Análisis de Encuestas Complejas (usando R)

Lectura 1 - Muestreo Complejo

Universidad de Santiago de Chile

Miguel Alvarado

October 7, 2020

Outline

Conceptos fundamentales

- El *muestreo* es un procedimiento que busca dar información precisa sobre la población y las subpoblaciones que la conforman; esto a partir de un subconjunto de elementos seleccionados de la población objeto de estudio.
- Una *población finita* U es un conjunto finito de N elementos: $U = \{1, \ldots, k, \ldots, N\}$.
- La *muestra aleatoria* S es un subconjunto de elementos de la población U. En tanto, $s = \{1, \ldots, k, \ldots, n\}$; es una realización de S, donde $(s \subseteq U)$ y n es el tamaño de la muestra.
- Al conjunto de todas las posibles muestras se denomina *soporte* Q: $Q = \{s_1, \ldots, s_q, \ldots, s_Q\}$

Conceptos fundamentales

- El *muestreo* es un procedimiento que busca dar información precisa sobre la población y las subpoblaciones que la conforman; esto a partir de un subconjunto de elementos seleccionados de la población objeto de estudio.
- Una *población finita* U es un conjunto finito de N elementos: $U = \{1, \ldots, k, \ldots, N\}$.
- La *muestra aleatoria* S es un subconjunto de elementos de la población U. En tanto, $s = \{1, \ldots, k, \ldots, n\}$; es una realización de S, donde $(s \subseteq U)$ y n es el tamaño de la muestra.
- Al conjunto de todas las posibles muestras se denomina *soporte* Q: $Q = \{s_1, \ldots, s_q, \ldots, s_Q\}$

En la inferencia clásica los valores observados son realizaciones de una variable aleatoria; sin embargo, en el proceso de estimación e inferencia en poblaciones finitas, el muestreo asume que los valores observados corresponden a valores fijos poblacionales.

Conceptos fundamentales

- Una muestra aleatoria es del tipo *probabilística* si: i) es posible construir (al menos definir teóricamente) un soporte Q y, ii) las probabilidades de selección de cada posible muestra $s \in Q$ son conocidas previo a la selección de la muestra.
- Un *marco muestral* es un dispositivo que permite identificar, seleccionar y ubicar a todos y cada uno de los objetos pertenecientes a la población objeto de estudio y que participarán en el proceso de selección muestral. En investigaciones por muestreo se consideran dos tipos de objetos: *Elementos* o *Conglomerados*.
- Un *diseño de muestreo* $p(\cdot)$ es una distribución de probabilidades definida sobre un soporte Q. Entonces, $\forall s \in Q$

$$
\begin{array}{rccc}
p(\cdot): & Q & \longrightarrow & (0\ ,\ 1] \\
 & s & \longrightarrow & Pr(S = s) = p(s)
\end{array}
$$

tal que:

1. $p(s) > 0, \quad \forall s \in Q$
2. $\sum_{s \in Q} p(s) = 1$

UdeSantiago de Chile

Conceptos fundamentales

Un diseño de muestreo es una distribución de probabilidades, pero no un procedimiento que selecciona la muestra *per se* ; sin embargo, permite conocer la probabilidad de inclusión del elemento k en la muestra S.

- Un *algoritmo de selección* es un procedimiento usado para seleccionar una muestra probabilística.
- Bajo un diseño de muestreo $p(\cdot)$, la *probabilidad de inclusión* π_k del k-ésimo elemento de la población en la muestra S, esta dada por:

$$\pi_k = Pr(k \in S) = \sum_{s \ni k} p(s)$$

  A π_k se suele denominar *probabilidad de inclusión de primer orden*.

- Bajo un diseño de muestreo $p(\cdot)$, la *probabilidad de inclusión* π_{kl} de los elementos $k \neq l$ de la población en la muestra S, esta dada por:

$$\pi_{kl} = Pr(k \in S \wedge l \in S) = \sum_{s \ni k \wedge l} p(s)$$

  A π_{kl} se denomina *probabilidad de inclusión de segundo orden*

Conceptos fundamentales

Las encuestas tienen como propósito entregar información acerca de una *característica de interés* y, la cual se encuentra asociada a cada elemento de la población y es un valor no aleatorio: y_k corresponde a la característica del k-ésimo elemento de la población.

- Un *parámetro* T, corresponde a una función de interés que toma por argumentos las características de interés de la población, esto es:

$$T = f(y_1, \ldots, y_k, \ldots, y_N)$$

Uno de los parámetros de interés más importantes corresponde al *total poblacional* de y, denotado por t_y y definido como:

$$t_y = \sum_{k \in U} y_k$$

Conceptos fundamentales

- Una *estadística* es una función G de la muestra aleatoria S y solo depende de los elementos pertenecientes a S. Cuando una estadística se usa para estimar un parámetro T se dice *estimador* y la realización del estimador se dice *estimación*.
- Las dos propiedades más comúnmente utilizadas de un estimador $\hat{T}$ de un parámetro de interés T son:
  1. El *sesgo*, denotado por $B(\hat{T})$, definido por:
  $$B(\hat{T}) \quad = \quad E(\hat{T}) - T$$
  2. El *error cuadrático medio*, denotado por $ECM(\hat{T})$, definido por:
  $$ECM(\hat{T}) \quad = \quad E\left[\hat{T} - T\right]^2 = Var(\hat{T}) + B(\hat{T})^2$$

Si el *sesgo* de un estimador es nulo se dice que el estimado es *insesgado* y cuando esto ocurre el *error cuadrático medio* se convierte en la *varianza* del estimador

Conceptos fundamentales

Uno de los estimadores más utilizados en muestreo probabilístico corresponde al estimador de Horvitz–Thompson.

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k$$

  Donde π_k corresponde a la *probabilidad de inclusión* del k-ésimo elemento en la muestra y a d_k se suele denominar como *ponderador básico* o *ponderador de diseño* y corresponde al inverso de la probabilidad de inclusión π_k.

- Resultado: Si todas las probabilidades de inclusión de primer orden son mayores que cero ($\pi_k > 0, \forall k$), el Estimador de Horvitz–Thompson es insesgado para el total poblacional de la característica de interés y. Por tanto:

$$E\left(\hat{t}_{y,\pi}\right) = t_y$$

Diseño Muestral Complejo: Características

Un diseño muestral se dice *complejo* si este considera alguna de las siguientes características.

- Estratificación.
- Conglomeración.
- Probabilidades de selección desiguales.
- Ajustes por no-respuesta.
- Ajustes con información exógena.