Universidad Nacional de Tres de Febrero

Maestría en Generación de Información Estadística
Teoría y Técnicas de Muestreo
TP Final Muestreo Polietápico

Augusto E. Hoszowski

Ejercicio IA

El conjunto de mesas electorales de la elección Octubre 2023 será nuestro universo bajo estudio. Se desea estimar el total de votos a Unión por la Patria, Juntos por el Cambio, La Libertad Avanza y FIT a *Presidente y Vice* a nivel nacional y proporción de votos respecto al total de votos *positivos* mediante una muestra aleatoria de mesas electorales.

Se compararán dos diseños, ambos bietápicos; con los circuitos electorales como Unidades de Primera Etapa (UPEs) y las mesas electorales como Unidades de Segunda Etapa (USEs). Los diseños son los siguientes:

Diseño A

Primera etapa de selección
- Muestral aleatoria, estratificada, de 80 UPEs
- Estratificando el marco de muestreo en estas zonas:

```
CABA
Partidos del Conurbano Bonaerense
Resto de Buenos Aires
Región Pampeana (Córdoba, Santa Fé, La Pampa, Entre Ríos)
NEA - NOA
Resto
```

- Asignación de la muestra por estrato en forma proporcional a la cantidad de mesas electorales en cada estrato
- Selección de circuitos en cada estrato mediante MAS

Segunda etapa de selección
En cada circuito se seleccionarán 12 mesas electorales mediante muestreo aleatorio simple (o todas las que haya de haber menos)

1. Cuántos circuitos y mesas electorales hay en cada estrato?. Presentarlo en una tabla
2. Presentar en un tabulado el total de votos a UxP, JxC, LLA y FIT a nivel nacional y la proporción de votos respecto al total de votos positivos
3. Seleccionar una muestra con este diseño
4. Declarar el diseño de muestreo a **survey** (declarar ambas etapas de selección) y con **survey** estimar los totales y proporciones pedidas, junto a sus CV, IC(90%) y deff (en el caso de proporciones no es necesario calcular el deff)
5. Presentar en dos tablas (una para los totales , otra para las proporciones) los resultados

6. Los intervalos de confianza al 90% contienen a los parámetros poblacionales?

Diseño B

Idem que A, salvo que en vez de seleccionar los circuitos mediante MAS, los seleccionamos mediante Madow, con probabilidad de selección proporcional a la cantidad de mesas electorales del colegio y ordenando los circuitos según jurisdicción, sección y código de circuito dentro de cada estrato.

1. Hay circuitos autorepresentados en algún estrato?
2. Seleccionar una muestra con este diseño
3. Declarar el diseño de muestreo a **survey** (ahora solo la primera etapa se declara, sin el fcp)
4. Con **survey** estimar los totales y proporciones pedidas, junto a sus CV, IC(90%) y deff
5. Presentar en dos tablas (una para los totales, otra para las proporciones) los resultados
6. Los intervalos de confianza contienen a los parámetros poblacionales?
7. Presentar en un gráfico de barras la comparación del CV estimado de las estimaciones del total de votos con uno y otro diseño. Idem para la estimación de la proporción de votos.

Ejercicio 1B

En este ejercicio trabajaremos con un extracto de la base censal del CEN2001, *pmeto*06*lm.dta* y *pmeto*06*lm.dta* (corresponde a un departamento de la provincia de Buenos Aires).

Deseamos estimar efecto diseño y CV de una serie de estimadores en un muestreo bietápico.

Parámetros a estimar:
- Proporción de hogares según tipo de vivienda
- Proporción de hogares según desague inodoro
- Proporción de hogares según tenencia de computadora
- Proporción de hogares según tenencia de teléfono fijo

h4-tipo-vivienda

```
0  casa-tipo-a
1  casa-tipo-b
2  rancho
3  casilla
4  departamento
5  pieza inquilinato
6  pieza hotel pension
7  local no habitacion
8  vivienda movil
9  en la calle
```

h15-desague-inodoro

```
1  red publica
2  camara septica y pozo
3  solo pozo
4  hoyo excavacion etc
```

h24k “Tiene computadora” 1 Si 2 No

h24g “Tiene telefono fijo” 1 Si 2 No

Compararemos dos diseños bietápicos, variando solo la cantidad de unidades de primera y segunda etapa a seleccionar

Diseño A y B

Unidad de Primera Etapa (UPE): segmento censal Unidad de Segunda etapa (USE): hogar

Estratificamos las UPEs en tres estratos **según el código de fracción** (la tabla es de un solo departamento):

Estrato A: 1; 2 - 11
Estrato B: 12 - 31
Estrato C : 32 -

Primera etapa de selección MAS de **n** conglomerados. Asignando la muestra en forma proporcional (a la cantidad de segmentos del estrato)

Segunda etapa de selección MAS de **m** hogares (o todos si hay menos de **m**)

En el diseño A: n=80, m=16

En el diseño B: n=240, m=4

1. Las tablas corresponden a qué jurisdicción y departamento?

Para cada diseño
2. Seleccionar una muestra con **sampling**. Los tamaño de muestra (en hogares) en cada diseño son similares?
3. Estimar con **survey** los parámetros, las estimaciones, cv y deff y presentar los resultados (para las proporciones no es necesario calcular los deff) 4. Para un mismo parámetro los CV son similares?
5. Comparar los tamaño de muestra, deff y CV (estimados) de los diseños A y B. Qué se concluye?

Nota I: En la tabla con las estimaciones, CV y deff incluir el parámetro, o sea el valor poblacional a estimar.

Nota II: En el CEN2010 no se imputaron las variables bajo estudio. Para que nuestro ejercicio sea más realista podemos imputar los valores *NA* de las variables bajo estudio mediante **hotdeck** (en la práctica podríamos ensayar otros métodos), por ejemplo con el paquete **VIM**. Los pasos a seguir serían:

a. Hallar el porcentaje de *missing* en cada variables, para verificar que ninguno es superior a, por ejemplo, 30%
b. Imputar los valores *NA* mediante hotdeck aleatorio (incluyendo la opción **domain_var**), utilizando como variables auxiliares **radio** y **frac**. Si quedan valores *NA*, repetir la imputación pero solo con **frac** como variable auxiliar.
c. Verificar que no quedan *NA* en las variables bajo estudio

Nota III: La variable fracción está en formato 'character'. Tenerlo en cuenta.

Ejercicio II

En la práctica muchas encuestas (Opinión Pública, Epidemiología, Paneles, etc.), relevadas por organismo privados u oficiales, emplean métodos no probabilísticos. Al decidir la aplicación de un método u otro intervienen cuestiones de costo, **precisón deseada**, información disponible, etc. Es un tema de investigación la comparación de estos dos métodos. En el artículo *Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day*, de G. Kalton hace una reseña de este problema. Hacer un resumen del artículo, o de las partes que les hayan parecido más interesantes (se puede obviar alguna parte muy técnica)