MUEST_ T6. 1. MUESTREO ESTRATIFICADO:

2. ESTIM. LINEALES, VARIANZAS y SUS ESTIMACIONES.

4. PRINCIPIOS BÁSICOS de la ESTRATIFICACIÓN.

5. CONSTRUCCIÓN y NÚMERO de ESTRATOS.

3. AFIJACIÓN de la MUESTRA con 1 CARACTERÍSTICA.

6. GANANCIA de PRECISIÓN

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, 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, $N_1 \dots N_h \dots N_L$

$\{u_{hi}\}$ | $h=1\dots L$ → nº estratos poblac.
$i=1\dots N_h$ → nº unidades por estrato

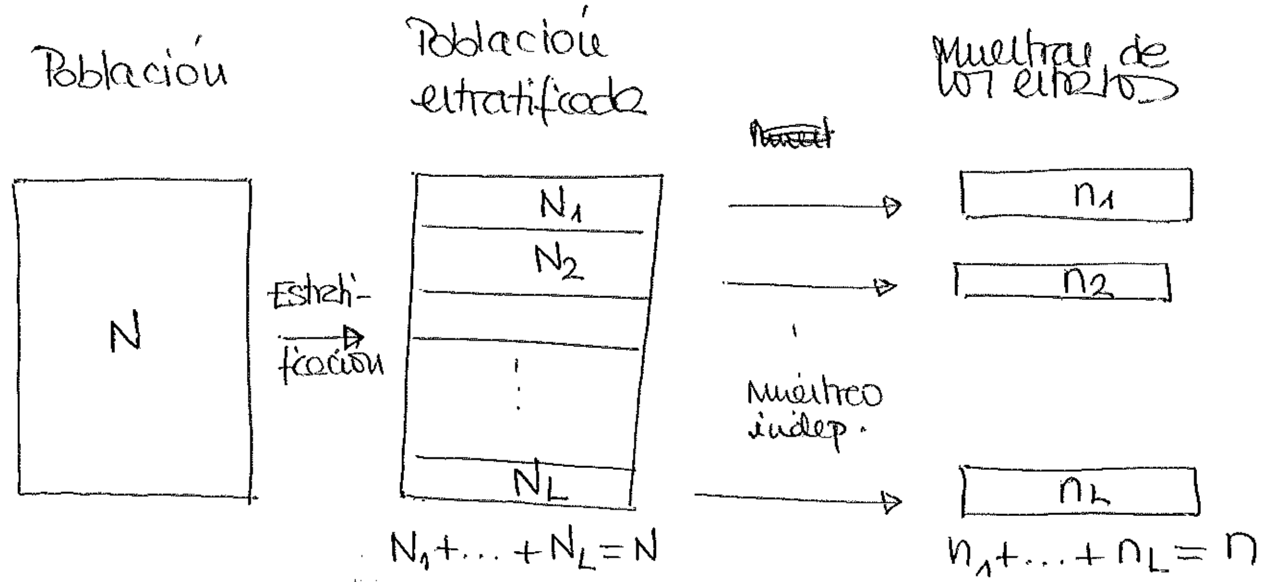
De cada estrato se selecciona una muestra aleatoria, el tipo de muestreo puede ser ≠ para cada estrato, de manera independiente.

$\{u_{hi}\}$ | $h=1\dots L$ → nº estrato
$i=1\dots n_h$ → nº unidades muestrales por estrato

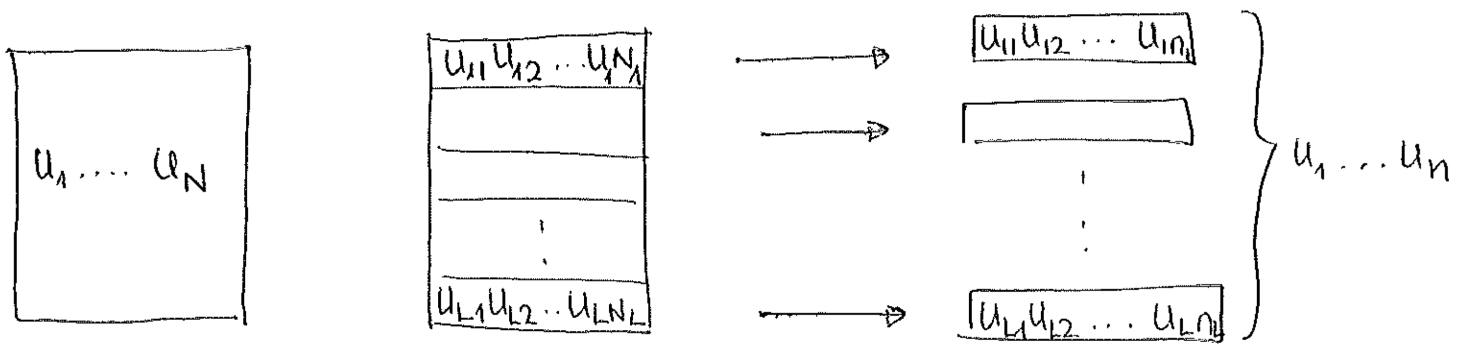
La homogeneidad / heterogeneidad se estudia a través de la varianza ≡ precisión del valor promedio.

Homogénea ≡ varianzas similares
Heterogénea ≡ varianzas 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

$i = 1 \dots N$

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

u_{hi}

$h = 1 \dots L$

$i = 1 \dots N_L$

$N_1 + \dots + N_L = N$

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

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

u_{hi}

$h = 1 \dots L$

$i = 1 \dots n_L$

$n_1 + \dots + n_L = n$

$\sum i = n_h$

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

Nótese que la estratificación es a nivel poblacional, y que en la muestra están 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 varianza en cada estrato sea menor que la varianza 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.

2 – ESTIMADORES LINEALES, VARIANZAS Y SUS ESTIMACIONES

Aunque se pueden utilizar distintos tipos de muestreo para cada estrato, vamos a suponer por comodidad que el muestreo se realiza de manera independiente y utilizando muestreo aleatorio con probabilidades iguales en cada estrato. Distinguimos los casos de SIN y CON reposición.

POBLACIÓN:

Dividimos la población en L estratos. Sobre cada unidad poblacional observamos la característica X. El parámetro a estimar es la suma poblacional de la característica.

$$\theta = \sum_{i=1}^{N} Y_i \longrightarrow \theta = \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi}$$

$h \to$ estrato
$i \to$ unidad dentro estrato

ESTIMADORES LINEALES:

Utilizamos como estimador lineal

$$\hat{\theta} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi} Y_{hi}$$

donde w_{hi} es el peso de muestreo de Y_{hi}, y representa el nº de unidades elementales de la unidad compuesta ó su ponderación ó importancia.

Para comprobar su insesgadez, nos ayudamos de la variable aleatoria auxiliar e_{hi}, $\equiv$ nº de veces que aparece u_{hi} en la muestra del estrato h.

Sin reposición:

muestral
lineal

Utilizamos como estimador insesgado la suma extendida a todos los estratos de los estimadores lineales insesgados de Horvitz y Thompson en cada estrato.

$$(SR) \longrightarrow \hat{\theta}_{st} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{Y_{hi}}{\pi_{hi}}$$, donde π_{hi} = probab. de que $u_{hi} \in S_h$.

Para demostrar la insesgadez, acudimos a e_{hi}.

$$e_{hi} = \begin{cases} 1 & \text{si } u_{hi} \in S_h \text{ con probab. } \pi_{hi} = \frac{n_h}{N_h} \\ 0 & \text{si } u_{hi} \notin S_h \text{ con probab. } 1-\pi_{hi} = 1 - \frac{n_h}{N_h} \end{cases}$$

probab. iguales

$e_{hi} \rightarrow Ber(\pi_{hi})$ con $E[e_{hi}] = \pi_{hi} = \frac{n_h}{N_h}$

$$E[\hat{\theta}] = E\left[\sum_{h=1}^{L}\sum_{i=1}^{n_h} Y_{hi} W_{hi}\right] = E\left[\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi} W_{hi} e_{hi}\right] =$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi} W_{hi} \underbrace{E[e_{hi}]}_{\pi_{hi}} = \theta \Longleftrightarrow W_{hi} = \frac{1}{\pi_{hi}}, \forall h,i.$$

La expresión del estimador se puede particularizar para los parámetros poblacionales más comunes:

Total poblacional: $\theta = X = \sum_{h,i} X_{hi} \longrightarrow \hat{X}_{st} = \sum_{h=1}^{L} \hat{X}_h$

El estimador del total es la suma de los estimadores del total en cada estrato.

$$\hat{X}_{st} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} \frac{X_{hi}}{\pi_{hi}} \overset{p.i.}{=} \sum_{h=1}^{L}\sum_{i=1}^{n_h} \frac{X_{hi}}{n_h/N_h} = \sum_{h=1}^{L} N_h \cdot \underbrace{\frac{1}{n_h}\sum_{i=1}^{n_h} X_{hi}}_{\hat{\bar{X}}_h = \bar{x}_h} =$$

$$= \sum_{h=1}^{L} N_h \cdot \bar{x}_h = \sum_{h=1}^{L} \hat{X}_h \Rightarrow \hat{X}_h = N_h \bar{x}_h$$

Media poblacional: Media ponderada de los estimadores de la media en cada estrato, siendo los coef. de ponderación los que marcan la importancia relativa de cada estrato en la poblac.

$$\theta = \overline{X} \Rightarrow Y_{hi} = \frac{X_{hi}}{N} \longrightarrow \hat{\overline{X}}_{st} = \sum_{h=1}^{L} W_h \overline{x}_h \quad / \; W_h = \frac{N_h}{N}.$$

$$\theta = \overline{X} \Rightarrow \hat{\overline{X}}_{st} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{X_{hi}}{N \pi_{hi}} = \sum_{h=1}^{L} \frac{1}{N} \sum_{i=1}^{n_h} \frac{X_{hi}}{n_h/N_h} =$$

$$= \sum_{h=1}^{L} \underbrace{\frac{N_h}{N}}_{W_h} \underbrace{\sum_{i=1}^{n_h} \frac{X_{hi}}{n_h}}_{\overline{x}_h} = \sum_{h=1}^{L} W_h \cdot \overline{x}_h \quad / \sum W_h = 1.$$

<u>Total de clase</u>: Suma de los estimadores del total de clase en cada estrato.

$$\theta = A \rightarrow Y_{hi} = A_{hi} \Rightarrow \hat{A}_{st} = \sum_{h=1}^{L} \hat{A}_h$$

$$\theta = A \Rightarrow \hat{A}_{st} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{A_{hi}}{n_h/N_h} =$$

$$= \sum_{h=1}^{L} N_h \underbrace{\sum_{i=1}^{n_h} \frac{A_{hi}}{n_h}}_{\hat{P}_h} = \sum_{h=1}^{L} N_h \cdot \hat{P}_h = \sum_{h=1}^{L} \hat{A}_h$$

<u>Proporción poblacional</u>: Media ponderada de los estimadores de la proporción en cada estrato, siendo los coef. de ponderación $W_h = \frac{N_h}{N}$ de suma unitaria

$$\theta = P \rightarrow Y_{hi} = \frac{A_{hi}}{N} \Rightarrow \hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{P}_h$$

$$\theta = P \Rightarrow \hat{P}_{st} = \sum_{h=1}^{L} \frac{1}{N} \sum_{i=1}^{n_h} \frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^{L} \frac{1}{N} \sum_{i=1}^{n_h} \frac{A_{hi}}{n_h/N_h} = \sum_{h=1}^{L} \frac{N_h}{N} \underbrace{\sum_{i=1}^{n_h} \frac{A_{hi}}{n_h}}_{\hat{P}_h}$$

$$= \sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h = \sum_{h=1}^{L} W_h \hat{P}_h$$

Con reposición:

La muestra estratificada de tamaño n se obtiene seleccionando n_h elementos $(h=1,\dots,L)$ de cada uno de los estratos de forma independiente, utilizando muestreo aleatorio simple con reposición.

Para el parámetro poblacional $\theta = \sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi}$, se utiliza como estimador la suma extendida a todos los estratos de los estimadores lineales insesgados de Hansen y Hurwitz en cada estrato.

$$\hat{\theta} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} \frac{Y_{hi}}{n_h P_{hi}} \qquad / \ \overset{\text{p. iguales}}{P_{hi} = \frac{1}{N_h}}$$

$P_{hi} \equiv$ probab. unitaria de que u_{hi} esté en la muestra del estrato h.

Al ser $P_{hi} = \frac{1}{N_h}$, se observa que $n_h P_{hi} = n_h/N_h = \pi_{hi}$, por lo que las expresiones del estimador con reposición y sin reposición coinciden, $\hat{\theta}_{HT} = \hat{\theta}_{HH}$, lo que se hace extensivo a los estimadores de los parámetros más comunes.

Para demostrar la insesgadez, acudimos a la v.a. auxiliar $e_{hi} \equiv$ nº veces que u_{hi} está en la muestra del estrato h.

$$e_{hi} \longrightarrow B(n_h, \overset{P_{hi}}{\frac{1}{N_h}}) \qquad E[e_{hi}] = \frac{n_h}{N_h}$$

$$E[\hat{\theta}] = E\left[\sum_{h=1}^{L}\sum_{i=1}^{n_h} Y_{hi} W_{hi}\right] = E\left[\sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi} W_{hi} e_{hi}\right] =$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{N_h} Y_{hi} W_{hi} \underbrace{E[e_{hi}]}_{= n_h/N_h} = \theta \iff W_{hi} = \frac{1}{\frac{n_h}{N_h}} = \frac{1}{n_h P_{hi}}$$

VARIANZAS de los ESTIMADORES

Al estudiar una muestra en lugar de toda la población, se produce un error de muestreo, medido con la desviación típica del estimador.
Un estimador será más preciso cuanto menor sea su varianza. En el caso de estimadores insesgados, la acuracidad y la precisión son lo mismo, pues el ECM del estimador coincide con su varianza.

Como el muestreo de cada estrato se realiza de forma independiente, la varianza del estimador es igual a la suma de las varianzas de cada estrato.

$$V(\hat{\theta}_{st}) = \sum_{h=1}^{L} V(\hat{\theta}_h)$$

Aunque las expresiones de los estimadores de Horvitz y Thompson coincidían con los estimadores de Hansen y Hurwitz, no ocurre así con sus varianzas, por lo que se hace ~~necesario~~ necesario distinguir los casos.

SIN reposición: Directamente para los parámetros

$$V(\hat{X}_{st}) = V\left(\sum_{h=1}^{L}\hat{X}_h\right) \overset{\text{indep.}}{=} \sum_{h=1}^{L} V(\hat{X}_h) = \sum_{h=1}^{L} N_h^2 (1-f_h)\cdot\frac{S_h^2}{n_h} \quad \text{(mirar T2)}$$

$$V(\hat{\bar{X}}_{st}) = V\left(\sum_{h=1}^{L} W_h \hat{\bar{X}}_h\right) = \sum_{h=1}^{L} W_h^2\, V(\hat{\bar{X}}_h) = \sum_{h=1}^{L} W_h^2\cdot(1-f_h)\frac{S_h^2}{n_h} \qquad W_h = \frac{N_h}{N}$$

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

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

donde: $f_h = \frac{n_h}{N_h}$; $S_h^2 \equiv$ cuasivarianza poblacional estrato h

$P_h Q_h \equiv$ varianza poblac. poblac. dicotóm. estrato h.

$(N_h-1)S_h^2 = N_h\sigma_h^2$

<u>CON reposición</u>: Tb. directamente

$$V(\hat{X}_{st}) = V\left(\sum_{h=1}^{L} \hat{X}_h\right) = \sum_{h=1}^{L} V(\hat{X}_h) = \sum_{h=1}^{L} N_h^2 \cdot \frac{\sigma_h^2}{n_h}$$

$$V(\hat{\bar{X}}_{st}) = V\left(\sum_{h=1}^{L} W_h \hat{\bar{X}}_h\right) = \sum_{h=1}^{L} W_h^2 \, V(\hat{\bar{X}}_h) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\sigma_h^2}{n_h}$$

$$V(\hat{A}_{st}) = V\left(\sum_{h=1}^{L} \hat{A}_h\right) = \sum_{h=1}^{L} V(\hat{A}_h) = \sum_{h=1}^{L} N_h^2 \cdot \frac{P_h Q_h}{n_h}$$

$$V(\hat{P}_{st}) = V\left(\sum_{h=1}^{L} W_h \hat{P}_h\right) = \sum_{h=1}^{L} W_h^2 \, V(\hat{P}_h) = \sum_{h=1}^{L} W_h^2 \cdot \frac{P_h Q_h}{n_h}$$

ESTIMACIÓN de las VARIANZAS

Los parámetros poblacionales de los que dependen las varianzas de los estimadores (S_h^2, $P_h Q_h$) suelen ser desconocidos, por lo que es preciso estimarlos.

Para características cuantitativas, la cuasivarianza muestral de cada estrato es un estimador insesgado de la cuasivarianza poblacional (SR) y de la varianza poblac. (CR)

Cuasiv. muestral $\longrightarrow \hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (X_{hi} - \bar{x}_h)^2 = \frac{n_h}{n_h - 1} \sigma_h^2$

Cuasiv. poblacional $\longrightarrow S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 = \frac{N_h}{N_h - 1} \sigma_h^2$.

En el caso de las proporciones, características cualitativas.

$\hat{S}_h^2 = \frac{n_h}{n_h - 1} \hat{P}_h \hat{Q}_h$ es un estimador insesgado de la cuasivarianza poblacional $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$. (SR) y de la varianza poblacional $\sigma_h^2 = P_h Q_h$ (CR)

Para obtener las expresiones de las estimaciones de las varianzas basta con sustituir cada parámetro poblacional por su estimador insesgado en la expresión de la varianza del estimador.

SIN reposición:

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 (1-f_h) \cdot \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{\bar{X}}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \cdot \frac{\hat{S}_h^2}{n_h}$$

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

$$\hat{V}(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 (1-f_h) \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}$$

CON reposición:

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 \cdot \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{\bar{X}}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \cdot \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}$$

$$\hat{V}(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}$$

3_ AFIJACIÓN de la MUESTRA con UNA CARACTERÍSTICA

Se llama afijación de la muestra al reparto del tamaño muestral entre todos los estratos.
Consiste en determinar el valor de n_h, $h=1\dots L$, de modo que $n_1+\dots+n_L=n$.
Los criterios de afijación más importantes son:

1_ Afijación UNIFORME o igual

Consiste en asignar el mismo número de unidades muestrales a cada estrato:

$$n_h = \frac{n}{L} = K \quad \forall h = 1\dots L \quad \longrightarrow f_h = \frac{K}{N_h}$$

aumentando ó disminuyendo en 1 unidad si n no es múltiplo de L.

Este tipo de afijación al dar ~~igual~~ la misma importancia a todos los estratos, favorece a los estratos pequeños y perjudica a los grandes en cuanto a precisión. Sólo es conveniente en estratos de tamaño similar.
No interviene en absoluto el hecho de que el muestreo sea CON o SIN reposición.

2_ Afijación PROPORCIONAL

Consiste en asignar a cada estrato un número de unidades muestrales proporcional a su tamaño

$$n_h = KN_h \Rightarrow n = \sum_h n_h = K\sum N_h = KN \Leftrightarrow K = \frac{n}{N}$$

$$n_h = \frac{n}{N}\cdot N_h \qquad f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

Directamente, se comprueba:

$f_h = K = f$.

$$W_h = \frac{N_h}{N} = \frac{n_h/K}{n/K} = \frac{n_h}{n}$$

En el caso de afijación proporcional, las expresiones de los estimadores del total poblacional y de la media poblacional quedan:

$$\hat{X}_{st} = \sum_{h=1}^{L} N_h \cdot \bar{x}_h = \sum_{h=1}^{L} \frac{n_h}{K} \bar{x}_h = \frac{1}{K} \sum_{h=1}^{L} \cancel{n_h} \cdot \frac{x_h}{\cancel{n_h}} = \frac{x}{K} = \frac{\text{Total muestral}}{\text{Fracción muestreo}}$$

$$\hat{\bar{X}}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \frac{n_h}{n} \bar{x}_h = \frac{1}{n} \sum_{h=1}^{L} \cancel{n_h} \cdot \frac{x_h}{\cancel{n_h}} = \frac{x}{n} = \frac{\text{Total muestral}}{\text{Tamaño muestra}}$$

- Las fracciones de muestreo en los estratos son iguales y coinciden con la fracción global de muestreo, igual a la de de proporcionalidad.
- Los coef. de ponderación W_h se obtienen exclusivamente a partir de la muestra. $W_h = \frac{N_h}{N} = \frac{n_h/K}{n/K} = \frac{n_h}{n}$
- Todas las unidades tienen la misma probabilidad de figurar en la muestra ⟶ muestras autoponderadas. ??

$$V(\bar{x}_{st}) = \frac{1-f}{n} \sum W_h S_h^2$$

3. Afijación de MÍNIMA VARIANZA (o de Neyman)

Consiste en determinar los tamaños muestrales de cada estrato para un tamaño muestral fijo n que minimicen la varianza del estimador.

En todos los casos, se resuelve aplicando el mét. de los multiplicadores de Lagrange al problema de optimización con restricciones:

$$\begin{cases} \min\limits_{n_h} \; V(\hat{\theta}_{st}) \\ \text{s.a.} \; \sum_{h=1}^{L} n_h = n \end{cases}$$

Al ser las expresiones de la varianza del estimador distintas para cada caso hay que resolver por separado, distinguiendo los casos SIN y CON reposición.

SIN reposición, para la media y la proporción

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

$$= \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} - \sum_{h=1}^{L} \underbrace{W_h}_{W_h=\frac{N_h}{N}} \cdot \frac{\cancel{n_h}}{\cancel{N_h}} \cdot \frac{S_h^2}{\cancel{n_h}} =$$

$$= \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} - \underbrace{\sum_{h=1}^{L} W_h \cdot \frac{S_h^2}{N}}_{\text{no depende de } n_h}$$

$$\left.\begin{array}{l} \min_{n_h} V(\bar{x}_{st}) \\ \text{s.a. } \sum_{h=1}^{L} n_h = n \end{array}\right\} \sim \begin{array}{l} \min_{n_h} \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} \\ \text{s.a. } \sum_{h=1}^{L} n_h = n \end{array}$$

1. Lagrangiano: $\phi(n_h, \lambda) = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} + \lambda\left(\sum_{h=1}^{L} n_h - n\right)$

2. $\left.\begin{array}{l} \frac{\partial \phi}{\partial n_h} = -W_h^2 \cdot \frac{S_h^2}{n_h^2} + \lambda = 0 \\ \text{para cada } h = 1 \ldots L \\ \frac{\partial \phi}{\partial \lambda} = \sum n_h - n = 0 \end{array}\right\} \Rightarrow \lambda = W_h^2 \cdot \frac{S_h^2}{n_h^2} = \left(\frac{N_h}{N}\cdot\frac{S_h}{n_h}\right)^2$

$$\sqrt{\lambda} = \frac{N_h}{N}\cdot\frac{S_h}{n_h}$$

$$N\sqrt{\lambda} = \frac{N_h S_h}{n_h} \Rightarrow \frac{n_h}{N_h S_h} = \frac{1}{N\sqrt{\lambda}} = \text{cte}$$

Desarrollando la igualdad para todo n_h y aplicando las propiedades de las proporciones:

$$\frac{n_h}{N_h S_h} = \frac{\sum n_h}{\sum N_h S_h} = \frac{n}{\sum N_h S_h} \Rightarrow \boxed{n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n \cdot \frac{W_h S_h}{\sum W_h S_h}}$$

- En el supuesto de que todos los estratos tengan la misma dispersión, $S_h = S$, la afijación mínima varianza coincide con la afijación proporcional.

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

- La afijación óptima resulta útil si hay grandes diferencias de variabilidad entre los estratos. En caso contrario, por su sencillez, se recomienda la afijación proporcional.

- El valor de la varianza mínima es:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

- Si se quiere la expresión de la varianza y la afijación para el estimador de la proporción basta sustituir en las fórmulas anteriores S_h^2 por $\frac{N_h(P_h Q_h)}{N_h - 1}$.

- En el caso del total poblacional y del total de clase, cambia la función objetivo, y aplicando multiplicadores de Lagrange se llega a la misma afijación que en el caso de la media, porque las funciones objetivo se diferencian en una cte $(N_h = W_h \cdot N)$.

- El valor de la varianza mínima para el estimador del total es

$$V(\hat{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} N_h S_h^2$$

CON reposición, para la media y la proporción

$$\left.\begin{array}{l} \text{MIN } V(\bar{x}_{st}) \\ \text{s.a. } \sum n_h = n \end{array}\right\} \text{, donde } V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\sigma_h^2}{n_h}$$

1_ Lagrangiano: $\phi(n_h, \lambda) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\sigma_h^2}{n_h} + \lambda\left(\sum n_h - n\right)$

2_ Derivar e igualar a cero:

$$\left.\begin{array}{l} \forall h \quad \frac{\partial \phi}{\partial n_h} = -W_h^2 \cdot \frac{\sigma_h^2}{n_h^2} + \lambda = 0 \\ \frac{\partial \phi}{\partial \lambda} = \sum n_h - n = 0 \end{array}\right\} \lambda = W_h^2 \cdot \frac{\sigma_h^2}{n_h^2} = \frac{N_h^2}{N^2} \cdot \frac{\sigma_h^2}{n_h^2}$$

$$\downarrow$$

$$\sqrt{\lambda} = W_h \cdot \frac{\sigma_h}{n_h} = \text{cte}$$

Aplicando las propiedades de las proporciones:

$$\frac{W_h \cdot \sigma_h}{n_h} = \frac{\sum W_h \sigma_h}{\sum n_h} = \frac{\sum W_h \sigma_h}{n} \Rightarrow n_h = n \cdot \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h} =$$

$$\boxed{n_h} = n \cdot \frac{\frac{N_h}{N}\sigma_h}{\sum_{h=1}^{L} \frac{N_h}{N}\sigma_h} = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}$$

- Los valores de n_h son proporcionales a los productos $N_h \cdot \sigma_h$. En el supuesto de que las dispersiones de los estratos coincidan, la afijación de mínima varianza coincide con la afijación proporcional.

- El valor de la varianza mínima es:
$$V(\bar{x}_{st}) = \frac{1}{n} \left(\sum_{h=1}^{L} W_h \sigma_h \right)^2$$

- Si se quiere la afijación de mínima varianza y la expresión de varianza mínima para el estimador de la proporción basta con sustituir σ_h^2 por $P_h Q_h$.

- En el caso del total poblacional, el tamaño muestral de cada estrato coincide con la media.

- ~~El valor de la varianza míni~~

4_ Afijación ÓPTIMA con costes

Consiste en determinar los tamaños muestrales para cada estrato, n_h, $h = 1 \dots L$, de forma que para un coste fijo C la varianza de los estimadores sea mínima.

El coste fijo C será la suma de los costes de cada unidad. $C = \sum_{h=1}^{L} c_h n_h$

Vuelve a ser un problema de optimización condicionado que se resuelve con los multiplicadores de Lagrange.

SIN reemplazamiento, para la media y la proporción

$$\left.\begin{array}{l} \text{MIN } V(\bar{x}_{st}) \\ \text{s.a. } \sum_{h=1}^{L} c_h n_h = C \end{array}\right\} \quad \text{donde } V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} - \underbrace{\sum_{h=1}^{L} W_h \cdot \frac{S_h^2}{N}}_{\text{no depende de } n_h}$$

por lo que $\text{MIN } V(\bar{x}_{st}) \sim \text{MIN}\left(\sum W_h^2 \frac{S_h^2}{n_h}\right)$

1_ Lagrangiano: $\phi(n_h, \lambda) = \sum_{h=1}^{L} W_h^2 \cdot \frac{S_h^2}{n_h} + \lambda\left(\sum_{h=1}^{L} c_h n_h - C\right)$

2_ Derivando e igualando a 0:

$$\left.\begin{array}{l} \frac{\partial \phi}{\partial n_h} = -W_h^2 \cdot \frac{S_h^2}{n_h^2} + \lambda c_h = 0 \\ \frac{\partial \phi}{\partial \lambda} = \sum c_h n_h - C = 0 \end{array}\right\} \quad \begin{array}{l} \lambda = \frac{W_h^2}{c_h} \cdot \frac{S_h^2}{n_h^2} = \frac{1}{c_h} \cdot \frac{N_h^2}{N^2} \cdot \frac{S_h^2}{n_h^2} \\ \sqrt{\lambda} = \frac{N_h}{N} \cdot \frac{S_h}{\sqrt{c_h} \cdot n_h} \end{array}$$

$$N\sqrt{\lambda} = \frac{N_h \cdot S_h / \sqrt{c_h}}{n_h} \Rightarrow$$

$$\Rightarrow \frac{n_h}{N_h S_h / \sqrt{c_h}} = \frac{1}{N\sqrt{\lambda}} = cte$$

Desarrollando la igualdad para todo h y aplicando las propiedades de las proporciones:

$$\frac{n_h}{N_h S_h / \sqrt{c_h}} = \frac{\sum n_h}{\sum N_h S_h / \sqrt{c_h}} = \frac{n}{\sum N_h S_h / \sqrt{c_h}} \Rightarrow$$

$$\Rightarrow n_h = n \cdot \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} = n \cdot \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}$$

_ Los valores de n_h son proporcionales a los productos $N_h S_h / \sqrt{c_h}$, y en el supuesto de que el coste ~~C~~

$$\boxed{- V_{min\,C}(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right) - \frac{1}{N} \sum_{h=1}^{L} W_h \cdot S_h^2}$$

sea constante en todos los estratos, $c_h = k, \forall h$, entonces la afijación óptima coincide con la de mínima varianza. Si además, las dispersiones de los estratos son iguales, la afijación óptima coincidirá con la afijación proporcional.

~~- El valor de la varianza mínima en este caso es:~~

- Si se quiere la afijación óptima para el estimador de la proporción basta con sustituir S_h^2 por $\frac{N_h}{N_h-1} \cdot P_h Q_h$

- En el caso del total poblacional, el tamaño muestral de cada estrato coincide con el obtenido para la media, al diferenciarse $V(\hat{X}_{st})$ de $V(\hat{\bar{X}}_{st})$ en una cte de proporcionalidad.

~~En~~

CON reemplazamiento, para la media y la proporción

$$\left.\begin{array}{l} \text{MIN } V(\bar{x}_{st}) \\ \text{s.a. } \sum c_h n_h = C \end{array}\right\} \text{ ~~donde } V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2~~$$

Aplicando multiplicadores de Lagrange se llega a:

$$n_h = n \cdot \frac{N_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h \sigma_h / \sqrt{c_h}}$$

- Se observa que n_h es proporcional a $N_h\sigma_h/\sqrt{c_h}$. Para coste de la afijación óptima coincide con la mínima. Si las varianzas coinciden además, ~~pt sea~~ igual que la propor.

- El valor de la varianza es $\frac{1}{n}\left(\sum W_h \sigma_h \text{ POR } \sqrt{c_h}\right)\left(\sum W_h \sigma_h / \sqrt{c_h}\right)$.

- Para el estimador del total, n_h coincide.

No está en Cobar Pérez
↓ rom/xc

4_ CONSTRUCCIÓN y NÚMERO de ESTRATOS

No es fácil dar reglas fijas con respecto al número de estratos y sus límites, pues dependen en cada caso de
- la estructura de la población
- la información disponible
- los objetivos concretos en cada caso.

Para que el muestreo estratificado dé buenos resultados y sus estimadores sean más precisos, los estratos deben estar constituidos por unidades lo más homogéneas posibles, y tienen que ser lo más heterogéneos posibles entre sí.

En general, la precisión aumenta cuanto mayor sea el número de estratos bien construidos, pero se complican los cálculos y la presencia de cada estrato ~~disminuye~~ en la muestra disminuye.

En todo caso, los estratos deben constituir una partición de la población: cada unidad poblacional debe pertenecer a uno y sólo uno de los estratos, y $\sum_{h=1} N_h = N$.

~~Deben de~~

A veces el orden en que aparecen las unidades en el marco implica una estratificación. Otras veces, la necesidad de efectuar estimaciones pequeñas. Generalmente la estratificación se hace por razones de eficiencia (mayor precisión y menor coste).

CONSTRUCCIÓN de los estratos

Cuando se estudia una sola característica de la población, la mejor ~~a elegir es~~ para estratificar es la distribución de frecuencias, pero suele ser desconocida. La siguiente mejor opción es la distrib. de frecuencias de otra variable altamente correlacionada con X.

Conocido el número de estratos L, el mínimo y el máximo valor de la característica en la población, se trata de encontrar los límites intermedios entre los estratos que minimicen la varianza del estimador.

En el caso concreto de la media muestral obtenida SIN reposición, su varianza es:

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

~~Para~~ Si la muestra ha sido obtenida con afijación óptima de Neyman, suponiendo la fracción de muestreo en cada estrato despreciable:

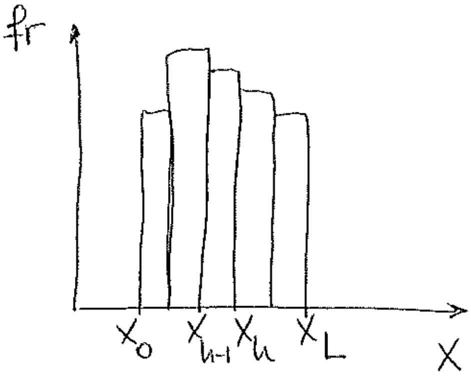
$$\left.\begin{array}{l} n_h = n \dfrac{W_h S_h}{\sum W_h S_h} \\ f_h = \dfrac{n_h}{N_h} \simeq 0 \end{array}\right| \longrightarrow V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2$$

por lo que minimizar la varianza del estimador equivale a minimizar $\sum_{h=1}^{L} W_h S_h$.

Dalenius y Hodges (1959) idearon un método rápido y aproximado basado en la cantidad

$$Z(x) = \int_{x_0}^{x} \sqrt{fr(t)}\, dt$$, donde $fr(x)$ = función de frecuencias de X

Según estos autores, si los estratos son muy numerosos y estrechos, la función de frecuencias debería ser aproximadamente constante (uniforme) dentro de cada estrato:



$$W_h = \int_{X_{h-1}}^{X_h} fr(t)\,dt \simeq fr_h (X_h - X_{h-1})$$

$$S_h \simeq \frac{1}{\sqrt{12}}(X_h - X_{h-1}) \quad , \text{ sp } N_h - 1 \simeq N_h$$

$$Z(X_h) - Z(X_{h-1}) = \int_{X_{h-1}}^{X_h} \sqrt{fr(t)}\,dt \simeq \sqrt{fr_h}\,(X_h - X_{h-1})$$

Sustituyendo estas aproximaciones en la aproximación a minimizar resulta:

$$\sum_{h=1}^{L} W_h S_h \simeq \sum_{h=1}^{L} fr_h \frac{(X_h - X_{h-1})^2}{\sqrt{12}} \simeq \sum_{h=1}^{L} \frac{(Z(X_h) - Z(X_{h-1}))^2}{\sqrt{12}}$$

que se minimiza al hacer $Z(X_h) - Z(X_{h-1})$ constante, por ser $(Z(X_L) - Z(X_0))$ una cantidad fija.

Así, la regla es elegir los X_h de modo que formen intervalos iguales en la escala $Z(X)\sqrt{fr(x)}$ acumulada.

Es importante puntualizar que resulta poco realista basar la estratificación en los valores de la caract. X.

En la práctica, se utiliza alguna variable altamente correlacionada con X como variable estratificadora a partir de la regresión lineal de X sobre ella.

NÚMERO de ESTRATOS

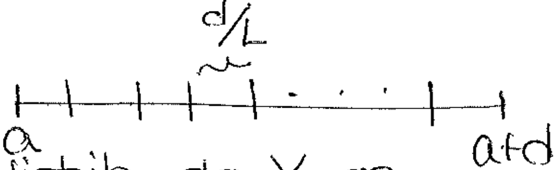
Si los estratos se construyen con los valores de X, el caso más sencillo es que X siga una distrib. Uniforme en el intervalo $(a, a+d)$ de modo que:

$$\sigma_x^2 = \frac{d^2}{12} \Rightarrow S_x^2 = \frac{d^2}{12} \quad , \text{sp. } N-1 \simeq N \text{ (N grande)}$$

Con una m.a.s. sin estratificar, la varianza del estimador media muestral sería (SIN reposición):

$$V(\bar{x}) = (1-f)\frac{S_x^2}{n} = \frac{d^2}{12n} \quad , \text{sp N grande} \Rightarrow 1-f \simeq 1 .$$

Si se divide la población en L estratos de igual tamaño, la cuasivarianza dentro de cada uno de ellos será: $S_{X_h}^2 = \frac{d^2}{12L^2}$



pues dentro de cada estrato la distrib. de X es uniforme en un intervalo de amplitud $\frac{d}{L}$.

Por tanto

$$\left.\begin{aligned} W_h &= \frac{d/L}{d} = \frac{1}{L} \\ n_h &= \frac{n}{L} \end{aligned}\right\} \Rightarrow V(\bar{x}_{st}) \simeq \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{1}{L^2} \cdot \frac{d^2}{12L^2} \cdot \frac{L}{n} =$$

$$= \frac{d^2}{12L^3 n} = \frac{V(\bar{x})}{L^3} \quad \leftarrow \text{cTx } L^2$$

Entonces la varianza de $\bar{x}_{st}$ decrece al aumentar el nº de estratos ⇒ mejora la precisión.

Cuando se utiliza una var. estratificadora, el aumento del nº de estratos sólo es beneficioso si las var. están altamente correlacionadas.

En cualquier caso, hay que tener en cuenta que el aumento del número de estratos implica un aumento del coste, salvo que se reduzca el tamaño de la muestra, lo que no siempre compensa.

5_ PRINCIPIOS BÁSICOS de la ESTRATIFICACIÓN

Si la estratificación es adecuada, debería obtenerse una varianza más pequeña para el estimador de la que se obtendría con un muestreo aleatorio simple.

Para ver qué se entiende por "estratificación adecuada", se acude a la descomposición de la Variabilidad Total de la población en la suma de la variabilidad dentro de cada estrato y la variabilidad entre los estratos.

$$\cancel{VT} = \sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X})^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h+\bar{X}_h-\bar{X})^2 =$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)^2 + \sum_{h=1}^{L}\sum_{i=1}^{N_h}(\bar{X}_h-\bar{X})^2 + 2\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)(\bar{X}_h-\bar{X})$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)^2 + \sum_{h=1}^{L}N_h(\bar{X}_h-\bar{X})^2 + 2\sum_{h=1}^{L}(\bar{X}_h-\bar{X})\underbrace{(X_h - \underbrace{N_h\bar{X}_h}_{X_h})}_{=0}$$

luego:

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X})^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}(X_{hi}-\bar{X}_h)^2 + \sum_{h=1}^{L}N_h(\bar{X}_h-\bar{X})^2$$

$$SCT = SCD + SCE$$

$$\cancel{VT = VD + VE}$$

que ~~también~~ se puede expresar en términos de las cuasivarianzas poblacionales y de las varianzas poblac:

$$VT = VD + VE$$

$$(N-1)S^2_T = (N-L)S^2_{\underset{w}{D}} + (L-1)S^2_{\underset{b}{E}}$$

Partiendo de la expresión de la Varianza de la media muestral en muestreo aleatorio simple sin reposición, y sustituyendo la cuasivarianza poblacional por la descomposición antes obtenida se llega a una expresión de donde se deduce el Principio Fundamental de la Estratificación.

$$V(\bar{X}) = (1-f)\frac{S^2}{n} = \frac{1-f}{n}\left[\underbrace{\frac{1}{N-1}\sum_h\sum_i (X_{hi}-\bar{X}_h)^2}_{\downarrow} + \frac{1}{N-1}\sum_h N_h(\bar{X}_h-\bar{X})^2\right] =$$

$$\frac{1}{N-1}\cdot\sum_{h=1}^{L}(N_h-1)S_h^2 = \underbrace{\frac{1}{N-1}\sum_h N_h S_h^2}_{\frac{N}{N-1}\sum_h \frac{N_h}{N}S_h^2} - \frac{1}{N-1}\sum_{h=1}^{L}S_h^2$$

$$= \frac{N-1+1}{N-1}\sum_h \frac{N_h}{N}S_h^2 - \frac{1}{N-1}\sum_h S_h^2 = \sum_h \frac{N_h}{N}S_h^2 + \frac{1}{N-1}\sum_h \frac{N_h}{N}S_h^2 - \frac{1}{N-1}\sum S_h^2$$

$$= \sum_h \frac{N_h}{N}S_h^2 + \frac{1}{N-1}\sum_h \underbrace{\left(\frac{N_h}{N}-1\right)}_{\frac{N_h-N}{N}}S_h^2 = \sum_{h=1}^{L}\frac{N_h}{N}S_h^2 - \frac{1}{N(N-1)}\sum_{h=1}^{L}(N-N_h)S_h^2 .$$

por lo que

$$V(\bar{X}) = \frac{1-f}{n}\left[\sum_{h=1}^{L}\overset{W_h}{\frac{N_h}{N}}S_h^2 - \frac{1}{N(N-1)}\sum_{h=1}^{L}(N-N_h)S_h^2 + \frac{1}{N-1}\sum_{h=1}^{L}N_h(\bar{X}_h-\bar{X})^2\right] =$$

Para simplificar esta expresión, utilizando afijación proporcional:

$$\left.\begin{array}{l} W_h = \dfrac{N_h}{N} \overset{\text{prop}}{=} \dfrac{n_h}{n} \\ f_h = \dfrac{n_h}{N_h} = \dfrac{n}{N} \end{array}\right\} \longrightarrow V_{prop}(\bar{X}_{st}) = \sum_{h=1}^{L} W_h \frac{S_h^2}{n}(1-f)$$

luego

$$\cancel{V(\bar{X}) = V_{prop}(\bar{X}_{st}) + \frac{1-f}{n(N-1)}\left[\frac{1}{N-1}\sum_{h=1}^{L}N_h(\bar{X}_h-\bar{X})^2 - \right.}$$

$$\underbrace{V(\bar{X})}_{V_{mas}} = \underbrace{V_{prop}(\bar{X}_{st})}_{V_{\text{estratif. af. prop}}} + \underbrace{\frac{1-f}{n(N-1)}\left[\underbrace{\sum_{h=1}^{L}N_h(\bar{X}_h-\bar{X})^2}_{\text{GRANDE}} - \underbrace{\frac{1}{N}\sum_h (N-N_h)S_h^2}_{\text{PEQUEÑO}}\right]}_{\text{cantidad que puede ser positiva o negativa, por eso no siempre es mejor}}$$

de donde se deduce el Principio Fundamental de la Estratificación:

"La ganancia en precisión con la estratificación será tanto mayor cuanto mayor sea el término entre corchetes, es decir, los estratos deben construirse de manera que:

- Las medias de los estratos difieran lo más posible
- Dentro de cada estrato exista una gran homogeneidad entre sus unidades.

6. GANANCIA de PRECISIÓN

En el supuesto de poblaciones grandes con estratos grandes ($N-1 \approx N$ y $N_h - 1 \approx N_h$) podemos descomponer la cuasivarianza poblacional en

$$S^2 \cong \sum_h W_h S_h^2 + \sum_h W_h (\bar{X}_h - \bar{X})^2$$

por lo que:

$$\underbrace{(1-f)\frac{S^2}{n}}_{V_{mas}(\bar{x})} = \underbrace{\frac{1-f}{n}\sum W_h S_h^2}_{V_{Af.Prop}(\bar{x}_{st})} + \underbrace{\frac{1-f}{n}\sum W_h (\bar{X}_h - \bar{X})^2}_{\geq 0}$$

Acabamos de demostrar que el muestreo estratificado con afijación proporcional es más preciso que el muestreo aleatorio simple (SR).

Serán iguales cuando las medias de los estratos sean iguales (pobl. homogénea), por lo que la ganancia en precisión será mayor cuanto más distintas entre sí sean las medias de los estratos.

Ahora vamos a comparar las precisiones de la afijación proporcional y la de mínima varianza:

$$V_{AfProp}(\bar{x}_{st}) - V_{MinVar}(\bar{x}_{st}) = \underbrace{\frac{1-f}{n}}_{\frac{1}{n}-\frac{1}{N}} \sum_{h=1}^{L} W_h S_h^2 - \left[\frac{1}{n}(\Sigma W_h S_h)^2 - \frac{1}{N}\Sigma W_h S_h^2\right]$$

$$= \frac{1}{n}\left[\Sigma W_h S_h^2 - (\underbrace{\Sigma W_h S_h}_{\bar{S}})^2\right] + \frac{1}{N}\left[\underbrace{\Sigma W_h S_h^2 - \Sigma W_h S_h^2}_{=0}\right] =$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 \geq 0$$

luego el muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo con afijación proporcional (se da la igualdad cuando las desviaciones típicas de todos los estratos son iguales).

La ganancia en precisión del muestreo estratificado con af. MV ~~será cua~~ respecto al m.e. af. prop. será tanto mayor cuanto más ~~distintos~~ heterogéneos entre sí sean los estratos.

En general: $\boxed{\text{Af. Min. Var} \gg \text{Af. Proporc} \gg \text{MAS}}$

- Los resultados son iguales para los estimadores de X, A y P, ya que las varianzas difieren en constantes de proporcionalidad.

- En el caso de muestreo CON reemplazamiento se llega a la misma conclusión, a partir de la expresión de la varianza poblacional.