejemplo MAS
- ejemplo BER

Estadística I_k

Ejemplo:

$$
\begin{aligned}
Cov(I_k, I_l) &= \sum_S p(s_i)\left(I_k(s_i) - \pi_k\right)\left(I_l(s_i) - \pi_l\right) \\
&= \sum_S p(s_i) I_k(s_i) I_l(s_i) - \pi_l \sum_S p(s_i) I_k(s_i) \\
&\quad - \pi_k \sum_S p(s_i) I_l(s_i) + \pi_k \pi_l \sum_S p(s_i) \\
&= \sum_{k,l \in s_i} p(s_i) - \pi_l \sum_{k \in s_i} p(s_i) - \pi_k \sum_{l \in s_i} p(s_i) + \pi_k \pi_l \\
&= \pi_{kl} - \pi_k \pi_l - \pi_k \pi_l + \pi_k \pi_l \\
&= \pi_{kl} - \pi_k \pi_l \\
&= \Delta_{kl}
\end{aligned}
$$

Diapositivas disponibles en GitHub.

Estadística n_s

Ejemplo:

$$
\begin{aligned}
E(n_s) &= \sum_S p(s_i) n_s \\
&= \sum_S p(s_i) \sum_U I_k \\
&= \sum_U \sum_S p(s_i) I_k \\
&= \sum_U \sum_{k \in s_i} p(s_i) \\
&= \sum_U \pi_k
\end{aligned}
$$

- ejemplo MAS y BER

Diapositivas disponibles en GitHub.

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.

Diapositivas disponibles en GitHub.

Estimador de Horvitz-Thompson (1952)

Para un universo U se desea estimar el total de una característic 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 **Factor de expansión**

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

Estimador de Horvitz-Thompson (1952) - Propiedades

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

Resultado

- $E(\hat{t}_{y,\pi}) = t_y$ demost.
- $V(\hat{t}_{y,\pi}) = \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$
- $\widehat{V}(\hat{t}_{y,\pi}) = \sum\sum_{s_i} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$

$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, además $E(\widehat{V}(\hat{t}_{y,\pi})) = V(\hat{t}_{y,\pi})$

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)

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

Muestreo Bernoulli

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

Ejercicio: Para los ejercicios hechos en clase de MAS y Bernoulli calcule $\hat{t}_{y,\pi}$ y $V_{Ber}(\hat{t}_{y,\pi})$, este último vía definición y por el estimador dado por la expresión. Para ambas expresiones compruebe el insesgamiento del estimador.

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.

UNIVERSIDAD EL BOSQUE

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

DISEÑOS CON INFORMACIÓN AUXILIAR

Diapositivas disponibles en GitHub.

Muestreo Proporcional al Tamaño $PPT(p_k, m)$

Se realizan m eventos aleatorios con probabilidad contante $1/i$ decir se selecciona uno de los N elementos, se realiza su medicion y_{k_1}, se repone en el universo y se selecciona de nuevo uno de los N elementos con medición y_{k_2} y así sucesivamente hasta completar m elementos. Sea:

- Z: *Número de veces que el elemento k aparece en la muestra*
- $Z = 0, 1, \ldots, m$
- $Z \sim Bin(m, 1/N)$
- $P(Z = r) = \binom{m}{r}(1/N)^r(1 - 1/N)^{m-r}$
- $P(Z = 0) = (1 - 1/N)^m$
- $\pi_k = P(Z \geq 1) = 1 - P(Z = 0) = 1 - (1 - \frac{1}{N})^m$

Diapositivas disponibles en GitHub.

Muestreo Proporcional al Tamaño $PPT(p_k, m)$

Una generalización conduce al diseño de Probabilidad Proporcional al Tamaño (PPT) cuando cada elemento se selecciona con probabilidad p_k, tal que $\sum_U p_k = 1$. Para este diseño se debe contar con información auxiliar disponible para todos los elementos de U, este diseño tienes varios beneficios entre los que se encuentra que reduce costos y que su uso es relativamente simple.

Algoritmo de Selección: Acumulativo Total

Ejemplo

Suponga que $U = 9$, para un estudio de mercados se desea seleccionar una muestra de 4 tiendas, pero basado en la información auxiliar se desea darle más peso a las tiendas que tienen una mayor rotación del producto. La información auxiliar para las 9 tiendas es:

Tienda	1	2	3	4	5	6	7	8	9
Ventas_Unid	32	96	65	140	22	78	65	47	106

π-Estimador

La muestra ordenada es de tamaño m mientras que la muestra no ordenada (sin repetición) es de tamaño n.

$$\hat{t}_{y,\pi} = \sum_{s_i} \frac{y_k}{\pi_k} = \sum_{s_i} \frac{y_k}{1-(1-p_k)^m}$$

El estimador cuenta con n sumandos, es decir, que los elementos repetidos solo cuentan una vez dentro del estimador.

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 $\rightarrow$ promedio.

Estimador de Hansen-Hurwitz (1943)

$$
\begin{aligned}
E\left(\hat{t}_{y,MCR}\right) &= E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}\right) \\
&= \frac{1}{m}E\left(\sum_{U}\frac{y_k}{p_k}Z_k\right) \\
&= \frac{1}{m}\sum_{U}\frac{y_k}{p_k}E\left(Z_k\right) \\
&= \frac{1}{m}\sum_{U}\frac{y_k}{p_k}mp_k \\
&= \sum_{U} y_k = t_y
\end{aligned}
$$

Estimador de Hansen-Hurwitz (1943)

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

- $V\left(\hat{t}_{y,MCR}\right) = \frac{1}{m}\sum_U p_k\left(\frac{y_k}{p_k} - t_y\right)^2$

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

Ejemplo:

Tu turno

Para el estudio de mercados suponga que $s_o = \{4; 7; 7; 9\}$; $p_k = \{0.215; 0.1; 0.1; 0.163\}$ y que los valores de las ventas semanales recolectados los establecimientos seleccionados fueron $y_k = \{12; 10; 10; 11\}$

- Usando el $\pi-$estimador, estime el total de ventas semanales (demanda) del producto en las nueve tiendas.
- Usando el $MCR-$estimador, estime el total de ventas semanales (demanda) del producto en las nueve tiendas y calcule el cve.

10:00

Muestreo $PPT(p_k, m)$: Notas

- Es fácil comprobar que $V\left(\hat{t}_{y,MCR}\right)$ si $y_k = cp_k, c$ constante. Es decir, si y_k es exactamente proporcional a p_k.

En la práctica eso es imposible !!!... primero es la probabilidad y luego la medición.

Muestreo $PPT(p_k, m)$: Notas

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

- En la práctica, sea x_k una variable auxiliar altamente correlacionada con y_k así $y_k/x_k \doteq c$ entonces se puede determinar

$$p_k = \frac{x_k}{\sum_U x_k} = \frac{x_k}{t_x}, k = 1, \ldots, N$$

Entre más esté correlacionado x_k con y_k entonces
$V\left(\hat{t}_{y,MCR}\right) \to 0$

Muestreo $PPT(p_k, m)$: Notas

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

- Si los p_k no varían casi, me voy con MAS pero si varían mucho me voy con PPT.

Ejercicio

Tu turno

Suponga que se tiene $U = \{1, 2, 3, 4, 5\}$ y que se conocen todos los valores de la variable de interés, que están dados por:

Y	79.0	76.00	54.0	39.00	12.0
pk	0.1	0.15	0.2	0.25	0.3

Para un muestreo con reemplazamiento con $m = 3$, si los valores aleatorios son $\zeta_k = (0.05, 058, 0.36)$, determine la muestra, calcule la estimación y el cve usando el MCR-estimador. ¿Qué tendría que hacer para determinar si es mejor la estrategia usando π-estimador o MCR-estimador?

12:00

Ejercicio

Tu turno

Resuelva los ejercicios: 3.1.1, 3.3, 3.4, 3.12, 3.14, 3.15 y 3.19 del libro de Gutiérrez, H.A. (2016)

UNIVERSIDAD EL BOSQUE

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

CONFIABILIDAD

Confiabilidad

Sea S el conjunto de todas las muestras posibles $s_1, \ldots, s_Q$; para el diseño $P(s_1), \ldots, P(s_Q)$ $\left[\sum_{s_i \in S} p(s_i) = 1\right]$; el parámetro θ $[\theta : U \to \mathbb{R}]$ y el estimador $\hat{\theta}(s_i)$.

Confiabilidad

Para cada muestra se construye el intervalo $(LI(s_i), LS(s_i))$ de manera que cada muestra tiene su propio intervalo y la longitud del intervalo puede variar dependiendo de la muestra.

Se define **confiabilidad** como la suma de las probabilidades de las muestras de las muestras cuyo intervalo de confianza cubre al parámetro:

$$\sum_{\theta \in (LI(s_i), LS(s_i))} P(s_i)$$

Cuando se habla de confiabilidad se habla de la confiabilidad del intervalo NO de la muestra.

Propuesta de intervalos

- $\hat{\theta} \pm z_{1-\alpha/2}\sqrt{V(\hat{\theta})}$, por TLC se tiene que:

$$P\left(\theta \in \hat{\theta} \pm z_{1-\alpha/2}\sqrt{V(\hat{\theta})}\right) = 1-\alpha, \text{ si } E(\hat{\theta}) = \theta.$$

Este intervalo presenta dos inconvenientes:

- Condición de estimador insesgado
- Se necesita la varianza del estimador y esa sólo se conoce en la teoría porque en la práctica solo se conoce cuando es censo y en ese caso no tiene mucho interés.

Propuesta de Intervalos

- $\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{V}(\hat{\theta})}$

Este intervalo presenta la cualidad de que en la práctica puede calcularse, lo que interesa es entonces es ¿Qué confiabilidad tiene?.

Puesto que los estimadores vistos son insesgados tanto para el parámetro de interés como para la varianza se espera que para $z_{1-\alpha/2} = 1.96$ en un 95% de las muestras el intervalo contenga el parámetro. Pero lo anterior no me garantiza que el parámetro esté en el intervalo de la muestra seleccionada.

Ejercicio

Tu turno

Calcule la confiabilidad para los dos tipos de intervalo usando los datos del gasto recogidos en clase. Hágalo para el caso MAS, de tarea en casa resuelva el diseño BER.

20:00

Sesgo Relativo

La confiabilidad es una medida que se ve directamente afectad
el sesgo[1] NO por el margen de error. Es decir, si $E(\hat{\theta}) \neq \theta$ la cobertura del intervalo decrece y por más de que $z_{1-\alpha/2} = 1.96$ la cobertura puede ser muy baja.

$$B_r(\hat{\theta}) = \frac{B(\hat{\theta})}{\sqrt{V(\hat{\theta})}} = \frac{E(\hat{\theta}) - \theta}{\sqrt{V(\hat{\theta})}}$$

1. medibles y no medibles

Sesgo Relativo

Tu turno

Para el ejercicio de clase INCLUYA sesgos de $0; 0.05; 0.1; 0.3; 0.5; 1; 1.5; 2$ y usando un nivel de confianza del 95% calcule y grafique la confiabilidad real. Use:

$$E(\hat{\theta}) - \theta = B_r(\hat{\theta})\sqrt{V(\hat{\theta})}$$

10:00

UNIVERSIDAD EL BOSQUE

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

GRACIAS!

Diapositivas disponibles en GitHub.

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

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/Semana2.html. 2024.