Resumen Tema 1: Introducción a la teoría de muestras. Elementos del problema de muestreo.

Conceptos:

- Población: Conjunto de unidades del que se desea obtener cierta información.
- Unidades: Personas, familias, viviendas, escuelas, fbricas, ...
- Información: Consumo medio por familia, número de personas en paro, ...
- Parmetros o características poblacionales: Valores resultados de medir o contar en cada unidad una o varias características (media, total, proporción, ...)
- Población objetivo: Población que se intenta investigar.
- Población investigada: Población con la que trabaja el investigador con duplicaciones, unidades extrañas y unidades para las que no se puede obtener información (no colaboran o ausentes).
- Marco: Conjunto de unidades a partir del cual se selecciona la muestra (listado de unidades que componen la población).
- Valores verdaderos: Valores que toman las variables que deseamos estudiar.
- Valores observados: Valores que no coinciden con los verdaderos por errores cometidos tanto en la obtención primaria como en las operaciones posteriores.
- Estimaciones: Valores aproximados de la población en su totalidad inferidos a partir de la muestra.
- Error debido al muestreo: Afectan a las estimaciones al inferirlas desde la muestra. Cuanto menor sea ste, diremos que mayor es la precisión de las estimaciones.
- Espacio muestral ($S(x)$): Conjunto de todas la muestras posibles extraidas con un procedimiento de muestreo dado.
- Muestreo: Procedimiento por el cual se extrae una muestra.

Tipos de muestreo

1- Probabilístico: Puede obtenerse de antemano cul es la probabilidad de obtener cada uno de las muestras que es posible seleccionar. Es necesario que la selección de cada elemento sea un experimento aleatorio. Se define una función de probabilidad en $S(x)$ tal que $\sum_{x \in S(x)} P(x) = 1$.

2- Muestreo intencional u opintico: En l es la persona que selecciona la muestra la que procura que sta sea representativa, pero la representatividad depende de su intención u opinión.

3- Muestreo sin norma: Se toma la muestra de cualquier manera por comodidad o capricho. La representatividad puede ser satisfactoria sólo si la población es homognea.

Los dos últimos tipos de muestreo carecen de base teórica satisfactoria.

Ms conceptos asociados a un muestreo

- N: Tamaño de la población bajo estudio.
- $U = \{u_1, \cdots, u_N\}$: Población finita constituida por unidades distintas e identificables.
- u_i: Unidades de la población o unidades de muestreo.
- Muestra (s): Conjunto de unidades de U. $s = (u_{i1}, \cdots, u_{in(s)})$ con $1 \leq i1 \leq \cdots \leq in(s) \leq N$.
- Tamaño muestral ($n(s)$): Número de unidades de la muestra.
- Diseño muestral (d): $d = (S_d, P_d)$ par donde S_d es un subconjunto del espacio muestral universal S y P_d es una función de probabilidad definida en S_d tal que: $P_d : S_d \to \mathbb{R}$

  $$i) P_d(s) > 0 \qquad \forall s \in S_d \qquad (0 \leq P_d(s) \leq 1) \qquad \sum_{S_d} P_d = 1$$

  $$ii) \forall u \in U \ \exists s \in S_d \ / \ u \in S$$

  d en U se dice que es uniforme si P_d es uniforme en S_d.

  d en U se dice que es de tamaño fijo n si $n(s) = n \ \forall s \in S_d$.

- π_i: Probabilidad de inclusión de primer orden bajo el diseño muestral d, donde i nos indica la unidad i-sima de la población.

  $$p(u_i \in s) = \pi_i(d) = \sum_{i \in s} P_d(s) \qquad \forall i$$

- $\pi_{i,j}$: Probabilidad de inclusión de segundo orden bajo el diseño muestral d.

  $$p(u_i, u_j \in s) = \pi_{i,j}(d) = \sum_{i,j \in s} P_d(s) \qquad \forall i, j$$

- Matriz de diseño de d de U: $\Pi(U) = (\pi_{ij})_{1 \leq i,j \leq N}$. Matriz cuadrada simtrica de orden $N \times N$. Adems se verifica que:

  $$\forall i, j \qquad \text{máx}\{0, \pi_i + \pi_j - 1\} \leq \pi_{i,j} \leq \text{mín}\{\pi_i, \pi_j\}$$

  .

- Variable indicadora: Para cada $i \in U$ población finita sobre la que est definido un diseño muestral d, se define la variable indicadora:

$$I_i(s) = \begin{cases} 1 & \text{si } i \in s \\ 0 & \text{en caso contrario.} \end{cases} \qquad \forall s \in S_d$$

  Se verifica que: $E[I_i(s)] = \pi_i \ \forall i$, $Cov[I_i(s), I_j(s)] = \pi_{ij} - \pi_i\pi_j \ \forall i, j$ y $V[I_i(s)] = \pi_i(1 - \pi_i) \ \forall i$.

- Estadístico: Función real de la muestra del diseño que se considere $d = (S - d, P_d)$ que depende de X sólo a travs de los x_i para los que $i \in S$: $e : S_d \times \mathbb{R}^N \to \mathbb{R}$

- Estimador: Es un estadístico, de un parmetro $\theta(x)$, cuyos valores se emplean para predecir el valor del parmetro de inters $\theta(x)$. Es una variable aleatoria cuya función de distribución asociada recibe el nombre de distribución en el muestreo del estimador.

Estimadores usuales

- Media muestral: $\overline{x}_s = \sum_{i \in s} \dfrac{x_i}{n(s)}$.

- Estimador de la razón: $\overline{x}_R = \overline{x}_s \dfrac{\overline{Y}}{\overline{y}_s}$.

- Estimador media de razones: $\overline{r}_{xy} = \dfrac{1}{n(s)} \displaystyle\sum_{i \in s} \frac{x_i}{y_i}$.

- Estimador de Horvitz-Thompson: $\overline{x}_{HT} = \dfrac{1}{N} \displaystyle\sum_{i \in s} \frac{x_i}{\pi_i}$.

- Estimador de diferencias: $\overline{x}_D = \overline{x}_s + (\overline{Y} - \overline{y}_s)$.

- Estimador de regresión: $\overline{x}_r = \overline{x}_s + b(\overline{Y} - \overline{y}_s)$.

Dado un estimador $e(s, x)$ de un parmetro $\theta(x)$ de una población finita U con diseño muestral $d = (S_d, P_d)$, se definen las siguientes funciones del estimador:

- Esperanza: $E[e(s, x)] = \sum_{s \in S_d} e(s, x) P_d(s)$.

- Sesgo: $B[e(s, x)] = E[e(s, x)] - \theta(x)$.

- Se dice que el estimador es insesgado si: $B[e(s, x)] = 0 \Rightarrow E[e(s, x)] = \theta(x)$.

- Varianza: $V[e(s, x)] = E[e(s, x) - E[e(s, x)]]^2$.

- Error cuadrtico medio: $ECM[e(s, x)] = E[e(s, x) - \theta(x)]^2 = V[e(s, x)] + B[e(s, x)]^2$.

- Error de muestreo: $\sqrt{ECM} (= \sqrt{V}$ si $B = 0)$.

- Coeficiente de variación: $CV[e(s,x)] = \dfrac{\sqrt{V[e(s,x)]}}{E[e(s,x)]}$.

- Intervalos de confianza (IC):

  - Desigualdad de Chebichef: Si $\widehat{\theta}$ es un estimador de θ

    $$\left(\widehat{\theta} - k\sqrt{V(\widehat{\theta})}, \widehat{\theta} + k\sqrt{V(\widehat{\theta})}\right)$$

    IC para θ con nivel de confianza $1-1/k^2$, es decir, la probabilidad de que el intervalo cubra el parmetro es de $1 - 1/k^2$.

  - Teorema Central de Límite (n grande > 35): Si $\widehat{\theta}$ es un estimador de θ

    $$\left(\widehat{\theta} - z_{\alpha/2}\sqrt{V(\widehat{\theta})}, \widehat{\theta} + z_{\alpha/2}\sqrt{V(\widehat{\theta})}\right)$$

    IC para θ con nivel de confianza $1 - \alpha$, donde $z_{\alpha/2}$ es el valor de una distribución normal que deja a la derecha un rea de $\alpha/2$.

- Errores:

  - Errores de muestreo: Son los errores debidos a que parte de la población no est en la muestra. $ECM(\widehat{\theta}) = V(\widehat{\theta}) + B^2$
  - Errores ajenos al muestreo:
    i) De observación: Se deben a la recogida, registro o procesamiento incorrecto de los datos.
       - Sobrecobertura (el marco contiene elementos que no pertenecen a la población investigada).
       - Medida (valor original-valor observado).
       - Procesamiento (debido a los procesos de entrada, edición, tabulación y anlisis de datos).

    ii) De no observación: Tienen lugar cuando no es posible obtener la información deseada para ciertos elementos o cuando es imposible incluir uno o ms elementos de la población en la muestra.
       - Cobertura (parte de la población investigada no est presente en el marco).
       - Falta de respuesta (ciertos elementos de la muestra no dan toda o parte de la información deseada).