MUEST-T7. MUESTREO ESTRATIFICADO:

TAMAÑOS MUESTRALES para MEDIAS y PROPORCIONES.
AFIJACIÓN con MÁS de una CARACTERÍSTICA.
AFIJACIÓN CON ALGUNA FRACCIÓN DE MUESTREO MAYOR QUE 1.
ESTRATIFICACIÓN A POSTERIORI.
APLICACIÓN a ESTIMACIÓN en SUBPOBLACIONES

1_ CONCEPTO de MUESTREO ESTRATIFICADO.

En el muestreo irrestricto aleatorio la población se considera homogénea respecto a la característica observada. Las estimaciones obtenidas son buenas a un coste bajo.

En el muestreo estratificado, la población se considera heterogénea respecto a la característica estudiada, y se subdivide en subpoblaciones lo más homogéneas posibles, con lo que se obtienen estimaciones más precisas con el mismo coste.

La población heterogénea con N unidades, $\{U_i\}$, $i=1\dots N$ se subdivide (partición) en L estratos de ≠ tamaños, $N_1 \dots N_h \dots N_L$

$\{U_{hi}\}$ | $h = 1 \dots L \rightarrow$ nº estratos pobl.
$i = 1 \dots N_h \rightarrow$ nº unidades/por estrato

De cada estrato se selecciona una muestra aleatoria, el tipo de muestreo puede ser ≠ para cada estrato, de manera independiente.

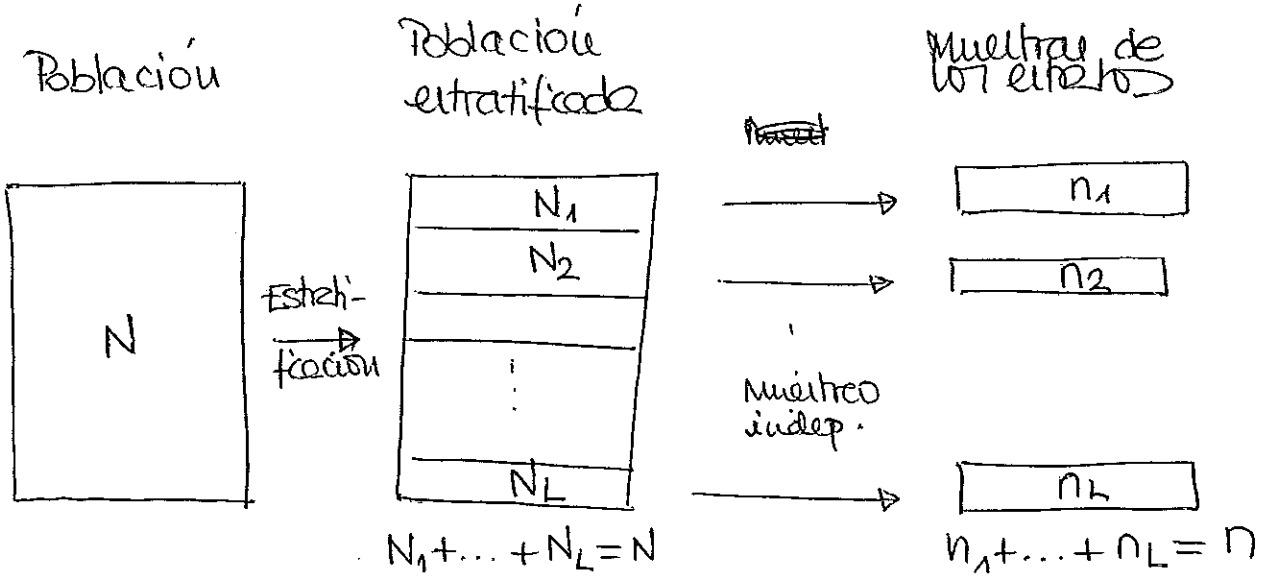
$\{U_{hi}\}$ | $h = 1 \dots L \rightarrow$ nº estrat.
$i = 1 \dots n_h \rightarrow$ nº unidades muestrales por estrato

---

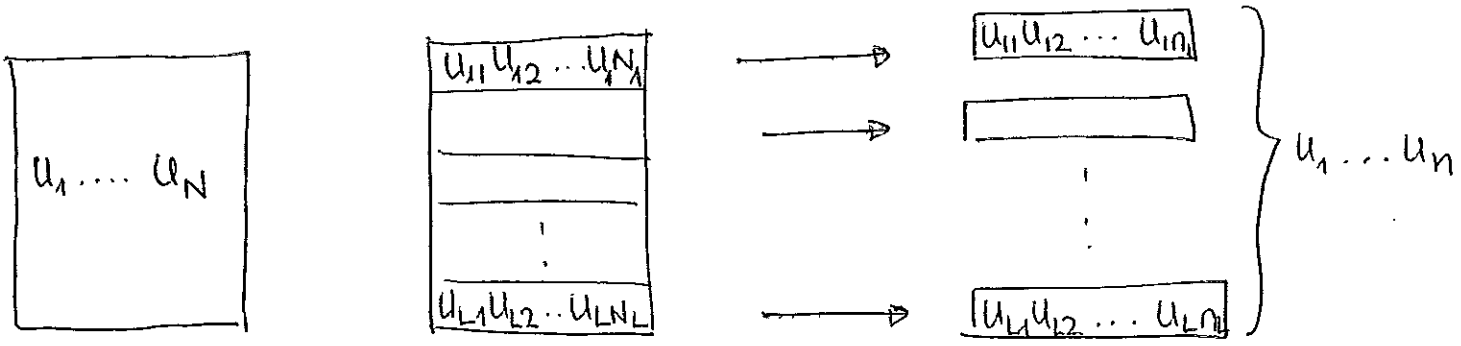
La homogeneidad / heterogeneidad se estudia a través de la variantza ≡ precisión del valor promedio.

Homogénea ≡ variantza mínimas
Heterogénea ≡ variantza distintas

El procedimiento del muestreo estratificado se puede resumir mediante el siguiente esquema:



que se puede detallar más especificando las unidades en cada grupo



u_i

$i = 1 \ldots N$

$\sum_{i=1}^{N} i = N$

u_{hi}

$h = 1 \ldots L$

$i = 1 \ldots N_L$

$N_1 + \ldots + N_L = N$

$\sum_{i=1}^{N_h} i = N_h$

$\sum_{h=1}^{L} \sum_{i=1}^{N_L} hi = N$

u_{hi}

$h = 1 \ldots L$

$i = 1 \ldots n_L$

$n_1 + \ldots + n_L = n$

$\sum_{L} i = n_h$

$\sum_{h=1}^{n_h} \sum_{i=1} hi = n$

Nótese que la estratificación es a nivel población, y en la muestra estarán presentes todos los estratos

El muestreo estratificado es aconsejable por los siguientes MOTIVOS:

- Permite obtener estimaciones para cada subpoblación.
- Puede generar ganancia en precisión. Al dividir la población heterogénea en estratos homogéneos, puede que la variante en cada estrato sea menor que la variante en toda la población.
- Se pueden utilizar distintos tipos de muestreo para cada estrato, lo que permite reducir el coste.
- Si existe una variable precisa para la estratificación, correlacionada con la variable, que permite dividir la población en estratos homogéneos.

Incorporar concepto de muestreo de T6

1- TAMAÑOS MUESTRALES PARA MEDIAS Y PROPORCIONES

~~a) SIN reposición~~

Si se quiere estimar el parámetro θ mediante un estimador lineal insesgado, $\hat{\theta}$, se puede calcular el tamaño muestral necesario que garantice un máximo error de muestreo dado.

Aunque las expresiones de los estimadores SIN y CON reposición, no ocurre así con su variancia, por lo que tendremos que tratar estos casos por separado.

MEDIA

a) SIN reposición

Para la media poblacional $\bar{X} = \frac{1}{N}\sum_{h}^{L}\sum_{i}^{N_h} X_{hi}$ ← ponderada

su estimador lineal insesgado es la suma extendida a todos los estratos de los estimadores lineales insesgados de Horvitz y Thompson en cada estrato:

$$\bar{x}_{st} = \sum_{h}^{L}\sum_{i}^{n_h}\frac{N_h}{N}\bar{x}_h = \sum_{h=1}^{L} W_h \bar{x}_h$$

← media ponderada de las medias muestrales de cada estrato, con $W_h = \frac{N_h}{N}$.

cuya variancia es:

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h)\cdot\frac{S_h^2}{n_h} \qquad f_h = \frac{n_h}{N_h}$$

- El error de muestreo es la desviación típica del estimador. A partir de su expresión se deduce el tamaño muestral:

$$E = \sqrt{V(\bar{x}_{st})} \Rightarrow E^2 = V(\bar{x}_{st})$$

$$E^2 = \sum_{h=1}^{L} W_h^2\cdot\frac{S_h^2}{n_h} - \sum_{h=1}^{L} W_h^2\cdot f_h\cdot\frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2\cdot\frac{S_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{L} W_h\cdot S_h^2$$

($W_h^2 = \frac{N_h^2}{N^2}$; $f_h = \frac{n_h}{N_h}$)

$$W_h = \frac{N_h}{N}$$

$$f_h = \frac{n_h}{N_h}$$

▸ Af. igual: $n_h = \frac{n}{L} \Rightarrow n = L\cdot n_h$

▸ Af. proporc: $n_h = \frac{n}{N}N_h = W_h n \Rightarrow n = \frac{n_h}{W_h}$

▸ Af. Neyman: $n_h = n\cdot\frac{W_h S_h}{\sum W_h S_h} \Rightarrow n =$

(CORREGIR)

de donde se puede despejar el valor de n_0.

La afijación que se utilice determinará el valor de los tamaños muestrales de cada estrato.

- Afijación uniforme: $n_h = \frac{n}{L}$, $\forall h = 1 \dots L$

$$E^2 = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n/L} - \sum_{h=1}^{L} \frac{N_h}{N^2} \cdot S_h^2 \Rightarrow n = \frac{L \sum_{h=1}^{L} W_h^2 S_h^2}{E^2 + \sum_{h=1}^{L} \frac{N_h}{N^2} S_h^2}$$

- Afijación proporcional: $n_h = W_h \cdot n \quad (W_h = \frac{N_h}{N})$

$$E^2 = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{W_h \cdot n} - \sum_{h=1}^{L} \frac{N_h}{N^2} S_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h S_h^2}{E^2 + \sum_{h=1}^{L} \frac{N_h}{N^2} S_h^2}$$

- Afijación mínima varianza: $n_h = n \frac{W_h S_h}{\sum W_h S_h}$

$$E^2 = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n \cdot \frac{W_h S_h}{\sum W_h S_h}} - \sum_{h=1}^{L} \frac{N_h}{N^2} S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} (W_h S_h)\right)^2}{E^2 + \sum_{h=1}^{L} \frac{N_h}{N^2} S_h^2}$$

b) CON reposición:

ATRÁS (folio anterior)

La varianza de la media muestral es: $E^2 = V(\bar{x}_{st}) = \frac{1}{n} \sum_{h=1}^{L} W_h \sigma_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{E^2}$ Af. prop.

* También se puede calcular el tamaño muestral para un error absoluto de muestreo y coeficiente de confianza dados, suponiendo que el estimador se distribuye con arreglo a una Normal. $E_\alpha = \lambda_\alpha \sigma(\hat{\theta}) \;/ P(z^* > \lambda_\alpha) = \frac{\alpha}{2}$;
Las fórmulas son iguales que las anteriores, sustituyendo E^2 por E^2/λ_α^2 (SIN y CON repos.)

~~* También se puede fijar el error relativo de muestreo:~~
$E = \frac{V(\bar{x}_{st})}{\sigma(\bar{x}_{st})}$

Media muestral

$$(CR)\ e^2 = V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \sigma_h^2 / n_h$$

Af. igual: $n_h = \frac{n}{L} \Rightarrow n = L \sum_{h=1}^{L} W_h^2 \sigma_h^2 / e^2$

Af. proporc: $n_h = \frac{n}{N} N_h \Rightarrow n = \frac{\sum W_h^2 \sigma_h^2 (N/N_h)}{e^2} = \frac{\sum W_h \sigma_h^2}{e^2}$

Af. mín-var: $n_h = n \frac{W_h \sigma_h}{\sum W_h \sigma_h} \Rightarrow n = \frac{(\sum W_h^2 \sigma_h^2)(\sum W_h \sigma_h)/W_h \sigma_h}{e^2} =$

$$= \frac{(\sum W_h \sigma_h)^2}{e^2}$$

$$e^2 = \sum W_h^2 \frac{\sigma_h^2}{n_h} \quad (*)$$

$$n_h = \frac{\sum W_h^2 \sigma_h^2}{e^2}$$

↑ Mejor

- Af. igual: $n_h = \frac{n}{L} \Rightarrow n = L n_h$
- Af. proporc: $n_h = \frac{n}{N} N_h \Rightarrow n = \frac{N}{N_h} \cdot n_h = \frac{1}{W_h} \cdot n_h$
- Af. mejorar: $n_h = n \cdot \frac{W_h \sigma_h}{\sum W_h \sigma_h} \Rightarrow n =$

Proporc: (SP): $(N_h - 1) S_h^2 = N_h \sigma_h^2 - N_h P_h Q_h$

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$$

(CR): $\sigma_h^2 \quad P_h Q_h$

FALTA HACERLO

PROPORCIÓN

Para la proporción poblacional $P = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} A_{hi}$, $A_{hi} = \begin{cases} 1 \\ 0 \end{cases}$

su estimador lineal insesgado es:

$\hat{P}_{st} = \sum_h W_h \hat{P}_h$ → media ponderada de la proporción en cada estrato, con $W_h = \frac{N_h}{N}$ y $\sum W_h = 1$.

con varianza

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \cdot \frac{N_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

La expresión de la varianza de la proporción es similar a la expresión de la varianza de la media muestral, teniendo en cuenta que $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$,

por lo que la obtención del tamaño muestral para la proporción se calcula sustituyendo S_h^2 por $\frac{N_h}{N_h - 1} P_h Q_h$ en las fórmulas anteriores,

CON reposición:

De manera análoga, sustituyendo σ_h^2 por $P_h Q_h$ en las fórmulas del tamaño de muestreo para la media.

FALTA considerar $e_d = \lambda_d \sigma(\hat{\theta})$

2. AFIJACIÓN de la MUESTRA con MÁS de UNA CARACTERÍSTICA

Es lógico que, teniendo en cuenta el coste (económico, personal y de tiempo) asociado a una encuesta, se pretenda obtener del colectivo información acerca de más de una característica.

Surge el problema de que la mejor ~~asignación~~ afijación para una característica no tiene por qué ser la mejor afijación para las demás.

El problema se resuelve adoptando una solución de compromiso.

La solución más sencilla consiste en reducir las características utilizadas para la afijación a un número relativamente pequeño, considerando solamente las características más importantes. Entonces se calcula la afijación óptima para cada una de estas características por separado, y se comprueba el grado de desacuerdo entre ellas.

Si las características importantes están altamente correlacionadas, las afijaciones calculadas pueden diferir relativamente poco. El problema surge cuando las afijaciones individuales difieren tanto que no se pueda llegar a una solución de compromiso obvia.

En este último caso, existen varias alternativas, que se pueden agrupar en dos opciones:

- Media ponderada de varianzas
- Acotar la varianza de cada estimador

1) Se construye una media ponderada de las varianzas de los estimadores de las características "importantes" y se busca la afijación óptima con costes, que minimice dicha media ponderada.

$$\left.\begin{array}{l} \underset{n_1,\dots n_L}{\text{MIN}} \ \sum_{j=1}^{K} a_j V(\bar{x}_{st_j}) \\ \text{s.a.} \ \sum_{h=1}^{L} c_h \cdot n_h = C \end{array}\right\} \rightarrow \text{Solución por multiplicadores de Lagrange.}$$

Encontrar los tamaños muestrales de cada estrato que minimicen la media ponderada (a_j = coef. ponderación) de las varianzas individuales de los estimadores de las K características importantes asociados a un coste de muestreo fijo C (c_h = coste unitario del estrato (h)).

2) Se especifica la precisión deseada para cada estimador acotando su varianza, y se calcula la afijación más económica que satisfaga todas las restricciones.

Para K medias poblacionales:

$$\left.\begin{array}{l} \underset{n_1\dots n_L}{\text{MIN}} \ \sum_{h=1}^{L} c_h n_h \\ \text{s.a.} \ V(\bar{x}_{st_j}) \leq V_j^*, \ j = 1\dots K \end{array}\right\} \rightarrow \begin{array}{l} \text{Progr. no lineal} \\ \text{ó} \\ \text{Algoritmo de Bethel} \\ \text{(basado en multiplicadores} \\ \text{de Lagrange)} \end{array}$$

Encontrar los tamaños muestrales de cada estrato que minimicen el coste total de muestreo tq no se superen los límites de precisión indicados para cada estimador

3. AFIJACIÓN que requiere UNA FRACCIÓN DE MUESTREO MAYOR QUE LA UNIDAD

En el caso de que la fracción de muestreo, $f = \frac{n}{N}$, sea muy grande y de que ~~algunos estratos~~ haya mucha heterogeneidad entre los estratos, puede ocurrir que la afijación óptima recomiende tamaños muestrales superiores a los poblacionales en algunos estratos.

En este caso, lo mejor que se puede hacer es tomar muestras exhaustivas en esos estratos ($n_h = N_h$) y, repartir el resto del tamaño muestral en los demás.

Consideremos un ejemplo: se trata de estimar la media poblacional utilizando afijación de mínima varianza, SR:

$$\left.\begin{array}{l} \underset{n_1 \dots n_L}{\text{MIN}}\ V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h)\frac{S_h^2}{n_h} \\ \text{s.a. } n_1 + \dots + n_L = n \end{array}\right\} \xrightarrow{\text{solución}} n_h = n \cdot \frac{W_h S_h}{\sum W_h S_h}$$

Sp. $n_1 > N_1 \Rightarrow$ la asignación óptima revisada es:

$$\tilde{n}_1 = N_1$$
$$\tilde{n}_h = (n - N_1) \cdot \frac{W_h S_h}{\sum W_h S_h} \quad \text{siempre que } \tilde{n}_h \leq N_h \ \ h = 2 \dots L$$

Si $\tilde{n}_2 > N_2 \Rightarrow$

$$\tilde{\tilde{n}}_1 = N_1$$
$$\tilde{\tilde{n}}_2 = N_2$$
$$\tilde{\tilde{n}}_h = (n - N_1 - N_2)\frac{W_h S_h}{\sum W_h S_h}$$

Hasta conseguir que todos los tamaños muestrales no superen los tamaños poblacionales.

Nótese que en los estratos donde se ha realizado un muestreo exhaustivo ($n_h = N_h$), la media muestral de cada estrato coincide con la media poblacional del estrato, por lo que estos estratos no aportan dispersión al estimador $\bar{x}_{st}$.

Suponiendo que ello ocurriera en los k primeros estratos:

$\tilde{n}_1 = N_1, \tilde{n}_2 = N_2 \ldots \tilde{n}_K = N_K, \tilde{n}_h = (n - \sum_{h=1}^{K} N_h) \frac{W_h S_h}{\sum W_h S_h}, h > K$.

$$V(\bar{x}_{st}) = V\left(\sum_{h=1}^{L} W_h \bar{x}_h\right) = \sum_{h=1}^{L} W_h^2 V(\bar{x}_h) = \sum_{h=1}^{K} W_h^2 V(\bar{x}_h) + \sum_{h=K+1}^{L} W_h^2 V(\bar{x}_h) =$$

$$= 0 + \sum_{h=K+1}^{L} W_h^2 V(\bar{x}_h) = \sum_{h=K+1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{\tilde{n}_h}$$

donde $f_h = \frac{\tilde{n}_h}{N_h}$

(CR en anexo(o))

4_ ESTRATIFICACIÓN a POSTERIORI

Puede ocurrir que no se conozca el estrato al que pertenece una unidad hasta después de recoger los datos. Los tamaños poblacionales de los estratos se pueden obtener a partir de estadísticas oficiales y su importancia en la población → N_h y W_h conocidos.

Una opción es tomar una m.a.s. (n) y clasificar las unidades en estratos a posteriori.

Como estimador, se utiliza una media ponderada de las medias

$$\bar{x}_w = \sum_{h=1}^{L} W_h \bar{x}_h \quad , \text{ donde } \bar{x}_h \equiv \text{media de los unidades que caen en el estrato } h.$$

Ahora el nº de unidades de la mas que caen en el estrato h, m_h, es una var. aleatoria ≠ para cada muestra.

Si los tamaños m_h fueran todos fijos y positivos:

$$V(\bar{X}_w/m_h) = \sum_{h=1}^{L} W_h^2 \cdot (1-f_h) \cdot \frac{S_h^2}{m_h}$$

En el caso de que algún m_h fuese 0, habría que agrupar ese estrato con otro. Se perdería precisión, pero es poco probable.

Pero como m_h son variables aleatorias:

$$V(\bar{X}_w) = E\left[V(\bar{X}_w/m_h)\right] = \sum_{h=1}^{L} W_h^2 (1-f_h) \cdot S_h^2 \cdot E\left[\frac{1}{m_h}\right]$$

Según Stephen (1945), $E\left[\frac{1}{m_h}\right] = \frac{1}{nW_h} + \frac{1-W_h}{n^2 W_h^2}$

por lo que:

$$E[V(\bar{X}_w)] = \ldots = V_{Af.prop}(\bar{X}_{st}) + \frac{1}{n^2} \sum_{h=1}^{L} (1-W_h) \cdot S_h^2$$

(neo que (q-f))

Por tanto, el valor esperado de $V(\bar{X}_w)$ es igual a la varianza de la media muestral obtenida con afijación proporcional + un término positivo de penalización, por ser los m_h variables aleatorias.

Para muestras de gran tamaño*, esta penalización es despreciable y la postestratificación es casi tan precisa como la estratificación con afijación proporcional.

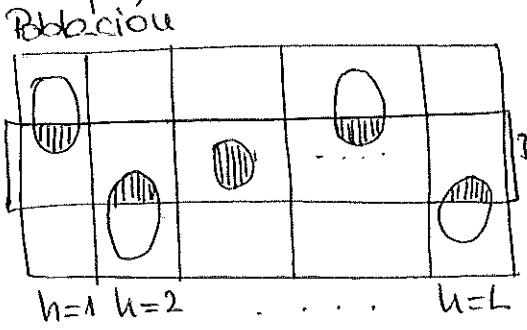
---

*Muestras grandes $\equiv n_h \geq 20$ unidades, $\forall h$.

5. APLICACIÓN AL MUESTREO EN SUBPOBLACIONES

Puede ocurrir que el dominio de estudio sea una subpoblación representada en todos los estratos. Por ejemplo: estratos: edades y subpoblaciones por sexo.

Esquemáticamente:



$N_{hj} \equiv$ nº unidades $<$ estrato h / dominio j

$\sum_{h=1}^{L} N_{hj} = N_j \equiv$ tamaño dominio.

Total: $X_{hj} = \sum_{i=1}^{N_{hj}} X_{hji}$

Medic: $\bar{X}_{hj} = \frac{1}{N_{hj}} \sum_{i=1}^{N_{hj}} X_{hji}$

Si se toma una muestra aleatoria estratificada de tamaño n_h en cada estrato:

Media muestral $\longrightarrow$ $\bar{x}_{hj} = \frac{1}{n_{hj}} \sum_{i=1}^{n_{hj}} x_{hji}$

MIRAR ATRÁS

El problema que tenemos es que los tamaños muestrales de cada estrato para el dominio, n_{hj}, son var. aleatorias, cuyo valor se desconoce hasta la observación de la muestra.

- Si los n_{hj} son conocidos, la media y el total se pueden estimar mediante sumas ponderadas:

$$\hat{\bar{X}}_j = \sum_{h=1}^{L} \frac{N_{hj}}{N_j} \bar{x}_{hj} = \sum_{h=1}^{L} W_{hj} \bar{x}_{hj}, \quad W_{hj} = \frac{N_{hj}}{N_j} \equiv \text{peso del estrato en el dominio}$$

$$\hat{X}_j = N_j \hat{\bar{X}}_j = \sum_{h=1}^{L} N_{hj} \bar{x}_{hj}$$

Como primero se observa la muestra y después se clasifican las unidades en dominios, se está aplicando una postestratificación de la muestra, luego la fórmula para la varianza del estimador se puede aplicar a este caso.

- Si los N_{hj} son desconocidos, se estima cada total de estrato del dominio y luego se suman para obtener el estimador del total del dominio j:

$$\hat{X}_j = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_{hj}} x_{hij}$$

Para calcular la varianza del estimador se acude a una var. auxiliar dicotómica con valor x_{hi} si u_{hi} pertenece al dominio:

$$x'_{hi} = \begin{cases} x_{hi} & \text{si } u_{hi} \in \text{dominio } j \\ 0 & \text{en otro caso} \end{cases}$$

de modo que

$$\sum_{i=1}^{n_{hj}} x_{hi} = \sum_{i=1}^{n_h} x'_{hi} \Rightarrow \hat{X}_j = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} x'_{hi} = \hat{X}_{st-j}$$

cuya varianza es:

$$V(\hat{X}_j) = \sum_{h=1}^{L} N_h^2 (1-f_h) \cdot \frac{S'^2_h}{n_h}, \quad S'_h \equiv \text{cuasiv. de } x' \text{ en estrato } h.$$

Para estimar la media poblacional del dominio j se requiere una estimación de N_j.

$$\hat{N}_j = \sum_{h=1}^{L} \frac{N_h}{n_h} n_{hj}, \text{ estimador insesgado de } N_j.$$