Organización de las Naciones Unidas para la Alimentación y la Agricultura

Banco Interamericano de Desarrollo

FLACSO CHILE FACULTAD LATINOAMERICANA DE CIENCIAS SOCIALES

POLITÉCNICA UNIVERSIDAD POLITÉCNICA DE MADRID

CURSO DE MUESTREO CON MARCOS MÚLTIPLES

TEMA 3.3

ESTIMACIÓN CON MARCOS SENCILLOS. MUESTREO ESTRATIFICADO SIN INFORMACIÓN AUXILIAR.

Luis Ambrosio Flores

2017

TABLA DE CONTENIDOS

1. Introducción

En el muestreo estratificado, la población de N elementos se divide en L subpoblaciones, llamadas estratos, de manera que los elementos de un mismo estrato sean lo más homogéneos posible entre sí, respecto del carácter estudiado, y lo más heterogéneos posible de los elementos de los restantes estratos. La subdivisión ha de hacerse sin solapamientos, de manera que se verifique $\sum_{h=1}^{L} N_h = N$, donde N_h denota el número de elementos del estrato $h; \{h = 1,2,....,L\}$.

Sea Y_{hi} el valor de la variable en estudio -*Y*- asociado al elemento $i-ésimo; \{i = 1,2,....,N_h\}$ del estrato $h; \{h = 1,2,....,L\}$. Nos interesamos en estimar la media de la variable *Y* en la población:

$$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h \qquad [3.3.1]$$

donde, $\bar{Y}_h$ es la media de la variable *Y* en el estrato h - $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi}$ - y $W_h = \frac{N_h}{N}$. El total de la variable es $N\bar{Y}$. Nótese que si la variable *Y* es cualitativa, de modo que toma el valor 1 si el elemento *i-ésimo* posee un determinado atributo y el valor 0 en otro caso, entonces $\bar{Y}$ coincide con la proporción de elementos que presentan dicho atributo y $N\bar{Y}$ coincide con el número total de elementos que poseen dicho atributo, en la población.

Para estimar $\bar{Y}$ se selecciona una muestra dentro de cada estrato, de manera independiente. Sea n_h el número de elementos de la muestra a seleccionar en el estrato $h; \{h = 1,2,....,L\}$. Los n_h elementos de la muestra pueden ser seleccionados con o sin reposición, y con probabilidades iguales o desiguales y el procedimiento puede diferir de uno a otro estrato. En este tema consideraremos el caso en el que procedimiento de selección de la muestra dentro de estratos es "aleatorio simple" (ver tema 3.1, epígrafe 1). En ese caso, todos los resultados establecidos en los temas 3.1 y 3.2 relativos a la estimación de medias, totales y proporciones, se tienen para cada estrato, considerado en sí mismo como una población, independiente de los restantes estratos.

2. Estimador de la media y estimador del total

El estimador

Para estimar la media $\bar{Y}$ se considera el estimador:

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_h \qquad [3.3.2]$$

donde $\hat{\bar{Y}}_h$ es el estimador de $\bar{Y}_h$, definido a partir de la expresión [3.1.2] aplicada al estrato $h; \{h = 1,2,....,L\}$, considerado como una población en sí mismo:

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

El definido en [3.1.2] es un estimador insesgado de la media de la población a la que se aplica, de modo que $\hat{\bar{Y}}_h$ es un estimador insesgado de $\bar{Y}_h$ y, por tanto, $\hat{\bar{Y}}_{str}$ es un estimador insesgado de $\bar{Y}$.

El estimador del total Y:

$$\hat{Y}_{str} = N \hat{\bar{Y}}_{str}$$

es insesgado por ser $\hat{\bar{Y}}_{str}$ un estimador insesgado de $\bar{Y}$

Varianza del estimador

Puesto que la muestra se selecciona dentro de cada estrato, de forma independiente de uno a otro estrato, las L variables $\left\{\hat{\bar{Y}}_h; h = 1,2,....., L\right\}$ se distribuyen en el muestreo de forma independiente y, por tanto la varianza del estimador de la media es:

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{L} W_h^2 V(\hat{\bar{Y}}_h) \qquad [3.3.3]$$

donde $V(\hat{\bar{Y}}_h)$ es la varianza del estimador $\hat{\bar{Y}}_h$ y se tiene a partir de la expresión [3.1.3], especificada para el estrato $h; \{h = 1,2,....,L\}$, considerado como una población en sí mismo:

$$V(\hat{\bar{Y}}_h) = (1 - f_h) \frac{S_h^2}{n_h}$$

donde:

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

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$$

de modo que,

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h} \qquad [3.3.4]$$

La varianza del estimador del total es:

$V(\hat{Y}_{str}) = N^2 V(\hat{\bar{Y}}_{str})$

Estimador de la varianza del estimador

Un estimador insesgado de $V(\hat{\bar{Y}}_h)$ se tiene a partir de la expresión [3.1.4], especificada para el estrato $h; \{h = 1,2,....,L\}$:

$$\hat{V}(\hat{\bar{Y}}_h) = (1 - f_s) \frac{\hat{S}_h^2}{n_h} \qquad [3.3.5]$$

donde,

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

$$\hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \bar{y}_h)^2$$

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

Así que un estimador insesgado de la varianza del estimador de la media -$V(\hat{\bar{Y}}_{str})$- es:

$$\hat{V}(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\hat{\bar{Y}}_h) \qquad [3.3.6]$$

donde $\hat{V}(\hat{\bar{Y}}_h)$ es el definido en [3.3.5].

Un estimador insesgado de la varianza del estimador del total -$V(\hat{Y}_{str})$- es:

$\hat{V}(\hat{Y}_{str}) = N^2 \hat{V}(\hat{\bar{Y}}_{str})$

3. Reparto proporcional y reparto óptimo. Varianzas

Se consideran dos criterios de reparto de la muestra total de tamaño n, entre los L estratos: (i) reparto proporcional al tamaño de los estratos y (ii) reparto óptimo con el criterio de minimizar la varianza del estimador.

Reparto proporcional

El número n_h de elementos a asignar al estrato $h; \{h = 1,2,....,L\}$ se determina de modo que sea proporcional al tamaño N_h del estrato:

$$n_h = \frac{N_h}{N} n = W_h n \qquad [3.3.7]$$

Con este tipo de reparto, la fracción de muestreo dentro de estratos es la misma e igual a la de la población:

$$f_h = f; \forall h\{h = 1,2,....,L\}; \quad f_h = \frac{n_h}{N_h}; \quad f = \frac{n}{N}$$

La varianza del estimador de la media correspondiente al reparto proporcional se tiene a partir de y [3.3.4], tras sustituir n_h por el definido en [3.3.7]:

$$V_{prop.}(\hat{\bar{Y}}_{str}) = (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 \qquad [3.3.8]$$

Y la varianza del estimador del total con reparto proporcional es:

$$V_{prop.}(\hat{Y}_{str}) = N^2(1-f)\frac{1}{n}\sum_{h=1}^{L} W_h S_h^2$$

Un estimador insesgado de la varianza del estimador de la media -$V_{prop.}(\hat{\bar{Y}}_{str})$- es,

$$\hat{V}_{prop.}(\hat{\bar{Y}}_{str}) = (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h \hat{S}_h^2 \qquad [3.3.9]$$

donde $\hat{S}_h^2$ es el definido en [3.3.5] .

Un estimador insesgado de la varianza del estimador del total -$V_{prop.}(\hat{Y}_{str})$- es,

$$\hat{V}_{prop.}(\hat{Y}_{str}) = N^2 \hat{V}_{prop.}(\hat{\bar{Y}}_{str})$$

Reparto óptimo

La varianza del estimador de la media en [3.3.4] se puede expresar como la suma de dos componentes, una - $V = V(\{n_h\})$ - dependiente de los tamaños de muestra $\{n_h\}; \{h = 1,2,....,L\}$ y la otra -V_0- independiente:

$$V(\hat{\bar{Y}}_{str}) = V_0 + V \qquad [3.3.10]$$

donde,

$$V = \sum_{h=1}^{L} \frac{V_h^2}{n_h}$$
$$V_0 = \sum_{h=1}^{L} \frac{V_h^2}{N_h}$$
$$V_h = W_h S_h$$

Sea c_h el coste unitario (por unidad de muestreo) de observar la variable en estudio en el estrato $h;\{h=1,2,....,L\}$. El coste total de la estimación, $C(\{n_h\})$, puede especificarse como la suma de un coste fijo -C_0-, independiente del tamaño de la muestra a observar, y de un coste variable con el tamaño de la muestra -$C=\sum_{h=1}^{L} c_h n_h$-:

$$C(\{n_h\}) = C_0 + C \qquad [3.3.11]$$

El número n_h óptimo de elementos a asignar al estrato $h;\{h=1,2,....,L\}$ es el que minimiza,

$V(\hat{\bar{Y}}_{str})(C(\{n_h\}) = (V_0 + V)(C_0 + C)$

y coincide con el que minimiza a VC cuando se fija la varianza del estimador -$V = V(\hat{\bar{Y}}_{str}) - V_0 = V_f$- o el coste de la estimación -$C = C(\{n_h\}) - C_0 = C_f$-. El n_h óptimo, cuando $C = C_f$, viene dado por la siguiente expresión (ver Apéndice A, expresión [3.3.A.6]):

$$n_h = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} C_f \qquad [3.3.12]$$

El VC mínimo que correspondiente a ese óptimo n_h es, por sustitución de [3.3.12] en [3.3.10] y [3.3.11]:

$$VC = \left[\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right]^2 \qquad [3.3.13]$$

Fijado el coste $C = C_f$, la V mínima es :

$$V = \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c_h})^2}{C_f} \qquad [3.3.14]$$

y, de [3.3.10], la $V(\hat{\bar{Y}}_{str})$ mínima es:

$$V_{opt.}(\hat{\bar{Y}}_{str}) = \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c_h})^2}{C_f} - \sum_{h=1}^{L} \frac{(W_h S_h)^2}{N_h} \qquad [3.3.15]$$

y fijada la varianza $V = V_f$, la C mínima es :

$$C = \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c_h})^2}{V_f} \qquad [3.3.16]$$

y el coste mínimo de las estimaciones es, por sustitución en [3.3.11]:

$$C\{n_h\} = C_0 + \frac{(\sum_{h=1}^{L} W_h S_h \sqrt{c_h})^2}{V_f}$$

Para el estimador del total , la varianza $V(\hat{Y}_{str})$ mínima es:

$$V_{opt.}(\hat{Y}_{str}) = N^2 V_{opt.}(\hat{\bar{Y}}_{str})$$

Un caso particular importante es cuando el coste unitario de observación de los elementos de la muestra, es igual en todos los estratos: $c_h = c; \forall h$, de modo que $C = \sum_{h=1}^{L} c_h n_h = c\sum_{h=1}^{L} n_h = cn$. En ese caso, [3.3.12] se reduce a

$$n_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} n \qquad [3.3.17]$$

y en [3.3.15] se tiene $C_f = cn$ de modo que la varianza se reduce a se reduce a:

$$V_{opt.}(\hat{\bar{Y}}_{str}) = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{n} - \sum_{h=1}^{L} \frac{(W_h S_h)^2}{N_h} \qquad [3.3.18]$$

Para el estimador del total se tiene,

$$V_{opt.}(\hat{Y}_{str}) = N^2 V_{opt.}(\hat{\bar{Y}}_{str})$$

Un estimador insesgado de la varianza del estimador de la media con reparto óptimo -$V_{opt.}(\hat{\bar{Y}}_{str})$- es,

$$\hat{V}_{opt.}(\hat{\bar{Y}}_{str}) = \frac{(\sum_{h=1}^{L} W_h \hat{S}_h)^2}{n} - \sum_{h=1}^{L} \frac{(W_h \hat{S}_h)^2}{N_h} \qquad [3.3.19]$$

donde $\hat{S}_h^2$ es el definido en [3.3.5], y un estimador insesgado de la varianza del estimador del total -$V_{opt.}(\hat{Y}_{str})$- es,

$$\hat{V}_{opt.}(\hat{Y}_{str}) = N^2 \hat{V}_{opt.}(\hat{\bar{Y}}_{str})$$

4. Eficiencia relativa del muestreo estratificado respecto del muestreo aleatorio simple. Estimación a partir de la muestra.

La eficiencia relativa entre dos planes de muestreo (muestreo aleatorio estratificado respecto del muestreo aleatorio simple) se define como el cociente entre las inversas de las varianzas del estimador correspondientes a uno y otro plan de muestreo:

$$ER_{str/r} = \frac{V(\hat{\bar{Y}}_r)}{V(\hat{\bar{Y}}_{str})} = \frac{V(\hat{Y}_r)}{V(\hat{Y}_{str})} \qquad [3.3.20]$$

Obsérvese que la eficiencia relativa para la estimación de la media y para la estimación del total coinciden.

a) Reparto proporcional

La varianza del estimador con muestreo aleatorio simple puede ser descompuesta de modo que una de sus componentes sea la varianza del estimador con muestreo aleatorio estratificado y reparto proporcional. En efecto, se verifica

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y})^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h)^2 + \sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2 \qquad [3.3.21]$$

donde,

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}Y_{hi}$$

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h}Y_{hi}$$

La expresión [3.3.21] puede escribirse en términos de las varianzas:

$$(N-1)S^2 = \sum_{h=1}^{L}(N_h-1)S_h^2 + \sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2 \qquad [3.3.22]$$

donde,

$$S^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y})^2$$

$$S_h^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(Y_{hi}-\bar{Y}_h)^2$$

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h}Y_{hi}$$

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h}Y_{hi}$$

Si el tamaño de los estratos es grande $\forall h$, de modo que se verifica $\frac{1}{N_h} \cong 0$, esto es, $N_h - 1 \cong N_h$ y $N - 1 \cong N$ entonces, de [3.3.22], se tiene:

$$S^2 \cong \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h (\overline{Y}_h - \overline{Y})^2 \qquad [3.3.23]$$

donde,

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

La varianza del estimador con muestreo aleatorio simple -$V_r(\hat{\overline{Y}})$- , definida en [3.1.3], se puede escribir de la siguiente forma, tras sustituir S^2 de [3.3.23]:

$$V_r(\hat{\overline{Y}}) \cong (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 + (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h (\overline{Y}_h - \overline{Y})^2$$

En esta última expresión, el primer sumando del segundo miembro es la varianza del estimador con muestreo aleatorio estratificado y reparto proporcional, definida en [3.3.8], de modo que :

$$V_r(\hat{\overline{Y}}) \cong V_{prop.}(\hat{\overline{Y}}_{str}) + (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h (\overline{Y}_h - \overline{Y})^2 \qquad [3.3.24]$$

Y ,de [3.3.18], la eficiencia relativa es:

$$ER_{str\_prop/r} \cong 1 + (1-f)\frac{1}{n}\frac{\sum_{h=1}^{L} W_h (\overline{Y}_h - \overline{Y})^2}{V_{prop.}(\hat{\overline{Y}}_{str})} \qquad [3.3.25]$$

Cabe destacar los siguientes resultados que se derivan de [3.3.25]:

- El muestreo aleatorio estratificado, con reparto proporcional, es más eficiente que el muestreo aleatorio simple: $ER_{str\_prop/r} \geq 1$. La igualdad se da sólo cuando: $\overline{Y}_h = \overline{Y}\ \forall h$.
- La eficiencia relativa del muestreo aleatorio estratificado, con reparto proporcional, respecto del muestreo aleatorio simple, es tanto mayor cuanto mayor es $\sum_{h=1}^{L} W_h (\overline{Y}_h - \overline{Y})^2$, esto es, cuanto mayor es la diferencia entre las medias de uno a otro estrato y, por tanto, cuanto menor es la variación dentro de estratos -$\sum_{h=1}^{L} W_h S_h^2$ -, según [3.3.23].

Reparto óptimo

Restando miembro a miembro [3.3.18] y [3.3.8], resulta:

$$V_{prop}(\hat{\bar{Y}}_{str}) - V_{opt}(\hat{\bar{Y}}_{str}) = \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 \qquad [3.3.26]$$

donde,

$$\bar{S} = \sum_{h=1}^{L} W_h S_h$$

De [3.3.26] se tiene:

$$V_{prop}(\hat{\bar{Y}}_{str}) = V_{opt}(\hat{\bar{Y}}_{str}) + \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 \qquad [3.3.27]$$

Por sustitución de $V_{prop}(\hat{\bar{Y}}_{str})$ de [3.3.27] en [3.3.24] , resulta:

$$V_r(\hat{\bar{Y}}) \cong V_{opt.}(\hat{\bar{Y}}_{str}) + \frac{1}{n}\sum_{h=1}^{L} W_h (S_h - \bar{S})^2 + (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h (\bar{Y}_h - \bar{Y})^2 \qquad [3.3.28]$$

Y ,de [3.3.18], la eficiencia relativa es:

$$ER_{str\_opt./r} \cong 1 + \frac{1}{n}\frac{\sum_{h=1}^{L} W_h (S_h - \bar{S})^2}{V_{opt.}(\hat{\bar{Y}}_{str})} + (1-f)\frac{1}{n}\frac{\sum_{h=1}^{L} W_h (\bar{Y}_h - \bar{Y})^2}{V_{opt.}(\hat{\bar{Y}}_{str})} \qquad [3.3.29]$$

Cabe destacar los siguientes resultados que se derivan de [3.3.29]:

- El muestreo aleatorio estratificado, con reparto óptimo, es más eficiente que el muestreo aleatorio simple: $ER_{str\_opt./r} \geq 1$. La igualdad se da sólo cuando: $\bar{Y}_h = \bar{Y}$ y $S_h = \bar{S}; \forall h$.
- La eficiencia relativa del muestreo aleatorio estratificado, con reparto óptimo, respecto del muestreo aleatorio simple, es tanto mayor cuanto mayor es $\sum_{h=1}^{L} W_h (\bar{Y}_h - \bar{Y})^2$ y mayor es $\sum_{h=1}^{L} W_h (S_h - \bar{S})^2$, esto es, cuanto mayores son las diferencias entre las medias y las varianzas de uno a otro estrato.

Reparto óptimo v.s. Reparto proporcional

La eficiencia relativa del muestreo aleatorio estratificado con reparto óptimo, respecto del muestreo aleatorio estratificado con reparto proporcional es:

$$ER_{str\_opt./str\_prop.} = \frac{V_{prop.}(\hat{\bar{Y}}_{str})}{V_{opt.}(\hat{\bar{Y}}_{str})} = \frac{V_{prop.}(\hat{Y}_{str})}{V_{opt.}(\hat{Y}_{str})} \qquad [3.3.30]$$

y se tiene dividiendo los dos miembros de [3.3.27] por $V_{opt.}(\hat{\bar{Y}}_{str})$,

$$ER_{str\_opt./str\_prop.} \cong 1 + \frac{1}{n}\frac{\sum_{h=1}^{L} W_h (S_h - \bar{S})^2}{V_{opt}(\hat{\bar{Y}}_{str})} \qquad [3.3.31]$$

Cabe destacar los siguientes resultados que se derivan de [3.3.31]:

- El muestreo aleatorio estratificado, con reparto óptimo, es más eficiente que el muestreo aleatorio estratificado con reparto proporcional: $ER_{str\_opt/str\_prop.} \geq 1$. La igualdad se da sólo cuando: $S_h = \bar{S}; \forall h$.
- La eficiencia relativa del muestreo aleatorio estratificado, con reparto óptimo, respecto del muestreo aleatorio estratificado con reparto proporcional, es tanto mayor cuanto mayor es $\sum_{h=1}^{L} W_h (S_h - \bar{S})^2$, esto es, cuanto mayores son las diferencias entre las varianzas de uno a otro estrato.

Estimación a partir de la muestra

Consideraremos dos casos : (i) la muestra es aleatoria estratificada y (ii) la muestra es aleatoria simple. En el primer caso se plantea el problema de encontrar estimadores insesgados de parámetros asociados a la distribución del estimador con muestreo aleatorio simple, cuando la muestra es estratificada. En el segundo caso, se plantea el problema de encontrar estimadores insesgados de parámetros asociados a distribución del estimador con muestreo estratificado, a partir de una muestra aleatoria simple.

(i) La muestra es aleatoria estratificada

Cuando el reparto es proporcional, la eficiencia relativa del muestreo aleatorio estratificado respecto del aleatorio simple se estima sustituyendo en [3.3.25] $\sum_{h=1}^{L} W_h (\bar{Y}_h - \bar{Y})^2$ y $V_{prop}(\hat{\bar{Y}}_{str})$ por estimadores insesgados de los mismos.

Un estimador insesgado $V_{prop}(\hat{\bar{Y}}_{str})$ se tiene en [3.3.9]. En cuanto a $\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2$, se tiene

$$\sum_{h=1}^{L} W_h(\bar{Y}_h - \bar{Y})^2 = \sum_{h=1}^{L} W_h \bar{Y}_h^2 - \bar{Y}^2 \qquad [3.3.32]$$

Por otra parte, la varianza de la media muestral en el estrato h-ésimo -$V(\bar{y}_h)$- es:

$$V(\bar{y}_h) = E\bar{y}_h^2 - \bar{Y}_h^2 \qquad [3.3.33]$$

de donde,

$$\bar{Y}_h^2 = E\bar{y}_h^2 - V(\bar{y}_h) \qquad [3.3.34]$$

de modo que un estimador insesgado de $\bar{Y}_h^2$ es:

$$\hat{\bar{Y}}_h^2 = \bar{y}_h^2 - \hat{V}(\bar{y}_h) \qquad [3.3.35]$$

donde,

$\hat{V}(\bar{y}_h) = (1 - f_h)\dfrac{\hat{S}_h^2}{n_h}$ es un estimador insesgado de $V(\bar{y}_h)$, esto es, $E\hat{V}(\bar{y}_h) = V(\bar{y}_h)$, por lo que se verifica $E(\hat{\bar{Y}}_h^2) = \bar{Y}_h^2$.

De manera análoga se tiene un estimador insesgado de $\bar{Y}^2$:

$$V(\hat{\bar{Y}}_{str}) = E\hat{\bar{Y}}_{str}^2 - \bar{Y}^2$$

de donde,

$$\bar{Y}^2 = E\hat{\bar{Y}}_{str}^2 - V(\hat{\bar{Y}}_{str})$$

de modo que un estimador insesgado de $\bar{Y}^2$ es:

$$\hat{\bar{Y}}^2 = \hat{\bar{Y}}_{str}^2 - \hat{V}(\hat{\bar{Y}}_{str}) \qquad [3.3.36]$$

donde,

$\hat{V}(\hat{\bar{Y}}_{str})$ es el estimador de $V(\hat{\bar{Y}}_{str})$ definido en [3.3.6], que es insesgado, esto es, $E\hat{V}(\hat{\bar{Y}}_{str}) = V(\hat{\bar{Y}}_{str})$, por lo que se verifica $E(\hat{\bar{Y}}^2) = \bar{Y}^2$ e $\hat{\bar{Y}}^2$ resulta ser un estimador insesgado de $\bar{Y}$.

Así que un estimador insesgado de $\sum_{h=1}^{L} W_h(\overline{Y}_h - \overline{Y})^2$ se tiene sin más que sustituir en [3.3.32], los parámetros desconocidos - $\overline{Y}_h^2$ e $\overline{Y}^2$ -por sus estimadores insesgados definidos en [3.3.35] y [3.3.36], respectivamente:

$$Est\{\sum_{h=1}^{L} W_h(\overline{Y}_h - \overline{Y})^2\} = \sum_{h=1}^{L} W_h \hat{\overline{Y}}_h^2 - \hat{\overline{Y}}^2 \qquad [3.3.37]$$

Finalmente, por sustitución en [3.3.25] se tiene un estimador de la eficiencia relativa del muestreo estratificado con reparto proporcional, respecto del muestreo aleatorio simple:

$$\hat{E}R_{str\_prop/r} \cong 1 + (1-f)\frac{1}{n}\frac{\sum_{h=1}^{L} W_h \hat{\overline{Y}}_h^2 - \hat{\overline{Y}}^2}{\hat{V}_{prop.}(\hat{\overline{Y}}_{str})} \qquad [3.3.38]$$

Cuando el reparto es óptimo, las varianzas poblacionales S_h^2 deben ser conocidas y en ese caso la eficiencia se tiene a partir de [3.3.29] donde sólo $\sum_{h=1}^{L} W_h(\overline{Y}_h - \overline{Y})^2$ sería desconocido. Un estimador de la eficiencia relativa se tendría sin más que sustituir $\sum_{h=1}^{L} W_h(\overline{Y}_h - \overline{Y})^2$ por su estimador definido en [3.3.37].

5. Error de muestreo. Intervalos de confianza

Si el tamaño de la muestra es grande en cada estrato $(n_h \geq 30)$, la distribución en el muestreo de $\hat{\overline{Y}}_h$ puede aproximarse a la de una Normal, por el teorema central del límite y, en consecuencia también la de $\hat{\overline{Y}}_{str}$. En ese caso, un intervalo de confianza aproximado para la media - $\overline{Y}$ -, con un nivel de confianza $(1-\alpha)$ es:

$$IC(\overline{Y},(1-\alpha)) \equiv \hat{\overline{Y}}_{str} \pm t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\overline{Y}}_{str})} \qquad [3.3.39]$$

donde $t_{1-\frac{\alpha}{2}}(n_{g.d.l.})$ es el cuantíl de orden $(1-\frac{\alpha}{2})$ en la distribución t-Student con $n_{g.d.l.}$ grados de libertad, y los estimadores $\hat{\overline{Y}}_{str}$ y $\hat{V}(\hat{\overline{Y}}_{str})$ se definen en [3.3.2] y [3.3.6], respectivamente.

Los grados de libertad de la distribución t-Student se estiman de la siguiente forma [Cochran(1977)]:

$$n_{g.d.l.} = \frac{(\sum_{h=1}^{L} g_h \hat{S}_h^2)^2}{\sum_{h=1}^{L} \frac{g_h^2 (\hat{S}_h^2)^2}{n_h - 1}}$$

donde $\hat{S}_h^2$ se define en [3.3.5] y $g_h = \frac{N_h (N_h - n_h)}{n_h}$

El cuantíl puede ser aproximado por el de la ley Normal -$U_{1-\frac{\alpha}{2}}$- cuando el número de grados es grande.

Un intervalo de confianza aproximado para el total -Y-, con un nivel de confianza $(1-\alpha)$ es:

$$IC(Y,(1-\alpha)) \equiv N[\hat{\bar{Y}}_{str} \pm t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}\ ]$$

b) Error absoluto y error relativo máximo probable

El error absoluto en la estimación de la media -$\bar{Y}$- es $\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|$ y el relativo es $\frac{\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|}{\bar{Y}}$. Con una probabilidad $(1-\alpha)$, se verifica:

$$\left|\hat{\bar{Y}}_{str} - \bar{Y}\right| \le t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})} \qquad [3.3.40]$$

El error absoluto máximo probable en la estimación de la media, con una probabilidad $(1-\alpha)$, es $t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}$ y el relativo $\frac{t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}}{\bar{Y}}$. Para valores grandes de $n_{g.d.l.}$, $t_{1-\frac{\alpha}{2}}(n_{g.d.l.}) \cong U_{1-\frac{\alpha}{2}}$ donde $U_{1-\frac{\alpha}{2}}$ es el cuantíl de orden $(1-\frac{\alpha}{2})$ en la distribución Normal -$N(0,1)$- .En ese caso, [3.3.40] se escribiría como:

$$\left|\hat{\bar{Y}}_{str} - \bar{Y}\right| \le U_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}$$

El error absoluto en la estimación del total -Y- es $\left|\hat{Y}_{str} - Y\right| = N\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|$. Con una probabilidad $(1-\alpha)$, se verifica:

$$\left|\hat{Y}_{str} - Y\right| \le N\, t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}$$

Esta última desigualdad coincide con la [3.3.40], dado que $\left|\hat{Y}_{str} - Y\right| = N\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|$.

El error absoluto máximo probable en la estimación del total, con una probabilidad $(1-\alpha)$, es $N\, t_{1-\frac{\alpha}{2}}(n_{g.d.l.})\sqrt{\hat{V}(\hat{\bar{Y}}_{str})}$ y el relativo coincide con el del estimador de la media. Para valores grandes de $n_{g.d.l.}$, $t_{1-\frac{\alpha}{2}}(n_{g.d.l.}) \cong U_{1-\frac{\alpha}{2}}$ donde $U_{1-\frac{\alpha}{2}}$ es el cuantíl de orden $(1-\frac{\alpha}{2})$ en la distribución Normal - $N(0,1)$ -.

6. Tamaño de la muestra en función de la precisión deseada

Es posible determinar el tamaño de muestra mínimo necesario para alcanzar una precisión deseada en las estimaciones. El nivel de precisión deseado, puede ser especificado fijando un límite de tolerancia –*V*- para la varianza del estimador de la media-$V(\hat{\bar{Y}}_{str})$-, o bien un límite –*d*- para el error absoluto de estimación de la media -$\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|$- o un límite –*r*- para el error relativo de estimación de la media, $\dfrac{\left|\hat{\bar{Y}}_{str} - \bar{Y}\right|}{\bar{Y}}$. Para la estimación del total, los límites de tolerancia correspondientes serían N^2V y Nd, para la varianza del estimador y el error absoluto, respectivamente, mientras que el error relativo es el mismo para la media y para el total - *r* -.

(i) Límite de tolerancia para la varianza del estimador : $V(\hat{\bar{Y}}_{str}) \le V$ **ó** $V(\hat{Y}_{str}) \le N^2V$

Para un reparto genérico de la muestra total - n - entre los estratos:- $n_h = w_h n$, con $0 < w_h < 1$-, se tiene, tras sustituir en [3.3.4]:

$$V(\hat{\bar{Y}}_{str}) \le V \Leftrightarrow \sum_{h=1}^{L} W_h^2 (1 - f_h)\frac{S_h^2}{n w_h} \le V$$

La desigualdad es la misma ya se especifique, como la anterior, en términos del estimador de la media o del estimador del total -$V(\hat{Y}_{str}) \le N^2V$ - . En uno u otro caso, para que la varianza del estimador no supere el límite deseado, el tamaño de muestra n debe ser:

$$n \ge \frac{\sum_{h=1}^{L} W_h^2 \frac{S_h^2}{w_h}}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.41]$$

Para el reparto óptimo se tiene $w_h = \dfrac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$, y por sustitución en [3.3.41] resulta:

$$n \geq \frac{(\sum_{h=1}^{L} W_h S_h)^2}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.42]$$

Para el reparto proporcional se tiene $w_h = W_h$, y por sustitución en [3.3.41] resulta:

$$n \geq \frac{\sum_{h=1}^{L} W_h S_h^2}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.43]$$

El valor de S_h^2, en general desconocido, puede ser aproximado a partir de la información auxiliar disponible , por ejemplo de trabajos previos sobre la misma población o debe ser estimado a partir de una encuesta piloto, haciendo uso del estimador definido en [3.3.5].

(ii) Límite de tolerancia d **para el error absoluto:** $\left|\hat{\bar{Y}}_{str} - \bar{Y}\right| \leq d$ **ó** $\left|\hat{Y}_{str} - Y\right| \leq Nd$

Para tamaños de muestra grandes, la distribución del estimador $\hat{\bar{Y}}_{Str}$ es aproximadamente Normal, entonces se verifica

$$\left|\hat{\bar{Y}}_{str} - \bar{Y}\right| \leq U_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\bar{Y}}_{str})} \qquad [3.3.44]$$

Para un reparto genérico de la muestra total -n- entre los estratos: $n_h = w_h n$, [3.3.44] puede escribirse así:

$$\left|\hat{\bar{Y}}_{str} - \bar{Y}\right| \leq U_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n w_h}}$$

Para que, con una probabilidad $(1-\alpha)$, el error absoluto máximo probable no supere el límite deseado -d si se establece sobre el estimador de la media y Nd si se establece sobre el estimador del total-, ha de verificarse:

$$U_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n w_h}} \leq d$$

y el tamaño de muestra n ha de ser:

$$n \geq \frac{U^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{w_h}}{d^2 + U^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.45]$$

Para el reparto óptimo se tiene $w_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$, y por sustitución en [3.3.45] resulta:

$$n \geq \frac{U^2_{1-\frac{\alpha}{2}} (\sum_{h=1}^{L} W_h S_h)^2}{d^2 + U^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.46]$$

Para el reparto proporcional se tiene $w_h = W_h$, y por sustitución en [3.3.45] resulta:

$$n \geq \frac{\mathrm{U}^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} W_h S_h^2}{d^2 + \mathrm{U}^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.47]$$

(iii) Límite de tolerancia r para el error relativo: $\frac{\left|\hat{\overline{Y}}_{str} - \overline{Y}\right|}{\overline{Y}} \leq r$ **ó** $\frac{\left|\hat{Y}_{str} - Y\right|}{Y} \leq r$

De [3.3.44] , tras dividir por $\overline{Y}$, se tiene, para un reparto genérico de la muestra total -n- entre los estratos - $n_h = w_h n$ -,

$$\frac{\left|\hat{\overline{Y}}_{str} - \overline{Y}\right|}{\overline{Y}} \leq \frac{1}{\overline{Y}} U_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n w_h}}$$

Para que, con una probabilidad $(1-\alpha)$, el error relativo máximo probable no supere el límite deseado r, ha de verificarse,

$$\frac{1}{\overline{Y}} U_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n w_h}} \leq r$$

esto es,

$$U_{1-\frac{\alpha}{2}} \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n w_h}} \leq r\overline{Y}$$

y el tamaño de muestra n se tiene a partir de [3.3.45], sin más que sustituir d por $r\overline{Y}$:

$$n \geq \frac{U^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{w_h}}{r^2\overline{Y}^2 + U^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.48]$$

Para el reparto óptimo se tiene $w_h = \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$, y por sustitución en [3.3.48] resulta:

$$n \geq \frac{U^2_{1-\frac{\alpha}{2}} (\sum_{h=1}^{L} W_h S_h)^2}{r^2\overline{Y}^2 + U^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.49]$$

Para el reparto proporcional se tiene $w_h = W_h$, y por sustitución en [3.3.48] resulta:

$$n \geq \frac{\mathrm{U}^2_{1-\frac{\alpha}{2}} \sum_{h=1}^{L} W_h S_h^2}{r^2\overline{Y}^2 + \mathrm{U}^2_{1-\frac{\alpha}{2}} \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \qquad [3.3.50]$$

El valor de $\overline{Y}$, en general desconocido, puede ser aproximado a partir de la información auxiliar disponible , o ser estimado a partir de una encuesta piloto, haciendo uso del estimador definido en [3.3.2].

7. Construcción de los estratos

Los límites entre estratos pueden ser establecidos de manera aproximadamente óptima [3.3.DALENIUS (1950)]. Para una sola variable de estratificación, existe un criterio de optimización bien establecido en la literatura: consiste en minimizar la varianza del estimador de la media de la variable de estratificación, para un tamaño de muestra dado. La variable de estratificación ideal es la variable en estudio. Sin embargo, en la práctica dicha variable será desconocida: es precisamente para estimar sus características por lo que nos interesemos en el diseño de una muestra. En su defecto, la mejor variable de estratificación es aquella más estrechamente relacionada (la de mayor coeficiente de correlación) con la variable en estudio.

Denotemos por Y la variable de estratificación. La varianza del estimador de la media se define en [3.3.4] y para un reparto óptimo de la muestra entre los estratos, ignorando el factor corrector de poblaciones finitas ($f_h \cong 0; \forall h$.), se reduce a :

$$V_{opt.}(\hat{\bar{Y}}_{str}) \cong \frac{1}{n}(\sum_{h=1}^{L} W_h S_h)^2$$

Se trata de determinar los valores $y_1 < y_2 <.....< y_{h-1} < y_h <.....< y_{L-1}$ de la variable de estratificación que, junto con los valores mínimo - y_0 - y máximo - y_L -, delimitan L estratos, de modo que $V(\hat{\bar{Y}}_{str})$ resulte mínima (Cochran (1977)). Supuesta la variable Y continua, y si denotamos por $f(y)$ su función de densidad de probabilidad, la proporción W_h de elementos de la población en el estrato $h; \{h = 1,2,...,L\}$ es:

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy = F(y_h) - F(y_{h-1})$$

donde $F(\cdot)$ denota la función de distribución.

Consideremos la distribución de la variable Y dentro del estrato $h; \{h = 1,2,...,L\}$,esto es, la distribución de Y truncada por la condición $y_{h-1} < y < y_h$. Su función de distribución es:

$$F(y / y_{h-1} < y < y_h) = \frac{F(y) - F(y_{h-1})}{F(y_h) - F(y_{h-1})}$$

y su función de densidad de probabilidad es:

$$f(y / y_{h-1} < y < y_h) = \frac{f(y)}{\int_{y_{h-1}}^{y_h} f(y)dy}$$

Su varianza es:

$$S_h^2 = V(y / y_{h-1} < y < y_h) = E\{y^2 / y_{h-1} < y < y_h\} - \left[E\{y / y_{h-1} < y < y_h\}\right]^2$$

donde,

$$E\{y / y_{h-1} < y < y_h\} = \frac{\int_{y_{h-1}}^{y_h} yf(y)dy}{\int_{y_{h-1}}^{y_h} f(y)dy}$$

$$E\{y^2 / y_{h-1} < y < y_h\} = \frac{\int_{y_{h-1}}^{y_h} y^2 f(y)dy}{\int_{y_{h-1}}^{y_h} f(y)dy}$$

El valor de y_h que minimiza a $V(\hat{\bar{Y}}_{str})$ es el mismo que minimiza a $\sum_{h=1}^{L} W_h S_h$ y ha de verificar la condición $\frac{\partial \sum_{h=1}^{L} W_h S_h}{\partial y_h} = 0$. Puesto que y_h aparece sólo como límite entre los estratos h y $h+1$, la condición necesaria se reduce a:

$$\frac{\partial (W_h S_h)}{\partial y_h} + \frac{\partial (W_{h+1} S_{h+1})}{\partial y_h} = 0$$

Consideremos el producto $W_h S_h^2$:

$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} y^2 f(y) dy - \frac{\left[\int_{y_{h-1}}^{y_h} y f(y) dy \right]^2}{\int_{y_{h-1}}^{y_h} f(y) dy}$$

y su derivada con respecto a y_h:

$$\frac{\partial W_h S_h^2}{\partial y_h} = y_h^2 f(y_h) - 2\mu_h y_h f(y_h) + (y_h)\mu_h^2$$

$$\mu_h = \int_{y_{h-1}}^{y_h} y f(y) dy$$

Por otra parte,

$$\frac{\partial W_h S_h^2}{\partial y_h} = S_h^2 \frac{\partial W_h}{\partial y} + 2 W_h S_h \frac{\partial S_h}{\partial y_h}$$

$$S_h^2 \frac{\partial W_h}{\partial y_h} = S_h^2 f(y_h)$$

esto es,

$$S_h^2 \frac{\partial W_h}{\partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h} = y_h^2 f(y_h) - 2\mu_h y_h f(y_h) + f(y_h)\mu_h^2$$

$$S_h^2 \frac{\partial W_h}{\partial y_h} = S_h^2 f(y_h)$$

y sumando miembro a miembro resulta:

$$2 S_h^2 \frac{\partial W_h}{\partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h} = \left[(y_h - \mu_h)^2 + S_h^2 \right] f(y_h)$$

de donde, al dividir por $2S_h$, se tiene:

$$\frac{\partial (W_h S_h)}{\partial y_h} = \frac{2 S_h^2 \frac{\partial W_h}{\partial \partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h}}{2 S_h} = \frac{1}{2} \frac{(y_h - \mu_h)^2 + S_h^2}{S_h} f(y_h)$$

y de manera análoga,

$$\frac{\partial(W_{h+1}S_{h+1})}{\partial y_h} = -\frac{1}{2}\frac{(y_h-\mu_{h+1})^2+S_{h+1}^2}{S_{h+1}}f(y_h)$$

La condición de mínimo puede entonces escribirse de la siguiente forma:

$$\frac{(y_h-\mu_h)^2+S_h^2}{S_h} = \frac{(y_h-\mu_{h+1})^2+S_{h+1}^2}{S_{h+1}} \quad ;h=1,2,\ldots..,L-1$$

En este sistema de ecuaciones, la solución en y_h depende de la media μ_h y de la varianza S_h^2 dentro de estratos, las cuales, a su vez, dependen de los límites y_h buscados, por lo que la obtención de soluciones plantea dificultades.

Aproximaciones prácticas

Se han propuesto diversas soluciones aproximadas al sistema de ecuaciones minimal de Dalenius (1950), una revisión de las cuales puede encontrarse en Kpedekpo (1973).

Dalenius y Hodges (1959) proponen la siguiente solución aproximada. Sea,

$$Z(y) = \int_{y_0}^{y}\sqrt{f(y)}dy$$

Supongamos que la distribución de la variable de estratificación es uniforme dentro de estratos, esto es, $f(y) = f_h;\ \forall y \left| y_{h-1} < y < y_h \right.$. En ese caso, se verifica:

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy = f_h\int_{y_{h-1}}^{y_h} dy = f_h(y_h-y_{h-1})$$

$$S_h^2 = \frac{(y_h-y_{h-1})^2}{12}$$

de donde,

$$\sum_{h=1}^{L} W_hS_h = \frac{1}{\sqrt{12}}\sum_{h=1}^{L} f_h(y_h-y_{h-1})^2$$

Por otra parte,

$$Z(y_h)-Z(y_{h-1}) = \int_{y_{h-1}}^{y_h}\sqrt{f(y)}dy = \sqrt{f_h}(y_h-y_{h-1})$$

de modo que por sustitución, resulta:

$$\sqrt{12}\sum_{h=1}^{L} W_hS_h = \sum_{h=1}^{L} f_h(y_h-y_{h-1})^2 = \sum_{h=1}^{L}(Z(y_h)-Z(y_{h-1}))^2$$

dado que $Z(y_L) - Z(y_0)$ es fija, el mínimo de $\sqrt{12}\sum_{h=1}^{L} W_h S_h$ se tiene cuando:

$$2(Z(y_h) - Z(y_{h-1})) - 2(Z(y_{h+1}) - Z(y_h)) = 0$$

esto es,

$$Z(y_h) - Z(y_{h-1}) = Z(y_{h+1}) - Z(y_h); \forall h$$

cuando $Z(y_h) - Z(y_{h-1})$ es constante:

$$Z(y_h) - Z(y_{h-1}) = \int_{y_{h-1}}^{y_h} \sqrt{f(y)}dy = constante$$

La regla práctica consiste en lo siguiente: sea $F(y)$ el valor acumulado de la raíz cuadrada de las frecuencias absolutas de los valores de la variable de estratificación inferiores o iguales a y, para construir un número de estratos L, los límites aproximadamente óptimos entre estratos son los valores $y_1 < y_2 <.....< y_{h-1} < y_h <.....< y_{L-1}$ de la variable de estratificación que satisfacen a la ecuación:

$$F(y) = \frac{hH}{L}; h = 1,2,...., L-1$$

en la que H es el valor de $F(y)$ correspondiente al máximo valor de la variable de estratificación, supuesto este finito. El extremo inferior del estrato $h = 1$ es el valor mínimo de la variable de estratificación y el superior del estrato $h = L$ es el valor máximo de la variable de estratificación.

Ekman (1959) propone tomar los límites entre estratos de modo que $W_h(y_h - y_{h-1})$ sea igual para todo h. Otras propuestas similares a ésta consisten en sustituir $(y_h - y_{h-1})$ por la media de la variable de estratificación en el estrato $h - \mu_h -$,o bien por la desviación típica - S_h -.

8. Número de estratos

Consideramos, en primer lugar, el caso en que la variable de estratificación es la variable en estudio y que su distribución es uniforme:

$$f(y) = \frac{1}{d}; d = y_{max.} - y_{min.}$$

Ignorando el factor corrector de poblaciones finitas ($f = \frac{n}{N} \cong 0$) y considerando que, por su distribución uniforme, la varianza de la variable en estudio es $S^2 = \frac{d^2}{12}$, la varianza del estimador de la media con muestreo aleatorio simple viene dada por:

$$V(\hat{\bar{Y}}_r) = \frac{d^2}{12n}$$

Supongamos que construimos L estratos de igual amplitud -d/L-. Asumimos distribución uniforme de la variable de estratificación dentro de cada estrato:

$$f(y) = \frac{1}{d/L}; \; y_{h-1} < y < y_h \quad \forall h = 1,2,....,L$$

de modo que la varianza de la variable de estratificación dentro de estratos es $S_h^2 = \frac{(d/L)^2}{12}$.

Supuesto $W_h = \frac{d/L}{d} = \frac{1}{L}$ y considerando un reparto proporcional, esto es, $n_h = \frac{n}{L}$, la varianza del estimador de la media con muestreo aleatorio estratificado es:

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} = \frac{d^2}{12nL^2}$$

Obsérvese que $V(\hat{\bar{Y}}_r) = \frac{d^2}{12n}$, de modo que la eficiencia relativa del muestreo aleatorio estratificado respecto del muestreo aleatorio simple es $\frac{V(\hat{\bar{Y}}_r)}{V(\hat{\bar{Y}}_{str})} = L^2$ y crece al aumentar el número de estratos.

Si consideramos como función de coste de las estimaciones:

$$C = C_{str} L + C_u n$$

donde C_{str} denota un coste unitario por estrato y C_u un coste unitario por elemento, el número de estratos óptimo, esto es, el que minimiza $V(\hat{\bar{Y}}_{str})$ fijado el coste C o minimiza este coste, fijada la varianza es:

$$L_{opt.} = \frac{2nC_u}{C_{str.}}$$

Consideremos ahora que la variable de estratificación es una variable X relacionada con la variable en estudio, de la forma especificada en el siguiente modelo:

$$y = \alpha + \beta x + e$$
$$E\left\{ y \middle/ x \right\} = \alpha + \beta x$$
$$Cov(x, e) = 0$$

La varianza del término de perturbación - S_e^2 - se considera igual en todos los estratos, así como los coeficientes α y β. La varianza de la variable en estudio en el estrato h es, según modelo, $S_{yh}^2 = \beta^2 S_{xh}^2 + S_e^2$. La varianza del estimador de la media con muestreo aleatorio estratificado y reparto proporcional es:

$$V(\hat{\bar{Y}}_{str}) = \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} = \frac{\beta^2}{n} \sum_{h=1}^{L} W_h S_{xh}^2 + \frac{S_e^2}{n}$$

Supongamos que la eficiencia relativa del muestreo aleatorio estratificado respecto del muestreo aleatorio simple, para la estimación de la media de la variable X es L^2, de modo que se verifica:

$$\frac{V(\hat{\bar{X}}_r)}{V(\hat{\bar{X}}_{str})} = L^2; V(\hat{\bar{X}}_r) = \frac{S_x^2}{n}; V(\hat{\bar{X}}_{str}) = \sum_{h=1}^{L} W_h \frac{S_h^2}{n}$$

y, por tanto:

$$V(\hat{\bar{Y}}_{str}) = \frac{\beta^2 S_x^2}{nL^2} + \frac{S_e^2}{n}$$

donde,

$$\beta = \rho \frac{S_y^2}{S_x^2}$$
$$S_e^2 = S_y^2 (1 - \rho^2)$$

siendo ρ el coeficiente de correlación entre X e Y. En consecuencia:

$$V(\hat{\bar{Y}}_{str}) = \frac{S_y^2}{n} \left[\frac{\rho^2}{L^2} + (1 - \rho^2) \right]$$

y la eficiencia relativa del muestreo estratificado respecto del aleatorio simple sería:

$$\frac{V(\hat{\bar{Y}}_r)}{V(\hat{\bar{Y}}_{str})} = \frac{1}{\left[\frac{\rho^2}{L^2} + (1 - \rho^2) \right]}$$

aumenta al aumentar el número de estratos. Sin embargo, dando valores a ρ y a L se puede observar que para valores mayores de $L = 6$ los aumentos de la eficiencia relativa son pequeños, y tanto menores cuanto menor sea ρ (Cohran (1977)).

Ejemplo 3.3.1

Ilustraremos con un ejemplo cada una de las etapas del proceso de diseño de una muestra aleatoria estratificada, descritas en los epígrafes 1 a 8 anteriores. Consideremos la población de 275 empresas del sector industrial de una determinada zona, para la que se ha construido el marco de muestreo que se recoge en el Apéndice B del Tema 3.1. El objetivo es estimar la media de la variable "número de empleados", por muestreo estratificado.

Construcción de los estratos

Empezaremos por tratar el problema de la construcción de los estratos. En el caso del ejemplo se dan las condiciones ideales para el tratamiento de este problema: se conocen el valor de la variable objetivo en todos y cada uno de los elementos de la población. Obviamente este caso no tiene interés práctico alguno: lo consideramos aquí sólo con fines ilustrativos y de comparación con los resultados de casos de interés práctico, que trataremos posteriormente.

Estratificación óptima. Variable de estratificación: "Número de empleados".

El proceso de construcción de los estratos se inicia agrupando los 275 elementos de la población en clases de igual amplitud, de modo que dentro de cada clase la distribución de la variable de estratificación sea aproximadamente uniforme. A este fin, se consideran clases de amplitud reducida: la amplitud de clases la fijamos en 5 empleados, hasta un número de empleados inferior a los 300. Entre 300 y 800 se consideran clases de amplitud 100 y a partir de 800 se consideran clases de amplitud 200. Con esta amplitud, algunas clases resultan vacías, particularmente en los tramos inferiores a los 250 empleados Cuando esto sucede, hemos optado por fundir la clase vacía con la siguiente, que pasa de este modo a duplicar su amplitud. En el cuadro 1 se muestra la distribución de los 275 elementos de la población entre las clases consideradas. Como puede observarse, la clase de 96 a 100 empleados, que resulta vacía, se ha fundido con la de 101 a 105 empleados, resultando una sola clase de 96 a 105 empleados. Así mismo, la clase de 181 a 200 resulta de fundir cuatro clases consecutivas, la última de las cuales es la que contiene los elementos del conjunto, estando las otras tres vacías.

Las clases de partida son, pues, de diferente amplitud. Cuando esto sucede, el valor acumulado de la raíz de las frecuencias $Cum\sqrt{f}$ se calcula multiplicando la $\sqrt{f}$ correspondiente a la clase en cuestión por $\sqrt{u}$,donde u es el cociente entre la amplitud de dicha clase y la de referencia, que tomamos como igual a 5.Este valor se suma al $Cum\sqrt{f}$ correspondiente a la clase anterior.

Cuadro 1 Elementos para la construcción de los estratos

Clase de Tamaño (Nº de empleados)	Nº de Empresas	Frecuencia (f)	$\sqrt{f}$	$Cum\sqrt{f}$
5 ó menos	27	0.09818	0.31334	0.31334
De 6 a 10	35	0.12727	0.35675	0.67009
De 11 a 15	37	0.13455	0.36680	1.03689
De 16 a 20	31	0.11273	0.33575	1.37264
De 21 a 25	16	0.05818	0.24121	1.61385
De 26 a 30	13	0.04727	0.21742	1.83127
De 31 a 35	10	0.03636	0.19069	2.02196
De 36 a 40	11	0.04000	0.20000	2.22196
De 41 a 45	3	0.01091	0.10445	2.32641
De 46 a 50	15	0.05455	0.23355	2.55996
De 51 a 55	7	0.02545	0.15954	2.7195
De 56 a 60	4	0.01455	0.12060	2.8401
De 61 a 65	6	0.02182	0.14771	2.98781
De 66 a 70	4	0.01455	0.12060	3.10841
De 71 a 75	2	0.00727	0.08528	3.19369
De 76 a 80	6	0.02182	0.14771	3.3414
De 81 a 85	3	0.01091	0.10445	3.44585
De 86 a 90	1	0.00364	0.06030	3.50615
De 91 a 95	4	0.01455	0.12060	3.62675
De 96 a 105	2	0.00727	0.08528	3.74735
De 106 a 110	2	0.00727	0.08528	3.83263
De 111 a 115	2	0.00727	0.08528	3.91791
De 116 a 125	1	0.00364	0.06030	4.00319
De 126 a 130	1	0.00364	0.06030	4.06349
De 131 a 140	1	0.00364	0.06030	4.14876
De 141 a 145	2	0.00727	0.08528	4.23404
De 146 a 150	3	0.01091	0.10445	4.33849
De 151 a 160	2	0.00727	0.08528	4.45909
De 161 a 175	2	0.00727	0.08528	4.60679
De 176 a 180	2	0.00727	0.08528	4.69207
De 181 a 200	2	0.00727	0.08528	4.86263
De 201 a 220	1	0.00364	0.06030	4.98323
De 221 a 230	1	0.00364	0.06030	5.06850
De 231 a 245	1	0.00364	0.06030	5.17294
De 246 a 250	1	0.00364	0.06030	5.23324
De 251 a 300	2	0.00727	0.08528	5.50291
De 301 a 400	2	0.00727	0.08528	5.88429
De 401 a 500	3	0.01091	0.10445	6.35140
De 501 a 600	3	0.01091	0.10445	6.81851
De 601 a 700	1	0.00364	0.06030	7.08818
De 701 a 800	1	0.00364	0.06030	7.35785
De 801 a 1000	1	0.00364	0.06030	7.62751
De 1001 a1200	1	0.00364	0.06030	8.00888

Los límites entre estratos se determinan sobre la escala de valores de las clases, agrupando clases de modo que en la escala de valores de la $Cum\sqrt{f}$ resulten intervalos de amplitud igual al cociente entre el total acumulado de $Cum\sqrt{f}$ y el número L de estratos. Para un número de estratos $L = 6$, la amplitud de cada uno de los seis intervalos debe ser

$$\frac{8.00888}{6} = 1.3348$$

Los límites de estrato son, pues, en la escala de valores de $Cum\sqrt{f}$ los siguientes

{1.33, 2.66, 3.99, 5.32, 6,65 y 7.98}

A estos valores corresponden, en la escala de valores de la variable de estratificación, los siguientes valores aproximados:

{20, 50, 115, 250 y 1000}

En consecuencia los estratos quedarían delimitados de la siguiente forma:

Estrato Nº	Empresas con un Número de empleados
1	20 ó menos
2	De 21 a 50
3	De 51 a 115
4	De 116 a 250
5	De 250 a 1000
6	Más de 1000

Las características de la estratificación resultante son las siguientes:

Cuadro 4.2 Características de los estratos

Estrato (h)	Número de Empresas (N_h)	Número medio de empleados $(\bar{Y}_h)$	Varianza dentro de estratos (S_h^2)
1	130	10.92	28.64
2	68	34.87	86.03
3	43	76.84	336.96
4	20	174.60	1387.73
5	13	526.69	38114.56
6	1	1001.00	0.00

Reparto de la muestra

Se considera una muestra de tamaño n, a repartir entre los estratos cuyas características se muestran en el cuadro 2. Los coeficientes de reparto correspondientes a cada estrato -calculados mediante [3.3.7], para el reparto proporcional y [3.3.17] para el reparto óptimo -, se recogen en el siguiente cuadro, junto con los tamaños de muestra que cabría asignar a cada estrato para $n = 15$.

Cuadro 3 Coeficientes de reparto y tamaño de la muestra: Ejemplo 3.3.1

Estrato (h)	*Coeficiente de Reparto* (w_h)		*Tamaño de la muestra* (n_h)			
	Proporcional	Óptimo	Proporcional	***	Óptimo	***
1	0.47273	0.13	7	7	2	2
2	0.24727	0.11	4	3	2	1
3	0.15636	0.15	2	2	2	2
4	0.07273	0.14	1	1	2	2
5	0.04727	0.47	1	1	7	7
6	0.00364	0.00	0	1	0	1
Total	1.00	1.00	15	15	15	15

*** Con uno u otro tipo de reparto, el tamaño de muestra que se le asignaría al estrato 6 es nulo. Esta solución debe ser rechazada porque es incompatible con un plan de muestreo estratificado: en este tipo de planes el tamaño mínimo de la muestra en un estrato cualquiera es 1, nunca cero. Una solución admisible consiste en asignar al estrato 6 el tamaño mínimo de muestra, esto es 1, y en asignar de nuevo los 14 elementos restantes de la muestra a los restantes 5 estratos: en esta operación los elementos del estrato 6 se consideran excluidos a efectos de cálculo del nuevo reparto, de modo que la población se reduce a los elementos de los estratos 1,2,3,4 y5. Los nuevos coeficientes de reparto proporcional apenas difieren de los antiguos debido a que el estrato 6, ahora excluido del reparto, es muy reducido: contiene sólo un elemento, de modo que en [3.3.7], el denominador pasa a ser 274, en lugar de 275, permaneciendo igual el numerador. Con reparto óptimo, los coeficientes de reparto se mantienen exactamente iguales debido a que la contribución del estrato 6 al denominador de [3.3.16], que ahora es nula por construcción, ya era nula en el reparto antiguo debido a que la varianza dentro de dicho estrato es nula, por constar de un solo elemento. Así que en el nuevo reparto lo que varía es el tamaño de la muestra a repartir, que es 14, en lugar de 15.

Sucede en ocasiones que la muestra asignada en el reparto a uno o más estratos iguala o supera al número de elementos de la población en el estrato. En ese caso, el tamaño de la muestra en el(los) estrato(s) se hace igual a al de la población y se procede a realizar un nuevo reparto en la forma indicada en el párrafo anterior.

Selección de la muestra

La muestra asignada a cada estrato se selecciona de forma independiente de uno a otro estrato. El procedimiento de selección puede ser con o sin reposición y con probabilidades iguales o desiguales. Consideraremos aquí el procedimiento de selección sin reposición y con probabilidades iguales descrito en el Tema 1, epígrafe 1, y lo aplicaremos a cada estrato como si de una población en sí mismo se tratara.

Empezaremos por numerar los elementos de cada estrato h, de 1 a N_h, donde N_h denota el número de elementos del estrato h. En el Apéndice B, se recoge el marco de muestreo del Apéndice B del tema 3.1, una vez estratificado y listo para su empleo con vistas a la selección de muestras aleatorias estratificadas.

Supongamos que se desea extraer una muestra de n=15 elementos, con el reparto óptimo (corregido) que se especifica en el cuadro 3, esto es: $n_1 = 2, n_2 = 1, n_3 = 2, n_4 = 2, n_5 = 7, n_6 = 1$. Para seleccionar los $n_1 = 2$ elementos del estrato 1, haremos uso de la tabla de números aleatorios (ver tema 3.1, tabla 1). Elijamos al azar un punto de arranque en la tabla: sea la intersección de la fila 26 y la columna 15 (número 6), el punto elegido. Puesto que el número $N_1 = 130$ de elementos del estrato 1 consta de 3 dígitos se requieren tres columnas (o filas) de la tabla para la selección de la muestra. Consideremos la columna a la que pertenece el punto de arranque y las dos inmediatas siguientes a su derecha, y recorrámosla de arriba hacia abajo para la selección de la muestra, descartando los números superiores a 130: los números así seleccionados resultan ser $\{091 \; y \; 081\}$. La muestra $n_2 = 1$ se selecciona, eligiendo al azar un nuevo punto de arranque en la tabla y procediendo de la misma forma que en el estrato 1: supongamos que el punto de arranque es el de intersección de la fila10 con la columna 23 (número 8); el número de elementos del estrato 2 es 68, de manera que consideraremos la columna a la que pertenece el punto de arranque y la inmediata siguiente a su derecha: recorriéndola de arriba hacia abajo el número seleccionado es $\{18\}$. De la misma forma, asumiendo que los punto de arranque aleatorio son (fila 4, columna 16), (40,12), y (34, 11), para la selección de la muestra en los estratos 3, 4 y 5, respectivamente, los números seleccionados son, respectivamente, $\{24,04\}, \{15,02\} \; y \; \{02,08,07,03,01,11,12\}$. Finalmente, el estrato 6 consta de un solo elemento, de modo que la muestra en ese estrato es ese elemento.

Estimación de la media

Los valores observados de la variable en estudio - número de empleados - en los individuos de la población seleccionados en la muestra, por orden de extracción, son los siguientes (Ver Apéndice A):

Estrato	Valores observados
1	{20, 15}
2	{31}
3	{63, 65}
4	{180, 147}
5	{297, 500, 953, 452, 417, 394, 791}
6	{1001}

La media se estima a partir de las observaciones muestrales, haciendo uso de la expresión [3.3.2]. Los elementos que se requieren en esta expresión son los siguientes:

Estrato (h)	N_h	$W_h = N_h/N$	$\hat{\bar{Y}}_h$	$W_h\hat{\bar{Y}}_h$
1	130	0.47273	17.50	8.2727
2	68	0.24727	31.00	7.6655
3	43	0.15636	64.00	10.0073
4	20	0.07273	163.50	11.8909
5	13	0.04727	543.43	25.6894
6	1	0.00364	1001.00	3.6400
Total	275	1.00	------	67.1657

Donde $\hat{\bar{Y}}_h$ es la media de las observaciones muestrales en el estrato "h". La estimación de la media según [3.3.2] , es el total de la última columna: $\hat{\bar{Y}}_{str} = 67.1657$

a) Varianza del estimador

Viene dada por la expresión [3.3.3]. Los elementos necesarios para su cálculo se recogen en el siguiente cuadro:

Estrato (h)	N_h	$W_h = N_h/N$	n_h	$f_h = n_h/N_h$	S_h^2	$W_h^2(1-f_h)\frac{S_h^2}{n_h}$
1	130	0.47273	2	0.01538	28.64	3.15045
2	68	0.24727	1	0.01470	86.03	5.18266
3	43	0.15636	2	0.04651	336.96	3.91589
4	20	0.07273	2	0.1	1387.73	3.30302
5	13	0.04727	7	0.53846	38114.56	5.61594
6	1	0.00364	1	1.00	0.00	0.00
Total	275	1.00	15	----------	----------	21.16796

La varianza del estimador es el total de la última columna del cuadro, esto es 21.16796. La desviación típica del estimador es de 4.60086 y el coeficiente de variación del estimador es del 6.88%. Para un mismo tipo de reparto, la varianza del estimador depende del tamaño de la muestra. La expresión [3.3.18] muestra esa dependencia, en el caso del reparto óptimo. Para un tamaño de muestra $n = 15$ (tasa de muestreo del 5.45%), la varianza del estimador calculada según [3.3.18] es 18.4459, la desviación típica 4.29487 y el coeficiente de variación 6.42079% (estos valores son ligeramente inferiores a los calculados según [3.3.3], debido a que en este último caso el reparto óptimo fue corregido (***) de modo que se correspondiera al de un plan de muestreo estratificado). Cuando el tamaño de la muestra se duplica $n = 30$ (tasa de muestreo del 10.9%), la varianza se reduce a 5.60467, la desviación típica a 2.36742 y el coeficiente de variación baja al 3.53927%.

f) Error de muestreo: Error absoluto y relativo máximo probable

Para el nivel de confianza usual del 95% -$(1-\alpha) = 0.95$-, el cuantíl de orden $1-\frac{\alpha}{2} = 0.975$ en la distribución Normal es 1.96. Así pues, según [3.3.50] cabe esperar, con una probabilidad del 95%, que con una muestra de tamaño $n = 15$ el error absoluto de la estimación -$\left|\hat{\overline{Y}}_{str} - \overline{Y}\right|$- sea inferior o igual a $1.96 \times 4.29487 = 8.4179$. Si aumentamos el tamaño de la muestra hasta los $n = 30$ elementos, manteniendo el nivel de confianza en el 95%, la desviación típica se reduce a 2.36742, de modo que el error absoluto de la estimación -$\left|\hat{\overline{Y}}_{str} - \overline{Y}\right|$- cabe esperar, con una probabilidad del 95%, que sea inferior o igual a $1.96 \times 2.36742 = 4.6401$. El error relativo máximo probable es del 12.58% para n=15 y del 6.94%, para n=30.

g) Tamaño de la muestra necesaria para alcanzar una precisión deseada

Si fijamos un límite de tolerancia para el error relativo r de, por ejemplo, el 10% -$r = 0.1$-, el tamaño de muestra mínimo necesario para que, con una probabilidad del 95%, el error relativo de la estimación no supere ese límite se tiene a partir de [3.3.49]. Los elementos necesarios para el calculo se recogen en el siguiente cuadro.

Estrato (h)	N_h	$W_h = N_h / N$	S_h^2	$W_h S_h$	$W_h S_h^2$
1	130	0.47273	28.64	2.52970	13.54
2	68	0.24727	86.03	2.29347	21.27
3	43	0.15636	336.96	2.86598	52.53
4	20	0.07273	1387.73	2.70925	100.93
5	13	0.04727	38114.56	9.22903	1801.78
6	1	0.00364	0.00	0.00	0.00
Total	275	1.00	----------	19.62743	1990.04

Por sustitución de estos elementos en [3.3.49], y sabiendo que $U_{1-\frac{\alpha}{2}} = 1.96$ y $\bar{Y} = 66.89$, resulta

$$n \geq \frac{1.96^2 \times (19.62743)^2}{0.1^2 \times 66.89^2 + 1.96^2 \times \frac{1}{275} \times 1990.04} \cong 20$$

esto es, la tasa de muestreo debería ser del 7.4%.

Eficiencia del muestreo estratificado respecto del aleatorio simple

Viene dada por [3.3.20]. La varianza del estimador de la media con muestreo aleatorio simple es, para un tamaño de muestra n=15, igual a 1113.57 (Ver ejemplo 3.1.2). Con el mismo tamaño de muestra y muestreo estratificado con reparto óptimo, la varianza del estimador es 21.16796. La eficiencia relativa del muestreo estratificado respecto del aleatorio simple es, pues, igual a

$$ER_{str/r} = \frac{1113.57}{21.16796} = 52.61$$

Para alcanzar con muestreo aleatorio simple la misma precisión que se tiene con el muestreo estratificado con reparto óptimo, el tamaño de la muestra aleatoria simple debe ser igual al de la muestra aleatoria estratificada multiplicada por 52.61. Esta eficiencia corresponde a un caso ideal: aquel en el la variable de estratificación es la variable objetivo. Ya hemos señalado que este caso no tiene interés práctico y se considera sólo con fines ilustrativos. En la práctica, la variable de estratificación tendrá que ser una distinta a la variable objetivo, más o menos correlacionada con ella, dependiendo la eficiencia de ese grado de correlación.

La eficiencia del reparto óptimo respecto del proporcional puede ser aproximada mediante [3.3.31]. Para $n = 15$, la varianza del estimador con muestreo estratificado y reparto óptimo resulta ser $V_{opt}(\hat{\bar{Y}}_{str}) = 21.16796$. A partir de los datos recogidos en el cuadro anterior se tiene $\bar{S} = 19.62743$ y $\sum_{h=1}^{6} W_h (S_h - \bar{S})^2 = 1603.41$. Por sustitución de estos datos en [3.3.31] se tiene:

$$ER_{str\_opt./str\_prop.} \cong 1 + \frac{1}{15}\frac{1603.41}{21.16796} \cong 6$$

La eficiencia es alta debido a que existe una gran variabilidad de la varianza intraestratos - S_h^2 -, de uno a otro estrato: los estratos son muy heterogéneos respecto de esta característica.

9. Postestratificación

Sucede en ocasiones que si bien existen criterios claros de estratificación de una población, que permiten asignar cada elemento a uno a otro estrato, inequívocamente; estos criterios no se aplican sino después de que la muestra ha sido seleccionada, de modo que es la muestra la que se estratifica, no la población (frecuentemente para reducir el coste de las estimaciones) . A este plan de muestreo consistente en seleccionar a priori una muestra aleatoria simple, que es estratificada a posteriori, se le denomina postestratificación. Se asume que la proporción de elementos de la población en cada estrato - W_h - es conocida.

El número n_h de elementos de la muestra aleatoria simple de tamaño n, que caen en el estrato h, es una variable aleatoria con distribución Hipergeométrica : $n_h \in H\left[N, n, p_h = W_h\right]$, donde N es el número de elementos de la población. El estimador de la media con muestreo aleatorio estratificado, definido en [3.3.2], sigue siendo insesgado. En efecto,

$$E\hat{\bar{Y}}_{str} = \sum_{h=1}^{L} W_h E\hat{\bar{Y}}_h$$

donde, la esperanza matemática - $E\hat{\bar{Y}}_h$ - se puede determinar en dos etapas, primero considerando fija n_h

- $E(\hat{\bar{Y}}_h | n_h)$ - y luego admitiendo que n_h varía en el muestreo. La esperanza de primer etapa es la del estimador de la media con muestreo aleatorio simple que, como se vio en el tema 3.1, es insesgado, esto es, se verifica $E(\hat{\bar{Y}}_h | n_h) = \bar{Y}_h$. Puesto que $\bar{Y}_h$ no depende de n_h, se verifica $\underset{n_h}{E}\, \bar{Y}_h = \bar{Y}_h$. De donde:

$$E\hat{\bar{Y}}_h = \underset{n_h}{E}\, E(\hat{\bar{Y}}_h | n_h) = \bar{Y}_h$$

$$E\hat{\bar{Y}}_{str} = \sum_{h=1}^{L} W_h \bar{Y}_h = \bar{Y}$$

y, por tanto, $\hat{\bar{Y}}_{str}$ es un estimador insesgado de $\bar{Y}$.

La varianza del estimador con muestreo postestratificado, se puede, asimismo, determinar en las mismas dos etapas que la esperanza matemática: primero considerando fijo un determinado reparto de la muestra n entre los estratos $\{n_h\}$ y admitiendo después que dicho reparto varía de una muestra a otra:

$$V_{post}(\hat{\bar{Y}}_{str}) = \underset{\{n_h\}}{E}\, V(\hat{\bar{Y}}_{str} | \{n_h\}) + \underset{\{n_h\}}{V}\, E(\hat{\bar{Y}}_{str} | \{n_h\})$$

Vimos que, para cualquier reparto $\{n_h\}$ de la muestra entre los estratos (de modo que la muestra no sea nula en ninguno de los estratos), $\hat{\bar{Y}}_{str}$ es un estimador insesgado de la media poblacional $\bar{Y}$, esto es, se verifica: $E(\hat{\bar{Y}}_{str}|\{n_h\}) = \bar{Y}$ y, puesto que $\bar{Y}$ no depende del reparto $\{n_h\}$, resulta:

$$V_{\{n_h\}} E(\hat{\bar{Y}}_{str}|\{n_h\}) = V_{\{n_h\}} \bar{Y} = 0, \quad \text{de donde:}$$

$$V(\hat{\bar{Y}}_{str}) = E_{\{n_h\}} V(\hat{\bar{Y}}_{str}|\{n_h\})$$

La varianza de $\hat{\bar{Y}}_{str}$ sí depende del reparto de la muestra, como puede observarse en [3.3.4], a partir de donde se tiene:

$$V_{post}(\hat{\bar{Y}}_{str}) = E_{\{n_h\}} V(\hat{\bar{Y}}_{str}|\{n_h\}) = \sum_{h=1}^{L} W_h^2 S_h^2 E(\frac{1}{n_h}) - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

donde,

$$E(\frac{1}{n_h}) \cong \frac{1}{nW_h}\left[1 + (1-f)\frac{1}{n}\frac{(1-W_h)}{W_h}\right]$$

de modo que,

$$V_{post}(\hat{\bar{Y}}_{str}) = (1-f)\frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 + (1-f)\frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2$$

donde el primer sumando del segundo miembro es la varianza de $\hat{\bar{Y}}_{str}$ con muestreo aleatorio estratificado y reparto proporcional, definida en [3.3.8]:

$$V_{post}(\hat{\bar{Y}}_{str}) = V_{prop}(\hat{\bar{Y}}_{str}) + (1-f)\frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2 \qquad [3.3.51]$$

El muestreo aleatorio postestratificado es menos preciso que el muestreo aleatorio estratificado. Las diferencias de precisión entre ambos se reducen a medida que el tamaño de la muestra crece: cuando $n \to \infty$ entonces $V_{post}(\hat{\bar{Y}}_{str}) \to V_{prop}(\hat{\bar{Y}}_{str})$.

10. Estimador de la proporción

Si la variable Y es cualitativa (toma sólo dos valores, el valor 1 si el elemento $i-esimo$ del estrato "h" presenta el atributo en cuestión y el valor cero en otro caso), su media coincide con la proporción y su total con el del número de elementos de la población que poseen un determinado atributo. En consecuencia, el estimador de la proporción es el de la media y el del número de efectivos que poseen un determinado atributo es el estimador del total.

Estimador de la proporción.

Se tiene como un caso particular del estimador definido en [3.3.2]. Escribiremos esa expresión cambiando sólo la notación para adecuarla a la usual para proporciones:

$$\hat{P}_{str} = \sum_{h=1}^{L} W_h \hat{P}_h \qquad [3.3.52]$$

donde $\hat{P}_h$ es el estimador de P_h, definido a partir de la expresión [3.1.22] aplicada al estrato $h; \{h = 1,2,....,L\}$, considerado como una población en sí mismo: $\hat{P}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$

donde $\sum_{i=1}^{n_h} Y_{hi}$ es el número de elementos de la muestra que poseen el atributo en cuestión y, por tanto, $\hat{P}_h$ es la proporción de elementos que, en la muestra, poseen dicho atributo. $\hat{\bar{Y}}_{str}$ es un estimador insesgado de $\bar{Y}$, como quiera que sea la variable Y -cualitativa o cuantitativa- , de modo que $\hat{P}_{str}$ es un estimador insesgado de P.

Varianza.

La varianza de $\hat{P}_{Str}$ es de la forma definida en [3.3.3]:

$$V(\hat{P}_{Str}) = \sum_{h=1}^{L} W_h^2 V(\hat{P}_h) \qquad [3.3.53]$$

donde $V(\hat{P}_h)$ se tiene a partir de [3.1.24], aplicada al estrato $h; \{h = 1,2,....,L\}$, considerado como una población en sí mismo:

$$V(\hat{P}_h) = (1 - f_h) \frac{1}{n_h} \frac{N_h P_h (1 - P_h)}{N_h - 1}$$

Estimador insesgado de la varianza.

Un estimador insesgado de $V(\hat{P}_{Str})$ es de la forma definida en [3.3.6]:

$$\hat{V}(\hat{P}_{Str}) = \sum_{h=1}^{L} W_h^2 \hat{V}(\hat{P}_h) \qquad [3.3.54]$$

donde $\hat{V}(\hat{P}_h)$ se tiene a partir de [3.1.25], aplicada al estrato $h; \{h = 1,2,....,L\}$, considerado como una población en sí mismo:

$$\hat{V}(\hat{P}_h) = (1 - f_h) \frac{p_h (1 - p_h)}{n_h - 1}$$

Reparto proporcional y reparto óptimo.

El reparto proporcional se define en [3.3.7] y la varianza correspondiente se tiene a partir de [3.3.8], sin más que sustituir S_h^2 por la definida en [3.1.23] aplicada a cada estrato como si de una población en sí misma se tratara. El reparto óptimo se tiene en [3.3.12] o [3.3.16], sin más que sustituir S_h como acabamos de indicar. La varianza correspondiente al reparto óptimo se tiene en [3.3.15] o [3.3.18].

Error de muestreo. Intervalo de confianza

Si el tamaño de la muestra es grande en cada estrato, ($n_h \geq 30$) , la distribución en el muestreo de $\hat{P}_h$ (la de $n_h \hat{P}_h$ es una Binomial de parámetro P_h), puede ser aproximada por una Normal de media P_h y varianza $V(\hat{P}_h)$. En ese caso, la distribución de $\hat{P}_{Str}$ puede ser aproximada por una Normal de media P y varianza $V(\hat{P}_{Str})$.

Un intervalo de confianza aproximado para P , con un nivel de confianza $(1-\alpha)$, es:

$$IC(P,(1-\alpha)) \equiv \hat{P}_{Str} \pm U_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{P}_{Str})} \qquad [3.3.55]$$

donde $U_{1-\frac{\alpha}{2}}$ es el cuantíl de orden $(1-\frac{\alpha}{2})$ en la distribución Normal - $N(0,1)$ -

Con una probabilidad $(1-\alpha)$, se verifica:

$$\left|\hat{P}_{Str} - P\right| \leq U_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{P}_{Str})}$$

El error absoluto máximo probable, con una probabilidad $(1-\alpha)$, es. $U_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{P}_{Str})}$.

Tamaño de la muestra en función de la precisión deseada

Es posible determinar el tamaño de muestra mínimo necesario para alcanzar una precisión deseada en las estimaciones. El nivel de precisión deseado, puede ser especificado fijando un límite de tolerancia –*V*- para la varianza del estimador $V(\hat{P})$, o bien un límite –*d*- para el error de estimación absoluto $\left|\hat{P}_{Str} - P\right|$ o un límite –*r*- para el error relativo $\frac{\left|\hat{P}_{Str} - P\right|}{P}$.

(i) Límite de tolerancia para la varianza del estimador : $V(\hat{P}_{Str}) \leq V$

El tamaño de la muestra necesaria se tiene en [3.3.42], para el reparto óptimo y en [3.3.43], para el proporcional, sin más que sustituir S_h^2 por la definida en [3.1.23] aplicada a cada estrato como si de una población en sí misma se tratara.

(ii) Límite de tolerancia d para el error absoluto: $\left|\hat{P}_{Str} - P\right| \leq d$ **.**

El tamaño de la muestra necesaria se tiene en [3.3.46], para el reparto óptimo y en [3.3.47], para el proporcional, sin más que sustituir S_h^2 por la definida en [3.1.23] aplicada a cada estrato como si de una población en sí misma se tratara.

(iii) Límite de tolerancia r para el error relativo: $\dfrac{\left|\hat{P}_{Str} - P\right|}{P} \leq r$

El tamaño de la muestra necesaria se tiene en [3.3.49], para el reparto óptimo y en [3.3.50], para el proporcional, sin más que sustituir S_h^2 por la definida en [3.1.23] aplicada a cada estrato como si de una población en sí misma se tratara.

El valor de P es desconocido. Puede ser aproximado a partir de la información auxiliar disponible , por ejemplo de trabajos previos sobre la misma población o debe ser estimado a partir de una encuesta piloto, haciendo uso del estimador definido en [3.1.25]. El valor máximo de S_h^2 corresponde a $P_h = 0.5$ .El valor de n correspondiente a este caso – el peor de los casos – es un límite máximo para el tamaño de muestra necesario para alcanzar la precisión requerida: en otros términos, con una probabilidad $(1-\alpha)$ la precisión conseguida con el valor de n correspondiente a $S_h{}^2 = 0.25\dfrac{N_h}{N_h - 1}$ es al menos igual a la requerida.

11.Reparto óptimo de la muestra para propósitos múltiples

En el epígrafe 3 b) nos interesamos en el problema del reparto óptimo de la muestra entre los estratos, para una sola variable en estudio. En este epígrafe consideramos el problema, más general, en que la muestra se diseña para estudiar múltiples variables.

Se consideran K variables en estudio, de las que se desea estimar la media y el total de cada una de ellas, a partir de una sola muestra aleatoria estratificada. Se considera una estratificación $\Im$ determinada de los elementos de la población en L estratos. Sea $\Re$ un reparto cualquiera de la muestra entre los estratos, para el que el número de efectivos correspondientes al estrato h es n_h (h = 1, 2, ...L).

La varianza del estimador $\hat{\bar{Y}}_{jstr}$ de la media $\bar{Y}_j$ de la variable $j; \{j = 1,2,...,K\}$ puede escribirse así:

$$V(\hat{\bar{Y}}_{jstr}) = V_{j0} + V_j \qquad [3.3.56]$$

donde:

$$V_j = \sum_{h=1}^{L} \frac{V_{jh}^2}{n_h}$$

$$V_{j0} = \sum_{h=1}^{L} V_{jh}^2 / N_h$$

$$V_{jh} = W_h S_{jh}$$

Sea C_{hj} el coste unitario (coste por unidad de muestreo) de observar la característica "j" (j = 1, 2,... K) en el estrato "h".

El coste total de estimación de $\overline{Y}_j$ puede especificarse así:

$$C_j(\{n_h\}) = C_{j0} + C_j = C_{j0} + \sum_{h=1}^{L} C_{jh} n_h \qquad [3.3.57]$$

donde C_{j0} es un coste fijo independiente de n_h.

El reparto óptimo para la estimación de $\overline{Y}_j$ es aquel n_h (h = 1, 2,.. L) que minimiza:

$$V(\hat{\overline{Y}}_{jstr}) C_j(\{n_h\}) = (V_j + V_{j0})(C_j + C_{j0}) \qquad [3.3.58]$$

cuando se fija el nivel de precisión $V(\hat{\overline{Y}}_{jstr}) = V_f^{(j)}$ o el coste de la estimación $C_j(\{n_h\}) = C_f^{(j)}$. Este n_h (h = 1, 2,.. L) coincide con el que minimiza a $V_j C_j$ y viene dado por la siguiente expresión:

$$n_{jh}^* = \frac{V_{jh} / \sqrt{C_{jh}}}{\sum_{h=1}^{L} V_{jh} \sqrt{C_{jh}}} C_{jf} \qquad [3.3.59]$$

cuando $C_j = C_j\{n_h\} - C_{j0} = C_{jf}$

El $V_j^* C_j^*$ mínimo que corresponde a este óptimo n_{jh}^* es por sustitución de [3.3.59] en [3.3.56] y en [3.3.57]:

$$V_j^* C_j^* = (\sum_{h=1}^{L} V_{jh} \sqrt{C_{hj}})^2 \qquad [3.3.60]$$

Fijando el coste $C_j = C_{jf}$, lo que implica $C_j^* = C_{jf}$, la varianza mínima correspondiente al reparto óptimo resulta ser:

$$V_j^* = (\sum_{h=1}^{L} V_{jh} \sqrt{C_{jh}})^2 / C_{jf} \qquad [3.3.61]$$

Y fijada la varianza $V_j = V_{jf}$, el coste mínimo correspondiente al reparto óptimo resulta ser:

$$C_j^* = (\sum_{h=1}^{L} V_{jh} \sqrt{C_{jh}})^2 / V_{jf}$$

Función de pérdida [Kish (1976)]

Sea $Q_j(\Re)$ el incremento relativo que experimenta $V_j C_j$ con el reparto $\Re = \{n_h; h = 1,2,....,L\}$ respecto del reparto óptimo $\Re_j^* = \{n_{jh}^*; h = 1,2,....,L\}$:

$$Q_j(\Re) = \frac{V_j C_j - V_j^* C_j^*}{V_j^* C_j^*} \qquad [3.3.62]$$

$Q_j(\Re)$ es una función de $\Re$, que asocia a cada posible reparto $\Re$ una medida de la pérdida de precisión en la estimación de la media $\overline{Y}_j$ de la variable "j" (j = 1, 2,....,K), fijado el coste de la estimación [3.3. $C_j = C_{jf} \Rightarrow C_j^* = C_{jf}$] o bien una medida del incremento del coste de la estimación, fijado el nivel de precisión [3.3. $V_j = V_{jf} \Rightarrow V_j^* = V_{jf}$]; respecto del reparto óptimo $\Re_j^*$. $Q_j(\Re)$, es no negativa y verifica $Q_j(\Re_{j'}) \geq 0; \forall j \neq j'$.

Se define la función de pérdida:

$$Q(\Re) = \sum_{j=1}^{P} \frac{V_j C_j}{V_j^* C_j^*} \qquad [3.3.63]$$

que fijado el coste $C_j = C_{jf}$ se reduce a:

$$Q(\Re / C_j = C_{jf}) = \sum_{j=1}^{P} \frac{V_j}{V_j^*} \qquad [3.3.64]$$

y fijada la precisión $V_j = V_{jf}$ se reduce a:

$$Q(\Re / V_j = V_{jf}) = \sum_{j=1}^{P} \frac{C_j}{C_j^*} \qquad [3.3.65]$$

Asignemos el peso J_j a la variable "j", para denotar la importancia relativa de la pérdida de precisión en la estimación de $\overline{Y}_j$ y hagamos $J_j = \frac{1}{V_j^*}$. La función de pérdida , fijado el coste de las estimaciones, en [3.3.40], puede entonces escribirse de la siguiente forma:

$$Q(\Re / C_j = C_{jf}) = \sum_{j=1}^{P} J_j V_j \qquad [3.3.66]$$

Obsérvese que, fijado el coste de la estimación, la función de pérdida es una medida de la varianza del conjunto de las K variables en estudio: es una suma ponderada de las varianzas de los estimadores de la media de cada variable.

Sustituyendo V_j de [3.3.56], se tiene:

$$Q(\Re / C_j = C_{jf}) = \sum_{h=1}^{L} \frac{Z_h^2}{n_h} \qquad [3.3.67]$$

donde,

$$Z_h^2 = \sum_{j=1}^{K} J_j V_{hj}^2$$

Por otra parte, el coste de la estimación sería, para la variable "j":

$$C_j = \sum_{h=1}^{L} C_{hj} n_h$$

y para el conjunto de las K variables,

$$C = \sum_{j=1}^{K} C_j = \sum_{h=1}^{L} C_h n_h$$

donde,

$$C_h = \sum_{j=1}^{K} C_{jh}$$

es el coste de observación de las "K" variables en estudio, por unidad de muestreo, en el estrato "h".

Reparto óptimo.-

El reparto óptimo para la estimación de la media de todas y cada una de las K variables en estudio -$\overline{Y}_j; j = 1,2,...,K$ - es aquel $\Re = \{n_h; h = 1,2,....,L\}$ que minimiza la función de pérdida cuando se fija el coste total de la estimación $C = C_f$ o la pérdida de precisión $Q(\Re) = Q_f$, esto es, el mínimo de

$$Q(\Re)C = (\sum_{h=1}^{L} \frac{Z_h^2}{n_h})(\sum_{h=1}^{L} C_h n_h) \qquad [3.3.68]$$

cuando se fija $C = C_f$ o $Q(\Re) = Q_f$.

El reparto óptimo es:

$$n_h^{**} = D\frac{Z_h}{\sqrt{C_h}} \qquad [3.3.69]$$

donde, fijado el coste total de las estimaciones $C = C_f$:

$$D = \frac{C_f}{\sum_{h=1}^{L} Z_h \sqrt{C_h}}$$

y fijada la pérdida de precisión $Q(\Re) = Q_f$:

$$D = \frac{\sum_{h=1}^{L} Z_h \sqrt{C_h}}{Q_f} \qquad [3.3.70]$$

La pérdida mínima de precisión para un coste dado $C = C_f$ es:

$$Q_{\text{mín imo}} = \frac{(\sum_{h=1}^{L} Z_h \sqrt{C_h})^2}{C_f} \qquad [3.3.71]$$

El coste mínimo necesario para un pérdida límite $Q = Q_f$ es:

$$C_{\text{mín imo}} = \frac{(\sum_{h=1}^{L} Z_h \sqrt{C_h})^2}{Q_f} \qquad [3.3.72]$$

Un caso particular importante es aquel en el que el coste por estrato es el mismo en todos los estratos

$$C_h = c_0 ; \forall h = 1,2,...,L$$

de manera que el tamaño de muestra que es posible observar con un coste fijo C_f es $n = {C_f}/{c_0}$.

El reparto óptimo correspondiente a este caso es, por sustitución en [3.3.69]:

$$n_h^{**} = \frac{Z_h}{\sum_{h=1}^{L} Z_h} n = \frac{\sqrt{\sum_{j=1}^{P} J_j V_{jh}^2}}{\sum_{h=1}^{L} \sqrt{\sum_{j=1}^{P} J_j V_{jh}^2}} n \qquad [3.3.73]$$

Tamaño de muestra para una precisión deseada

Consideremos la función de pérdida definida en [3.3.67] , en la que, tras sustituir el tamaño óptimo de [3.3.73], resulta:

$$n = \frac{(\sum_{h=1}^{L} Z_h)^2}{Q}$$

donde la función de pérdida Q es una suma ponderada de las varianzas de los estimadores de las medias:

$$Q = \sum_{j=1}^{K} J_j V(\hat{\bar{Y}}_{jstr})$$

Si fijamos un límite máximo a la varianza de cada estimador: $V(\hat{\bar{Y}}_{jstr}) \leq V_d(\hat{\bar{Y}}_{jstr})$, de modo que Q ha de ser menor o igual que $Q_d = \sum_{j=1}^{P} J_j V_d(\hat{\bar{Y}}_{str})$, el tamaño de muestra necesario para que $Q \leq Q_d$ es:

$$n \geq \frac{(\sum_{h=1}^{L} \sqrt{\sum_{j=1}^{K} J_j V_{hj}^2})^2}{Q_d}$$

[3.3.74]

el valor de n que resulta de [3.3.74], satisface la ecuación [3.3.72] para $C_h = c_0$ y $Q_f = Q_d$

Referencias

Cochran, W.G. (1977): Sampling techniques. Wiley. (Traducción: Técnicas de muestreo. C.E.C.S.A.)

Hansen, M. H., Hurwitz, W. N., Madow, W. G. (1953): Sample Survey Methods and Theory. Wiley.

Jessen R. J.(1978): Statistical Survey Techniques. Wiley.

Sukhatme, P.V. et. al.(1984): Sampling Theory of Surveys with Applications. Iowa State University Press.

Thompson (1992): Sampling. Wiley.

Yates F. R. S. (1981): Sampling methods for censuses and surveys. Academic Press.