Universidad Nacional de Tres de Febrero

Maestría en Generación de Información Estadística
Teoría y Técnicas de Muestreo
TP Ejercicios parciales

Augusto E. Hoszowski

Ejercicio I

Se desea estimar total de votos a FdT, JxC y FIT a *diputados nacionales* a nivel nacional y proporción de votos respecto al total de votos *positivos* mediante una muestra aleatoria estratificada de 400 colegios electorales para las elecciones de 14 Noviembre 2021. El marco de muestreo se construirá a partir de la tabla **MESAS ESCRUTADAS Cierre.csv** (tablas publicadas por la DNE).

Los votos se estimarán (a nivel nacional) para tres agrupamientos de listas:
- FdT (Frente de Todos) - JxC (Juntos por el Cambio) - FIT (Frente de Izquierda y de los Trabajadores)

Los colegios electorales se estratificarán en 6 estratos:
- CABA
- Partidos del Gran Buenos Aires
- Resto de Buenos Aires
- Región Pampeana (Córdoba, Santa Fé, La Pampa, Entre Ríos)
- NEA - NOA
- Resto

asignando la muestra mediante asignación proporcional a la cantidad de *electores* del estrato

Se desea comparar dos estrategias: - Estrategia 1: MAS en cada estrato
- Estrategia 2: Madow en cada estrato, con probabilidad de selección proporcional a la cantidad de mesas electorales del colegio, ordenando los estratos por jurisdicción, Sección y IdCircuito.

1. Qué cantidad mínima de variables identifica una mesa electoral? Y un colegio?
2. Hallar la proporción de votos a cada uno de los cuatro agrupamientos de partidos
3. Tabular la cantidad de colegios electorales, mesas electorales y electores por Estrato
4. Construir a partir del archivo dado una tabla de mesas electorales (lo necesitaremos más adelante), cada una con el total de votos a cada partido, el total de votos positivos y las variables de identificación.
5. A partir de la tabla de mesas electorales construir la tabla de colegios electorales, cada uno con el total de votos a cada partido, el total de votos positivos y las variables de identificación.
6. Tabular y graficar los tres totales y los tres porcentajes poblacionales a estimar
7. Seleccionar con **sampling** una muestra de colegios con cada estrategia
8. Calcular con **survey** las estimaciones pedidas, junto a sus CV, IC(90%) y deff.
9. Presentar en un cuadro y gráfico los resultados
10. Con alguna de las dos estrategias se puede determinar con un 95% de confianza quien sacó más votos?

Ejercicio II

Se seleccionó una muestra aleatoria estratificada de 700 hogares en una localidad, para estimar datos de pobreza. La localidad se estratificó en tres zonas. En cada zona se seleccionó una muestra aleatoria simple de hogares. Y en cada hogar seleccionado se encuestó a todos los miembros del hogar (suponemos que no hay no respuesta). La siguiente tabla muestra los resultados de la encuesta:

Zona	Hogares en el marco	Hogares encuestados	Hogares pobres en la muestra	Pob pobre encuestada	Pob encuestada
A	25000	200	70	260	820
B	65000	150	80	400	700
C	20000	250	22	60	600

1. Presentar en una tabla o gráfico la proporción de hogares pobres por zona
2. Hallar el factor de expansión de cada hogar de la muestra y de cada persona encuestada
3. Las tres zonas presentan un perfil diferencial en términos de pobreza?
4. Estimar, con la muestra seleccionada, el total de hogares pobres y la proporción de hogares pobres, el CV y deff correspondientes. Dar un IC(90%) para cada estimación.
5. Porqué, siendo que las tres zonas son diferentes respecto a la variable bajo estudio (proporción de hogares pobres), el deff es claramente mayor a 1 en la estimación del total y proporción de hogares pobres?
6. Estimar el total de personas pobres y la proporción de personas pobres (recordar que población pobre es la que habita en hogares pobres). Se puede con los datos disponibles estimar el CV y deff de estas estimaciones?
7. En base a los resultados de la encuesta, cómo debería haber sido la distribución de la muestra por zona si el objetivo era estimar el total de hogares pobres en la localidad? (obtener la asignación de Neyman)

Ejercicio III

El siguiente universo artificial de N=8 unidades se estratifica en dos estratos, 1 y 2. Se desea estimar mediante muestreo estratificado con MAS en cada estrato la media de Y y el correspondiente CV, con esta asignación de muestra:

En el estrato 1 (autorepresentado), se selecciona esa única unidad
En el estrato 2 , nh=1
En el estrato 3, n=3

Seleccionar con **sampling** una muestra ese diseño y estimar luego con **survey** la media de Y y el correspondiente CV

Estrato	Unidad	Y
1	1	18
2	2	9
2	3	10
3	4	5
3	5	6
3	6	2
3	7	4
3	8	6

Nota: Habrá que decirle a *survey* cómo tratar el estrato 2. Por ejemplo, con la opción

options(survey.lonely.psu='fail')

survey señala error: encuentra un estrato con un solo PSU (en este caso, el estrato 2) Buscar las otras opciones en https://r-survey.r-forge.r-project.org/survey/exmample-lonely.html

Ejercicio IV

La tabla *tabla_deptos_2022.xlsx* será nuestro universo. Se desea estimar el total de cierta variable Y continua (población 2022), que tiene una distribución asimétrica. El máximo de estratos que se pueden definir es H=4. Se desea un CV menor al 5%. Debemos estratificar el marco de muestreo mediante Y y deerminar la cantidad de muestra (mediante MAS) a seleccionar en cada estrato, para cumplir con un CV $< 5\%$.

Hay que definir entonces los tres *cortes* que definenden los cuatro estratos. Y la cantidad de muestra a seleccionar en cada estrato. Utilizaremos para ello el paquete *stratification*. Y el comando *strata.LH*.

- *strata.LH* nos da la 'mejor' estratificación, tamaño de muestra y asignación
- Para suponer *asignación óptima*, se indica **alloc = c(0.5,0.5,0)**)
- *strata.LH* nos da los puntos de corte que definen los estratos y el tamaño de muestra en cada uno

1. Estratificar el marco
2. En una tabla presentar los N_h y n_h de cada estrato
3. Seleccionar una muestra con **sampling** (MAS en cada estrato)
4. Declarar el diseño a **survey**
5. Estimar el total de **Y** con **survey**, con el correspondiente CV y deff. El intervalo de confianza al 95% contiene al parámetro?

Nota: El paquete supone MAS en cada estrato, en la práctica podemos seleccionar con Madow En la práctica **Y** es desconocido, pero a veces se dispone de una variable auxiliar muy asociada, por ejemplo una medición de Y en algún censo anterior. El paquete *stratification* permite modelizar esto se puede investigar cómo hacerlo.