Proyecto Práctico - Muestreo – Maestría en Ciencia de Datos – ITAM – marzo/2018 – Emilio López Escobar

Ejercicio 1:

Actividades:

1. Hacer los puntos 2 a 5 utilizando los siguientes tamaños de muestra n:
    - n = 6,500 secciones (secciones = renglones en la tabla provista)
    - n = 500 secciones
    - n = 250 secciones
    - n = 100 secciones (tamaño de muestra más común en la práctica: TV, prensa, etc.)
    - n = 50 secciones
2. Extraer una muestra de tamaño n bajo muestreo aleatorio simple (sin remplazo). Se sugiere salvar las muestras extraídas para que no se re-extraigan o guardar la semilla.
3. Estimar el **total de votos** por cada uno de los candidatos de 2012 (EPN, AMLO, JVM, GQT, NoReg, y Nulos - notar que hay coaliciones, tendrá que crear nuevas columnas en los datos – creará un total de 6 columnas, que son las que utilizará -, para más información ver: https://es.wikipedia.org/wiki/Elecciones_federales_en_M%C3%A9xico_de_2012)
4. Estimar la varianza del estimador en cada estimación (utilizando un estimador que corresponda, si hay opciones, utilice uno y justifique).
5. Estimar por intervalos de confianza al 95%, estimar el coeficiente de variación y el efecto de diseño para la estimación que hizo de cada candidato.

6. Repetir el punto 1 a 5 pero ahora extrayendo las muestras con muestreo con probabilidades proporcionales al listado nominal (variable LISTA NOMINAL).

7. Discuta ampliamente sus resultados, de ser posible utilice gráficos resumen (esto es lo más importante de este proyecto práctico, la discusión). Utilice las siguientes **preguntas guía** para orientar su discusión:
    - ¿funciona mejor el muestreo aleatorio simple o el utilizar probabilidades desiguales? ¿por qué?
    - ¿qué fracción de muestreo está utilizando con cada tamaño de muestra? ¿importa el tamaño de la población? ¿cómo se compara el tamaño de muestra con el tamaño de la población?
    - ¿qué tan veraces son sus conclusiones utilizando el CVE, comparando con el verdadero valor de lo que está estimando?
    - ¿confía en sus estimaciones? ¿qué tan amplios son sus intervalos de confianza? En sus estimaciones, ¿qué candidato es más sensible al cambio en los tamaños de muestra? ¿qué candidato se estima mejor utilizando muestreo aleatorio simple y cuál utilizando probabilidades desiguales? ¿por qué?
    - ¿qué opina de utilizar 100 secciones para estimar bajo muestreo aleatorio simple o bajo muestreo con probabilidades desiguales?
    - Comente, discuta sus hallazgos.

Población:

Se utiliza una tabla (base de datos) con los resultados del 2012 por sección electoral.

Parámetro a estimar:

Total de votos por cada candidato (incluyendo No Registrados y Nulos).

Estimador puntual:

- El visto en clase para estimar un total.

Diseño de muestreo:

- Muestreo aleatorio simple con distintos tamaños de muestra (primera parte: incisos 1 a 5).
- Muestreo con probabilidades desiguales proporcionales al listado nominal (segunda parte: inciso 6).
- Variables de interés: Construidas según la candidatura (4 candidaturas más No Reg y Nulos – en total 6 variables de interés).
- Tamaños de muestra: 6500, 500, 250, 100 y 50.

Método de selección de las muestras:

- Máxima entropía o de rechazo de Hájek (1964), disponible en el paquete R “sampling” (UPmaxentropy).

Probabilidades de inclusión de segundo orden:

- Aproximadas por Hájek (1964), utilice la versión poblacional donde no se estima al término d. Si se complica por cuestiones de cómputo (poca memoria), utilice la versión muestral (aquella donde se estima al término d). Se sugiere usar el paquete samplingVarEst para ello.

Equipos:

- máximo 3 (definir su equipo a más tardar el martes 13, mandar esta info por e-mail).

Entregables:

- Se esperan 3 cosas:
    1. Tablas de resultados de sus estimaciones (entendibles y comentadas).
    2. Si es creativo, se esperan gráficos (entendibles y comentados).
    3. Reporte breve donde comenten sus hallazgos.
- En total los entregables de este ejercicio máximo deben ser 4 páginas tamaño carta. No desperdicien el tiempo de sus lectores, comuniquen directamente, comenten hallazgos relevantes. No entreguen código.

Sugerencias:

- Eviten irse a lo mínimo. Si se esfuerzan poco no esperen una gran nota de calificación del proyecto.
- Todo esto ya se hizo en el script que viene en sus notas pero para otros datos. Adapten ese script.
- Revisen los manuales de los paquetes R: sampling y samplingVarEst. Ahí está todo lo que necesitan.
- Cualquier duda, díganme: emilio@numerika.mx. Hagan sus preguntas con tiempo**. No se resuelven dudas desde un día antes a la entrega**.

Entrega:

- Lunes 2 de abril hasta las 9:30 horas (la mañana). (estricto, para ser justos con todos). Se enviará e-mail de confirmación. Enviar la tarea vía e-mail: emilio@numerika.mx (poner el asunto: Tarea ITAM MCDatos).

No lo dejen para lo último.

Información adicional sobre los datos proporcionados:

- El número de la sección (variable SECCION), es un número entero cuya numeración re-inicia dentro de cada entidad federativa ¡CUIDADO!.
- La BD está a nivel sección. Las filas son secciones. Las muestras que se extraerán serán de secciones.
- Algunos detalles de la BD:

ENTIDAD : Entidad (clave)
NOMBRE ENTIDAD : Nombre de la entidad
NUM FILA : Es un consecutivo (no tiene uso para este ejercicio)
DISTRITO : Distrito federal electoral (no tiene uso para este ejercicio)
MUNICIPIO : Municipio
SECCION : Número de la sección electoral
CASILLAS : Total de casillas que hay en la sección electoral (no tiene uso para este ejercicio)
PAN :
PRI :
PRD :
PVEM :
PT :
MOVIMIENTO CIUDADANO :
NUEVA ALIANZA :
COALICIÓN PRI PVEM :
COALICIÓN PRD PT MC :
COALICIÓN PRD PT :
COALICIÓN PRD MC :
COALICIÓN PT MC :
NO REGISTRADOS : E.g. "Cantinflas"
NULOS : Blanco, tachado, etc. no válido
TOTAL : Total de votos emitidos (no tiene uso para este ejercicio)
LISTA NOMINAL : Listado nominal, i.e. la cantidad de electores que podrían votar en esa sección

Ejercicio 2:

Objetivo: Utilizar una muestra no probabilística asociada a una encuesta on-line para estimar el **porcentaje** de intención de voto (Nota: se trata de **datos ficticios**).

Actividades:

1. Utilizar el método raking para calibrar los factores de expansión de la base de datos muestral provista para que cumpla la distribución de las siguientes variables de manera simultánea.

SEL : SOCIOECONOMIC LEVEL (AMAI MEXICO 8X7)	
SEL	**%**
AB	3.90
C+	9.30
C	10.70
C-	12.80
D+	19.00
D	31.80
E	12.50
TOTAL	**100.00**

TEL : TELEPHONE SERVICE	
TEL	**%**
LANDLINE	2.50
MOBILE	53.77
BOTH	35.61
NEITHER	8.12
TOTAL	**100.00**

GA5 * SEX	SEX		TOTAL
GA5	**M**	**F**	
18 TO 24	9.21	9.45	**18.66**
25 TO 34	10.79	11.87	**22.66**
35 TO 44	9.94	11.03	**20.97**
45 TO 59	10.46	11.75	**22.21**
60 +	7.17	8.33	**15.50**
TOTAL	**47.57**	**52.43**	**100.00**

   Nota: El método raking se puede usar con números absolutos o con porcentajes. Utilice porcentajes.

2. Posterior a haber calibrado los factores de expansión con el método raking, calcule un factor de corrección por redondeo RND = Suma de los factores de expansión sin calibrar / Suma de los factores de expansión calibrados. Este valor RND lo multiplicará a sus factores de expansión calibrados para obtener sus factores de expansión calibrados finales. Con esta corrección evitará desvíos acumulados por efectos de redondeo.
3. Estimar de manera puntual la proporción de voto para los candidatos utilizando los factores de expansión sin calibrar y utilizando los factores de expansión calibrados finales. Compare las estimaciones.

4. Discuta ampliamente sus resultados, de ser posible utilice gráficos resumen (esto es lo más importante de este proyecto práctico, la discusión). Utilice las siguientes **preguntas guía** para orientar su discusión:
   - ¿qué candidatos subieron su porcentaje de intención de voto al calibrar sus datos muestrales no probabilísticos? ¿qué candidatos bajaron? ¿por qué cree que pasa esto? ¿confía en sus estimaciones?
   - ¿valdrá la pena calcular intervalos de confianza (justifique)?
   - ¿cómo estaba la distribución de las variables SEL, TEL, GA5 y SEX en su base de datos muestral antes de calibrar? ¿cómo está la distribución de tales variables después de calibrar? ¿porqué cree que estaba con tal distribución antes de calibrar?
   - ¿cómo mejoraría sus estimaciones? ¿por qué otras variables valdría la pena calibrar adicionalmente?
   - ¿qué tan pesados son sus individuos más pesados con respecto al resto de individuos en su muestra?
   - Comente, discuta sus hallazgos.

Población de interés:

Personas mayores de 18 años con Internet en su hogar de acuerdo con información de ENDUTIH INEGI 2016, EIC INEGI 2015.

Parámetro a estimar:

Porcentaje de intención de voto por cada candidato, incluyendo nulos y blancos.

Estimador puntual:

- Estimador de una razón (numerador: total de intención de voto por el candidato en cuestión, denominador: total de intención de voto), i.e. estimador de una proporción con N desconocido. O bien, un estimador de una proporción suponiendo N conocido con el valor 40,855,000 (que es la suma de los factores de expansión naturales y que en este ejercicio se podría asumir como algo dado, ya que lo que interesa es la calibración más que la estimación).

Diseño de muestreo:

- No probabilístico
- Tamaño de muestra: 500.

Método de selección de las muestras:

- Pop-up survey.

Equipos:

- máximo 3 (definir su equipo a más tardar el martes 13, mandar esta info por e-mail).

Entregables:

- Se esperan 3 cosas:
    1. Estadísticos descriptivos y/o gráficos de las variables faltantes RND y EFC en su base de datos muestral.
    2. Tablas de estimación puntual de proporción de intención de voto con y sin calibración.
    3. Reporte breve donde comenten sus hallazgos.
- En total los entregables de este ejercicio máximo deben ser 3 páginas tamaño carta. No desperdicien el tiempo de sus lectores, comuniquen directamente, comenten hallazgos relevantes.

Sugerencias:

- Eviten irse a lo mínimo. Si se esfuerzan poco no esperen una gran nota de calificación del proyecto.
- Todo esto ya se hizo, aunque de manera rústica y simplificada en el pizarrón.
- Cualquier duda, díganme: emilio@numerika.mx. Hagan sus preguntas con tiempo**. No se resuelven dudas desde un día antes a la entrega**.

Entrega:

- Lunes 2 de abril hasta las 9:30 horas (la mañana). (estricto, para ser justos con todos). Se enviará e-mail de confirmación. Enviar la tarea vía e-mail: emilio@numerika.mx (poner el asunto: Tarea ITAM MCDatos).

No lo dejen para lo último.

Información adicional sobre los datos proporcionados:

VARIABLE	DESCRIPTION
SEL	SOCIOECONOMIC LEVEL (AMAI MEXICO 8X7)
TEL	TELEPHONE SERVICE
SEX	SEX
GA5	GROUP OF AGE (5 CATEGORIES)
EFN	EXPANSION FACTOR (WITHOUT CALIBRATION) (PEOPLE 18+ WITH INTERNET AT HOME - ENDUTIH INEGI 2016, EIC INEGI 2015)
RND	ROUND ADJUST
EFC	EXPANSION FACTOR (WITH CALIBRATION: SEL, TEL, SEX, GA5) (PEOPLE 18+ WITH INTERNET AT HOME - ENDUTIH INEGI 2016, EIC INEGI 2015) (ROUND ADJUSTED)
VOT	VOTE