# TALLER
# DISEÑOS DE MUESTREO ESTADÍSTICO

**Profesor: Giovany Babativa**

1. (Ejer. 1.7 Lohr) Para cada una de las siguientes encuestas, describa la población objetivo, el marco de muestreo, la unidad de muestreo y la unidad de observación.

   *a*) Un estudiante desea estimar el porcentaje de fondos cuyas acciones aumentaron de precio la semana pasada. Usando una lista que salió en el periódico Portafolio, el alumno eligió uno de cada 10 fondos y con estos calculó el porcentaje de aquellos donde el precio de la acción había aumentado.

   *b*) Se extrae una muestra de 8 arquitectos de una ciudad con 45 arquitectos. Para formar la muestra de la encuesta, cada arquitecto fue contactado por teléfono por orden de aparición en el directorio telefónico. Los primeros ocho que acordaron ser entrevistados conformaron la muestra.

   *c*) Para estimar cuántos libros de la biblioteca deben ser encuadernados de nuevo, un bibliotecario utiliza una tabla de números aleatorios para elegir al azar 100 posiciones en los estantes de la biblioteca. Luego, camina hasta cada posición, busca el libro que se encuentra en ese punto y registra si el libro debe encuadernarse o no.

   *d*) Se realiza un estudio para determinar el peso promedio de las vacas en una región del país. Usando una lista de las granjas de la región, se eligen 50 de ellas. Luego se registra el peso de cada vaca de las 50 granjas elegidas.

   *e*) En un juicio sobre marcas registradas, un demandante que afirme que otra compañía infringe sus marcas debe mostrar con frecuencia que las marcas tienen un *significado secundario* en el mercado; es decir, los usuarios potenciales del producto asocian las marcas registradas con el demandante aún cuando no esté presente el nombre de la compañía. Estos estudios son denominados de *confusión de marca*. Un estudio de confusión de marca entre los empaques de *Something Special* y *Haig Supreme* aplicado con una encuesta realizada a 500 personas consumidoras de whisky, mostró que cuando se mostraba la botella de *Haig Supreme* (Sin etiqueta) más del 50 % de los consumidores lo asociaron con *Something Special*.

2. Considere un marco de lista de $N$ familias, si se extrae una muestra de tamaño $n$ aplicando MAS. Calcule la probabilidad de que:

   a. La familia $k$ quede incluida en la muestra.

   b. Las familias $k$ y $l$ queden incluidas en la muestra.

   c. Las familias $k, l$ y $m$ queden incluidas en la muestra.

   d. ¿Cuál es la probabilidad de selección de la familia $k$?

3. (Ejer. 2.1 Lohr) Sean $N = 6$ y $n = 3$. Para estudiar las distribuciones del muestreo suponga que se conocen todos los valores de la variable de interés en la población.

$$y_1 = 98 \quad y_2 = 102 \quad y_3 = 154$$
$$y_4 = 133 \quad y_5 = 190 \quad y_6 = 175$$

El parámetro de interés es $\overline{y}_U$, la media de la población. Se proponen dos planes de muestreo.

- Plan 1: se cuenta con ocho muestras posibles

| Muestra | $S$ | $P(S)$ |
|---|---|---|
| 1 | {1,3,5} | $\frac{1}{8}$ |
| 2 | {1,3,6} | $\frac{1}{8}$ |
| 3 | {1,4,5} | $\frac{1}{8}$ |
| 4 | {2,4,6} | $\frac{1}{8}$ |
| 5 | {2,3,5} | $\frac{1}{8}$ |
| 6 | {2,3,6} | $\frac{1}{8}$ |
| 7 | {2,4,5} | $\frac{1}{8}$ |
| 8 | {2,4,6} | $\frac{1}{8}$ |

- Plan 2: se tienen tres muestras posibles

| Muestra | $S$ | $P(S)$ |
|---|---|---|
| 1 | {1,4,6} | $\frac{1}{4}$ |
| 2 | {2,3,6} | $\frac{1}{2}$ |
| 3 | {1,3,5} | $\frac{1}{4}$ |

a. ¿Cuál es el valor de $\overline{y}_U$?.

b. Sea $\overline{y}_s$ la media de los valores de la muestra. Para cada plan de muestreo, determine:

   I. $E(\overline{y}_s)$
   II. $V(\overline{y}_s)$
   III. $B(\overline{y}_s)$
   IV. $ECM(\overline{y}_s)$

c. Entre los dos planes de muestreo ¿Cuál es el mejor y por qué?

4. Sean $N = 8$ y $n = 4$. Suponga que los valores de la variable de interés son

$$y_1 = 98 \quad y_2 = 102 \quad y_3 = 154 \quad y_4 = 133$$
$$y_5 = 190 \quad y_6 = 175 \quad y_7 = 185 \quad y_8 = 105$$

Considere la siguiente estrategia de muestreo

| $S$ | $P(S)$ |
|---|---|
| {1,3,5,6} | $\frac{1}{8}$ |
| {2,3,7,8} | $\frac{1}{4}$ |
| {1,4,6,8} | $\frac{1}{8}$ |
| {2,4,6,8} | $\frac{3}{8}$ |
| {4,5,7,8} | $\frac{1}{8}$ |

a. Determine la probabilidad de inclusión $\pi_k$ para cada elemento.

b. ¿Cuál es la distribución muestral de $\widehat{t}_{y\pi}$.

c. ¿Cuál es la distribución muestral de $\overline{y}_s$.

5. (Ejer. 2.5 Lohr) Mayr et al.(1994) tomaron una muestra mediante un diseño $MAS$ de tamaño $n = 240$ niños con edad entre 2 y 6 años, ellos encontraron la siguiente distribución de frecuencias para la variable: edad en que los niños empezaron a caminar sin ayuda.

| Edad (Meses) | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # de niños | 13 | 35 | 44 | 69 | 36 | 24 | 7 | 3 | 2 | 5 | 1 | 1 |

*a*) Construya un histograma de la distribución de la edad al comenzar a caminar. ¿Cree Ud. que la variable sigue una distribución normal? ¿La distribución de la edad promedio en que los niños comienzan a caminar sigue una distribución normal? Explique.

*b*) Determine la media $\overline{y}_s$, el coeficiente de variación y un intervalo de confianza del 95 % para la edad promedio en que los niños comienzan a caminar solos.

*c*) Suponga que en el 2009 se desea hacer otro estudio en una región diferente y se desea que el intervalo de confianza del 95 % para la edad promedio en que los niños comienzan a caminar solos tenga un margen de error de 5 %. Basandose en la información auxiliar proporcionada ¿que tamaño de muestra se necesitaría usando un diseño $MAS$?

6. Suponga una población artificial donde se conocen todos los valores de $y_i$ para cada una de las $N = 8$ unidades de la población. Sea $U = 1, 2, 3, 4, 5, 6, 7, 8$ y $y_i = \{20, 32, 16, 27, 15, 22, 18, 31\}$. Usando el procedimiento ilustrado en clase, para el intervalo de confianza:

$$IC(S) = \left[\widehat{t} - 1{,}96\sqrt{V\left(\widehat{t}\right)}, \widehat{t} + 1{,}96\sqrt{V\left(\widehat{t}\right)}\right]$$

a. Determine la confiabilidad exacta del intervalo basado en una muestra aleatoria simple (MAS) sin reemplazo, de tamaño 4. ¿Es igual al 95 % la confiabilidad?

b. En la práctica la varianza es desconocida y es estimada. Si en la formula del intervalo de confianza el término $V\left(\widehat{t}\right)$ es sustituido por $\widehat{V}\left(\widehat{t}\right)$, determine la confiabilidad del intervalo basado en los mismos criterios del inciso anterior.

7. (Ejer. 2.10 Lohr) Una carta publicada en una revista indicaba lo siguiente: "*He observado que en los últimos números que no hay ganadores del sur en los concursos. Ustedes siempre dicen que los ganadores se eligen al azar. ¿significa esto que ustedes venden menos en el sur?*". En respuesta, los editores realizaron una muestra con un diseño MAS de 1000 datos de los últimos 10000 concursos y encontraron que 175 provenían del sur.

a. Determine un intervalo de confianza para los datos provenientes del sur.

b. De acuerdo con el reporte de censos de los editores, el 30.9 % de los suscriptores vive en estados considerados del sur. ¿Hay alguna evidencia en el intervalo de confianza de que el porcentaje de premios entregados en el sur difiere del porcentaje de suscriptores que viven en esa región del país?.

8. (ejer. 2.3 Särndal) Considere una población $U$ acompañada de tres subpoblaciones diferentes $U_1, U_2, U_3$ de tamaños $N_1 = 600, N_2 = 300$ y $N_3 = 100$ respectivamente. Así $U$ es de tamaño $N = 1000$, la inclusión o no inclusión en la muestra $s$ depende de un experimento bernoulli que da al elemento $k$ la probabilidad $\pi_k$ de quedar en la muestra. Los experimentos son independientes.

a. Sea

$$\pi_k = 0{,}1 \qquad \text{para todo} \quad k \in U_1$$

$$\pi_k = 0{,}2 \qquad \text{para todo} \quad k \in U_2$$
$$\pi_k = 0{,}8 \qquad \text{para todo} \quad k \in U_3$$

Encuentre el valor esperado y la varianza del tamaño de la muestra $n_s$ bajo este diseño.

b. Suponga que $\pi_k$ es constante para todo $k \in U$. Determine esta constante de tal manera que el valor esperado del tamaño de la muestra coincida con la esperanza obtenida en el item a. Obtenga la varianza del tamaño muestral y comparela con la varianza del caso a.

9. (ejer. 2.5 Särndal) Considere una población de tamaño $N = 3$, $U = \{1, 2, 3\}$. Sea $s_1 = \{1, 2\}$, $s_2 = \{1, 3\}$, $s_3 = \{2, 3\}$, $s_4 = \{1, 2, 3\}$ con $P(s_1) = 0{,}4$, $P(s_2) = 0{,}3$, $P(s_3) = 0{,}2$ y $P(s_4) = 0{,}1$.

   a. Calcule todos los $\pi_k$ y $\pi_{kl}$.

   b. Encuentre el valor de $E(n_s)$ de dos formas: por cálculo directo usando la definición y por uso de la fórmula que expresa a $E(n_s)$ como función de $\pi_k$.

10. (ejer. 2.6 Särndal) Considere la población y el diseño del ejercicio anterior. Sean los valores de la variable estudiada $Y = \{y_1 = 16, y_2 = 21, y_3 = 18\}$.

    a. Usando la definición, calcular el valor esperado y la varianza del $\pi$-estimador.

    b. Calcular la varianza del $\pi$-estimador usando la fórmula:

$$\sum\sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

    c. Calcule el coeficiente de variación del $\pi$-estimador.

    d. Calcule la varianza estimada $\hat{V}(\hat{t}_{y\pi})$ para cada una de las cuatro muestras posibles. ¿Es este un estimador insesgado de la varianza real?.

11. Suponga que se tiene $U = \{1, 2, 3, 4, 5\}$ y que se conocen todos los valores de la variable de interés, que están dados por:

| $Y$ | $P(S)$ |
|---|---|
| 79 | 0.1 |
| 76 | 0.15 |
| 54 | 0.2 |
| 39 | 0.25 |
| 12 | 0.3 |

Para una muestra de tamaño 3 con reemplazamiento, calcule (en excel):

a. Todas las muestra posibles (muestra ordenada).

b. Para cada muestra, determine la muestra no ordenada.

c. Para cada muestra, calcule la estimación de $t_y$ usando como estimadores $\hat{t}_{ymcr}$ y $\hat{t}_{y\pi}$.

d. Determine el valor esperado, varianza, varianza estimada y valor esperado de la varianza estimada para los dos estimadores.

e. Para cada muestra, calcule un intervalo de confianza del 90 %, 95 % y 99 % para los dos estimadores.

f. Determine la probabilidad de cobertura (confiabilidad) de cada uno de los intervalos hallados en el item anterior.

g. Concluya.

12. (ejer. 2.12 Särndal) Estimar el ingreso promedio por hogar ($\sum_U y_k/N$) para una población de $N = 200$ hogares, un listado de las 600 personas que pertenecen a los 200 hogares fue usado así: Una muestra MCR de tamaño $m = 10$ personas fue seleccionada. Los hogares de las personas de la muestra fueron identificados y la información sobre el ingreso promedio del hogar ($y_k/x_k$) fue recolectada, donde $y_k$ es el ingreso total del hogar en dolares y $x_k$ es el número de personas en el hogar. Los resultados son:

| Extracción $i$ | Ingreso promedio x familia $(y_{k_i}/x_{k_i})$ |
|---|---|
| 1 | 7.000 |
| 2 | 8.000 |
| 3 | 6.000 |
| 4 | 5.000 |
| 5 | 9.000 |
| 6 | 4.000 |
| 7 | 7.000 |
| 8 | 8.000 |
| 9 | 4.000 |
| 10 | 2.000 |

Calcule una estimación del ingreso promedio por hogar basada en el estimador M.C.R. y halle el cve correspondiente. Concluya.

13. (ejer. 3.1 Särndal) Suponga que se necesita extraer una muestra Bernoulli de una población conformada por 124 paises con el fin de estimar el total de la variable $Y$ : Importaciones de 1983 en millones de dolares, con un error estándar relativo del 10 %. Use como información auxiliar $\sum_U y_k = 1{,}81 \times 10^6$ y $\sum_U y_k^2 = 1{,}69 \times 10^{11}$, suponiendo que se usa el $\pi$-estimador determine el tamaño de muestra esperada. El error estándar relativo se define por $\sqrt{V(\hat{t})}/t$.