Resumen Tema 2: Muestreo aleatorio simple. Muestreo con probabilidades desiguales.

M.A.S.: Muestreo aleatorio simple con probabilidades iguales sin reemplazo.

Hipótesis: Marco perfecto, sin omisiones ni duplicados y elementos bien definido.
Se trata del muestreo más básico y simple que se puede realizar. No se presupone información a priori y sirve para comparar con otro tipos de muestreo. Si otro muestreo no es mejor teóricamente con igual costes, suele rechazarse el otro por su mayor complejidad.

Es el método más importante en poblaciones finitas. Consiste en:

- Se seleccionan n unidades distintas de entre N unidades poblacionales de la población U, con el diseño muestral $d = (S_d, P_d)$.
- S_d: El espacio muestral esta formado por todos las muestras de tamaño n que se pueden obtener de U.
- P_d: La distribución de probabilidad que se toma es la distribución uniforme.

$$M.A.S.(N,n) \qquad P_d : S_d \to [0,1]$$
$$P_d(s) = \frac{1}{\binom{N}{n}}$$

Para realizar un M.A.S. se deben ordenar todos las posibles muestras de n unidades distintas de la población (poco práctico si N es grande), y seleccionar una de ellas. Debido a que dicho proceso puede ser excesivamente laborioso se hace un esquema alternativo para realizar el muestreo que respeta las probabilidades de cada muestra de ser elegida:

Proceso para realizar un M.A.S.

1- Partimos de U población con N unidades.

2- Extraemos sucesivamente e independientemente las unidades con probabilidades iguales, en cada extracción, a $\dfrac{1}{N-t}$ para $t = 0, 1, \cdots, n-1$.

3- Una vez seleccionada una unidad de la población esta se excluye para que todas las sean distintas.

Con este método se verifica que $p(s)$ con $s \in S_d$ es $\binom{N}{n}^{-1}$ ya que hay $\binom{N}{n}$ elementos en S_d.

Probabilidades de inclusión

- π_i: $p(u_i \in \text{muestra elegida}) = \sum_{j=1}^{n} p(u_i \text{ sea elegida en la extracción j-ésima}) =$
$$= \frac{1}{N} + \frac{N-1}{N}\frac{1}{N-1} + \cdots + \frac{N-1}{N}\frac{N-2}{N-1}\cdots\frac{N-n+1}{N-n+2}\frac{1}{N-n+1} = \frac{1}{N} + \cdots + \frac{1}{N} = \frac{n}{N}$$

- $\underline{\pi_{ij}}$: $\dfrac{\text{Casos favorables}}{\text{Casos posibles}} = \dfrac{\text{número de muestras que contienen a } u_i, u_j}{\text{número total de muestras}} = \dfrac{\binom{N-2}{n-2}}{\binom{N}{n}} = \dfrac{n(n-1)}{N(N-1)}$

Parámetros	**Estimadores insesgados**	**Varianzas**	**Estimadores insesgados varianza**
Total $X = \sum_{i=1}^{N} x_i$	$\widehat{X} = N\overline{x}$	$V(\widehat{X}) = N^2 \dfrac{(1-f)}{n} S^2$	$\widehat{V}(\widehat{X}) = \dfrac{N^2(1-f)}{n} s^2$
Media $\overline{X} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$	$\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i\in s} x_i$	$V(\overline{x}) = \dfrac{N-n}{N-1}\dfrac{\sigma^2}{n}$	$\widehat{V}(\overline{x}) = \dfrac{(1-f)}{n} s^2$
Proporción $P = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} A_i$	$p = \dfrac{1}{n}\displaystyle\sum_{i\in s} A_i$	$V(p) = \dfrac{N-n}{N-1}\dfrac{\text{PQ}}{n}$	$\widehat{V}(p) = \dfrac{N-n}{N-1}\dfrac{\text{pq}}{n-1}$

Donde

- $A_i = \begin{cases} 1 & \text{si } u_i \in U \\ 0 & \text{en caso contrario} \end{cases}$
- $\pi_i = \dfrac{n}{N}$, $\pi_{ij} = \dfrac{n}{N}\dfrac{(n-1)}{(N-1)}$,
- $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j = \dfrac{-f(1-f)}{N-1} < 0$,
- $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i\in s}(x_i - \overline{x})^2$, que es la cuasivarianza muestral. Si S^2, es decir la cuasivarianza poblacional, es grande debemos tomar una muestra de tamaño n grande para que V sea pequeña y el error muestral sea pequeño. Si S^2 es pequeña, nos basta con tomar una muestra de tamaño n pequeña para que V sea adecuada ya que $V(\overline{x}) = \dfrac{(1-f)}{n} S^2$
- $f = \dfrac{n}{N} \in [0,1]$ que es la fracción de muestreo. f vale 0 cuando no hay representación de la población en la muestra y vale 1 cuando toda la población está en la muestra.
- $1 - f = \dfrac{N-n}{N}$ es el coeficiente de corrección por finitud.

Estimación de los Intervalos de Confianza

- Desigualdad de Chebichef: IC para θ con nivel de confianza $1 - 1/k^2$

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

- Teorema Central de Límite (n grande > 35): IC para θ con nivel de confianza $1-\alpha$.

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

Determinación del tamaño muestral para un error máximo admisible $e = |\widehat{\theta} - \theta|$ con nivel de confianza p_k: $n = \dfrac{n_o}{\dfrac{n_0}{N} + 1}$ donde

- Media poblacional $\overline{X}$: $n_0 = \dfrac{k^2 S^2}{e^2}$,
- Total poblacional X: $n_0 = \dfrac{N^2 k^2 S^2}{e^2}$,
- Media poblacional P:$n_0 = \dfrac{k^2 PQ}{e^2}$.

Debe tomarse como tamaño muestral n el valor entero más próximo por exceso al obtenido en la fórmula. Para el M.A.S. con reemplazamiento se verifica $n = n_0$.
A saber:

- $n \leq n_0$, luego el M.A.S. sin reposición tendrá menor tamaño que el M.A.S. con reposición para un mismo e.
- n aumenta se e disminuye.
- Si el nivel de confianza aumenta, es decir $1-\alpha$ aumenta, entonces n aumenta.
- Si S^2 aumenta, entonces n aumenta para un mismo e. Luego si la población es homogénea, basta un tamaño muestral n pequeño para estimar con una precisión aceptable.

Como el tamaño muestral depende de S^2 poblacional o del P poblacional, si dichos valores son desconocidos se pueden aproximar:

- Si se tiene un valor de S^2 de encuestas anteriores se puede usar.
- Si se conoce un intervalo de variación para S^2, se toma el valor máximo del mismo dando un tamaño muestral n mayor.
- Si se conoce el rango de variación de la variable, se puede aproximar S por el cociente $\dfrac{\text{rango}}{4}$.
- Si no se sabe nada y queremos un valor de n según un valor de P, se toma $P = 0.5$ que dará el tamaño máximo.
- Si no se está en ninguno de los casos anteriores, se toma una muestra piloto para estimar S^2 y luego se calcula el tamaño adecuado para la muestra.

- Procedimiento alternativo (Stein): Se toma una muestra preliminar de tamaño n_i. Se calcula s_1^2 para estimar S^2. Por último se toma una muestra adicional de tamañ o $n - n_1$ donde n es el calculado con la fórmula antes indicada usando la estimación obtenida con la primera muestra.

M.A.S.: Muestreo aleatorio simple con probabilidades iguales con reemplazo.

Hipótesis: Marco perfecto, sin omisiones ni duplicados y elementos bien definido.

Este método consiste en:

- Se seleccionan n unidades de entre N unidades poblacionales de la población U, con el diseño muestral $d = (S_d, P_d)$.
- S_d: El espacio muestral esta formado por todos las muestras de tamaño n que se pueden obtener de U, pudiéndose repetir los elementos en cada muestra.
- P_d: La distribución de probabilidad que se toma es la distribución uniforme.

$$M.A.S.(N,n) \qquad P_d : S_d \to [0,1]$$
$$P_d(s) = \frac{1}{N^n}$$

Para realizar un M.A.S. con reemplazo se deben ordenar todos las posibles muestras de n unidades de la población (poco práctico si N es grande), y seleccionar una de ellas. Debido a que dicho proceso puede ser excesivamente laborioso se hace un esquema alternativo para realizar el muestreo que respeta las probabilidades de cada muestra de ser elegida:

Proceso para realizar un M.A.S.

1- Partimos de U población con N unidades.

2- Extraemos sucesivamente e independientemente las unidades con probabilidades iguales, en cada extracción, a $\frac{1}{N}$.

3- Una vez seleccionada una unidad de la población esta se repone a la población.

Con este método se verifica que $p(s)$ con $s \in S_d$ es N^{-n} ya que hay N^n elementos en S_d.

Probabilidades de inclusión

- $\pi_i = \frac{n}{N}$
- $\pi_{ij} = \frac{n}{N}\,\frac{n}{N}$

Parámetros	**Estimadores insesgados**	**Varianzas**	**Estimadores insesgados varianza**
Total $X = \sum_{i=1}^{N} x_i$	$\widehat{X} = N\overline{x}$	$V(\widehat{X}) = N^2\dfrac{\sigma^2}{n}$	$\widehat{V}(\widehat{X}) = \dfrac{N^2 s^2}{n}$
Media $\overline{X} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i$	$\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i\in s} x_i$	$V(\overline{x}) = \dfrac{\sigma^2}{n}$	$\widehat{V}(\overline{x}) = \dfrac{s^2}{n}$
Proporción $P = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} A_i$	$p = \dfrac{1}{n}\displaystyle\sum_{i\in s} A_i$	$V(p) = \dfrac{\mathrm{PQ}}{n}$	$\widehat{V}(p) = \dfrac{\mathrm{pq}}{n-1}$

Estimación de los Intervalos de Confianza

- Desigualdad de Chebichef: IC para θ con nivel de confianza $1 - 1/k^2$

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

- Teorema Central de Límite (n grande > 35): IC para θ con nivel de confianza $1 - \alpha$.

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

Determinación del tamaño muestral para un error máximo admisible $e = |\widehat{\theta} - \theta|$ con nivel de confianza p_k: $n = \dfrac{n_o}{\dfrac{n_0}{N} + 1}$ donde

- Media poblacional $\overline{X}$: $n = \dfrac{z_{\alpha/2}^2 S^2}{e^2}$,
- Total poblacional X: $n_0 = \dfrac{N^2 z_{\alpha/2}^2 S^2}{e^2}$,
- Media poblacional P: $n_0 = \dfrac{z_{\alpha/2}^2 PQ}{e^2}$.

Debe tomarse como tamaño muestral n el valor entero más próximo por exceso al obtenido en la fórmula.
Si el valor de σ^2 es desconocido se debe actuar como en el M.A.S., tomando el valor de estimaciones anteriores, o el valor máximo con que se pueda acotar, y para P se debe usar el valor 0.5 si se desconoce.

Muestreo con probabilidades desiguales con reemplazo. Estimador de Hansen Hurwitz.

Hipótesis: Marco perfecto, sin omisiones ni duplicados y elementos bien definido.

Este método consiste en: Se seleccionan n unidades de entre N unidades poblacionales de la población U, con el diseño muestral $d = (S_d, P_d)$.

El espacio muestral esta formado por todos las muestras de tamaño n que se pueden obtener de U, pudiéndose repetir los elementos en cada muestra, donde la probabilidad de cada u_i es p_i distinta para cada unidad poblacional.

Probabilidad de inclusión: π_i=$p(u_i \in s) = 1 - (1 - p_i)^n$

Estimadores insesgados	**Varianzas**	**Estimadores insesgados varianza**
$\widehat{X}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{p_i}$	$V(\widehat{X}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{x_i}{p_i} - X\right)^2 p_i$	$\widehat{V}(\widehat{X}_{HH}) = \frac{\sum_{i=1}^{n}\left(\frac{x_i}{p_i}\right)^2 - n\widehat{X}_{HH}^2}{n(n-1)}$
$\widehat{\overline{x}}_{HH} = \frac{1}{N}\sum_{i=1}^{n}\frac{x_i}{p_i}$	$V(\widehat{\overline{x}}_{HH}) = \frac{1}{N^2}\frac{1}{n}\sum_{i=1}^{N}\left(\frac{x_i}{p_i} - X\right)^2 p_i$	$\widehat{V}(\widehat{\overline{x}}_{HH}) = \frac{1}{N^2}\frac{\sum_{i=1}^{n}\left(\frac{x_i}{p_i}\right)^2 - n\widehat{X}_{HH}^2}{n(n-1)}$
$\widehat{P}_{HH} = \frac{1}{N}\sum_{i=1}^{n}\frac{A_i}{p_i}$	$V(\widehat{P}_{HH}) = \frac{1}{N^2}\frac{1}{n}\sum_{i=1}^{N}\left(\frac{A_i}{p_i} - NP\right)^2 p_i$	$\widehat{V}(\widehat{P}_{HH}) = \frac{1}{N^2}\frac{\sum_{i=1}^{n}\left(\frac{A_i}{p_i}\right)^2 - nN^2\widehat{P}_{HH}^2}{n(n-1)}$

Si los p_i son proporcionales a los x_i, entonces el estimador de Hansen-Hurwitz coincide con el verdadero valor del parámetro y la varianza se anularía, serían parámetros perfectos. Por ello se suelen usar probabilidades proporcionales a una variable conocida y_i relacionada con los x_i.

Si $y_i = M_i$, es decir, tomamos como valor de la variable conocida el tamaño de la unidad i, entonces $M = \sum_{i=1}^{N} M_i$ y $p_i = \frac{M_i}{M}$.

Proceso para seleccionar una muestra:

1- Método de las probabilidades acumuladas. En primer lugar se calcula una columna de valores B_i acumulando los valores p_i, es decir $B_i = \sum_{j=1}^{i} p_j = p_1 + p_2 + \cdots + p_i$. A continuación se selecciona un valor aleatorio e entre $(0, 1)$ y se comprueba entre que dos valores de B_i se encuentra dicho valor, $B_{i-1} < e \leq B_i$. Se debe elegir para la muestra la unidad u_i y repetir el proceso hasta alcanzar el tamaño muestral deseado.

2- Método de Lahiri. Para comenzar se debe calcular el máximo de los p_i en la población, $q = \underset{i \in U}{\text{máx}}\, p_i$. A continuación se generan dos número aleatorios $i \in U[1, N]$ y $r \in U[0, q]$. Si $r \leq p_i$ se debe seleccionar la unidad u_i de la población y si $r > p_i$ no se selecciona ninguna unidad. El proceso se repite hasta obtener alguna unidad, y luego hasta completar el tamaño muestral.

Muestreo con probabilidades desiguales sin reemplazo. Estimador de Horvitz Thompson.

Hipótesis: Marco perfecto, sin omisiones ni duplicados y elementos bien definido.

Este método consiste en: Se seleccionan n unidades de entre N unidades poblacionales de la población U, con el diseño muestral $d = (S_d, P_d)$.

El espacio muestral esta formado por todos las muestras de tamaño n que se pueden obtener de U, no pudiéndose repetir los elementos en cada muestra, donde la probabilidad de cada u_i es p_i distinta para cada unidad poblacional.

Probabilidad de inclusión: π_i=$p(u_i \in s)$ no puede calcularse. Si es conocido podemos usarlo en las fórmulas, o en caso contrario puede tomarse $\pi_i = n\dfrac{y_i}{\sum\limits_N y_i}$ siendo Y una variable auxiliar conocida relacionada con la variable X.

Estimadores insesgados	**Varianzas**	**Estimadores insesgados varianza**
$\widehat{X}_{HT} = \sum\limits_{i=1}^{n} \dfrac{x_i}{\pi_i}$	$V(\widehat{X}_{HT}) = \sum\limits_{i=1}^{N} \left(\dfrac{x_i}{\pi_i}\right)^2 \pi_i(1-\pi_i) + \sum\limits_{i \neq j=1}^{N} \dfrac{x_i}{\pi_i}\dfrac{x_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$	$\widehat{V}(\widehat{X}_{HT}) = \sum\limits_{i,j \in s} \dfrac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\dfrac{x_i}{\pi_i}\dfrac{x_j}{\pi_j} = \dfrac{1}{n-1} \times \left\{ \sum\limits_{i<j=1}^{n} \left[1 - (\pi_i + \pi_j) + \sum\limits_{i=1}^{N} \dfrac{\pi_i^2}{n}\right] \left(\dfrac{x_i}{\pi_i} - \dfrac{x_j}{\pi_j}\right)^2 \right\}$
$\widehat{\widehat{x}}_{HT} = \dfrac{1}{N}\sum\limits_{i=1}^{n} \dfrac{x_i}{\pi_i}$	$V(\widehat{\widehat{x}}_{HT}) = \dfrac{1}{N^2} \sum\limits_{i,j=1}^{N} \dfrac{x_i x_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i\pi_j)$	$\widehat{V}(\widehat{\widehat{x}}_{HT}) = \dfrac{1}{N^2}\dfrac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\dfrac{x_i}{\pi_i}\dfrac{x_j}{\pi_j}$
$\widehat{P}_{HT} = \dfrac{1}{N}\sum\limits_{i=1}^{n} \dfrac{A_i}{\pi_i}$	$V(\widehat{P}_{HT}) = \dfrac{1}{N^2} \sum\limits_{i,j=1}^{N} \dfrac{A_i A_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i\pi_j)$	$\widehat{V}(\widehat{P}_{HT}) = \dfrac{1}{N^2}\dfrac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\dfrac{A_i}{\pi_i}\dfrac{A_j}{\pi_j}$

Proceso para seleccionar una muestra:

1- Aproximación por una extracción con reemplazo.

2- Método de Lahiri. Para comenzar se debe buscar una constante C tal que $Max_s \sum\limits_{i \in s} y_i < C$, siendo y_i los valores de la variable auxiliar. A continuación se genera un número

aleatorios $e \in U[0, C]$ y una muestra de n unidades con probabilidades iguales y sin reemplazo (M.A.S.). Si $\sum_{i \in s} y_i \geq e$ la muestra elegida es válida. En caso contrario hay que repetir el proceso hasta conseguir una muestra adecuada.

Es un método costoso y no se sabe de antemano cuantas operaciones hay que hacer.

3- Método de las extracciones sistemáticas. En primer lugar se calcula una columna de valores C_i acumulando los valores π_i, es decir $C_i = \sum_{j=1}^{i} \pi_j = \pi_1 + \pi_2 + \cdots + \pi_i$. A continuación se selecciona un valor aleatorio e entre $(0, 1)$ y se comprueba entre que dos valores de C_i se encuentra dicho valor, $C_{i-1} < e \leq C_i$. Se debe elegir como primera unidad para la muestra la unidad u_i y repetir el proceso n veces sumando 1 cada vez a e.

Es un método complicado de implementar.