# Teoría de Muestreo

Yves Tillé
Groupe de Statistique, Université de Neuchâtel
Espace de l'Europe 4, Case postale 1825, 2002 Neuchâtel , Suisse
email : yves.tille@unine.ch

18 de enero de 2005

# Índice general

# Capítulo 1

# Población, diseño muestral, y estimación

## 1.1. Población finita

El objetivo es estudiar una población finita

$$U = \{1, \ldots, N\}$$

de tamaño $N$.
La variable de interés $y$ toma el valor $y_k, k \in U$.

Vamos a estudiar una función de interés de los $y_k$,

$$\theta = f(y_1, \ldots, y_k, \ldots, y_N).$$

El total y la media

$$Y = \sum_{k \in U} y_k, \text{ e } \overline{Y} = \frac{1}{N} \sum_{k \in U} y_k.$$

La varianza

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \overline{Y})^2.$$

La cuasivarianza

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \overline{Y})^2.$$

## 1.2. Plan de muestreo

Una muestra $s$ es un subconjunto de la población

$$s \subset U.$$

Un diseno muestral $p(s)$ es una distribución de probabilidad sobre todas las muestras posibles

$$\sum_{s \subset U} p(s) = 1.$$

La muestra aleatoria $S$ toma el valor $s$ con la probabilidad

$$Pr(S = s) = p(s).$$

Las variables indicadoras son definidas por :

$$I_k = \begin{cases} 1 & \text{si la unidad } k \in S \\ 0 & \text{si la unidad } k \notin S \end{cases}$$

La probabilidad de inclusión

$$\pi_k = E(I_k) = Pr(k \in S) = \sum_{s \ni k} p(s).$$

La probabilidad de inclusión de segundo orden:

$$\pi_{k\ell} = E(I_k I_\ell) = Pr(k \text{ y } \ell \in S) = \sum_{s \ni k,\ell} p(s).$$

Además

$$\Delta_{k\ell} = Cov(I_k, I_\ell) \begin{cases} \pi_k(1-\pi_k) & \text{si } k = \ell \\ \pi_{k\ell} - \pi_k \pi_\ell & \text{si } k \neq \ell \end{cases}$$

Si el diseño muestral es de tamaño fijo, entonces

$$\sum_{k \in U} \pi_k = n$$

$$\sum_{\ell \in U} \pi_{k\ell} = n\pi_k \qquad (\text{con } )\pi_{kk} = \pi_k.$$

## 1.3. El estimador de Horvitz-Thompson

El estimador de Horvitz-Thompson viene dado por

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k},$$

e

$$\widehat{\overline{Y}}_\pi = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

El estimador de Horvitz-Thompson es insesgado, si $\pi_k > 0, k \in U$,

$$\begin{aligned} E\left(\widehat{Y}_\pi\right) &= E\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) \\ &= E\left(\sum_{k \in U} \frac{y_k}{\pi_k} I_k\right) \\ &= \sum_{k \in U} \frac{y_k}{\pi_k} E\left(I_k\right) \\ &= \sum_{k \in U} \frac{y_k}{\pi_k} \pi_k \\ &= \sum_{k \in U} y_k \\ &= Y. \end{aligned}$$

La varianza del estimador de Horvitz-Thompson es

$$Var\left(\hat{Y}_\pi\right) = \sum_{k \in U} \frac{y_k^2}{\pi_k^2} \pi_k(1-\pi_k) + \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}. \tag{1.1}$$

Se puede demostrar que con un tamaño fijo de muestra

$$Var\left(\hat{Y}_\pi\right) = \frac{-1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell}. \tag{1.2}$$

La varianza puede estimarse por

$$\widehat{Var}\left(\hat{Y}_\pi\right) = \sum_{k\in S} \frac{y_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}} \frac{y_k y_\ell}{\pi_k \pi_\ell}\frac{\Delta_{k\ell}}{\pi_{kl}}. \tag{1.3}$$

Si el plan es de tamaño fijo,

$$\widehat{Var}\left(\hat{Y}_\pi\right) = \frac{-1}{2}\sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{1.4}$$

## 1.4. Estimación de $N$

Como $N$ es un total

$$N = \sum_{k\in U} 1,$$

podemos estimar $N$ por el estimador de Horvitz-Thompson

$$\widehat{N}_\pi = \sum_{k\in S}\frac{1}{\pi_k}.$$

## 1.5. Mala propiedad del estimador de Horvitz-Thompson

El estimador de Horvitz-Thompson tiene una mala propiedad, cuando la variable es constante, $y_k = C$

$$\widehat{\overline{Y}}_\pi = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{k\in S}\frac{C}{\pi_k} = C\frac{1}{N}\sum_{k\in S}\frac{1}{\pi_k} = C\frac{\widehat{N}_\pi}{N}$$

## 1.6. El problema de los elefantes de Basu (1971)

The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh ? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take 50 y (where y is the present weight of Sambo) as an estimate of the total weight Y = Y1 + Y2 + . . . + Y50 of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive samplings plan. "How can you get an unbiased estimate of Y this way ?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why ? The estimate ought to be 50y of course," says the owner. Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is 100y/99 and not 50y." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been 4900y, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became teacher of statistics!).

# Capítulo 2

# Muestreo simple

## 2.1. Muestreo simple sin reemplazamiento (o muestro aleatorio simple m.a.s.)

**Definición 1** *Un diseño muestral es aleatorio simple si todas las muestras de mismo tamaño tienen la misma probabilidad de ser seleccionadas.*

Existe solamente un solo plan simple de tamaño fijo.

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{si } \#s = n \\ 0 & \text{en caso contrario ,} \end{cases}$$

donde

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

$$\pi_k = \sum_{s \ni k} p(s) = \sum_{s \ni k} \binom{N}{n}^{-1} = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}, \text{ para todo } k \in U.$$

Probabilidades de inclusión del segundo orden :

$$\pi_{k\ell} = \sum_{s \ni k,\ell} p(s) = \sum_{s \ni k,\ell} \binom{N}{n}^{-1} = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)},$$

para todos $k \neq \ell \in U$. Luego tenemos,

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k\pi_\ell = \dfrac{n(n-1)}{N(N-1)} - \dfrac{n^2}{N^2} = -\dfrac{n(N-n)}{N^2(N-1)} & \text{si } k \neq \ell \\ \pi_k(1-\pi_k) = \dfrac{n}{N}\left(1 - \dfrac{n}{N}\right) = \dfrac{n(N-n)}{N^2} & \text{si } k = \ell. \end{cases} \tag{2.1}$$

$$\widehat{\overline{Y}}_\pi = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{k\in S} y_k\frac{N}{n} = \frac{1}{n}\sum_{k\in S} y_k.$$

$$\widehat{Y}_\pi = \sum_{k\in S}\frac{y_k}{\pi_k} = \sum_{k\in S} y_k\frac{N}{n} = \frac{N}{n}\sum_{k\in S} y_k = N\widehat{\overline{Y}}_\pi.$$

## 2.2. La varianza del plan simple sin reemplazamiento

$$
\begin{aligned}
Var\left[\widehat{Y}_\pi\right] &= \frac{-1}{2}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2\Delta_{k\ell} && (2.2)\\
&= \frac{1}{2}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\left(\frac{y_k N}{n}-\frac{y_\ell N}{n}\right)^2\frac{n(N-n)}{N^2(N-1)} && (2.3)\\
&= \frac{N(N-n)}{n}\frac{1}{2N(N-1)}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\left(y_k-y_\ell\right)^2 && (2.4)\\
&= N^2\frac{N-n}{N}\frac{S_y^2}{n}. && (2.5)
\end{aligned}
$$

**Teorema 1** *En un m.a.s., la cuesivarianza de la población es*

$$S_y^2=\frac{1}{N-1}\sum_{k\in U}(y_k-\overline{Y})^2,$$

*y puede estimarse por*

$$s_y^2=\frac{1}{n-1}\sum_{k\in S}(y_k-\widehat{\overline{Y}}_\pi)^2$$

**Demostración**

$$
\begin{aligned}
E(s_y^2) &= E\left\{\frac{1}{n-1}\sum_{k\in S}(y_k-\widehat{\overline{Y}}_\pi)^2\right\}\\
&= E\left\{\frac{1}{2n(n-1)}\sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}}(y_k-y_\ell)^2\right\}\\
&= \frac{1}{2n(n-1)}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}(y_k-y_\ell)^2E\left(I_kI_\ell\right)\\
&= \frac{1}{2n(n-1)}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}(y_k-y_\ell)^2\frac{n(n-1)}{N(N-1)}\\
&= \frac{1}{2N(N-1)}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}(y_k-y_\ell)^2\\
&= S_y^2.
\end{aligned}
$$

□

## 2.3. Algoritmo de selección-rechazo

Fan, Muller y Rezucha (1962)y Bebbington (1975)

## 2.4. Planes simples con reemplazamiento

Selección con reemplazamiento de manera independiente $\widetilde{S}$ las unidades de la muestra son

$$\widetilde{y}_1,...,\widetilde{y}_i,...,\widetilde{y}_m$$

ALGORÍTMO 1

Cuadro 2.1: Método de selección-rechazo

| | |
|---|---|
| Definición $k, j$ : entero; $u$ : real; | |
| $k = 0$; | |
| $j = 0$; | |
| Repetir mientras $j < n$ | $u =$ variable aleatoria uniforme a$[0, 1[$; |
| | Si $u < \frac{n-j}{N-k}$ \| seleccionar la unidad $k+1$; $j = j+1$; |
| | sino pasar la unidad $k+1$; |
| | $k = k+1$. |

Los $\widetilde{y}_i$ son $m$ variables aleatorias de varianza

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in U}(y_k - \overline{Y})^2.$$

Se puede estimar $\overline{Y}$ sin sesgo por

$$\widehat{\overline{Y}}_{CR} = \frac{1}{m}\sum_{i=1}^{m}\widetilde{y}_i = \frac{1}{m}\sum_{k\in\widetilde{S}} y_k.$$

La varianza de $\widehat{\overline{Y}}_{CR}$ es

$$Var(\widehat{\overline{Y}}_{CR}) = \frac{1}{m^2}\sum_{i=1}^{m} Var(\widetilde{y}_i) = \frac{1}{m^2}\sum_{i=1}^{m}\sigma_y^2 = \frac{\sigma_y^2}{m}. \tag{2.6}$$

y puede estimarse por

$$\widetilde{s}_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\widetilde{y}_i - \widehat{\overline{Y}}_{CR})^2.$$

La varianza del estimador de la media puede estimarse por

$$\widehat{Var}(\widehat{\overline{Y}}_{CR}) = \frac{\widetilde{s}_y^2}{m}.$$

## 2.5. Comparación de los planes simples

Cuadro 2.2: Planes simples

| Plan simple | Sin reemplazamiento | Con reemplazamiento |
|---|---|---|
| Tamaño de la muestra | $n$ | $m$ |
| Estimador de la media | $\widehat{\overline{Y}}_{SR} = \frac{1}{n}\sum_{k\in S} y_k$ | $\widehat{\overline{Y}}_{CR} = \frac{1}{m}\sum_{k\in \widetilde{S}} y_k$ |
| Varianza del estimador | $Var\left(\widehat{\overline{Y}}_{SR}\right) = \frac{(N-n)}{nN} S_y^2$ | $Var\left(\widehat{\overline{Y}}_{CR}\right) = \frac{\sigma_y^2}{m}$ |
| Esperanza de la varianza | $E\left(s_y^2\right) = S_y^2$ | $E\left(\widetilde{s}_y^2\right) = \sigma_y^2$ |
| Estimador de la varianza | $\widehat{Var}\left(\widehat{\overline{Y}}_{SR}\right) = \frac{(N-n)}{nN} s_y^2$ | $\widehat{Var}\left(\widehat{\overline{Y}}_{CR}\right) = \frac{\widetilde{s}_y^2}{m}$ |

**Ejercicio 1**
Seleccione una muestra de tamaño 4 en una población de tamaño 10 según un plan simple sin reemplazamiento con el método de selección-rechazo. Use las realizaciones siguientes de una variable aleatoria uniforme [0, 1] :

$$0,375489 \quad 0,624004 \quad 0,517951 \quad 0,0454450 \quad 0,632912$$
$$0,246090 \quad 0,927398 \quad 0,32595 \quad 0,645951 \quad 0,178048.$$

# Capítulo 3

# Estratificación

## 3.1. Población y estratos

Población $U = \{1, ..., k, ..., N\}$ dividida en $H$ subconjuntos, $U_h, h = 1, .., H$, llamados estratos

$$\bigcup_{h=1}^{H} U_h = U \text{ y } U_h \bigcap U_i = \emptyset, h \neq i.$$

Siendo $N_h$ el tamaño del estrato $U_h$.

$$\sum_{h=1}^{H} N_h = N.$$

El objetivo es estimar

$$Y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} Y_h,$$

donde

$$Y_h = \sum_{k \in U_h} y_k.$$

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y}_h,$$

donde $\overline{Y}_h$ es la media calculada en el estrato $h$

$$\overline{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k.$$

Además, $\sigma^2_{yh}$ representa la varianza del estrato $h$

$$\sigma^2_{yh} = \frac{1}{N_h} \sum_{k \in U_h} \left(y_k - \overline{Y}_h\right)^2$$

y $S^2_{yh}$ la cuasivarianza

$$S^2_{yh} = \frac{N_h}{N_h - 1} \sigma^2_{yh}.$$

La varianza total $\sigma^2_y$ se logra por

$$\sigma^2_y = \frac{1}{N} \sum_{k \in U} (y_k - \overline{Y})^2 = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma^2_{yh} + \frac{1}{N} \sum_{h=1}^{H} N_h (\overline{Y}_h - \overline{Y})^2. \tag{3.1}$$

Esta igualdad es la descomposición clásica de la varianza, que se escribe

$$\sigma^2_y = \sigma^2_{y(intra)} + \sigma^2_{y(inter)}$$

donde $\sigma^2_{y(intra)}$ es la varianza intra-estratos

$$\sigma^2_{y(intra)} = \frac{1}{N}\sum_{h=1}^{H} N_h \sigma^2_{yh}$$

y $\sigma^2_{y(inter)}$ es la varianza inter-estratos

$$\sigma^2_{y(inter)} = \frac{1}{N}\sum_{h=1}^{H} N_h (\overline{Y}_h - \overline{Y})^2.$$

## 3.2. Muestra, probabilidad de inclusión , estimación

Un diseño muestral es estratificado si,
- en cada estrato, se selecciona una muestra simple aleatoria de tamaño fijo $n_h$
- la selección de una muestra en un estrato es independiente de selección de las muestras de los otros estratos.

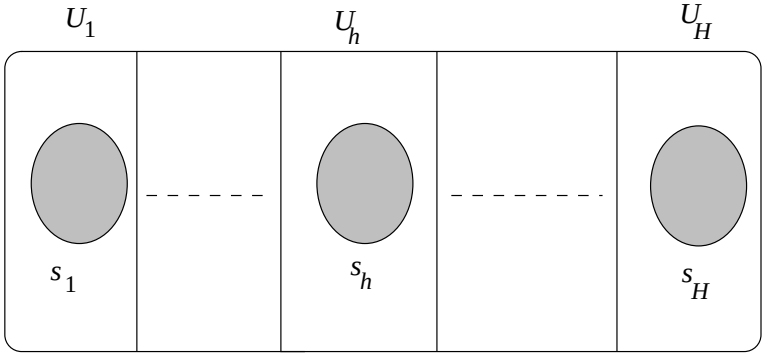
$S_h$ representa la muestra aleatoria seleccionada en el estrato $h$ con el plan $p_h(.)$, donde $p_h(s_h) = Pr(S_h = s_h)$. La muestra aleatoria total es

$$S = \bigcup_{h=1}^{H} S_h.$$

Además, de manera general $s$ representa un valor posible de $S$ donde

$$s = \bigcup_{h=1}^{H} s_h.$$

Figura 3.1: Plan estratificado




El diseño muestral global es $p(.)$ donde

$$p(s) = Pr(S = s).$$

Por la independencia de las selecciones en cada estrato, tenemos

$$p(s) = \prod_{h=1}^{H} p_h(s_h), s = \bigcup_{h=1}^{H} s_h.$$

$n_h$ representa el tamaño de la muestra en el estrato $h$, tenemos

$$\sum_{h=1}^{H} n_h = n,$$

donde $n$ es el tamaño de la muestra.

## 3.3. Probabilidad de inclusión

Si la unidad $k$ está en el estrato $h$,

$$\pi_k = \frac{n_h}{N_h}, k \in U_h.$$

Para calcular las probabilidades de inclusión de segundo orden, tenemos que separar dos casos :

- En el caso donde las unidades $k$ y $\ell$ están en el mismo estrato

$$\pi_{k\ell} = \frac{n_h(n_h-1)}{N_h(N_h-1)}, k \text{ y } \ell \in U_h.$$

- Si dos individuos $k$ y $\ell$ están en dos estratos distintos,

$$\pi_{k\ell} = \frac{n_h n_i}{N_h N_i}, k \in U_h \text{ y } \ell \in U_i.$$

Se logra

$$\Delta_{k\ell} = \begin{cases} \frac{n_h}{N_h}\frac{N_h-n_h}{N_h} & \text{si } \ell = k, k \in U_h \\ -\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)} & \text{si } k \text{ y } \ell \in U_h, k \neq \ell \\ 0 & \text{si } k \in U_h \text{ y } \ell \in U_i, h \neq i. \end{cases} \tag{3.2}$$

El $\pi$-estimador

$$\widehat{Y}_{estrat} = \sum_{k\in S}\frac{y_k}{\pi_k} = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h} y_k = \sum_{h=1}^{H}\widehat{Y}_h,$$

y

$$\widehat{\overline{Y}}_{strat} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in S_h} y_k = \frac{1}{N}\sum_{h=1}^{H} N_h\widehat{\overline{Y}}_h.$$

donde $\widehat{Y}_h$ es el estimador del total del estrato $h$

$$\widehat{Y}_h = \frac{N_h}{n_h}\sum_{k\in S_h} y_k.$$

e $\widehat{\overline{Y}}_h$ es la media de la muestra en el estrato $h$

$$\widehat{\overline{Y}}_h = \frac{1}{n_h}\sum_{k\in S_h} y_k.$$

Como la selecciones son independientes entre los estratos y que los planes son simples en los estratos :

$$Var\left(\widehat{Y}_{strat}\right) = Var\left(\sum_{h=1}^{H}\widehat{Y}_h\right) = \sum_{h=1}^{H} Var\left(\widehat{Y}_h\right) = \sum_{h=1}^{H} N_h\frac{N_h-n_h}{n_h}S_{yh}^2. \tag{3.3}$$

La varianza de este estimador puede estimarse sin sesgo por

$$\widehat{Var}\left(\widehat{Y}_{strat}\right) = \sum_{h=1}^{H} N_h\frac{N_h-n_h}{n_h}s_{yh}^2, \tag{3.4}$$

donde

$$s_{yh}^2 = \frac{1}{n_h-1}\sum_{k\in S_h}(y_k - \widehat{\overline{Y}}_h)^2, h = 1, ..., H.$$

## 3.4. Plan estratificado con afijación proporcional

Un plan estratificado tiene una afijación proporcional, si

$$\frac{n_h}{N_h} = \frac{n}{N}, h = 1, ..., N.$$

Suponemos que $n_h = nN_h/N$ son enteros. El estimador del total es

$$\widehat{Y}_{prop} = \sum_{h=1}^{H} \widehat{Y}_h = \frac{N}{n} \sum_{k \in S} y_k,$$

y el estimador de la media

$$\widehat{\overline{Y}}_{prop} = \frac{1}{N} \sum_{h=1}^{H} N_h \widehat{\overline{Y}}_h = \frac{1}{n} \sum_{k \in S} y_k,$$

donde $\widehat{\overline{Y}}_h$ es la media de la muestra en el estrato $h$ e $\widehat{Y}_h$ es el estimador del total en el estrato $h$

$$\widehat{\overline{Y}}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

La varianza del estimador del total se simplifica

$$Var(\widehat{Y}_{prop}) = \frac{N-n}{n} \sum_{h=1}^{H} N_h S_{yh}^2, \tag{3.5}$$

y la varianza del estimador de la media viene dada por :

$$Var(\widehat{\overline{Y}}_{prop}) = \frac{N-n}{nN^2} \sum_{h=1}^{H} N_h S_{yh}^2. \tag{3.6}$$

Si $N$ es grande, $S_{yh}^2 \approx \sigma_{yh}^2$.

$$Var(\widehat{\overline{Y}}_{prop}) \approx \frac{N-n}{nN^2} \sum_{h=1}^{H} N_h \sigma_{yh}^2 = \frac{N-n}{N} \frac{\sigma_{y(intra)}^2}{n}. \tag{3.7}$$

Comparación del plan estratificado con el muestro aleatorio simple.

$$Var(\widehat{\overline{Y}}_{srs}) \approx \frac{N-n}{N} \frac{\sigma_y^2}{n}. \tag{3.8}$$

La varianza del estimador de la media puede estimarse por :

$$\widehat{Var}(\widehat{\overline{Y}}_{prop}) = \frac{N-n}{nN^2} \sum_{h=1}^{H} N_h s_{yh}^2, \tag{3.9}$$

donde

$$s_{yh}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \widehat{\overline{Y}}_h)^2, h = 1, ..., H.$$

## 3.5. Plan estratificado óptimo para el total

Neyman (1934)
Se busca la afijación para los tamaños en la muestra $n_1, ..., n_h, ..., n_H$ que maximiza la varianza del estimador de Horvitz-Thompson para un tamaño de muestreo fijo.
Tenemos que minimizar

$$Var(\widehat{Y}_{strat}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} S_{yh}^2, \tag{3.10}$$

en $n_1, ..., n_h, ..., n_H$ sujeta a que

$$\sum_{h=1}^{H} n_h = n. \tag{3.11}$$

Podemos escribir la ecuación de Lagrange

$$\mathcal{L}(n_1, ..., n_H, \lambda) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} S_{yh}^2 + \lambda \left( \sum_{h=1}^{H} n_h - n \right).$$

Anulamos las derivadas parciales respecto a los $n_h$ e a $\lambda$, se logra

$$\frac{\partial \mathcal{L}}{\partial n_h} = -\frac{N_h^2}{n_h^2} S_{yh}^2 + \lambda = 0, h = 1, ..., H, \tag{3.12}$$

y

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{h=1}^{H} n_h - n = 0. \tag{3.13}$$

Luego

$$n_h = \frac{N_h}{\sqrt{\lambda}} S_{yh}, h = 1, ..., H. \tag{3.14}$$

y

$$\sum_{h=1}^{H} n_h = n = \frac{\sum_{h=1}^{H} N_h S_{yh}}{\sqrt{\lambda}}.$$

Obtenemos

$$\sqrt{\lambda} = \frac{\sum_{h=1}^{H} N_h S_{yh}}{n}. \tag{3.15}$$

y finalmente

$$n_h = \frac{n N_h S_{yh}}{\sum_{h=1}^{H} N_h S_{yh}}, h = 1, ..., H. \tag{3.16}$$

**Nota**

- Hay un problema de redondeo,
- Se puede obtener $n_h > N_h$.

## 3.6. Nota sobre la optimalidad en estratificación

Sea una población dividida en dos estratos $H = 2$ donde queremos estimar la diferencia $D = \overline{Y}_1 - \overline{Y}_2$.
El estimador

$$\widehat{D} = \widehat{\overline{Y}}_1 - \widehat{\overline{Y}}_2.$$

Como las selecciones de las muestras son independientes entre los estratos

$$Var\left(\widehat{D}\right) = Var\left(\widehat{\overline{Y}}_1\right) + Var\left(\widehat{\overline{Y}}_2\right) = \frac{N_1 - n_1}{n_1 N_1} S_{y1}^2 + \frac{N_2 - n_2}{n_2 N_2} S_{y2}^2. \tag{3.17}$$

Se minimiza (3.17) sujeta a que $n_1 + n_2 = n$ y se logra

$$n_h = \frac{S_{yh}}{\sqrt{\lambda}}, h = 1, 2,$$

donde $\lambda$ es el multiplicador de Lagrange. Como $n_1 + n_2 = n$, se logra

$$n_h = \frac{n S_{yh}}{S_{y1} + S_{y2}}, h = 1, 2.$$

## 3.7. Optimalidad y coste

El problema es estimar un total $Y$ para un coste fijado $C$. Minimizamos la expresión (3.10) sujeta a que

$$\sum_{h=1}^{H} n_h C_h = C,$$

donde $C_h$ es el coste de la entrevista en el estrato $h$.
Obtenemos

$$\begin{cases} n_h = \dfrac{N_h S_{yh}}{\sqrt{\lambda C_h}}, h = 1, ..., H, \\ \displaystyle\sum_{h=1}^{H} n_h C_h = C, \end{cases}$$

donde $\lambda$ es el multiplicador de Lagrange, y

$$n_h = \frac{C N_h S_{yh}}{\sqrt{C_h} \sum_{\ell=1}^{H} N_\ell S_{y\ell} \sqrt{C_\ell}}.$$

## 3.8. Tamaño de muestra mínimo

Otra manera de abordar el problema es buscar la afijación que da el tamaño de muestra mínimo para una varianza fijada.
Sea

$$a_h = n_h/n, h = 1, ..., H,$$

entonces

$$\sum_{h=1}^{H} a_h = 1.$$

De (3.10),

$$Var(\widehat{Y}_{strat}) = \sum_{h=1}^{H} N_h \frac{N_h - n a_h}{n a_h} S_{yh}^2. \tag{3.18}$$

Buscamos entonces un valor mínimo de (3.18) en $a_1, ..., a_H$, para un valor fijado $Var(\widehat{Y}_{strat})$ representado por $V$. Sustituyendo (3.18) en $Var(\widehat{Y}_{strat})$ por $V$, se logra

$$V = \frac{1}{n} \sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2 - \sum_{h=1}^{H} N_h S_h^2,$$

lo que se puede escribir

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{V + \sum_{h=1}^{H} N_h S_h^2}. \tag{3.19}$$

Entonces minimizamos

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{V + \sum_{h=1}^{H} N_h S_h^2}. \tag{3.20}$$

en $a_1, ..., a_H$, sujeta a que

$$\sum_{h=1}^{H} a_h = 1,$$

y después de algunos cálculos, tenemos

$$a_h = \frac{N_h S_{yh}}{\sum_{\ell=1}^{H} N_\ell S_{y\ell}}. \tag{3.21}$$

Se logra el mismo tipo de afijación . Finalmente se puede fijar el tamaño de la muestra

$$n^* = \frac{\left(\sum_{h=1}^{H} N_h S_{yh}\right)^2}{V + \sum_{h=1}^{H} N_h S_{yh}^2}.$$

**Ejercicio 2**
Queremos estimar medias para las empresas de un departamento. Las empresas son clasificadas según el volumen de negocios y son clasificadas en tres clases. Los datos de un censo son los siguientes :

| Volumen de negocios | Número de empresas |
|---|---|
| de 0 a 1 | 1000 |
| de 1 a 10 | 100 |
| de 10 a 100 | 10 |

Se quiere seleccionar una muestra de 111 empresas. Si se supone que la distribución es uniforme en cada estrato, calcule la varianza del estimador de la media del volumen de negocios para un plan con representación proporcional y para un plan estratificado óptimo.

# Capítulo 4

# Planes con conglomerados, multi-etápicos, y multi-fases

## 4.1. Planes con conglomerados

### 4.1.1. Notación y definición

La población $U = \{1, ..., k, ..., N\}$ se divide en $M$ subconjuntos, $U_i, i = 1, .., M$, llamados conglomerados

$$\bigcup_{i=1}^{M} U_i = U \text{ y } U_i \cap U_j = \emptyset, i \neq j.$$

El número $N_i$ de unidades del conglomerado $i$ se llama el tamaño del conglomerado :

$$\sum_{i=1}^{M} N_i = N,$$

donde $N$ es el tamaño de la población $U$. El total puede escribirse

$$Y = \sum_{k \in U} y_k = \sum_{i=1}^{M} \sum_{k \in U_i} y_k = \sum_{i=1}^{M} Y_i$$

y la media

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{1}{N} \sum_{i=1}^{M} \sum_{k \in U_i} y_k = \frac{1}{N} \sum_{i=1}^{M} N_i \overline{Y}_i,$$

e $Y_i$ es el total del conglomerado $i$ e $\overline{Y}_i$ la media del conglomerado $i$ :

$$Y_i = \sum_{k \in U_i} y_k, i = 1, ..., M,$$

$$\overline{Y}_i = \frac{1}{N_i} \sum_{k \in U_i} y_k, i = 1, ..., M.$$

Además, $\sigma_{yi}^2$ representa la varianza del conglomerado $i$

$$\sigma_{yi}^2 = \frac{1}{N_i} \sum_{k \in U_i} \left(y_k - \overline{Y}_i\right)^2$$

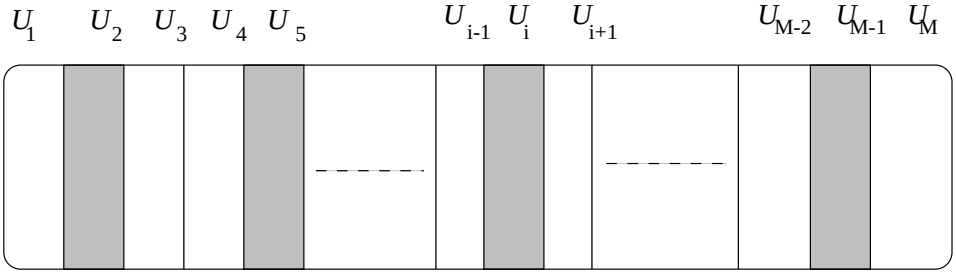
y $S_{yi}^2$ la varianza corregida

$$S_{yi}^2 = \frac{N_i}{N_i - 1} \sigma_{yi}^2.$$

Un plan es por conglomerados si

- se selecciona una muestra de conglomerados $s_I$ con un plan $p_I(s_I)$, $S_I$ representa la muestra aleatoria tal que $Pr(S_I = s_I) = p_I(s_I)$ y $m = \#S_I$, el número de conglomerados seleccionados.
- Todas las unidades de los conglomerados seleccionados son observadas :

Figura 4.1: Plan con conglomerados




La muestra aleatoria completa viene dada por

$$S = \bigcup_{i \in S_I} U_i.$$

El tamaño de $S$ es

$$n = \sum_{i \in S_I} N_i.$$

El tamaño de la muestra es generalmente aleatorio.
La probabilidad de seleccionar un conglomerado es

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I), i = 1, ..., M,$$

La probabilidad de seleccionar dos conglomerados distintos es

$$\pi_{Iij} = \sum_{s_I \ni i,j} p_I(s_I), i = 1, ..., M, j = 1, ..., M, i \neq j.$$

Si la unidad $k$ está en el conglomerado $i$, tenemos

$$\pi_k = \pi_{Ii}, k \in U_i.$$

Para las probabilidades de inclusión del segundo orden hay que separar dos casos :

- Si $k$ y $\ell$ están en el mismo conglomerado $i$,

$$\pi_{k\ell} = \pi_{Ii}, k \text{ y } \ell \in U_i.$$

- Si $k$ y $\ell$ no están en el mismo conglomerado respectivamente $i$ y $j$,

$$\pi_{k\ell} = \pi_{Iij}, k \in U_i \text{ y } \ell \in U_j, i \neq j.$$

Las condiciones de Sen-Yates-Grundy no se verifican.
En efecto, si $k$ y $\ell \in U_i$, entonces

$$\pi_k \pi_\ell - \pi_{k\ell} = \pi_{Ii}^2 - \pi_{Ii} = -\pi_{Ii}(1 - \pi_{Ii}).$$

El estimador de Horvitz-Thompson del total y de la media son

$$\widehat{Y}_\pi = \sum_{i \in S_I} \frac{Y_i}{\pi_{Ii}},$$

y

$$\widehat{\overline{Y}}_\pi = \frac{1}{N}\sum_{i\in S_I}\frac{N_i\overline{Y}_i}{\pi_{Ii}}.$$

La varianza

$$Var(\widehat{Y}_\pi) = \sum_{i=1}^{M}\frac{Y_i^2}{\pi_{Ii}}(1-\pi_{Ii}) + \sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}(\pi_{Iij}-\pi_{Ii}\pi_{Ij}). \tag{4.1}$$

Si el número de conglomerados es fijo,

$$Var(\widehat{Y}_\pi) = \frac{1}{2}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\left(\frac{Y_i}{\pi_{Ii}}-\frac{Y_j}{\pi_{Ij}}\right)^2(\pi_{Ii}\pi_{Ij}-\pi_{Iij}). \tag{4.2}$$

Estimación de la varianza

$$\widehat{Var}(\widehat{Y}_\pi)_1 = \sum_{i\in S_I}\frac{Y_i^2}{\pi_{Ii}^2}(1-\pi_{Ii}) + \sum_{i\in S_I}\sum_{\substack{j\in S_I\\j\neq i}}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}\frac{\pi_{Iij}-\pi_{Ii}\pi_{Ij}}{\pi_{Iij}}. \tag{4.3}$$

Cuando el número de conglomerados seleccionados es fijo, se puede construir otro estimador de esta varianza mediante (4.2)

$$\widehat{Var}(\widehat{Y}_\pi)_2 = \frac{1}{2}\sum_{i\in S_I}\sum_{\substack{j\in S_I\\j\neq i}}\left(\frac{Y_i}{\pi_{Ii}}-\frac{Y_j}{\pi_{Ij}}\right)^2\frac{\pi_{Ii}\pi_{Ij}-\pi_{Iij}}{\pi_{Iij}}.$$

Una aproximación practica (pero sesgada) es

$$\widehat{Var}(\widehat{Y}_\pi)_3 = \sum_{i\in S_I}\frac{c_{Ii}}{\pi_{Ii}^2}\left(Y_i-\widehat{Y_i^*}\right)^2, \tag{4.4}$$

donde

$$\widehat{Y_i^*} = \pi_{Ii}\frac{\sum_{j\in S}c_{Ij}Y_j/\pi_{Ij}}{\sum_{j\in S}c_{Ij}},$$

y donde

$$c_{Ii} = (1-\pi_{Ii})\frac{m}{m-1}.$$

En los planes por conglomerados, el estimador de Horvitz-Thompson tiene una mala propiedad.
Si la variable es constante ($y_k = C, \text{para todos } k \in U$), se logra

$$\widehat{\overline{Y}}_\pi = C\frac{1}{N}\sum_{i\in S_I}\frac{N_i}{\pi_{Ii}}.$$

En este caso, es preferible usar el razón de Hájek :

$$\widehat{\overline{Y}}_R = \left(\sum_{i\in S_I}\frac{N_i}{\pi_{Ii}}\right)^{-1}\left(\sum_{i\in S_I}\frac{Y_i}{\pi_{Ii}}\right).$$

### 4.1.2. Selección de los conglomerados con probabilidades iguales

Un plan clásico es seleccionar los conglomerados por un m.a.s. de tamaño $m$.

$$\pi_{Ii} = \frac{m}{M}, i = 1, ..., M,$$

y

$$\pi_{Iij} = \frac{m(m-1)}{M(M-1)}, i = 1, ..., M.$$

El tamaño de la muestra es aleatorio. Su esperanza es

$$E\left(n_S\right) = E\left(\sum_{i\in S_I} N_i\right) = \sum_{i\in U_I} N_i \frac{m}{M} = \frac{Nm}{M},$$

lo que permite construir el estimador de Horvitz-Thompson del total :

$$\widehat{Y} = \frac{M}{m}\sum_{i\in S_I} Y_i$$

y de la media

$$\widehat{\overline{Y}}_\pi = \frac{M}{Nm}\sum_{i\in S_I} N_i \overline{Y}_i.$$

La varianza es

$$Var(\widehat{Y}) = \frac{M-m}{M-1}\frac{M}{m}\sum_{i=1}^{M}\left(Y_i - \frac{Y}{M}\right)^2, \tag{4.5}$$

y puede estimarse sin sesgo por

$$\widehat{Var}(\widehat{Y}) = \frac{M-m}{m-1}\frac{M}{m}\sum_{i\in S_I}\left(Y_i - \frac{\widehat{Y}}{M}\right)^2. \tag{4.6}$$

### 4.1.3. El plan sistemático

El plan sistemático puede verse como un plan con conglomerados donde se selecciona un solo conglomerado.

## 4.2. Plan bietápico

### 4.2.1. Población, unidades primarias y secundarias

Sea la población $U = \{1, ..., k, ..., N\}$ compuesta de $M$ subpoblaciones, $U_i, i = 1, ..., M$, llamadas unidades primarias. Cada unidad $U_i$ se compone de $N_i$ unidades secundarias , tenemos

$$\sum_{i=1}^{M} N_i = N,$$

donde $N$ es el tamaño de la población $U$.

De manera general, un plan bietápico se define de la manera siguiente :

- Una muestra de unidades primarias es seleccionada con un plan $p_I(s_I)$. $S_I$ representa la muestra aleatoria tal que $Pr(S_I = s_I) = p_I(s_I)$ y $m = \#S_I$ ;
- Si una unidad primaria $U_i$ se selecciona en la primera etapa $U_i$, se selecciona una muestra $s_i$ de unidades secundarias con el plan $p_i(s_i)$. $S_i$ representa la muestra aleatoria de unidades primarias seleccionadas de manera que $Pr(S_i = s_i) = p_i(s_i)$ y $n_i = \#S_i$.

  Los planes bietápicos tienen que tener las dos propiedades de invarianza y de independencia. La invarianza significa que los planes $p_i(s_i)$ de la segunda etapa no dependen de lo que pasó en la primera etapa, entonces $Pr(S_i = s_i) = Pr(S_i = s_i|S_I)$. La independencia significa que las selecciones de la segunda etapa son independientes las unas de las otras (como en estratificación).

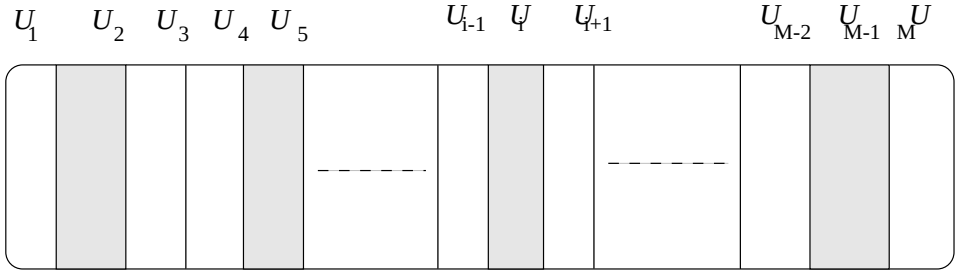
La muestra aleatoria viene dada por

$$S = \bigcup_{i\in S_I} S_i.$$

Para la variable $y$, el total se escribe

$$Y = \sum_{k\in U} y_k = \sum_{i=1}^{M}\sum_{k\in U_i} y_k = \sum_{i=1}^{M} Y_i,$$

Figura 4.2: Plan bietápico




donde $Y_i$ es el total calculado en la unidad primaria $i$

$$Y_i = \sum_{k \in U_i} y_k, i = 1, ..., M.$$

Del mismo modo, la media calculada en la población se escribe

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{1}{N} \sum_{i=1}^{M} \sum_{k \in U_i} y_k = \frac{1}{N} \sum_{i=1}^{M} N_i \overline{Y}_i,$$

donde $\overline{Y}_i$ es la media calculada en la unidad primaria $i$

$$\overline{Y}_i = \frac{1}{N_i} \sum_{k \in U_i} y_k, i = 1, ..., M.$$

$\sigma_{yi}^2$ es la varianza en $U_i$

$$\sigma_{yi}^2 = \frac{1}{N_i} \sum_{k \in U_i} \left( y_k - \overline{Y}_i \right)^2,$$

y $S_{yi}^2$ la varianza corregida

$$S_{yi}^2 = \frac{N_i}{N_i - 1} \sigma_{yi}^2.$$

La muestra aleatoria total està formada par :

$$S = \bigcup_{i \in S_I} S_i.$$

El tamaño de $S$ es

$$n = \sum_{i \in S_I} n_i.$$

El tamaño de $S$ es generalmente aleatorio.
Podemos definir
- $\pi_{Ii}$ : La probabilidad de seleccionar la unidad primaria $U_i$.
- $\pi_{Iij}$, La probabilidad de inclusión del segundo orden para dos unidades primarias $U_i$ y $U_j$.
Estas probabilidades vienen del plan $p_I(s_I)$. Al final, tenemos

$$\Delta_{Iij} = \begin{cases} \pi_{Iij} - \pi_{Ii}\pi_{Ij} & \text{si } i \neq j \\ \pi_{Ii}(1 - \pi_{Ii}) & \text{si } j = i. \end{cases} \tag{4.7}$$

- $\pi_{k|i}$, la probabilidad de seleccionar la unidad $k$ dado que $i$ ha sido seleccionada
- $\pi_{k\ell|i}$ la probabilidad de seleccionar conjuntamente $k$ y $\ell$ dado que $i$ ha sido seleccionada.

$$\Delta_{k\ell|i} = \begin{cases} \pi_{k\ell|i} - \pi_{k|i}\pi_{\ell|i} & \text{si } k \neq \ell \\ \pi_{k|i}(1 - \pi_{k|i}) & \text{si } k = \ell \end{cases}, i = 1, ..., M. \tag{4.8}$$

La probabilidad de inclusión de la unidad es

$$\pi_k = \pi_{Ii}\pi_{k|i}, k \in U_i.$$

Para las probabilidades de inclusión del segundo orden, hay que separar dos casos :

- Si dos unidades $k$ y $\ell$ están en la misma unidad primaria $U_i$,

$$\pi_{k\ell} = \pi_{Ii}\pi_{k\ell|i}.$$

- Si dos unidades $k$ y $\ell$ están en dos unidades primarias distintas , $U_i$ y $U_j$,

$$\pi_{k\ell} = \pi_{Iij}\pi_{k|i}\pi_{\ell|j}.$$

### 4.2.2. El estimador de Horvitz-Thompson

El estimador de Horvitz-Thompson del total es

$$\widehat{Y}_\pi = \sum_{i\in S_I}\sum_{k\in S_i}\frac{y_k}{\pi_{Ii}\pi_{k|i}} = \sum_{i\in S_I}\frac{\widehat{Y}_i}{\pi_{Ii}},$$

donde $\widehat{Y}_i$ es el estimador de Horvitz-Thompson de $Y_i$

$$\widehat{Y}_i = \sum_{k\in S_i}\frac{y_k}{\pi_{k|i}},$$

y el estimador de Horvitz-Thompson de la media por

$$\widehat{\overline{Y}}_\pi = \frac{1}{N}\sum_{i\in S_I}\sum_{k\in S_i}\frac{y_k}{\pi_{Ii}\pi_{k|i}}.$$

**Teorema 2** *En un plan bietápico*

$$Var(\widehat{Y}_\pi) = V_{UP} + V_{US},$$

*donde $V_{UP}$ es la parte que se refiere a las unidades primarias*

$$V_{UP} = \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}\Delta_{Iij},$$

*$V_{US}$ es la parte que se refiere a las unidades secundarias*

$$V_{US} = \sum_{i=1}^{M}\frac{Var(\widehat{Y}_i)}{\pi_{Ii}},$$

*y*

$$Var(\widehat{Y}_i) = \sum_{k\in U_i}\sum_{\ell\in U_i}\frac{y_ky_\ell}{\pi_{k|i}\pi_{\ell|i}}\Delta_{k\ell|i}, i = 1, ..., M. \tag{4.9}$$

**Demostración**
La varianza se divide en dos partes :

$$Var\left[\widehat{Y}_\pi\right] = VarE\left[\widehat{Y}_\pi|S_I\right] + EVar\left[\widehat{Y}_\pi|S_I\right].$$

La varianza de la esperanza condicional es

$$VarE\left[\widehat{Y}_\pi|S_I\right] = VarE\left[\sum_{i\in S_I}\widehat{Y}_i\middle| S_I\right].$$

Por la propiedad de invarianza

$$E\left[\sum_{i\in S_I}\widehat{Y}_i\middle| S_I\right] = \sum_{i\in S_I}E\left[\widehat{Y}_i\middle| S_I\right] = \sum_{i\in S_I}E\left[\widehat{Y}_i\right] = \sum_{i\in S_I}\frac{Y_i}{\pi_{Ii}}.$$

Luego

$$VarE\left[\widehat{Y}_\pi|S_I\right] = Var\left[\sum_{i\in S_I}\frac{Y_i}{\pi_{Ii}}\right] = \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}\Delta_{Iij}.$$

La esperanza de la varianza condicional es

$$EVar\left[\widehat{Y}_\pi|S_I\right] = EVar\left[\sum_{i\in S_I}\frac{\widehat{Y}_i}{\pi_{Ii}}\middle| S_I\right].$$

Por las propiedades de invarianza y de independencia

$$Var\left[\sum_{i\in S_I}\frac{\widehat{Y}_i}{\pi_{Ii}}\middle| S_I\right] = \sum_{i\in S_I} Var\left[\frac{\widehat{Y}_i}{\pi_{Ii}}\middle| S_I\right] = \sum_{i\in S_I}\frac{Var\left[\widehat{Y}_i\right]}{\pi_{Ii}^2}.$$

Luego,

$$EVar\left[\widehat{Y}_\pi|S_I\right] = E\left[\sum_{i\in S_I}\frac{Var\left[\widehat{Y}_i\right]}{\pi_{Ii}^2}\right] = \sum_{i=1}^{M}\frac{Var\left[\widehat{Y}_i\right]}{\pi_{Ii}},$$

donde $Var\left[\widehat{Y}_i\right]$ es dado en (4.9). □

**Teorema 3** *En un plan bietápico*

$$\widehat{Var}_1(\widehat{Y}_\pi) = \hat{V}_{UP} + \hat{V}_{US}$$

*es un estimador insesgado de* $Var(\widehat{Y}_\pi)$*, donde* $\hat{V}_{UP}$ *es la parte de la varianza que se refiere a las unidades primarias*

$$\hat{V}_{UP} = \sum_{i\in S_I}\sum_{j\in S_I}\frac{\widehat{Y}_i\widehat{Y}_j}{\pi_{Ii}\pi_{Ij}}\frac{\Delta_{Iij}}{\pi_{Iij}},$$

*(con* $\pi_{Iii} = \pi_{Ii}$*,)* $\hat{V}_{US}$ *es la parte de la varianza que se refiere a las unidades secundarias*

$$\hat{V}_{US} = \sum_{i\in S_I}\frac{\widehat{Var}(\widehat{Y}_i)}{\pi_{Ii}},$$

*y*

$$\widehat{Var}(\widehat{Y}_i) = \sum_{k\in S_i}\sum_{\ell\in S_i}\frac{y_k y_\ell}{\pi_{k|i}\pi_{\ell|i}}\frac{\Delta_{k\ell|i}}{\pi_{k\ell|i}},$$

*con* $\pi_{kk|i} = \pi_{k|i}$*.*

**Demostración**
Como

$$E\left[\widehat{Y}_i\widehat{Y}_j|S_I\right] = \begin{cases} Var(\widehat{Y}_i) + Y_i^2 & \text{si } i = j \\ Y_iY_j & \text{si } i \neq j, \end{cases}$$

$$\begin{aligned} E[\hat{V}_{UP}] &= EE\left[\sum_{i\in S_I}\sum_{j\in S_I}\frac{\widehat{Y}_i\widehat{Y}_j}{\pi_{Ii}\pi_{Ij}}\frac{\Delta_{Iij}}{\pi_{Iij}}\middle| S_I\right] \\ &= E\left[\sum_{i\in S_I}\sum_{j\in S_I}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}\frac{\Delta_{Iij}}{\pi_{Iij}} + \sum_{i\in S_I}\frac{Var(\widehat{Y}_i)}{\pi_{Ii}^2}(1-\pi_{Iij})\right] \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_{Ii}\pi_{Ij}}\Delta_{Iij} + \sum_{i=1}^{M}Var(\widehat{Y}_i)\left(\frac{1}{\pi_{Iij}} - 1\right). \end{aligned}$$

De otra parte

$$\begin{aligned} E[\hat{V}_{US}] &= EE\left[\sum_{i\in S_I}\frac{\widehat{Var}(\widehat{Y}_i)}{\pi_{Ii}}\middle| S_I\right] \\ &= E\left[\sum_{i\in S_I}\frac{Var(\widehat{Y}_i)}{\pi_{Ii}}\right] \\ &= \sum_{i=1}^{M} Var(\widehat{Y}_i) \\ &= \sum_{i=1}^{M}\frac{Var(\widehat{Y}_i)}{\pi_{Ii}} + \sum_{i=1}^{M} Var(\widehat{Y}_i)\left(1-\frac{1}{\pi_{Ii}}\right). \end{aligned}$$

Entonces tenemos

$$E[\hat{V}_{UP}] + E[\hat{V}_{US}] = Var[\widehat{Y}_\pi].$$

□

Es importante ver que $\hat{V}_{UP}$ es un estimador sesgado de $V_{UP}$ y que $\hat{V}_{US}$ es un estimador sesgado de $V_{US}$. El estimador $\hat{V}_{UP}$ sobrestima $V_{UP}$ y prácticamente $\hat{V}_{UP}$ es a veces más grande $\hat{V}_{US}$.
De nuevo, hay un estimador más práctico, pero sesgado

$$\widehat{Var}_2(\widehat{Y}_\pi) = \sum_{i\in S_I}\frac{c_{Ii}}{\pi_{Ii}^2}\left(\widehat{Y}_i - \widehat{\widehat{Y}_i^*}\right)^2 + \sum_{i\in S_I}\frac{1}{\pi_{Ii}}\sum_{k\in S_i}\frac{c_{k|i}}{\pi_{k|i}^2}\left(y_k - \widehat{y}_k^*\right)^2, \tag{4.10}$$

donde

$$\begin{aligned} \widehat{\widehat{Y}_i^*} &= \pi_{Ii}\frac{\sum_{j\in S} c_{Ij}\widehat{Y}_j/\pi_{Ij}}{\sum_{j\in S} c_{Ij}}, \\ c_{Ii} &= (1-\pi_{Ii})\frac{m}{m-1}, \\ \widehat{y}_k^* &= \pi_{k|i}\frac{\sum_{k\in S_i} c_{k|i}y_k/\pi_{k|i}}{\sum_{k\in S_i} c_{k|i}}, \\ c_{k|i} &= (1-\pi_{k|i})\frac{n_i}{n_i-1}. \end{aligned}$$

### 4.2.3. Selección de las unidades primarias con probabilidades iguales

En las dos etapas se usa un plan simple.
Las probabilidades de inclusión para la primera etapa

$$\pi_{Ii} = \frac{m}{M}, i = 1, ..., M$$

y

$$\pi_{Iij} = \frac{m(m-1)}{M(M-1)}, i = 1, ..., M, j = 1, ..., M, i \neq j.$$

Para la segunda etapa $n_i$, La probabilidad de inclusión para todo el diseño muestral

$$\pi_k = \frac{mn_i}{MN_i}.$$

El estimador de Horvitz-Thompson es

$$\widehat{Y}_\pi = \frac{M}{m}\sum_{i\in S_I}\sum_{k\in S_i}\frac{N_i y_k}{n_i},$$

y su estimador de varianza se simplifica

$$\widehat{Var}(\widehat{Y}_\pi) = M^2\frac{M-m}{Mm}s_I^2 + \frac{M}{m}\sum_{i\in S_I} N_i^2\frac{N_i-n_i}{n_i N_i}s_i^2,$$

donde

$$s_I^2 = \frac{1}{m-1}\sum_{i\in S_I}\left(\widehat{Y}_i - \frac{\widehat{Y}_\pi}{M}\right)^2,$$

y

$$s_i^2 = \frac{1}{n_i-1}\sum_{k\in S_i}\left(y_k - \frac{\widehat{Y}_i}{N_i}\right)^2.$$

Se puede coger tamaños de muestras de unidades secundarios proporcionales a los tamaños de la población

$$n_i = n_0\frac{N_i}{N},$$

Se logra

$$\pi_{k|i} = \frac{n_0}{N}, k \in U_i.$$

Al final, la probabilidad de inclusión para todo el diseño muestral

$$\pi_k = \frac{n_0 m N_i}{MN}.$$

Este plan tiene problemas importantes. El tamaño de la muestra $n_S$ es aleatorio, y es de media

$$E(n_S) = E\left(\sum_{k\in S_I} n_i\right) = E\left(\sum_{k\in S_I} n_0\frac{N_i}{N}\right) = \sum_{k\in U_I} n_0\frac{N_i}{N}\frac{m}{M} = \frac{mn_0N_i}{N}.$$

### 4.2.4. Plan bietápico autoponderado

En la primera etapa, se selecciona las unidades primarias con probabilidades de inclusión proporcionales al tamaño de las unidades primarias

$$\pi_{Ii} = \frac{N_i}{N}m,$$

Se supone $\pi_{Ii} < 1$.
En la segunda etapa se selecciona unidades secundarias según un plan aleatorio simple sin reemplazamiento con un tamaño de muestra $n_i = n_0$ constante (en cada unidad primaria).

$$\pi_{k|i} = \frac{n_0}{N_i}.$$

La probabilidad de inclusión es

$$\pi_k = \pi_{Ii}\pi_{k|i} = \frac{N_i}{N}\frac{mn_0}{N_i} = \frac{mn_0}{N}, k \in U_i.$$

Las probabilidades de inclusión son constantes para todas las unidades primarias de la población.
El plan es de tamaño fijo.
El estimador de Horvitz-Thompson del total es :

$$\widehat{Y}_\pi = \frac{N}{n}\sum_{k\in S} y_k.$$

## 4.3. Planes multi-etápicos

Suponemos que tenemos $M$ unidades primarias y que el primer diseño muestral consiste en seleccionar $m$ unidades primarias con probabilidades de inclusión $\pi_{Ii}$ para $i = 1, ..., M$. También $S_I$ representa la muestra aleatoria de unidades primarias seleccionadas. Suponemos que en cada unidad primaria, se puede calcular el estimador de Horvitz-Thompson $\widehat{Y}_i$ del total $Y_i$ para las $m$ unidades primarias seleccionadas. El estimador de Horvitz-Thompson del total viene dado por

$$\widehat{Y}_\pi = \sum_{k\in S_I}\frac{\widehat{Y}_i}{\pi_{Ii}}.$$

Usando exactamente el mismo desarrollo que por los planes bietápicos, la varianza del estimador de Horvitz-Thompson es

$$Var(\widehat{Y}_\pi) = \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{Y_i Y_j}{\pi_{Ii}\pi_{Ij}}\Delta_{Iij}, + \sum_{i=1}^{M} \frac{Var(\widehat{Y}_i)}{\pi_{Ii}},$$

y puede estimarse sin sesgo por

$$\widehat{Var}_1(\widehat{Y}_\pi)_2 = \sum_{i\in S_I}\sum_{j\in S_I} \frac{\widehat{Y}_i\widehat{Y}_j}{\pi_{Ii}\pi_{Ij}}\frac{\Delta_{Iij}}{\pi_{Iij}} + \sum_{i\in S_I} \frac{\widehat{Var}(\widehat{Y}_i)}{\pi_{Ii}}$$

o por

$$\widehat{Var}_2(\widehat{Y}_\pi)_2 = \sum_{i\in S_I} \frac{c_{Ii}}{\pi_{Ii}^2}\left(\widehat{Y}_i - \widehat{\widehat{Y_i^*}}\right)^2 + \sum_{i\in S_I} \frac{\widehat{Var}(\widehat{Y}_i)}{\pi_{Ii}},$$

donde

$$\widehat{Y_i^*} = \pi_{Ii}\frac{\sum_{j\in S} c_{Ij}Y_j/\pi_{Ij}}{\sum_{j\in S} c_{Ij}},$$

y donde

$$c_{Ii} = (1-\pi_{Ii})\frac{m}{m-1}.$$

Conclusión : la expresión es recursiva
El estimador de Horvitz-Thompson y su estimador de varianza se escribe como una función del estimador del total y del estimador de la varianza calculada al nivel inferior.

$$\widehat{Y}_\pi = T\left(\widehat{Y}_i, \pi_{Ii}, i\in S_I\right),$$

y

$$\widehat{Var}(\widehat{Y}_\pi) = Q\left(\widehat{Y}_i, \widehat{Var}(\widehat{Y}_i), \pi_{Ii}, i\in S_I\right).$$

El estimador $\widehat{Var}(\widehat{Y}_i)$ puede igualmente escribirse como una función de las etapas inferiores.

## 4.4. Muestreo en dos fases

- En la primera fase, se selecciona una muestra con cualquier plan $p_I(S_a)$ de tamaño $n$ (eventualmente con un plan multi-etápico).
- En la segunda fase, se selecciona una muestra $S_b$ según otro diseño muestral en $S_a$ con un plan $p(s_b|S_a) = Pr(S_b = s_b|S_a)$.
El plan de segundo etapa puede depender de lo que pasó en la primera etapa. Tenemos

$$\pi_{ak} = Pr\left(k\in S_a\right),$$

$$\pi_{ak\ell} = Pr\left(k \text{ y } \ell \in S_a\right), k\neq\ell, \text{ con } \pi_{akk} = \pi_{ak},$$

$$\Delta_{ak\ell} = \begin{cases} \pi_{ak\ell} - \pi_{ak}\pi_{a\ell}, & k\neq\ell \\ \pi_{ak}(1-\pi_{ak}), & k=\ell \end{cases}$$

además

$$\pi_{bk} = Pr\left(k\in S_b|S_a\right),$$

$$\pi_{bk\ell} = Pr\left(k \text{ y } \ell \in S_b|S_a\right), k\neq\ell, \text{ con } \pi_{bkk} = \pi_{bk},$$

$$\Delta_{bk\ell} = \begin{cases} \pi_{bk\ell} - \pi_{bk}\pi_{b\ell}, & k\neq\ell \\ \pi_{bk}(1-\pi_{bk}), & k=\ell \end{cases}$$

Los $\pi_{bk}$, $\pi_{bk\ell}$ y $\Delta_{bk\ell}$ son variables aleatorias que dependen de $S_a$.
La probabilidad de inclusión de la unidad $k$ es

$$\pi_k = \pi_{ak}E(\pi_{bk}).$$

Pero esta probabilidad no puede ser calculada. Se estima el total por

$$\widehat{Y}_E = \sum_{k \in S_b} \frac{y_k}{\pi_{ak}\pi_{bk}},$$

que no es el estimador de Horvitz-Thompson, en efecto no se divide los $y_k$ por las probabilidades de inclusión. Este estimador se llama : estimador por expansión. Es insesgado. En efecto,

$$E\left(\widehat{Y}_E\right) = EE\left(\sum_{k \in S_b} \frac{y_k}{\pi_{ak}\pi_{bk}} \middle| S_a\right) = E\left(\sum_{k \in S_b} \frac{y_k}{\pi_{ak}}\right).$$

La varianza del estimador por expansión Särndal y Wretman (1987).

**Teorema 4**

$$Var\left(\widehat{Y}_E\right) = \sum_{k \in U}\sum_{\ell \in U} \frac{y_k y_\ell}{\pi_{ak}\pi_{a\ell}}\Delta_{ak\ell} + E\left(\sum_{k \in S_a}\sum_{\ell \in S_a} \frac{y_k y_\ell}{\pi_{ak}\pi_{bk}\pi_{a\ell}\pi_{b\ell}}\Delta_{bk\ell}\right),$$

*Esta varianza puede estimarse por*

$$\widehat{Var}_a\left(\widehat{Y}_E\right) = \sum_{k \in S_b}\sum_{\ell \in S_b} \frac{y_k y_\ell}{\pi_{ak}\pi_{a\ell}} \frac{\Delta_{ak\ell}}{\pi_{ak\ell}\pi_{bk\ell}} + \sum_{k \in S_b}\sum_{\ell \in S_b} \frac{y_k y_\ell}{\pi_{ak}\pi_{bk}\pi_{a\ell}\pi_{b\ell}} \frac{\Delta_{bk\ell}}{\pi_{bk\ell}}. \tag{4.11}$$

# Capítulo 5

# Muestreo con probabilidades desiguales

Brewer y Hanif, 1983, Gabler, 1990

## 5.1. Información auxiliar y probabilidades de inclusión

Variable auxiliar $x$ conocida sobre $U$.
$x$ es aproximadamente proporcional a $y$.
Selección de las unidades con probabilidades de inclusión proporcional a $x$.
Varianza

$$Var\left(\hat{Y}_\pi\right) = \frac{1}{2}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2(\pi_k\pi_\ell-\pi_{k\ell}). \tag{5.1}$$

## 5.2. Cálculo de las probabilidades de inclusión

Calculamos

$$\pi_k = \frac{x_k n}{\sum_{\ell\in U} x_\ell},\ \text{para todo } k\in U.$$

Algunos de los $\pi_k$ pueden tener $\pi_k > 1$.
Estas unidades son incluidas en la muestra con una probabilidad 1.
Se vuelve a empezar el cálculo sobre las unidades que quedan.
Al final, tenemos dos grupos :
- un primer grupo de unidades con probabilidades de inclusión iguales a 1
- un segundo grupo con probabilidades de inclusión $0 < \pi_k < 1$ y proporcional a $x_k$.

El problema es seleccionar $n$ unidades con probabilidades de inclusión fijadas con

$$0<\pi_k<1,\ \text{para todo } k\in U,\ \text{tal que}\ \sum_{k\in U}\pi_k = n. \tag{5.2}$$

**Ejemplo 1.** Si $N = 6$, $n = 3$, $x_1 = 1$, $x_2 = 9$, $x_3 = 10$, $x_4 = 70$, $x_5 = 90$, $x_6 = 120$, tenemos

$$X = \sum_{k\in U} x_k = 300,$$

y entonces

$$\frac{nx_1}{X} = \frac{1}{100}, \frac{nx_2}{X} = \frac{9}{100}, \frac{nx_3}{X} = \frac{1}{10}, \frac{nx_4}{X} = \frac{7}{10}, \frac{nx_5}{X} = \frac{9}{10}, \frac{nx_6}{X} = \frac{6}{5} > 1.$$

La unidad 6 es seleccionada (con una probabilidad 1). Luego, volvemos a calcular las probabilidades de inclusión

$$\sum_{k\in U\setminus\{6\}} x_k = 180,$$

y entonces

$$\frac{(n-1)x_1}{\sum_{\ell\in U\setminus\{6\}} x_\ell} = \frac{1}{90}, \frac{(n-1)x_2}{\sum_{\ell\in U\setminus\{6\}} x_\ell} = \frac{1}{10}, \frac{(n-1)x_3}{\sum_{\ell\in U\setminus\{6\}} x_\ell} = \frac{1}{9},$$

$$\frac{(n-1)x_4}{\sum_{\ell\in U\setminus\{6\}} x_\ell} = \frac{7}{9}, \frac{(n-1)x_5}{\sum_{\ell\in U\setminus\{6\}} x_\ell} = 1.$$

Las probabilidades de inclusión son

$$\pi_1 = \frac{1}{90}, \pi_2 = \frac{1}{10}, \pi_3 = \frac{1}{9}, \pi_4 = \frac{7}{9}, \pi_5 = 1, \pi_6 = 1.$$

Dos unidades son seleccionadas con una probabilidad 1. El problema es reducido a la selección de una unidad en una subpoblación de tamaño 4.

## 5.3. Muestreo con probabilidades desiguales con reemplazamiento

Hansen y Hurwitz (1943).
Sea

$$p_k = \frac{x_k}{\sum_{\ell\in U} x_\ell}, k \in U,$$

y

$$v_k = \sum_{\ell=1}^{k} p_\ell, \text{ con } v_0 = 0.$$

- $u$ es una variable continua, uniforme en $[0, 1[$
- se selecciona la unidad $k$ tal que $v_{k-1} \leq u < v_k$.
- esta operación es repetida $m$ veces de manera independiente.

$\tilde{y}_i$ es la $i$ésima unidad seleccionada en la muestra
$Y$ es estimado por el estimador de Hansen-Hurwitz

$$\widehat{Y}_{HH} = \frac{1}{m}\sum_{i=1}^{m} \frac{\tilde{y}_i}{p_i}.$$

Como

$$E\left[\frac{\tilde{y}_i}{p_i}\right] = \sum_{k\in U} \frac{y_k}{p_k} p_k = Y,$$

$\widehat{Y}_{HH}$ es un estimador insesgado $Y$. En efecto,

$$E\left(\widehat{Y}_{HH}\right) = \frac{1}{m}\sum_{i=1}^{m} E\left(\frac{\tilde{y}_i}{p_i}\right) = \frac{1}{m}\sum_{i=1}^{m} Y = Y.$$

Varianza :

$$Var[\widehat{Y}_{HH}] = \frac{1}{m}\left(\sum_{k\in U}\frac{y_k^2}{p_k} - t_y^2\right) = \frac{1}{m}\sum_{k\in U} p_k \left(\frac{y_k}{p_k} - Y\right)^2, \tag{5.3}$$

y puede estimarse por

$$\widehat{Var}[\widehat{Y}_{HH}] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{\tilde{y}_i}{p_i} - \widehat{Y}_{yHH}\right)^2.$$

## 5.4. Plan de Poisson

Cada unidad de $U$ es seleccionada de manera independiente con una probabilidad de inclusión $\pi_k$.

$$\pi_{k\ell} = \pi_k \pi_\ell,$$

$\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell = 0$, para todos $k \neq \ell$.
El diseño muestral viene dado por

$$p(s) = \left\{ \prod_{k \in s} \pi_k \right\} \times \left\{ \prod_{k \in U \setminus s} (1 - \pi_k) \right\}, \text{ para todos } s \subset U. \tag{5.4}$$

En un plan de Poisson, $\Delta_{k\ell} = 0$ cuando $k \neq \ell$,
la varianza del estimador puede ser calculada simplemente

$$Var\left[\widehat{Y}_\pi\right] = \sum_{k \in U} \frac{\pi_k (1 - \pi_k) y_k^2}{\pi_k^2}, \tag{5.5}$$

y puede estimarse por

$$\widehat{Var}\left[\widehat{Y}_\pi\right] = \sum_{k \in Se} \frac{(1 - \pi_k) y_k^2}{\pi_k^2}. \tag{5.6}$$

## 5.5. Muestreo de entropía máxima con tamaño fijo

Buscamos un diseño muestral con la entropía máxima sobre el conjunto de todas las muestras de $U$ de tamaño fijo $n$.

$$\mathcal{S}_n = \{s | \# s = n\}.$$

El problema es maximizar

$$I(p) = -\sum_{s \in \mathcal{S}_n} p(s) \log p(s),$$

sujeta a que

$$\sum_{\substack{s \ni k \\ s \in \mathcal{S}_n}} p(s) = \pi_k, \text{ y } \sum_{s \in \mathcal{S}_n} p(s) = 1. \tag{5.7}$$

Existe una solución pero es complicada.

$$p(s) \quad = \quad \frac{\exp \sum_{k \in s} \lambda_k}{\sum_{s \in \mathcal{S}_n} \exp \sum_{k \in s} \lambda_k}$$

Un algoritmo (ver Chen y Dempster, y Deville) permite calcular los $\pi_k$ a partir de los $\lambda_k$.

## 5.6. El diseño muestral sistemático

Madow (1949)
Método con tamaño fijo.
Tenemos $0 < \pi_k < 1, k \in U$ con

$$\sum_{k \in U} \pi_k = n.$$

Sea

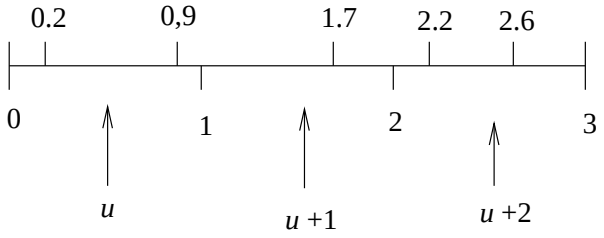
$$V_k = \sum_{\ell=1}^{k} \pi_\ell, \text{ para todos } k \in U, \text{ con } V_o = 0. \tag{5.8}$$

Una variable uniforme es generada en [0, 1].
- la primera unidad seleccionada $k_1$ es tal que $V_{k_1 - 1} \leq u < V_{k_1}$,

- la segunda unidad seleccionada es tal que $V_{k_2-1} \leq u+1 < V_{k_2}$ y
- la $j$ésima unidad seleccionada es tal que $V_{k_j-1} \leq u+j-1 < V_{k_j}$.

**Ejemplo 2.** $N = 6$ y $n = 3$
$\pi_1 = 0,2$, $\pi_2 = 0,7$, $\pi_3 = 0,8$, $\pi_4 = 0,5$, $\pi_5 = \pi_6 = 0,4$,
$V_1 = 0,2$, $V_2 = 0,9$, $V_3 = 1,7$, $V_4 = 2,2$, $V_5 = 2,6$, $V_6 = 3$.
$u = 0,3658$,
Las unidades 2, 3 y 5 son seleccionadas.

Figura 5.1: Muestreo sistemático




El algoritmo puede también ser presentado de la manera siguiente :
Primero, se selecciona la unidad $k$ tal que los intervalos $[V_{k-1} - u, V_k - u[$ contienen un número entero.

**Algoritmo de muestreo sistemático**

| | |
|---|---|
| Definición $a, b, u$ real; $k$ entero; | |
| $u =$ un número aleatorio uniforme en [0,1]; | |
| $a = -u$; | |
| Repetir para $k = 1, .., N$ | $b = a$; |
| | $a = a + \pi_k$; |
| | si $\lfloor a \rfloor \neq \lfloor b \rfloor$ seleccionar $k$. |

El problema es que la mayoría des las probabilidades de inclusión son iguales a cero.
La matriz de probabilidades de inclusión viene dada por :

$$\begin{bmatrix} - & 0 & 0,2 & 0,2 & 0 & 0 \\ 0 & - & 0,5 & 0,2 & 0,4 & 0,3 \\ 0,2 & 0,5 & - & 0,3 & 0,4 & 0,2 \\ 0,2 & 0,2 & 0,3 & - & 0 & 0,3 \\ 0 & 0,4 & 0,4 & 0 & - & 0 \\ 0 & 0,3 & 0,2 & 0,3 & 0 & - \end{bmatrix}$$

## 5.7. El método de escisión

### 5.7.1. Escisión en dos partes

La técnica básica es muy simple : cada $\pi_k$ se separa en dos partes $\pi_k^{(1)}$ y $\pi_k^{(2)}$ que verifican :

$$\pi_k = \lambda \pi_k^{(1)} + (1-\lambda)\pi_k^{(2)}; \tag{5.9}$$

$$0 \leq \pi_k^{(1)} \leq 1 \text{ y } 0 \leq \pi_k^{(2)} \leq 1, \tag{5.10}$$

$$\sum_{k \in U} \pi_k^{(1)} = \sum_{k \in U} \pi_k^{(2)} = n, \tag{5.11}$$

donde $\lambda$ puede elegirse libremente con $0 < \lambda < 1$. El método consiste en seleccionar $n$ unidades con probabilidades desiguales

$$\begin{cases} \pi_k^{(1)}, k \in U, & \text{con una probabilidad } \lambda \\ \pi_k^{(2)}, k \in U, & \text{con una probabilidad } 1-\lambda. \end{cases}$$

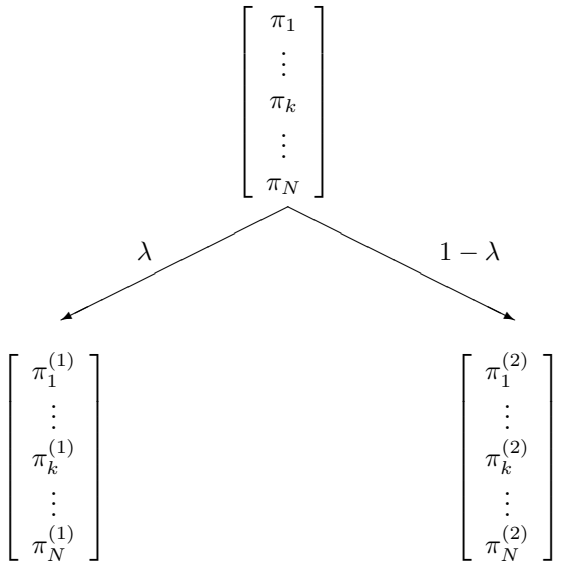



Figura 5.2: Escisión en dos partes

El problema es reducido a otro problema de muestreo con probabilidades desiguales. Si la escisión es tal que uno o algunos de los $\pi_k^{(1)}$ y de los $\pi_k^{(2)}$ son iguales a 0 o 1, el problema de muestreo será más simple en la próxima etapa porque la escisión es aplicada a una población más pequeña.

### 5.7.2. Escisión en $M$ partes

El método puede ser generalizado a una técnica de escisión en $M$ vectores de probabilidades de inclusión. Primero, construimos los $\pi_k^{(j)}$ y los $\lambda_j$ de manera que

$$\sum_{j=1}^{M} \lambda_j = 1,$$

$$0 \leq \lambda_j \leq 1 \ (j = 1, ..., M),$$

$$\sum_{j=1}^{M} \lambda_j \pi_k^{(j)} = \pi_k,$$

$$0 \leq \pi_k^{(j)} \leq 1 \ (k \in U, j = 1, ..., M),$$

$$\sum_{k \in U} \pi_k^{(j)} = n \ (j = 1, ..., M).$$

El método consiste en seleccionar uno de los vectores $\pi_k^{(j)}$ con probabilidades $\lambda_j$ $(j = 1, ..., M)$. De nuevo, los $\pi_k^{(j)}$ son tales que el problema de muestreo será más simple en la próxima etapa.

### 5.7.3. Plan con un soporte mínimo

$(\pi_{(1)}, ..., \pi_{(k)}, ..., \pi_{(N)})$ representa el vector de probabilidades de inclusión. Luego, definimos

$$\begin{aligned} \lambda &= \text{mín}\{1 - \pi_{(N-n)}, \pi_{(N-n+1)}\}, \\ \pi_{(k)}^{(1)} &= \begin{cases} 0 & \text{si } k \leq N - n \\ 1 & \text{si } k > N - n, \end{cases} \end{aligned}$$

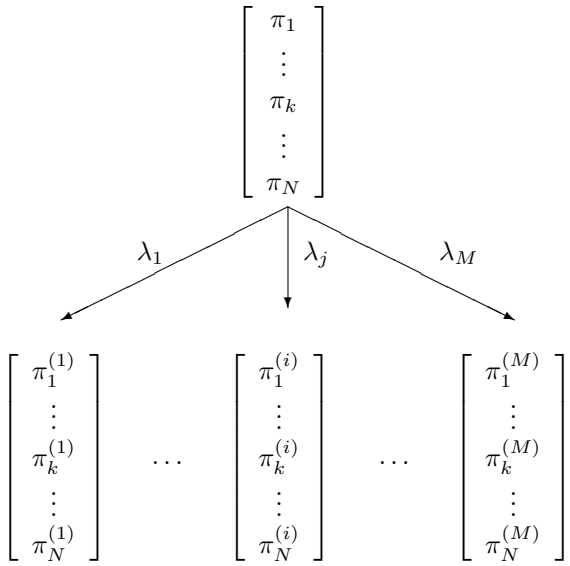



Figura 5.3: Escisión en $M$ partes

$$\pi_{(k)}^{(2)} = \begin{cases} \dfrac{\pi_{(k)}}{1-\lambda} & \text{if } k \leq N-n \\ \dfrac{\pi_{(k)}-\lambda}{1-\lambda} & \text{if } k > N-n. \end{cases}$$

**Ejemplo 1.** Suponemos que $N = 6, n = 3$, $\pi_1 = 0{,}07$, $\pi_2 = 0{,}17$, $\pi_3 = 0{,}41$, $\pi_4 = 0{,}61$, $\pi_5 = 0{,}83$, $\pi_6 = 0{,}91$. En este caso, la solución se encuentra en 4 etapas. El vector de probabilidades de inclusión se separa en dos partes dadaos en las columnas 2 y 3 de la Tabla 1. Con la probabilidad $\lambda = 0{,}59$, la muestra $\{4, 5, 6\}$ es seleccionada y con probabilidad $1 - \lambda = 0{,}41$, otro diseño muestral se aplica con probabilidades de inclusión dadas por (0.171, 0.415, 1, 0.049, 0.585, 0.780). En la etapa 2, la escisión se aplica al vector y, en 4 etapas la muestra es seleccionada. El diseño muestral es el siguiente $p(\{4, 5, 6\}) = 0, 59$; $p(\{3, 5, 6\}) =$

Cuadro 5.1: Plan con soporte mínima

| | Etapa 1 | | Etapa 2 | | Etapa 3 | | Etapa 4 | |
|---|---|---|---|---|---|---|---|---|
| $\pi_k$ | $\lambda = 0, 59$ | | $\lambda = 0, 585$ | | $\lambda = 0, 471$ | | $\lambda = 0, 778$ | |
| $0, 07$ | 0 | $0, 171$ | 0 | $0, 412$ | 0 | $0, 778$ | 1 | 0 |
| $0, 17$ | 0 | $0, 415$ | 0 | 1 | 1 | 1 | 1 | 1 |
| $0, 41$ | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $0, 61$ | 1 | $0, 049$ | 0 | $0, 118$ | 0 | $0, 222$ | 0 | 1 |
| $0, 83$ | 1 | $0, 585$ | 1 | 0 | 0 | 0 | 0 | 0 |
| $0, 91$ | 1 | $0, 780$ | 1 | $0, 471$ | 1 | 0 | 0 | 0 |

$(1 - 0, 59) \times 0, 585 = 0, 24$; $p(\{2, 3, 6\}) = (1 - 0, 59 - 0, 24) \times 0, 471 = 0, 08$; $p(\{1, 2, 3\}) = (1 - 0, 59 - 0, 24 - 0, 08) \times 0, 778 = 0, 07$; $p(\{2, 3, 4\}) = 1 - 0, 59 - 0, 24 - 0, 08 - 0, 7 = 0, 02$.

El diseño muestral viene dado por $p(\{4, 5, 6\}) = 0{,}59$, $p(\{3, 5, 6\}) = (1-0{,}59)\times 0{,}585 = 0{,}24$, $p(\{2, 3, 6\}) = (1 - 0{,}59 - 0{,}24) \times 0{,}471 = 0{,}08$, $p(\{1, 2, 3\}) = (1 - 0{,}59 - 0{,}24 - 0{,}08) \times 0{,}778 = 0{,}07$, $p(\{2, 3, 4\}) = (1 - 0{,}59 - 0{,}24 - 0{,}08 - 0{,}7) = 0{,}02$.

### 5.7.4. Escisión en planes simples

Este método permite separar el vector de probabilidades de inclusiones en dos partes. Definimos

$$\lambda = \text{mín}\left\{\pi_{(1)}\frac{N}{n}, \frac{N}{N-n}(1-\pi_{(N)})\right\}, \tag{5.12}$$

y calculamos, para $k \in U$,

$$\pi_{(k)}^{(1)} = \frac{n}{N}, \pi_{(k)}^{(2)} = \frac{\pi_k - \lambda\frac{n}{N}}{1-\lambda}.$$

Si $\lambda = \pi_{(1)}N/n$, entonces $\pi_{(1)}^{(2)} = 0$; si $\lambda = (1-\pi_{(N)})N/(N-n)$, entonces $\pi_{(N)}^{(2)} = 1$. En la próxima etapa, el problema se reduce en la selección de una muestra de tamaño $n-1$ o $n$ en una población de tamaño $N-1$. En $N-1$ etapas, el problema es reducido.

**Ejemplo 2** Con los mismos $\pi_k$ que en el ejemplo 1, el resultado del método viene dado en la Tabla 2. El

Cuadro 5.2: Descomposición en planes simples

| | Etapa 1 | | Etapa 2 | | Etapa 3 | | Etapa 4 | | Etapa 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\pi_k$ | $\lambda = 0,14$ | | $\lambda = 0,058$ | | $\lambda = 0,173$ | | $\lambda = 0,045$ | | $\lambda = 0,688$ | |
| 0,07 | 0,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0,17 | 0,5 | 0,116 | 0,600 | 0,086 | 0,5 | 0 | 0 | 0 | 0 | 0 |
| 0,41 | 0,5 | 0,395 | 0,600 | 0,383 | 0,5 | 0,358 | 0,667 | 0,344 | 0,5 | 0 |
| 0,61 | 0,5 | 0,628 | 0,600 | 0,630 | 0,5 | 0,657 | 0,667 | 0,656 | 0,5 | 1 |
| 0,83 | 0,5 | 0,884 | 0,600 | 0,901 | 0,5 | 0,985 | 0,667 | 1 | 1 | 1 |
| 0,91 | 0,5 | 0,977 | 0,600 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

problema consiste finalmente en seleccionar uno de los 6 planes simples definidos en las columnas de la Tabla 3. $\lambda_1 = 0{,}14$, $\lambda_2 = (1-0{,}14)\times 0{,}058 = 0{,}050$, $\lambda_3 = (1-0{,}14)\times(1-0{,}058)\times 0{,}173 = 0{,}14$, $\lambda_4 = (1-0{,}14)\times(1-0{,}058)\times(1-0{,}173)\times 0{,}045 = 0{,}03$, $\lambda_5 = (1-0{,}14)\times(1-0{,}058)\times(1-0{,}173)\times(1-0{,}045)\times 0{,}688 = 0{,}44$, $\lambda_6 = (1-0{,}14)\times(1-0{,}058)\times(1-0{,}173)\times(1-0{,}045)\times(1-0{,}688) = 0{,}200$.

Cuadro 5.3: Escisión en $N$ planes simples

| $k$ | $\lambda_1 = 0,14$ | $\lambda_2 = 0,050$ | $\lambda_3 = 0,14$ | $\lambda_4 = 0,03$ | $\lambda_5 = 0,44$ | $\lambda_6 = 0,200$ |
|---|---|---|---|---|---|---|
| 1 | 0,5 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0,5 | 0,6 | 0,5 | 0 | 0 | 0 |
| 3 | 0,5 | 0,6 | 0,5 | 0,667 | 0,5 | 0 |
| 4 | 0,5 | 0,6 | 0,5 | 0,667 | 0,5 | 1 |
| 5 | 0,5 | 0,6 | 0,5 | 0,667 | 1 | 1 |
| 6 | 0,5 | 0,6 | 1 | 1 | 1 | 1 |

### 5.7.5. El método del pivote

Solamente dos probabilidades de inclusión son modificadas : $i$ y $j$.
Si $\pi_i + \pi_j > 1$, entonces

$$\lambda = \frac{1-\pi_j}{2-\pi_i-\pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U\backslash\{i,j\} \\ 1 & k = i \\ \pi_i + \pi_j - 1 & k = j, \end{cases}$$

$$\pi_k^{(2)} = \begin{cases} \pi_k & k \in U\backslash\{i,j\} \\ \pi_i + \pi_j - 1 & k = i \\ 1 & k = j. \end{cases}$$

Por otra parte, si $\pi_i + \pi_j < 1$, entonces

$$\lambda = \frac{\pi_i}{\pi_i + \pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \backslash \{i,j\} \\ \pi_i + \pi_j & k = i \\ 0 & k = j, \end{cases}$$

$$\pi_k^{(2)} = \begin{cases} \pi_k & k \in U \backslash \{i,j\} \\ 0 & k = i \\ \pi_i + \pi_j & k = j. \end{cases}$$

### 5.7.6. Método de Brewer

Brewer y Hanif, 1983, método 8, p. 26.
Brewer, 1975.
*draw by draw procedure*

$$\lambda_j = \left\{ \sum_{z=1}^{N} \frac{\pi_z(n - \pi_z)}{1 - \pi_z} \right\}^{-1} \frac{\pi_j(n - \pi_j)}{1 - \pi_j}.$$

Luego, calculamos

$$\pi_k^{(j)} = \begin{cases} \dfrac{\pi_k(n-1)}{n - \pi_j} & \text{si } k \neq j \\ 1 & \text{si } k = j. \end{cases}$$

La validez del método se deriva del resultado siguiente :

**Teorema 5**

$$\sum_{j=1}^{N} \lambda_j \pi_k^{(j)} = \pi_k,$$

*para todo* $k = 1, \ldots, N$,

## 5.8. Varianza en planes con probabilidades desiguales

Aproximación de la varianza

$$Var(\widehat{Y}_\pi) = \sum_{k \in U} \frac{b_k}{\pi_k^2} \left(y_k - y_k^*\right)^2.$$

con

$$y_k^* = \pi_k \frac{\sum_{\ell \in U} b_\ell y_\ell / \pi_\ell}{\sum_{\ell \in U} b_\ell}$$

$$b_k = \frac{N\pi_k(1 - \pi_k)}{(N-1)}.$$

Estimación de la aproximación de la varianza

$$\widehat{Var}(\widehat{Y}_\pi) = \sum_{k \in S} \frac{c_k}{\pi_k^2} \left(y_k - \hat{y}_k^*\right)^2.$$

con

$$y_k^* = \pi_k \frac{\sum_{\ell \in S} c_\ell y_\ell / \pi_\ell}{\sum_{\ell \in S} b_\ell}$$

$$c_k = \frac{n\pi_k(1 - \pi_k)}{(n-1)}.$$

# Capítulo 6

# Muestreo equilibrado

## 6.1. Introducción

Thionet (1953)
Royall y Herson (1973)
Deville, Grosbras y Roth (1988),
Ardilly (1991),
Hedayat y Majumdar (1995)
Brewer (1999)

**Definición 2** *Un diseño muestral $p(s)$ es equilibrado sobre las variables $x_1, ..., x_p$, si verifica las ecuaciones de equilibrio dadas por*

$$\widehat{\mathbf{X}}_\pi = \mathbf{X}, \tag{6.1}$$

*lo que se puede también escribir*

$$\sum_{k \in s} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

*para toda $s \in \mathcal{S}$ tal que $p(s) > 0$, y para todos $j = 1, ..., p$, o con otras palabras*

$$Var\left(\widehat{\mathbf{X}}_\pi\right) = 0.$$

**Ejemplo 3.** Un muestreo de tamaño fijo es equilibrado sobre la variable $x_k = \pi_k, k \in U$. En efecto,

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} 1 = \sum_{k \in U} \pi_k = n.$$

**Ejemplo 4.** Un plan estratificado es equilibrado sobre las variables

$$\delta_{kh} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{si } k \notin U_h. \end{cases}$$

**Ejemplo 5.** $N = 10, n = 7, \pi_k = 7/10, k \in U$,
$x_k = k, k \in U$.

$$\sum_{k \in S} \frac{k}{\pi_k} = \sum_{k \in U} k,$$

lo que da que

$$\sum_{k \in S} k = 55 \times 7/10 = 38{,}5,$$

ES IMPOSIBLE: Problema de redondeo.

## 6.2. Representación por un cubo

Representación geométrica de un diseño muestral.

$$\mathbf{s} = (I[1 \in s] \ ... \ I[k \in s] \ ... \ I[N \in s])',$$

donde $I[k \in s]$ toma el valor 1 si $k \in s$ y 0 sino.
Geométricamente, cada vector $\mathbf{s}$ es un vértice de un $N$-cubo.

$$E(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

donde $\boldsymbol{\pi} = [\pi_k]$ es el vector de probabilidad de inclusión.

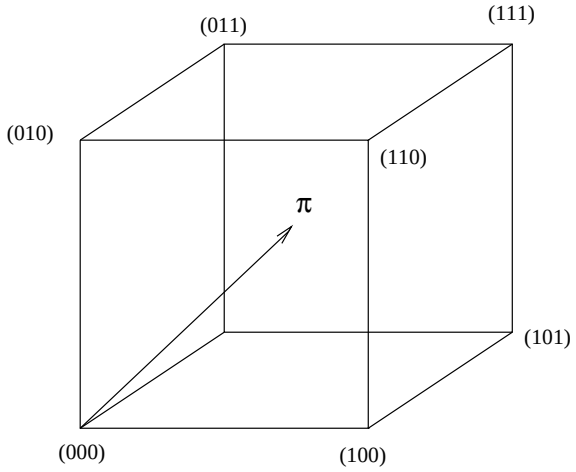



Figura 6.1: Muestras posibles en una población de tamaño $N = 3$

## 6.3. Muestras equilibradas

Método del cubo

1. fase de vuelo,
2. fase de aterrizaje.

Las ecuaciones de equilibrio (6.1) pueden también ser escritas

$$\sum_{k \in U} \mathbf{a}_k c_k = \sum_{k \in U} \mathbf{a}_k \pi_k \qquad (6.2)$$
$$c_k \in \{0, 1\}, k \in U,$$

donde $\mathbf{a}_k = \mathbf{x}_k / \pi_k, k \in U$. (6.2) define un subespacio en $\mathbb{R}^N$ de dimensión $N - p$.

**El problema**
Se elige un vértice del $N$-cubo (una muestra) que queda en el subespacio $Q$.

- Si $C$ representa el $N$-cubo en $\mathbb{R}^N$. Los vértices del $N$-cubo son las muestras de $U$, la intersección entre $C$ y $Q$ es no-vacio, porque $\boldsymbol{\pi}$ es en el interior de $C$ y pertenecen a $Q$.
- La intersección entre el $N$-cubo està un subespacio lineal define un poliedro convexo $K$ que es definido por
$$K = C \cap Q = \left\{[0, 1]^N \cap (\boldsymbol{\pi} + \text{Ker } \mathbf{A})\right\}$$
y tiene la dimensión $N - p$.

**Ejemplo 6.**
$\pi_1 + \pi_2 + \pi_3 = 2$.
$x_k = \pi_k, k \in U$ y $\sum_{k \in S} c_k = 2$.

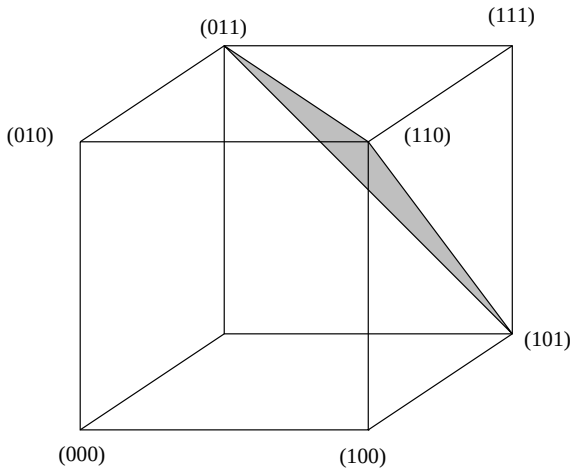



Figura 6.2: Plan de tamaño fijo

**Ejemplo 7.**

- $6 \times \pi_2 + 4 \times \pi_3 = 5$.
- $x_1 = 0, x_2 = 6 \times \pi_2$ y $x_3 = 4 \times \pi_3$.
- $6c_2 + 4c_3 = 5$.

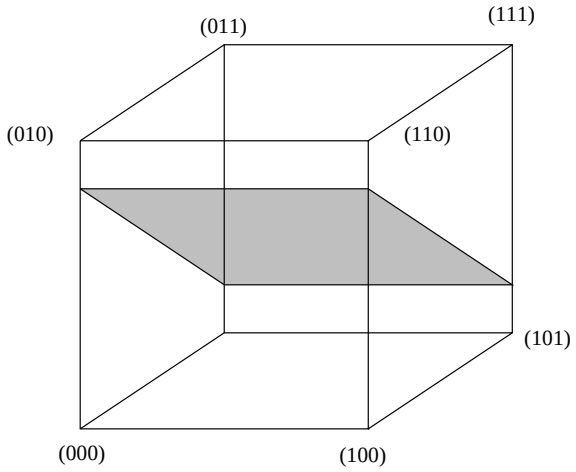



Figura 6.3: Los vértices de $K$ no son vértices del cubo

**Ejemplo 8.**
$\pi_1 + 3 \times \pi_2 + \pi_3 = 4$.

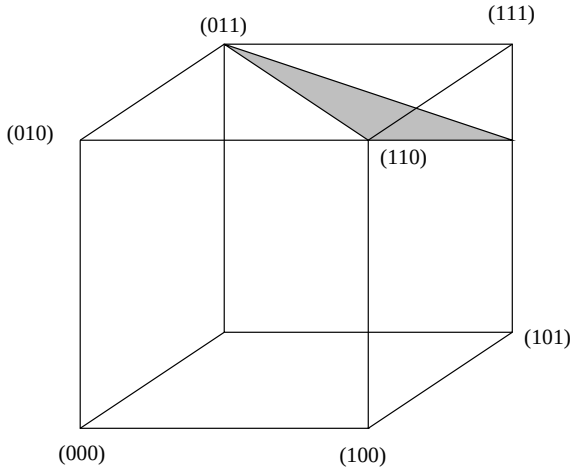



Figura 6.4: Algunos vértices de $K$ son vértices del cubo y otros no le son

$x_1 = \pi_1, x_2 = 3 \times \pi_2$ y $x_3 = \pi_3$.
$c_1 + 3c_2 + c_3 = 4$.

## 6.4. La martingala equilibrada

**Definición 3** *Un proceso aleatorio discreto* $\boldsymbol{\pi}(t) = [\pi_k(t)], t = 0, 1, ...$ *en* $\mathbb{R}^N$ *se llama una martingala equilibrada para un vector de probabilidades de inclusión* $\boldsymbol{\pi}$ *y para las variables auxiliares* $x_1, ..., x_p$, *si*

1. $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$,
2. $E[\boldsymbol{\pi}(t)|\boldsymbol{\pi}(t-1), ...., \boldsymbol{\pi}(0)] = \boldsymbol{\pi}(t-1), t = 1, 2, ...$
3. $\boldsymbol{\pi}(t) \in K = \left\{[0,1]^N \cap (\boldsymbol{\pi} + Ker\ \mathbf{A})\right\}$, *donde* $\mathbf{A}$ *es una matriz* $p \times N$ *dada por* $\mathbf{A} = (\mathbf{x}_1/\pi_1 \ldots \mathbf{x}_k/\pi_k \ldots \mathbf{x}_N/\pi_N)$.

## 6.5. Implementación de la fase de vuelo

Primero, inicializamos por $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Luego, En la etapa $t = 1, ...., T$,

1. Definimos un vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ tal que
   *(i)* $\mathbf{u}(t)$ es en el núcleo (kernel)de la matriz $\mathbf{A}$,
   *(ii)* $u_k(t) = 0$ si $\pi_k(t)$ es entero.
2. Calculamos $\lambda_1^*(t)$ y $\lambda_2^*(t)$, el valor más grande tal que
   $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1$,
   $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1$.
3. Elegimos
$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t-1) + \lambda_1^*(t)\mathbf{u}(t) & \text{con una probabilidad } q_1(t) \\ \boldsymbol{\pi}(t-1) - \lambda_2^*(t)\mathbf{u}(t) & \text{con una probabilidad } q_2(t), \end{cases}$$
   donde
   $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$
   $q_2(t) = \lambda_1^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$.

## 6.6. Método simple.

Definimos un vector $\mathbf{v}(t) = [v_k(t)]$.

$$u_k(t) = \begin{cases} v_k(t) - \mathbf{a}_k' \left( \sum_{\ell \in U_{t-1}} \mathbf{a}_\ell \mathbf{a}_\ell' \right)^- \sum_{\ell \in U_{t-1}} \mathbf{a}_\ell v_\ell(t) & k \in U_{t-1} \\ 0 & k \notin U_{t-1}, \end{cases}$$

donde $U_t = \{k \in U | 0 < \pi_k(t) < 1\}$ y $\left( \sum_{\ell \in U_{t-1}} \mathbf{a}_\ell \mathbf{a}_\ell' \right)^-$ es una generalización de $\sum_{\ell \in U_{t-1}} \mathbf{a}_\ell \mathbf{a}_\ell'$.

## 6.7. Implementación de la fase de aterrizaje

Sea $T$ la ultima etapa de la fase 1,
y notamos por $\boldsymbol{\pi}^* = [\pi_k^*] = \boldsymbol{\pi}(T)$. Sea también $U^* = \{k \in U | 0 < \pi_k^* < 1\}$,
El problema es buscar un plan de muestreo que da una muestra $s \subset U$ tal que

$$\sum_{k \in s} \mathbf{a}_k \approx \sum_{k \in U} \mathbf{a}_k \pi_k^* = \sum_{k \in U} \mathbf{a}_k \pi_k,$$

lo que es equivalente a buscar un diseño muestral que da una muestra $s^* \subset U^*$ tal que

$$\sum_{k \in s^*} \mathbf{a}_k \approx \sum_{k \in U^*} \mathbf{a}_k \pi_k^*,$$

donde $s^* = U^* \cap s$.
Como $q = \#U^*$ es inferior o igual a $p$,

**Solución**
Aplicación del algoritmo del símplex sobre el programa lineal,

$$\min_{p(.)} \sum_{s^* \subset U^*} C(s^*) p(s^*),$$

sujeto a que

$$\sum_{s^* \subset U} p(s^*) = 1,$$

$$\sum_{s^* \ni k} p(s^*) = \pi_k, k \in U,$$

$$0 \leq p(s^*) \leq 1, s^* \subset U,$$

donde $C(s^*)$ es el coste asociado a la muestra $s^*$. Este coste aumenta si las ecuaciones de equilibrio (6.1) no se verifican.

## 6.8. Varianza en un plan equilibrado

$$Var\left(\widehat{Y}_{bal}\right) = Var\left(\widehat{E}_{poiss}\right) = \frac{N}{N-p} \sum_{k \in U} \frac{E_k^2}{\pi_k^2} \pi_k \left(1 - \pi_k\right),$$

donde

$$E_k = y_k - \mathbf{x}_k' \mathbf{B}.$$

# Capítulo 7

# Estimación con informaciones auxiliares y planes simples

Un estimador de La clase de los estimadores lineales es de la forma

$$\widehat{Y}_w = w_0(S) + \sum_{k \in S} w_k(S) y_k, \tag{7.1}$$

donde los pesos $w_k(S)$ pueden depender de la información auxiliar disponible y de los datos observados.

## 7.1. Postestratificación

### 7.1.1. El problema y la notación

Holt y Smith (1979), Jagers (1986); Jagers, Oden y Trulsson (1985).
La variable auxiliar es cualitativa y puede coger $H$ valores distintos.

Partición de la población $U = \{1, ..., k, ..., N\}$ en $H$ subconjuntos, $U_h, h = 1, .., H$, llamados postestratos, tal que

$$\bigcup_{h=1}^{H} U_h = U \text{ y } U_h \bigcap U_i = \emptyset, h \neq i.$$

El número de elementos del postestrato es $N_h$.

$$\sum_{h=1}^{H} N_h = N.$$

El total en la población puede escribirse :

$$Y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} N_h \overline{Y}_h,$$

y la media :

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y}_h,$$

donde $\overline{Y}_h$ representa la media del postestrato $h$

$$\overline{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, h = 1, .., H.$$

Además, $\sigma^2_{yh}$ representa la varianza del postestrato $h$

$$\sigma^2_{yh} = \frac{1}{N_h} \sum_{k \in U_h} \left(y_k - \overline{Y}_h\right)^2,$$

y $S_{yh}^2$ la varianza corregida

$$S_{yh}^2 = \frac{N_h}{N_h - 1}\sigma_{yh}^2.$$

La varianza total $\sigma_y^2$ se obtiene a partir de la formula clásica de descomposición de varianza.

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in U}(y_k - \overline{Y})^2 = \frac{1}{N}\sum_{h=1}^{H} N_h\sigma_{yh}^2 + \frac{1}{N}\sum_{h=1}^{H} N_h(\overline{Y}_h - \overline{Y})^2. \tag{7.2}$$

### 7.1.2. El estimador postestratificado

Se selecciona en la población una muestra aleatoria $S$ con un muestreo aleatorio simple .
Las frecuencias de los postestratos $n_h$ son variables aleatorias que tienen una distribución geométrica.
Como $\pi_k = n/N, k \in U$, el estimador de Horvitz-Thompson de $Y$ viene dado por

$$\widehat{Y}_\pi = \frac{N}{n}\sum_{k\in S} y_k = \frac{N}{n}\sum_{\substack{h=1\\ n_h>0}}^{H} n_h\widehat{\overline{Y}}_h;$$

donde $\widehat{\overline{Y}}_h$ es la media de la muestra en el postestrato $h$

$$\widehat{\overline{Y}}_h = \frac{1}{n_h}\sum_{k\in S_h} y_k.$$

El estimador postestratificado se define por :

$$\widehat{Y}_{post} = \sum_{\substack{h=1\\ n_h>0}}^{H} N_h\widehat{\overline{Y}}_h.$$

Es necesario el conocimiento de las frecuencias de la población $N_h$ para calcular este estimador .
Los postestratos tienen que ser bastante grandes

$$n\frac{N_h}{N} \geq 30,$$

lo que hace que sea muy improbable tener $n_h$ nulos.

### 7.1.3. Propiedad del estimador

$$\begin{aligned} E(\widehat{Y}_{post}|n_h, h=1,...,H) &= \sum_{\substack{h=1\\ n_h>0}}^{H} N_h E(\widehat{\overline{Y}}_h|n_h, h=1,...,H) \\ &= \sum_{\substack{h=1\\ n_h>0}}^{H} N_h\overline{Y}_h \\ &= Y - \sum_{\substack{h=1\\ n_h=0}}^{H} N_h\overline{Y}_h. \end{aligned} \tag{7.3}$$

El sesgo se escribe

$$\begin{aligned} EE(\widehat{Y}_{post}|n_h, h=1,...,H) &= E\left(Y - \sum_{\substack{h=1\\ n_h=0}}^{H} N_h\overline{Y}_h\right) \\ &= Y - \sum_{h=1}^{H} N_h\overline{Y}_h Pr[n_h = 0]. \end{aligned}$$

Como los $n_h$ tienen una distribución geométrica,

$$Pr[n_h = r] = \frac{\binom{N_h}{r}\binom{N-N_h}{n-r}}{\binom{N}{n}}, r = 0, ..., n,$$

se obtiene

$$Pr[n_h = 0] = \frac{(N-N_h)^{[n]}}{N^{[n]}},$$

donde

$$N^{[n]} = \frac{N!}{(N-n)!} = N \times (N-1) \times ... \times (N-n+2) \times (N-n+1),$$

lo que da finalmente

$$E(\widehat{Y}_{post}) = Y - \sum_{h=1}^{H} N_h \overline{Y}_h \frac{(N-N_h)^{[n]}}{N^{[n]}} \approx Y.$$

La varianza

$$\begin{aligned} Var(\widehat{Y}_{post}) &= VarE(\widehat{Y}_{post}|n_h, h = 1, ..., H) \\ &+ EVar(\widehat{Y}_{post}|n_h, h = 1, ..., H). \end{aligned}$$

Por (7.3),

$$VarE(\widehat{Y}_{post}|n_h, h = 1, ..., H) \approx 0,$$

y entonces

$$Var(\widehat{Y}_{post}) \approx EVar(\widehat{Y}_{post}|n_h, h = 1, ..., H). \tag{7.4}$$

Condicionalmente a los $n_h$, el plan es m.a.s. en cada postestrato.
La varianza condicional es entonces la misma que para un plan estratificado

$$Var(\widehat{Y}_{post}|n_h, h = 1, ..., H) = \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \frac{N_h - n_h}{n_h} S_{yh}^2. \tag{7.5}$$

La varianza no-condicional es

$$\begin{aligned} Var(\widehat{Y}_{post}) &= E\left\{ \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \frac{N_h - n_h}{n_h} S_{yh}^2 \right\} \\ &\approx \sum_{h=1}^{H} N_h \left\{ N_h E\left(\frac{1}{n_h}\right) - 1 \right\} S_{yh}^2. \end{aligned} \tag{7.6}$$

Tenemos que calcular la esperanza de $n_h^{-1}$. Si

$$\epsilon = 1 - \frac{n_h}{E(n_h)} = 1 - \frac{N n_h}{n N_h},$$

tenemos

$$E\left(\frac{1}{n_h}\right) = \frac{1}{E(n_h)} E\left(\frac{1}{1-\epsilon}\right).$$

Cuando $n$ es grande, podemos considerar que $\epsilon$ está cerca de cero y usar un desarrollo en serie.

$$E\left(\frac{1}{n_h}\right) \approx \frac{1}{E(n_h)} E(1 + \epsilon + \epsilon^2).$$

Como

$$E(n_h) = n\frac{N_h}{N} \text{ y } Var(n_h) = n\frac{N_h}{N}\frac{N-N_h}{N}\frac{N-n}{N-1},$$

se obtiene

$$\begin{aligned} E\left(\frac{1}{n_h}\right) &\approx \frac{1}{E(n_h)} E\left\{1+\left(1-\frac{n_h N}{nN_h}\right)+\left(1-\frac{n_h N}{nN_h}\right)^2\right\} \\ &= \frac{N}{N_h n}\left\{1+0+\frac{N^2 Var(n_h)}{n^2 N_h^2}\right\} \\ &= \frac{N}{N_h n}+\frac{(N-N_h)N}{N_h^2}\frac{N-n}{n^2(N-1)}. \end{aligned} \tag{7.7}$$

Usando el resultado (7.7) en la expresión (7.6),

$$Var(\widehat{Y}_{post}) \approx \frac{N-n}{n}\sum_{h=1}^{H} N_h S_{yh}^2 + \frac{(N-n)N^2}{n^2(N-1)}\sum_{h=1}^{H}\frac{N-N_h}{N}S_{yh}^2. \tag{7.8}$$

Esta varianza se compone de dos partes. La primera es igual a la varianza del estimador de Horvitz-Thompson para el plan estratificado con afijación proporcional.

$$\begin{aligned} \frac{Var(\widehat{Y}_{post})}{Var(\widehat{Y}_{prop})} &= \left\{\frac{N-n}{n}\sum_{h=1}^{H} N_h S_{yh}^2\right\}^{-1} \\ &\quad \times\left\{\frac{N-n}{n}\sum_{h=1}^{H} N_h S_{yh}^2 + \frac{(N-n)N^2}{n^2(N-1)}\sum_{h=1}^{H}\frac{N-N_h}{N}S_{yh}^2\right\} \\ &= 1+\frac{N}{n(N-1)}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_{yh}^2\right)^{-1}\sum_{h=1}^{H}\frac{N-N_h}{N}S_{yh}^2 \\ &= 1+O(n^{-1}). \end{aligned}$$

## 7.2. Estimación de calibración sobre márgenes

### 7.2.1. El problema

Sean dos variables auxiliares cualitativas.
La primera variable permite dividir la población en $H$ subconjuntos $U_{1.},...,U_{h.},...,U_{H.}$, y la segunda en $I$ subconjuntos $U_{,1},...,U_{.i},...,U_{.I}$.

| | | | | | |
|---|---|---|---|---|---|
| $U_{11}$ | $\dots$ | $U_{1i}$ | $\dots$ | $U_{1I}$ | $U_{1.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $U_{h1}$ | $\dots$ | $U_{hi}$ | $\dots$ | $U_{hI}$ | $U_{h.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $U_{H1}$ | $\dots$ | $U_{Hi}$ | $\dots$ | $U_{HI}$ | $U_{H.}$ |
| $U_{,1}$ | $\dots$ | $U_{.i}$ | $\dots$ | $U_{.I}$ | $U$ |

$N_{hi} = \#U_{hi}, h = 1,...,H, i = 1,...,I$, (desconocidos)
$N_{h.} = \#U_{h.}, h = 1,..,H$, (conocidos)
$N_{.i} = \#U_{.i}, i = 1,...,I$, (conocidos)
Sea una muestra aleatoria simple de tamaño fijo.
El objetivo es entonces estimar el total

$$Y = \sum_{k\in U} y_k. \tag{7.9}$$

El estimador lineal es

$$\widehat{Y}_w = \sum_{k\in S} w_k(S) y_k, \tag{7.10}$$

donde los pesos $w_k(S)$ dependen de los $n_{hi}$ y de los totales marginales de la población $N_{h.}$ y $N_{h.}$.

Cuadro 7.1: Frecuencias según dos variables

| | | | | | |
|---|---|---|---|---|---|
| $n_{11}$ | $\dots$ | $n_{1i}$ | $\dots$ | $n_{1I}$ | $n_{1.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $n_{h1}$ | $\dots$ | $n_{hi}$ | $\dots$ | $n_{hI}$ | $n_{h.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $n_{H1}$ | $\dots$ | $n_{Hi}$ | $\dots$ | $n_{HI}$ | $n_{H.}$ |
| $n_{,1}$ | $\dots$ | $n_{.i}$ | $\dots$ | $n_{.I}$ | $n$ |

Estimador "calado" sobre los márgenes.
Idea de calibración

$$N_{h.} = \sum_{k \in U} z_k,$$

donde $z_k$ es igual a 1 si $k \in U_{h.}$ y 0 sino.
El estimador $\widehat{N}_{h.}$ Se dice que es de calibración sobre $N_{h.}$ si

$$\widehat{N}_{h.} = \sum_{k \in S} w_k(S) z_k = N_{h.}.$$

### 7.2.2. Calibración sobre márgenes

Deming y Stephan (1940) y Stephan (1942). Frielander, (1961), Ireland y Kullback, (1968), Fienberg, (1970), Thionet, (1959 et 1976), Froment y Lenclud, (1976) y Durieux y Payen, (1976).
"método iterativo del cociente"
*Iterative Proportional Fitting Procedure (IPFP).*
"calibración sobre márgenes"

Calibración, tabla de partida

| | | | | | |
|---|---|---|---|---|---|
| $a_{11}$ | $\dots$ | $a_{1i}$ | $\dots$ | $a_{1I}$ | $a_{1.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_{h1}$ | $\dots$ | $a_{hi}$ | $\dots$ | $a_{hI}$ | $a_{h.}$ |
| $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_{H1}$ | $\dots$ | $a_{Hi}$ | $\dots$ | $a_{HI}$ | $a_{H.}$ |
| $a_{,1}$ | $\dots$ | $a_{.i}$ | $\dots$ | $a_{.I}$ | $a_{..}$ |

Buscamos una tabla que està próxima a la tabla de los $a_{hi}$ Con las márgenes $b_{h.}, h = 1, ..., H$, y $b_{.i}, i = 1, ..., I$.
Inicialización

$$b_{hi}^{(0)} = a_{hi}, h = 1, ..., H, i = 1, ..., I.$$

Luego se repite los dos afijaciones siguientes para $j = 1, 2, 3, ....$

$$b_{hi}^{(2j-1)} = b_{hi}^{(2j-2)} \frac{b_{h.}}{b_{h.}^{(2j-2)}}, h = 1, ..., H, i = 1, ..., I,$$

$$b_{hi}^{(2j)} = b_{hi}^{(2j-1)} \frac{b_{.i}}{b_{.i}^{(2j-1)}}, h = 1, ..., H, i = 1, ..., I,$$

donde

$$b_{h.}^{(2j-2)} = \sum_{i=1}^{I} b_{hi}^{(2j-2)}, h = 1, ..., H,$$

y

$$b_{.i}^{(2j-1)} = \sum_{h=1}^{H} b_{hi}^{(2j-1)}, h = 1, ..., H.$$

El algoritmo puede verse como un problema de optimización donde se minimiza la entropía.
Se busca la tabla de los $b_{hi}$ que minimiza

$$\sum_{h=1}^{H}\sum_{i=1}^{I} b_{hi}\log\frac{b_{hi}}{a_{hi}},$$

sujeta a que

$$\sum_{i=1}^{I} b_{hi} = b_{h.}, h = 1, ..., H, \tag{7.11}$$

y

$$\sum_{h=1}^{H} b_{hi} = b_{.i}, i = 1, ..., H. \tag{7.12}$$

Tenemos la ecuación de Lagrange

$$\begin{aligned}
&\mathcal{L}(b_{hi}, \lambda_h, \mu_i) \\
&= \sum_{h=1}^{H}\sum_{i=1}^{I} b_{hi}\log\frac{b_{hi}}{a_{hi}} + \sum_{h=1}^{H}\lambda_h\left(\sum_{i=1}^{I} b_{hi} - b_{h.}\right) \\
&\quad + \sum_{i=1}^{I}\mu_i\left(\sum_{h=1}^{H} b_{hi} - b_{.i}\right).
\end{aligned}$$

Anulando las derivadas de $\mathcal{L}$ con respecto a los $b_{hi}$, tenemos :

$$\log\frac{b_{hi}}{a_{hi}} + 1 + \lambda_h + \mu_i = 0, \tag{7.13}$$

Si $\alpha_h = \exp(-1/2 - \lambda_h)$ y $\beta_i = \exp(-1/2 - \mu_i)$, de (7.13), podemos escribir

$$b_{hi} = a_{hi}\alpha_h\beta_i, h = 1, ..., H, i = 1, ..., I. \tag{7.14}$$

### 7.2.3. Estimación de calibración

Estimador de calibración

$$\widehat{Y}_C = \sum_{h=1}^{H}\sum_{i=1}^{I}\widehat{N}_{Chi}\hat{\overline{Y}}_{hi},$$

donde

$$\hat{\overline{Y}}_{hi} = \frac{1}{n_{hi}}\sum_{k\in S_{hi}} y_k$$

y $S_{hi} = U_{hi}\bigcap S$.
El estimador es entonces linear y puede escribirse

$$\widehat{Y} = \sum_{k\in S} w_k(S)y_k,$$

donde

$$w_k(S) = \frac{\widehat{N}_{Chi}}{n_{hi}}, k \in U_{hi}.$$

Los pesos $w_k(S)$ son funciones no-lineales de los $n_{hi}$ y de los márgenes conocidos.

## 7.3. La variable auxiliar es cuantitativa

### 7.3.1. El problema

Supongamos que el total $X$ de la variable auxiliar $x$ es conocido,

$$X = \sum_{k \in U} x_k,$$

donde $x_1, ..., x_k, ..., x_N$ son los $N$ valores tomados por la variable $x$ sobre las unidades de $U$. Queremos estimar

$$Y = \sum_{k \in U} y_k,$$

### 7.3.2. Notación

$\overline{X} = N^{-1}X$ y $\overline{Y} = N^{-1}Y$ representan las medias de las variables $x$ y $y$ en la población. Las varianzas corregidas son

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \overline{Y})^2$$

y

$$S_x^2 = \frac{1}{N-1} \sum_{k \in U} (x_k - \overline{X})^2.$$

y

$$S_{xy} = \frac{1}{N-1} \sum_{k \in U} (x_k - \overline{X})(y_k - \overline{Y}),$$

la covarianza entre las dos variables.
En un plan simple, los estimadores de Horvitz-Thompson s de los totales son

$$\widehat{Y}_\pi = \frac{N}{n} \sum_{k \in S} y_k = N\widehat{\overline{Y}},$$

y

$$\widehat{X}_\pi = \frac{N}{n} \sum_{k \in S} x_k = N\widehat{\overline{X}},$$

donde

$$\widehat{\overline{Y}} = \frac{1}{n} \sum_{k \in S} y_k$$

y

$$\widehat{\overline{X}} = \frac{1}{n} \sum_{k \in S} x_k.$$

Tenemos igualmente

$$s_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \widehat{\overline{Y}})^2,$$

$$s_x^2 = \frac{1}{n-1} \sum_{k \in S} (x_k - \widehat{\overline{X}})^2$$

y

$$s_{xy} = \frac{1}{n-1} \sum_{k \in S} (x_k - \widehat{\overline{X}})(y_k - \widehat{\overline{Y}}).$$

### 7.3.3. Estimación de diferencia

El estimador de diferencia viene dado por

$$\widehat{Y}_D = \widehat{Y}_\pi + X - \widehat{X}_\pi.$$

Es un estimador linear con :
$w_k(S) = N/n, k \in U$, y
$w_0(S) = X - \widehat{X}_\pi$.
El estimador de diferencia verifica :

$$E(\widehat{Y}_D) = E(\widehat{Y}_\pi) + E(X) - E(\widehat{X}_\pi) = Y + X - X = Y.$$

Como este estimador es insesgado, su error cuadrático medio es igual a su varianza

$$\begin{aligned} Var(\widehat{Y}_D) &= Var(\widehat{Y}_\pi) + Var(\widehat{X}_\pi) - 2Cov(\widehat{X}_\pi, \widehat{Y}_\pi) \\ &= \frac{N(N-n)}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right). \end{aligned}$$

Esta varianza puede estimarse sin sesgo por

$$\widehat{Var}(\widehat{Y}_D) = \frac{N(N-n)}{n}\left(s_y^2 + s_x^2 - 2s_{xy}\right).$$

### 7.3.4. Estimación de razón

El estimador de razón (en ingles *ratio estimator*) es definido mediante

$$\widehat{Y}_{regr} = \frac{X\widehat{Y}_\pi}{\widehat{X}_\pi}.$$

Este estimador es lineal con

$$w_k(S) = \frac{XN}{\widehat{X}_\pi n}.$$

Para calcular el sesgo, tenemos que calcular la esperanza de

$$\widehat{Y}_R - Y = X\frac{\widehat{Y}_\pi}{\widehat{X}_\pi} - Y = X\frac{\widehat{Y}_\pi - r\widehat{X}_\pi}{\widehat{X}_\pi},$$

donde

$$r = \frac{Y}{X}.$$

Si

$$\epsilon = \frac{\widehat{X}_\pi - X}{X},$$

se puede escribir

$$\widehat{Y}_R - Y = \frac{\widehat{Y}_\pi - r\widehat{X}_\pi}{1+\epsilon}.$$

Con un desarrollo de $\widehat{Y}_R - Y$, se logra

$$\begin{aligned} \widehat{Y}_R - Y &= (\widehat{Y}_\pi - r\widehat{X}_\pi)(1 - \epsilon + \epsilon^2 - \epsilon^3 + ...) \\ &\approx (\widehat{Y}_\pi - r\widehat{X}_\pi)(1-\epsilon) \\ &\approx (\widehat{Y}_\pi - r\widehat{X}_\pi)\left(1 - \frac{\widehat{X}_\pi - X}{X}\right). \end{aligned} \tag{7.15}$$

Si se supone que $\epsilon$ es pequeño cuando $n$ es grande, se logra una aproximación del sesgo de este estimador

$$\begin{aligned} E(\widehat{Y}_R - Y) &\approx E(\widehat{Y}_\pi - r\widehat{X}_\pi)\left(1 - \frac{\widehat{X}_\pi - X}{X}\right) \\ &\approx E(\widehat{Y}_\pi - r\widehat{X}_\pi) \\ &\quad - \frac{E\left\{\widehat{Y}_\pi(\widehat{X}_\pi - X)\right\} - rE\left\{\widehat{X}_\pi(\widehat{X}_\pi - X)\right\}}{X} \\ &\approx \frac{rVar(\widehat{X}_\pi) - Cov(\widehat{X}_\pi, \widehat{Y}_\pi)}{X} \\ &\approx \frac{N(N-n)}{n}\frac{rS_x^2 - S_{xy}}{X}. \end{aligned}$$

La esperanza del estimador de razón es entonces dada por

$$E(\widehat{Y}_R - Y) + Y \approx Y + \frac{N-n}{n}\frac{rS_x^2 - S_{xy}}{X}.$$

El sesgo es despreciable cuando $n$ es grande.

### 7.3.5. Precisión del estimador de razón

$$ECM(\widehat{Y}_R) = E(\widehat{Y}_R - Y)^2.$$

En una primera aproximación, por (7.15),

$$\begin{aligned} ECM(\widehat{Y}_R) &\approx E(\widehat{Y}_\pi - r\widehat{X}_\pi)^2 \\ &\approx E\left\{(\widehat{Y}_\pi - Y) - r(\widehat{X}_\pi - X)\right\}^2 \\ &\approx Var(\widehat{Y}_\pi) + r^2 Var(\widehat{X}_\pi) - 2rCov(\widehat{X}_\pi, \widehat{Y}_\pi) \\ &\approx \frac{N(N-n)}{n}\left(S_y^2 + r^2 S_x^2 - 2rS_{xy}\right). \end{aligned}$$

Este error cuadrático medio puede estimarse por

$$\widehat{ECM}(\widehat{Y}_R) = \frac{N(N-n)}{n}\left(s_y^2 + \hat{r}^2 s_x^2 - 2\hat{r}s_{xy}\right),$$

donde

$$\hat{r} = \frac{\widehat{Y}_\pi}{\widehat{X}_\pi}.$$

### 7.3.6. Estimación de regresión

El estimador de regresión viene dado por

$$\widehat{Y}_{regr} = \widehat{Y}_\pi + \hat{b}(X - \widehat{X}_\pi),$$

donde

$$\hat{b} = \frac{s_{xy}}{s_x^2},$$

$$s_{xy} = \frac{1}{n-1}\sum_{k\in S}(x_k - \widehat{\overline{X}})y_k.$$

El estimador es linear con $w_0(S) = 0$ y

$$w_k(S) = \frac{N}{n} + \frac{1}{n-1}(X - \widehat{X}_\pi)\frac{(x_k - \widehat{\overline{X}})}{s_x^2}, k \in U.$$

No es posible calcular exactamente la esperanza matemática y la varianza del estimador de regresión. Pero,

$$\widehat{Y}_{regr} = \widehat{Y}_\pi + b(X - \widehat{X}_\pi) + (\hat{b} - b)(X - \widehat{X}_\pi), \tag{7.16}$$

donde

$$b = \frac{S_{xy}}{S_x^2}.$$

Si se suprime el último termino de la expresión (7.16),

$$\widehat{Y}_{regr} \approx \widehat{Y}_\pi + b(X - \widehat{X}_\pi).$$

Tenemos entonces

$$E(\widehat{Y}_{regr}) \approx E\left\{\widehat{Y}_\pi + b(X - \widehat{X}_\pi)\right\} \approx Y$$

y

$$\begin{aligned} ECM(\widehat{Y}_{regr}) &\approx E\left\{\widehat{Y}_\pi + b(X - \widehat{X}_\pi) - Y\right\}^2 \\ &\approx \frac{N(N-n)}{n}\left(S_y^2 - 2bS_{xy} + b^2 S_x^2\right) \\ &\approx \frac{N(N-n)}{n} S_y^2\left(1 - \rho^2\right), \end{aligned}$$

donde

$$\rho = \frac{S_{xy}}{S_x S_y}.$$

Este error cuadrático medio puede estimarse por

$$\widehat{ECM}(\widehat{Y}_{regr}) = \frac{N(N-n)}{n} s_y^2\left(1 - \hat{\rho}^2\right),$$

donde

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}.$$

La estimación de regresión puede ser generalizada al uso de varias variables auxiliares

### 7.3.7. Discusión de los tres métodos

Cuadro 7.2: Métodos de estimación

| Estimador | Definición | $\left\{\frac{N(N-n)}{n}\right\}^{-1} \times$ ECM |
|---|---|---|
| estimador HT | $\widehat{Y}_\pi = \frac{N}{n}\sum_{k\in S} y_k$ | $S_y^2$ |
| de diferencia | $\widehat{Y}_D = \widehat{Y}_\pi + X - \widehat{X}_\pi$ | $S_y^2 + S_x^2 - 2S_{xy}$ |
| de razón | $\widehat{Y}_{regr} = \widehat{Y}_\pi X/\widehat{X}_\pi$ | $S_y^2 + r^2 S_x^2 - 2rS_{xy}$ |
| de regresión | $\widehat{Y}_{RY} = \widehat{Y}_\pi + \hat{b}(X - \widehat{X}_\pi)$ | $S_y^2\left(1 - \rho^2\right)$ |

### 7.3.8. Comparación del estimador de diferencia y del estimador de Horvitz-Thompson

$$\begin{aligned} & Var(\widehat{Y}_\pi) - Var(\widehat{Y}_D) \\ &= \frac{N(N-n)}{n} S_y^2 - \frac{N(N-n)}{n} \left\{ S_y^2 + S_x^2 - 2S_{xy} \right\} \\ &= \frac{N(N-n)}{n} \left\{ 2S_{xy} - S_x^2 \right\}. \end{aligned}$$

El estimador de diferencia es entonces mejor que el estimador de Horvitz-Thompson cuando

$$2S_{xy} - S_x^2 > 0;$$

lo que puede escribirse de la forma

$$b > \frac{1}{2}.$$

### 7.3.9. Comparación del estimador de razón y del estimador de Horvitz-Thompson

$$\begin{aligned} & ECM(\widehat{Y}_\pi) - ECM(\widehat{Y}_{regr}) \\ &\approx \frac{N(N-n)}{n} S_y^2 - \frac{N(N-n)}{n} \left\{ S_y^2 + r^2 S_x^2 - 2rS_{xy} \right\} \\ &\approx \frac{N(N-n)}{n} \left\{ 2rS_{xy} - r^2 S_x^2 \right\}. \end{aligned}$$

El estimador de razón es generalmente mejor que el estimador de Horvitz-Thompson cuando

$$2rS_{xy} - r^2 S_x^2 > 0,$$

es decir, cuando

$$b > \frac{r}{2} \text{ si } r > 0 \text{ y } b < \frac{r}{2} \text{ si } r < 0.$$

### 7.3.10. Comparación del estimador de razón y del estimador de diferencia

$$\begin{aligned} & ECM(\widehat{Y}_D) - ECM(\widehat{Y}_{regr}) \\ &\approx \frac{N(N-n)}{n} (S_y^2 + S_x^2 - 2S_{xy}) - \frac{N(N-n)}{n} \left\{ S_y^2 + r^2 S_x^2 - 2rS_{xy} \right\} \\ &\approx \frac{N(N-n)}{n} \left\{ 2(1-r)S_{xy} - (1-r^2)S_x^2 \right\}. \end{aligned}$$

El estimador de razón es generalmente preferible cuando

$$2(1-r)S_{xy} - (1-r^2)S_x^2 > 0,$$

es decir cuando

$$2(1-r)b > (1-r^2).$$

### 7.3.11. Comparación del estimador de regresión con los otros estimadores

$$ECM(\widehat{Y}_\pi) - ECM(\widehat{Y}_{regr}) \approx \rho^2 ECM(\widehat{Y}_\pi) \geq 0,$$

$$
\begin{aligned}
& ECM(\widehat{Y}_D) - ECM(\widehat{Y}_{RY}) \\
& \quad \approx \quad \frac{N(N-n)}{n}(S_y^2 + S_x^2 - 2S_{xy}) - \frac{N(N-n)}{n}(1-\rho^2)S_y^2 \\
& \quad \approx \quad \frac{N(N-n)}{n}(S_x - \frac{S_{xy}}{S_x})^2 \geq 0,
\end{aligned}
$$

y

$$
\begin{aligned}
& ECM(\widehat{Y}_R) - ECM(\widehat{Y}_{RY}) \\
& \quad \approx \quad \frac{N(N-n)}{n}(S_y^2 + r^2 S_x^2 - 2rS_{xy}) - \frac{N(N-n)}{n}(1-\rho^2)S_y^2 \\
& \quad \approx \quad \frac{N(N-n)}{n}(rS_x - \frac{S_{xy}}{S_x})^2 \geq 0.
\end{aligned}
$$

# Capítulo 8

# Estimación con informaciones auxiliares y planes complejos

## 8.1. El problema y la notación

El objetivo es siempre estimar el total

$$Y = \sum_{k \in U} y_k,$$

La información auxiliar viene dada por : $J$ variables auxiliares

$$x_1, ..., x_j, ..., x_J.$$

El valor tomado por la variable $x_j$ sobre la unidad $k$ es $x_{kj}$.
$\mathbf{x}_k = (x_{k1}, ..., x_{kj}, ..., x_{kJ})'$ es el vector de los valores tomados por las $J$ variables auxiliares sobre $k$.

$$\overline{X}_j = \frac{1}{N} \sum_{k \in U} x_{kj} = \text{ y } X_j = \sum_{k \in U} x_{kj}, j = 1, ..., J,$$

o

$$\overline{\mathbf{X}} = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k = \text{ y } \mathbf{X} = \sum_{k \in U} \mathbf{x}_k.$$

El vector $\mathbf{X}$ es conocido sobre la población total.
Dos estimadores

- El estimador de regresión :
  Särndal, Swensson y Wretman (1992).
- Calibración
  Deville y Särndal (1992) y Deville, Särndal y Sautory (1993).

Estimadores de la clase de los estimadores lineales

$$\widehat{Y}_G = \sum_{k \in S} w_k(S) y_k.$$

## 8.2. El estimador de regresión

Si existe una relación lineal entre los $\mathbf{x}_k$ e $y_k$, se busca el vector de coeficientes de regresión $\mathbf{b} \in \mathbb{R}^J$ que minimiza:

$$\sum_{k \in U} c_k \left(y_k - \mathbf{x}_k' \mathbf{b}\right)^2, \tag{8.1}$$

donde $c_k$ es un coeficiente de ponderación estrictamente positivo que permite dar una importancia particular a cada unidad. Anulando la derivada de (8.1) con respecto a $\mathbf{b}$, buscamos

$$\sum_{k \in U} c_k \mathbf{x}_k \left(y_k - \mathbf{x}_k' \mathbf{b}\right) = 0,$$

lo que da

$$\sum_{k\in U} c_k \mathbf{x}_k y_k = \sum_{k\in U} c_k \mathbf{x}_k \mathbf{x}_k' \mathbf{b}.$$

Si

$$\mathbf{T} = \sum_{k\in U} c_k \mathbf{x}_k \mathbf{x}_k'$$

y

$$\mathbf{t} = \sum_{k\in U} c_k \mathbf{x}_k y_k$$

y que suponemos que $\mathbf{T}$ es inversible, logramos el coeficiente de regresión :

$$\mathbf{b} = \mathbf{T}^{-1}\mathbf{t}.$$

Como $\mathbf{T}$ y $\mathbf{t}$ son totales podemos estimarlos por estimadores de Horvitz-Thompson :

$$\widehat{\mathbf{T}} = \sum_{k\in S} \frac{c_k \mathbf{x}_k \mathbf{x}_k'}{\pi_k}$$

y

$$\widehat{\mathbf{t}} = \sum_{k\in S} \frac{c_k \mathbf{x}_k y_k}{\pi_k},$$

que no tienen sesgo para $\mathbf{T}$ y $\mathbf{t}$. Luego se estima $\mathbf{b}$ por

$$\widehat{\mathbf{b}} = \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}}.$$

Atención $\widehat{\mathbf{b}}$ no es insesgado para $\mathbf{b}$.
El estimador de regresión :

$$\widehat{Y}_{greg} = \widehat{Y}_\pi + (\mathbf{X} - \widehat{\mathbf{X}}_\pi)'\widehat{\mathbf{b}}. \tag{8.2}$$

### 8.2.1. Otra presentación del estimador de regresión

$\widehat{Y}_{greg}$ es un estimador lineal

$$\widehat{Y}_{greg} = \sum_{k\in S} w_k(S) y_k,$$

donde

$$w_k(S) = \frac{1}{\pi_k}\left\{1 + (\mathbf{X} - \widehat{\mathbf{X}}_\pi)'\widehat{\mathbf{T}}^{-1} c_k \mathbf{x}_k\right\}.$$

Otra presentación viene dada por

$$\begin{aligned} \widehat{Y}_{greg} &= \mathbf{X}'\widehat{\mathbf{b}} + \widehat{Y}_\pi - \widehat{\mathbf{X}}_\pi'\widehat{\mathbf{b}} \\ &= \mathbf{X}'\widehat{\mathbf{b}} + \sum_{k\in S} \frac{e_k}{\pi_k}, \end{aligned}$$

donde

$$e_k = y_k - \mathbf{x}_k'\widehat{\mathbf{b}}.$$

Los $e_k$ son los residuos : las diferencias entre los valores observados y los valores predichos. En algunos casos

$$\sum_{k\in S} \frac{e_k}{\pi_k}$$

es nulo; el estimador de regresión tiene entonces una forma más simple.

**Teorema 6** *Una condición suficiente para que*

$$\sum_{k\in S} \frac{e_k}{\pi_k} = 0,$$

*es que existe un vector* $\boldsymbol{\lambda}$ *tal que*

$$\boldsymbol{\lambda}'\mathbf{x}_k = \frac{1}{c_k}, \text{ para todo } k \in U.$$

**Demostración**

$$\begin{aligned}
\sum_{k\in S}\frac{e_k}{\pi_k} &= \sum_{k\in S}\frac{1}{\pi_k}\left(y_k-\mathbf{x}_k'\widehat{\mathbf{b}}\right)\\
&= \sum_{k\in S}\frac{1}{\pi_k}\left(y_k-c_k\boldsymbol{\lambda}'\mathbf{x}_k\mathbf{x}_k'\widehat{\mathbf{b}}\right)\\
&= \widehat{Y}_\pi-\sum_{k\in S}\frac{c_k\boldsymbol{\lambda}'\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}}\\
&= \widehat{Y}_\pi-\boldsymbol{\lambda}'\sum_{k\in S}\frac{c_k\mathbf{x}_k y_k}{\pi_k}\\
&= \widehat{Y}_\pi-\widehat{Y}_\pi\\
&= 0.
\end{aligned}$$

□

En el caso donde la condición suficiente del teorema 6 es verificada, el estimador de regresión puede escribirse simplemente

$$\widehat{Y}_{greg}=\mathbf{X}'\widehat{\mathbf{b}}.$$

### 8.2.2. Calibración del estimador de regresión

Un estimador se llama de calibración sobre un total de una variable auxiliar si es exactamente igual a este total. Suponemos que calculemos el estimador de regresión para la variable auxiliar $x_j$. El coeficiente de regresión es

$$\widehat{\mathbf{b}}=\left(\sum_{k\in S}\frac{c_k\mathbf{x}_k\mathbf{x}_k'}{\pi_k}\right)^{-1}\sum_{k\in S}\frac{c_k\mathbf{x}_k x_{kj}}{\pi_k}=(0\ldots0\ 1\ 0\ldots0)'.$$

El estimador es

$$\widehat{X}_{j,greg}=\widehat{X}_{j\pi}+(\mathbf{X}-\widehat{\mathbf{X}}_\pi)'\widehat{\mathbf{b}}=\widehat{X}_{j\pi}+(X_j-\widehat{X}_{j\pi})=X_j.$$

### 8.2.3. Estimación de razón

El estimador de razón se logra usando una sola variable auxiliar y cogiendo $\mathbf{x}_k=x_k, c_k=1/x_k$. Tenemos

$$\widehat{b}=\frac{\sum_{k\in S}y_k/\pi_k}{\sum_{k\in S}x_k/\pi_k}=\frac{\widehat{Y}_\pi}{\widehat{X}_\pi},$$

y por (8.2)

$$\widehat{Y}_{greg}=\widehat{Y}_\pi+(X-\widehat{X}_\pi)\frac{\widehat{Y}_\pi}{\widehat{X}_\pi}=X\frac{\widehat{Y}_\pi}{\widehat{X}_\pi},\tag{8.3}$$

que es el estimador de razón.

### 8.2.4. Plan simple y estimación de regresión

El estimador de regresión en el m.a.s. se logra cogiendo $\pi_k=n/N, c_k=1$, y $\mathbf{x}_k=(1,x_k)'$. Luego tenemos

$$\widehat{\mathbf{T}}=\begin{bmatrix} N & \widehat{X}_\pi\\ \widehat{X}_\pi & \dfrac{N}{n}\displaystyle\sum_{k\in S}x_k^2\end{bmatrix},$$

$$\widehat{\mathbf{T}}^{-1}=\frac{n}{N^2s_x^2(n-1)}\begin{bmatrix}\dfrac{N}{n}\displaystyle\sum_{k\in S}x_k^2 & -\widehat{X}_\pi\\ -\widehat{X}_\pi & N\end{bmatrix},$$

$$\widehat{\mathbf{t}} = \begin{bmatrix} \widehat{Y}_\pi \\ \dfrac{N}{n}\sum_{k\in S} x_k y_k \end{bmatrix},$$

$$\widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}} = \frac{n}{N^2 s_x^2 (n-1)} \begin{bmatrix} \widehat{Y}_\pi \dfrac{N}{n}\sum_{k\in S} x_k^2 - \widehat{X}_\pi \dfrac{N}{n}\sum_{k\in S} x_k y_k \\ N^2 s_{xy} \end{bmatrix},$$

donde

$$s_x^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k\in S} x_k^2 - \frac{\widehat{X}_\pi^2}{N^2}\right),$$

y

$$s_{xy} = \frac{n}{n-1}\left(\frac{1}{n}\sum_{k\in S} x_k y_k - \frac{\widehat{X}_\pi \widehat{Y}_\pi}{N^2}\right).$$

Como $\mathbf{X} - \widehat{\mathbf{X}}_{x\pi} = (0, X - \widehat{X}_\pi)'$, tenemos finalmente

$$\widehat{Y}_{greg} = \widehat{Y}_\pi + \left(\mathbf{X} - \widehat{\mathbf{X}}_\pi\right)' \widehat{\mathbf{T}}^{-1}\widehat{\mathbf{t}} = \widehat{Y}_\pi + \left(X - \widehat{X}_\pi\right)\frac{s_{xy}}{s_x^2},$$

que es el estimador de regresión clásico.

## 8.3. Estimación de calibración

### 8.3.1. El método

Deville y Särndal (1992) y Deville, Särndal y Sautory (1993).
La información auxiliar usada es un vector de totales conocidos $\mathbf{X}$.
El método de calibración consiste en buscar nuevos coeficientes de ponderación.
Estimador lineal que se escribe

$$\widehat{Y}_G = \sum_{k\in S} w_k(S) y_k,$$

donde los $w_k(S), k \in S$ son los pesos que dependen de la muestra.
Propiedad de calibración

$$\widehat{\mathbf{X}} = \sum_{k\in S} w_k \mathbf{x}_k = \mathbf{X}. \tag{8.4}$$

Como existe une infinidad de pesos $w_k$ que verifican la relación (8.4), vamos a buscar pesos próximos a los pesos $\pi_k^{-1}$ del estimador de Horvitz-Thompson , lo que va dar un pequeño sesgo.
Definimos

$$d_k = \frac{1}{\pi_k}, k \in S,$$

El objetivo consiste entonces en la búsqueda a los pesos $w_k$ próximos de los $d_k$ que verifican la calibración.
Pseudo-distancia $G_k(w_k, d_k)$, (la simetría no es requerida.)
$G_k(w_k, d_k)$ es positiva, derivable con respecto a $w_k$ estrictamente convexa, tal que $G_k(d_k, d_k) = 0$.
Los pesos $w_k, k \in S$, se logran minimizando

$$\sum_{k\in S} \frac{G_k(w_k, d_k)}{q_k}$$

sujeto a que la calibración sea verificada.
$q_k$ coeficientes de ponderación

$$\begin{aligned} &\mathcal{L}(w_k, k \in S, \lambda_j, j = 1, ..., J) \\ &= \sum_{k\in S} \frac{G_k(w_k, d_k)}{q_k} - \sum_{j=1}^{J} \lambda_j \left\{\sum_{k\in S} w_k x_{kj} - X_j\right\}, \end{aligned}$$

donde los $\lambda_j$ son los multiplicadores de Lagrange.

$$\frac{\partial \mathcal{L}(w_k, k \in S, \lambda_j, j = 1, ..., J)}{\partial w_k} = \frac{g_k(w_k, d_k)}{q_k} - \sum_{j=1}^{J} \lambda_j x_{kj} = 0, \tag{8.5}$$

donde

$$g_k(w_k, d_k) = \frac{\partial G_k(w_k, d_k)}{\partial w_k}.$$

Como $G_k(., d_k)$ es estrictamente convexo y positivo y

$$G_k(d_k, d_k) = 0,$$

$g_k(., d_k)$ es estrictamente creciente y $g_k(d_k, d_k) = 0$.
Los pesos

$$w_k = d_k F_k \left( \sum_{j=1}^{J} \lambda_j x_{kj} \right), \tag{8.6}$$

donde $d_k F_k(.)$ es la función inversa de $g_k(., d_k)/q_k$.
La función $F_k(.)$ es estrictamente creciente y $F_k(0) = 1$, y $F_k'(.)$ la derivada de $F_k(.)$ es entonces estrictamente positiva.
Además, suponemos que $F_k'(0) = q_k$.
Ecuaciones de calibración :

$$\sum_{k \in S} d_k x_{kj} F_k \left( \sum_{i=1}^{J} \lambda_i x_{ki} \right) = X_j, j = 1, ..., J, \tag{8.7}$$

que permite obtener los $\lambda_j$.
Con una escritura matricial

$$\sum_{k \in S} d_k \mathbf{x}_k F_k \left( \mathbf{x}_k' \boldsymbol{\lambda} \right) = \mathbf{X}, \tag{8.8}$$

donde $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_j, ..., \lambda_J)'$.
Al final,una vez calculado $\boldsymbol{\lambda}$, podemos calcular el estimador de calibración :

$$\widehat{Y}_{CAL} = \sum_{k \in S} d_k y_k F_k \left( \mathbf{x}_k' \boldsymbol{\lambda} \right). \tag{8.9}$$

### 8.3.2. Elección de la pseudo-distancia

$$G^{\alpha}(w_k, d_k) = \begin{cases} \dfrac{\frac{w_k^{\alpha}}{d_k^{\alpha-1}} + (\alpha - 1) d_k - \alpha w_k}{\alpha(\alpha - 1)} & \alpha \in \mathbb{R} \backslash \{0, 1\} \\ w_k \log \dfrac{w_k}{d_k} + d_k - w_k & \alpha = 1 \\ d_k \log \dfrac{d_k}{w_k} + w_k - d_k & \alpha = 0. \end{cases}$$

Si derivamos $G^{\alpha}(w_k, d_k)$ con respecto a los $w_k$, logramos

$$g^{\alpha}(w_k, d_k) = \begin{cases} \dfrac{1}{(\alpha - 1)} \left( \dfrac{w_k^{\alpha-1}}{d_k^{\alpha-1}} - 1 \right) & \alpha \in \mathbb{R} \backslash \{1\} \\ \log \dfrac{w_k}{d_k} & \alpha = 1. \end{cases}$$

La inversa de $g^{\alpha}(w_k, d_k)/q_k$ con respecto a $w_k$ es :

$$d_k F_k^{\alpha}(u) = \begin{cases} d_k \sqrt[\alpha - 1]{1 + q_k u(\alpha - 1)} & \alpha \in \mathbb{R} \backslash \{1\} \\ d_k \exp q_k u & \alpha = 1. \end{cases}$$

Según los diferentes valores de $\alpha$, logramos varias pseudo-distancias. Les distancias más usadas son los casos $\alpha = 2$ (khi-cuadrado) y $\alpha = 1$ (entropía).

Cuadro 8.1: Pseudo-distancias para la calibración

| $\alpha$ | $G^\alpha(w_k, d_k)$ | $g^\alpha(w_k, d_k)$ | $F_k^\alpha(u)$ | *Tipo* |
|---|---|---|---|---|
| 2 | $\frac{(w_k-d_k)^2}{2d_k}$ | $\frac{w_k}{d_k} - 1$ | $1 + q_k u$ | Khi-cuadrado |
| 1 | $w_k \log \frac{w_k}{d_k} + d_k - w_k$ | $\log \frac{w_k}{d_k}$ | $\exp(q_k u)$ | Entropía |
| 1/2 | $2(\sqrt{w_k} - \sqrt{d_k})^2$ | $2\left(1 - \sqrt{\frac{d_k}{w_k}}\right)$ | $(1 - q_k u/2)^{-2}$ | Distancia de Hellinger |
| 0 | $d_k \log \frac{d_k}{w_k} + w_k - d_k$ | $1 - \frac{d_k}{w_k}$ | $(1 - q_k u)^{-1}$ | Entropía Inversa |
| $-1$ | $\frac{(w_k-d_k)^2}{2w_k}$ | $\left(1 - \frac{d_k^2}{w_k^2}\right)/2$ | $(1 - 2q_k u)^{-1/2}$ | Khi-cuadrado inverso |

### 8.3.3. El método lineal

Un caso particular importante se logra usando como pseudo-distancia una función de tipo chi-cuadrado (caso $\alpha = 2$) :

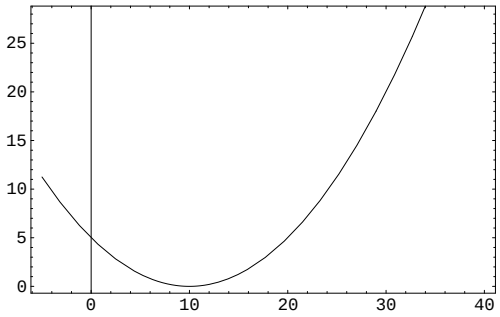



Figura 8.1: Método lineal : función $G(w_k, d_k)$ con $q_k = 1$ y $d_k = 10$

$$G^2(w_k, d_k) = \frac{(w_k - d_k)^2}{2d_k},$$

obtenemos una función lineal

$$F_k(u) = 1 + q_k u.$$

Las ecuaciones de calibración son

$$X_j = \sum_{k \in S} x_{kj} d_k \left(1 + q_k \sum_{i=1}^{J} \lambda_i x_{ki}\right), j = 1, ..., J,$$

o, con escritura matricial,

$$\mathbf{X} = \widehat{\mathbf{X}}_\pi + \sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k' \boldsymbol{\lambda}$$

donde $\boldsymbol{\lambda}' = (\lambda_1, ..., \lambda_j, ..., \lambda_J)$. Si la matriz $\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k'$ es inversible, podemos calcular $\boldsymbol{\lambda}$ :

$$\boldsymbol{\lambda} = \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k'\right)^{-1} \left(\mathbf{X} - \widehat{\mathbf{X}}_\pi\right). \tag{8.10}$$

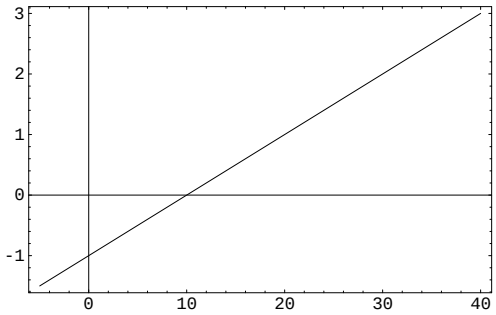



Figura 8.2: Método lineal: función $g(w_k, d_k)$ con $q_k = 1$ y $d_k = 10$

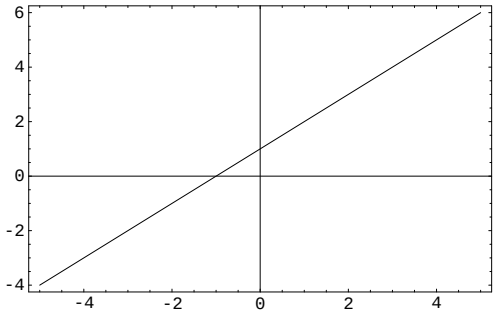



Figura 8.3: Método lineal: función $F_k(u)$ con $q_k = 1$

Luego, podemos calcular los $w_k$ mediante

$$\begin{aligned} w_k &= d_k \left(1 + q_k \sum_{j=1}^{J} \lambda_j x_{kj}\right) \\ &= d_k \left(1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k\right) \\ &= d_k \left\{1 + q_k \left(\mathbf{X} - \widehat{\mathbf{X}}_\pi\right)' \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k\right\}. \end{aligned}$$

Al final, el estimador se escribe

$$\begin{aligned} \widehat{Y}_L &= \sum_{k \in S} w_k y_k \\ &= \widehat{Y}_\pi + \left(\mathbf{X} - \widehat{\mathbf{X}}_\pi\right)' \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k'\right)^{-1} \sum_{k \in S} \mathbf{x}_k d_k q_k y_k. \end{aligned}$$

Si cogemos $q_k = c_k, k \in U$.

### 8.3.4. El método del "raking ratio"

El método del "raking ratio" que incluye el estimador de calibración sobre márgenes se logra usando una pseudo-distancia de tipo "Entropía" (caso $\alpha = 1$) :

$$G^1(w_k, d_k) = w_k \log \frac{w_k}{d_k} + d_k - w_k.$$

Obtenemos

$$F_k(u) = \exp q_k u.$$

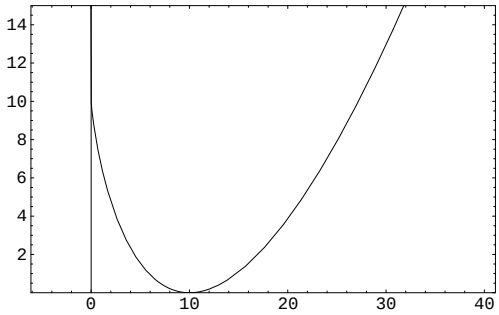



Figura 8.4: “Raking ratio": función $G(w_k, d_k)$ con $q_k = 1$ y $d_k = 10$

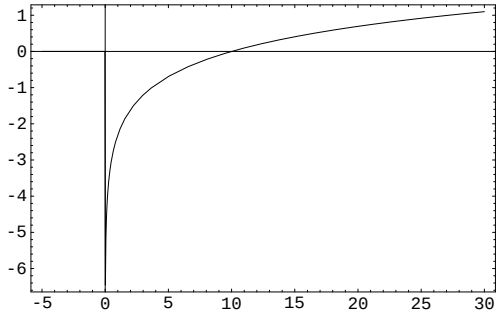



Figura 8.5: “Raking ratio": función $g(w_k, d_k)$ con $q_k = 1$ y $d_k = 10$

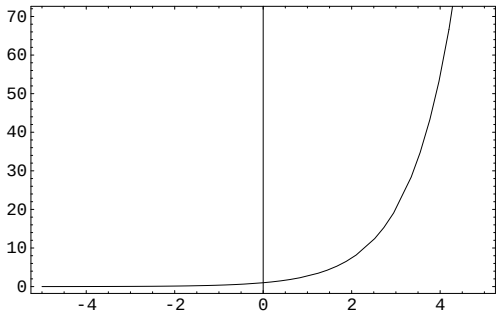



Figura 8.6: “Raking ratio": función $F_k(u)$ con $q_k = 1$

Los pesos son siempre positivos.
El estimador viene dado por

$$w_k = d_k \exp\left(q_k \sum_{j=1}^{J} \lambda_j x_{kj}\right),$$

donde les $\lambda_j$ son calculados por la ecuación

$$\sum_{k \in S} d_k x_{kj} \exp\left(q_k \sum_{i=1}^{J} \lambda_i x_{ki}\right) = X_j, j = 1, ..., J.$$

Caso particular : la calibración sobre márgenes.
En este caso, los $x_{ki}$ son iguales a 1 o 0 según que la unidad $i$ esté o no en la subpoblación $U_i \subset U$. Si, además, $q_k = 1, k \in U$, tenemos

$$w_k = d_k \prod_{i | U_i \ni k} \beta_i$$

donde $\beta_j = \exp \lambda_j$. Los $\beta_j$ son calculados mediante la ecuación

$$\sum_{k \in S} d_k x_{kj} \prod_{i|U_i \ni k} \beta_i = X_j, j = 1, ..., J.$$

### 8.3.5. El método logit

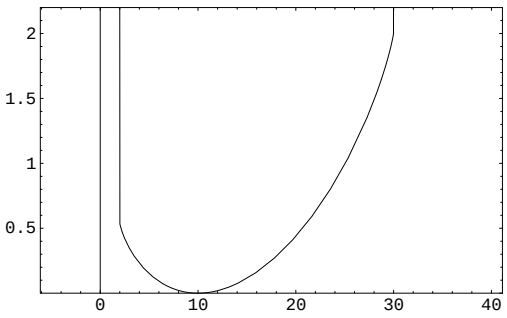



Figura 8.7: Método logístico: función $G(w_k, d_k)$ con $q_k = 1$, $d_k = 10$, $L = 0,2$ y $H = 3$

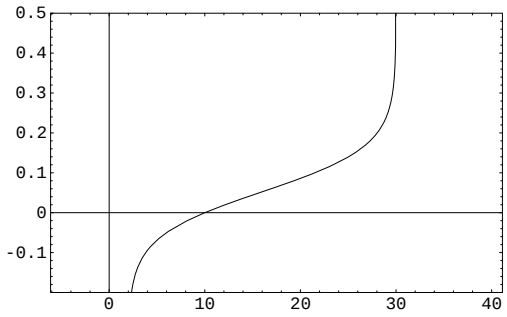



Figura 8.8: Método logístico: función $g(w_k, d_k)$ con $q_k = 1$, $d_k = 10$, $L = 0,2$ y $H = 3$

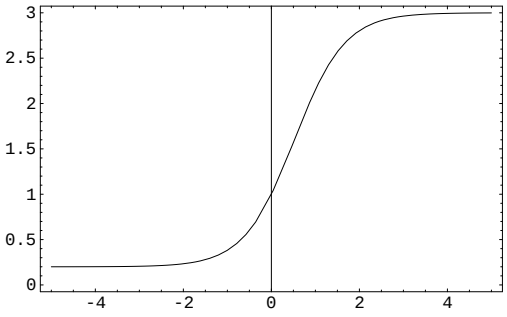



Figura 8.9: Método logístico : función $F_k(u)$ con $q_k = 1$, $L = 0,2$, y $H = 3$

A veces se quiere que los $w_k$ no sean demasiado variable.
Es posible imponer que los pesos $w_k$ se encuentren entre dos valores $Ld_k$ y $Hd_k$ $(L < 1 < H)$ usando una función de tipo logit

$$G(w_k, d_k) = \begin{cases} a_k \log \frac{a_k}{1-L} + b_k \log \frac{b_k}{H-1} \quad \frac{1}{A} & Ld_k < w_k < Hd_k \\ \infty & \text{en otro caso,} \end{cases}$$

donde

$$a_k = \frac{w_k}{d_k} - L, b_k = H - \frac{w_k}{d_k}, A = \frac{H-L}{(1-L)(H-1)}.$$

Obtenemos

$$F_k(u) = \frac{L(H-1) + H(1-L)\exp(Aq_k u)}{H - 1 + (1-L)\exp(Aq_k u)}.$$

Tenemos $F_k(-\infty) = L, F_k(\infty) = H$. Los pesos obtenidos estan entonces siempre en el intervalo $[Ld_k, Hd_k]$.

### 8.3.6. El método lineal truncado

Más simplemente, para restringir el intervalo de soluciones, podemos usar una función del tipo

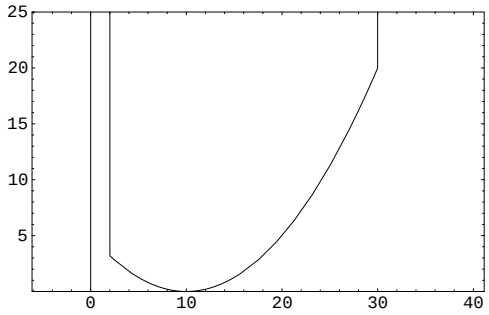



Figura 8.10: Método lineal truncado : función $G(w_k, d_k)$ con $q_k = 1$, $d_k = 10$, $L = 0,2$ y $H = 3$

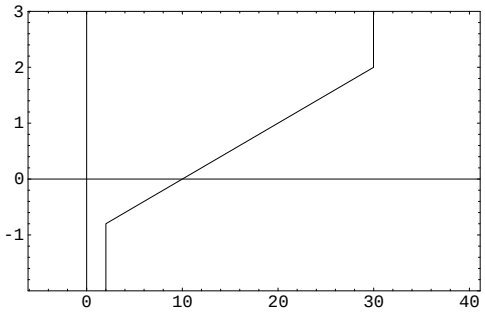



Figura 8.11: Método lineal truncado: función $g(w_k, d_k)$ con $q_k = 1$, $d_k = 10$, $L = 0,2$ y $H = 3$

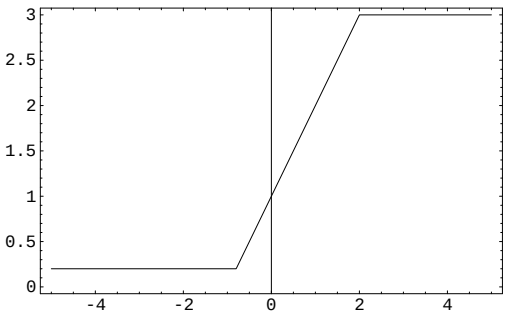



Figura 8.12: Método lineal truncado: función $F_k(u)$ con $q_k = 1$, $L = 0,2$, y $H = 3$

$$G(w_k, d_k) = \begin{cases} \dfrac{(w_k - d_k)^2}{d_k} & Ld_k < w_k < Hd_k \\ \infty & \text{sino.} \end{cases}$$

Obtenemos una función lineal truncada

$$F_k(u) = \begin{cases} 1 + q_k u & \text{si} \quad (L-1)/q_k \leq u \leq (H-1)/q_k \\ H & \text{si} \quad u > (H-1)/q_k \geq H \\ L & \text{si} \quad u < (L-1)/q_k \leq L \end{cases}$$

# Capítulo 9

# Estimación de la varianza por linealización

Las funciones de interés estimadas por muestreo son a veces funciones mas complejas que simples totales, por ejemplo coeficientes de regresión, de correlación, varianza, índices de desigualdades. Además, se usa generalmente una información auxiliar para la calibración de los estimadores, lo que da una forma mas compleja a los estimadores.

Es posible aproximar la varianza por las técnicas de linealización para estimar la precisión de estos estimadores. Las técnicas de linealizsación han sido introducida por Woodruff (1971). Las aplicaciones en la teoría de muestreo han sido desarrolladas por Binder (1983), Binder y Patak (1994), Wolter (1985), Deville (1999).

## 9.1. Orden de magnitud en probabilidad

Las técnicas de linealización estàn basadas en los métodos de desarrollo en Serie de Taylor. El desarrollo se hace con respeto a una variable aleatoria. Para tratar estos problemas vamos a introducir los ordenes de magnitud en probabilidad

**Definición 4** *Una sucesión de números* $f_n, n = 1, 2, ...$ *es dice que es de orden de magnitud inferior a* $h_n > 0, n = 1, 2, ...,$ *si*

$$\lim_{n\to\infty} \frac{f_n}{h_n} = 0.$$

*Se escribe*

$$f_n = o\left(h_n\right).$$

**Definición 5** *Una sucesión de números* $f_n, n = 1, 2, ...$ *està acotada por* $h_n > 0, n = 1, 2, ...,$ *si existe* $M > 0$ *tal que*

$$\mid f_n \mid \leq M h_n,$$

*para todo* $n = 1, 2, ....$ *Se escribe*

$$f_n = O\left(h_n\right).$$

Se puede también definir el orden de magnitud en probabilidad.

**Definición 6** *Una sucesión de variables aleatorias* $X_n$ *converge en probabilidad hacia una variable aleatoria* $X$ *si, para todo* $\epsilon > 0$,

$$\lim_{n\to\infty} Pr\left[\mid X_n - X \mid > \epsilon\right] = 0.$$

*Se escribe*

$$p \lim_{n\to\infty} X_n = X,$$

*o mas simplemente*

$$X_n \overset{P}{\longrightarrow} X.$$

La convergencia en probabilidad permite introducir la noción de orden de magnitud aleatoria :

**Definición 7** *Sea $X_n$ una sucesión de variables aleatorias, $X_n$ se dice que es inferior en probabilidad a $h_n > 0$, si*

$$p \lim_{n\to\infty} \frac{X_n}{h_n} = 0.$$

*Se escribe*

$$X_n = o_p(h_n).$$

**Definición 8** *Sea $X_n$ una sucesión de variables aleatorias, $X_n$ se dice que està acotada por $h_n > o$ en probabilidad por $h_n > 0$ si para todo $\epsilon > 0$, existe un número $M_\epsilon > 0$ tal que*

$$Pr[\mid X_n \mid \geq M_\epsilon h_n] \leq \epsilon,$$

*para todo $n = 1, 2, 3, ...$ se escribe*

$$X_n = O_p(h_n).$$

**Teorema 7** *Sean $X_n$ y $Y_n$ dos sucesiones de variables aleatorias, tales que*

$$X_n = o_p(h_n) \text{ e } Y_n = o_p(g_n),$$

*si $a$ es un real $\alpha > 0$, entonces*
*(i) $aX_n = o_p(h_n)$,*
*(ii) $\mid X_n \mid^\alpha = o_p(h_n^\alpha)$,*
*(iii) $X_nY_n = o_p(h_ng_n)$,*
*(iv) $X_n + Y_n = o_p(\text{máx}(h_n, g_n))$.*

**Demostración**
(i) Si $X_n = o_p(h_n)$ e $Y_n = o_p(g_n)$, entonces

$$\lim_{n\to\infty} Pr\left[\left|\frac{X_n}{h_n}\right| > \epsilon\right] = 0, \tag{9.1}$$

y

$$\lim_{n\to\infty} Pr\left[\left|\frac{Y_n}{g_n}\right| > \epsilon\right] = 0.$$

para todo $\epsilon > 0$. Lo que implica que

$$aX_n = o_p(h_n).$$

(ii) Como

$$Pr\left[\left|\frac{X_n}{h_n}\right| > \epsilon\right] = Pr\left[\left|\frac{X_n}{h_n}\right|^\alpha > \epsilon^\alpha\right],$$

se obtiene $\mid X_n \mid^\alpha = o_p(h_n^\alpha)$.
(iii) Luego, tenemos que, para todo $\epsilon > 0$,

$$\begin{aligned} Pr\left[\left|\frac{X_n}{h_n}\right| > \epsilon\right] + Pr\left[\left|\frac{Y_n}{g_n}\right| > \epsilon\right] &\geq Pr\left[\left|\frac{X_n}{h_n}\right| > \epsilon \text{ donde } \left|\frac{Y_n}{g_n}\right| > \epsilon\right] \\ &\geq Pr\left[\left|\frac{X_nY_n}{h_ng_n}\right| > \epsilon^2\right], \end{aligned}$$

lo que implica que

$$\lim_{n\to\infty} Pr\left[\left|\frac{X_nY_n}{h_ng_n}\right| > \epsilon^2\right] = 0,$$

y

$$X_nY_n = O_p(h_ng_n).$$

(iv) Al final, $X_n + Y_n = o_p(\text{máx}(h_n, g_n))$ es obvio. □

**Teorema 8** *Sean $X_n$ y $Y_n$ los dos sucesiones de variables aleatorias, tales que*

$$X_n = O_p(h_n) \text{ e } Y_n = O_p(g_n),$$

*si a es un real y $\alpha > 0$,*

$$\begin{aligned} aX_n &= O_p(h_n), \\ \mid X_n \mid^{\alpha} &= O_p(h_n^{\alpha}), \\ X_n Y_n &= O_p(h_n g_n), \\ X_n + Y_n &= O_p\left(\text{máx}(h_n, g_n)\right). \end{aligned}$$

Los demostraciones son similares a la precedente. □

**Teorema 9 Desigualdad de Bienaymé-Tchebychev (caso discreto)** *Sean $\alpha > 0$ y $X$ una variable aleatoria discreta tal que $E[X^{\alpha}] < \infty$, entonces para todo $\epsilon > 0$ y para todo $A \in R$,*

$$Pr\left[\mid X - A \mid \geq \epsilon\right] \leq \frac{E\left[\mid X - A \mid^{\alpha}\right]}{\epsilon^{\alpha}}.$$

**Demostración**
Si se nota $X_1, ..., X_i, ..., X_I$, a los valores posibles $X$, se puede escribir

$$\begin{aligned} E\left[\mid X - A \mid^{\alpha}\right] &= \sum_{i=1}^{I} \mid X_i - A \mid^{\alpha} Pr[X = X_i] \\ &= \sum_{\substack{i=1 \\ |X_i - A| < \epsilon}}^{I} \mid X_i - A \mid^{\alpha} Pr[X = X_i] \\ &\quad + \sum_{\substack{i=1 \\ |X_i - A| \geq \epsilon}}^{I} \mid X_i - A \mid^{\alpha} Pr[X = X_i] \\ &\geq \epsilon^{\alpha} \sum_{\substack{i=1 \\ |X_i - A| \geq \epsilon}}^{I} Pr[X = X_i] \\ &= \epsilon^{\alpha} Pr\left[\mid X - A \mid \geq \epsilon\right]. \end{aligned}$$

□

**Teorema 10** *Sea $X_n$ una sucesión de variables aleatorias tal que*

$$E\left[X_n^2\right] = O(h_n),$$

*entonces $X_n = O_p(\sqrt{h_n})$.*

**Demostración**

Como $E\left[X_n^2\right] = O(h_n)$ entonces existe un $M_A > 0$ tal que

$$E\left[X_n^2\right] \leq M_A h_n$$

para todo $n$. Por otro lado, con $\alpha = 2$, $A = 0$, y $\epsilon = \sqrt{M_B h_n}$, por la desigualdad de Bienaymé-Tchébichev, se tiene

$$Pr\left[\mid X_n \mid \geq \sqrt{M_B h_n}\right] \leq \frac{E\left[X_n^2\right]}{M_B h_n}.$$

Si tomamos, $M_B \geq M_A \alpha$, se tiene

$$\frac{E\left[X_n^2\right]}{M_B h_n} = \frac{M_A}{M_B} \leq \alpha,$$

lo que da

$$Pr\left[\mid X_n \mid \geq \sqrt{M_B h_n}\right] \leq \alpha,$$

y entonces $X_n = O_p(\sqrt{h_n})$. □

**Teorema 11** *Sea $X_n$ una sucesión de variables aleatorias tales que*

$$E\left[(X_n - E[X_n])^2\right] = O(h_n),$$

*y que*

$$E\left[X_n\right] = O(\sqrt{h_n}),$$

*entonces $X_n = O_p(\sqrt{h_n})$*

**Demostración**
Como

$$E\left[X_n^2\right] = E\left[(X_n - E[X_n])^2\right] + E[X_n]^2 = O(h_n),$$

el resultado viene del teorema 11. □

**Ejemplo 9.** Sea $X_1, ..., X_n$, $n$ variables independientes con la misma distribución de media $\mu$ y con desviación tipica $\sigma$. La variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$$

tiene varianza

$$Var\left[\bar{X}_n\right] = \frac{\sigma^2}{n},$$

y entonces $\bar{X}_n = O_p(n^{-1/2})$.

**Teorema 12** *Sea $X_n$ una sucesión de variables aleatorias tales que $X_n = x_0 + O_p(h_n)$, $f(x)$ una función derivable $\alpha$ veces con derivadas continuas en el punto $x = x_0$, y $h_n$ una sucesión de números positivos tales que* $\lim_{n\to\infty} h_n = 0$,

$$f(X_n) = f(x_0) + \sum_{i=1}^{\alpha-1}(X_n - x_0)^i \frac{f^{(i)}(x_0)}{i!} + O_p(h_n^\alpha),$$

*donde $f^{(i)}(x_0)$ es la $i$-ésima derivada de $f(x)$ calculada en el punto $x = x_0$.*

**Demostración**
Con un desarrollo en Serie de Taylor, tenemos

$$f(X_n) = f(x_0) + \sum_{i=1}^{\alpha-1}(X_n - x_0)^i \frac{f^{(i)}(x_0)}{i!} + (X_n - x_0)^\alpha \frac{f^{(\alpha)}(b)}{\alpha!},$$

donde $b$ varia entre $x_0$ y $X_n$. Puesto que $f^{(\alpha)}(.)$ es una función continua, $f^{(\alpha)}(b)$ està acotado, en probabilidad, i.e. $f^{(\alpha)}(b) = O_p(1)$. Se obtiene

$$(X_n - x_0)^\alpha \frac{f^{(\alpha)}(b)}{\alpha!} = O_p(h_n^\alpha).$$

□

**Teorema 13** *Sean $X_{1n}, ..., X_{jn}, ..., X_{pn}$, $p$ sucesiones de variables aleatorias tales que $X_{jn} = x_{j0} + O_p(h_n)$, $j = 1, ..., p$, $f(x_1, ..., x_p)$ una función continua cuyas derivadas parciales existen y son continuas en los puntos $x_j = x_{j0}$, y $h_n$ una sucesión de números positivos tales que* $\lim_{n\to\infty} h_n = 0$, *entonces*

$$\begin{aligned} f(X_{1n}, ..., X_{pn}) &= f(x_{10}, ..., x_{p0}) \\ &+ \sum_{j=1}^{p}(X_{jn} - x_{j0}) \left.\frac{\partial f(x_1, ..., x_p)}{\partial x_j}\right|_{x_j = x_{j0}} \\ &+ O_p(h_n^2). \end{aligned}$$

**Demostración**
Aplicando un desarrollo en Serie de Taylor, se logra

$$\begin{aligned} & f(X_{1n},...,X_{pn}) \\ = \; & f(x_{10},...,x_{p0}) \\ + \; & \sum_{j=1}^{p}(X_{jn}-x_{j0})\left.\frac{\partial f(x_1,...,x_p)}{\partial x_j}\right|_{x_j=x_{j0}} \\ + \; & \sum_{j=1}^{p}\sum_{i=1}^{p}\frac{(X_{jn}-x_{j0})(X_{in}-x_{i0})}{2!}\left.\frac{\partial^2 f(x_1,...,x_p)}{\partial x_j \partial x_i}\right|_{x_j=b_j}, \end{aligned}$$

donde los $b_j$ están entre $X_{jn}$ y $x_{j0}$. Como en el teorema precedente, $(X_{jn}-x_{j0})(X_{in}-x_{i0}) = O_p\left(h_n^2\right)$ que multiplica una cantidad acotada en probabilidad. □

# 9.2. Aproximación de la varianza por linealización

## 9.2.1. Linealisación de una función de totales

El objetivo es estimar la varianza de una función de $p$ totales cuyas derivadas existen y son continuas hasta el orden dos

$$\theta = f(Y_1,...,Y_j,...Y_p).$$

Para estimar esta función, se utiliza el estimador por substitución

$$\widehat{\theta} = f(\widehat{Y}_1,...,\widehat{Y}_j,...\widehat{Y}_p),$$

donde los $\widehat{Y}_j$ son los estimadores (eventualmente sesgados) de los $Y_j$. Generalmente los $\widehat{Y}_j$ son los estimadores de Horvitz-Thompson, pero pueden también ser estimadores de razón, de regresión o más complejos.

**Definición 9** *Si $N^{-\alpha}\theta$ está acotado para todo valor de $N$, entonces $\theta$ se dice que es de grado $\alpha$.*

Por ejemplo $R = Y/X$ es de grado 0, $Y$ es de grado 1, $\mathbf{y} = Y/N$ es de grado 0 (porque es una razón de dos funciones de grado 1) y $Var[\widehat{Y}_\pi]$ es de grado 2.

En muestreo, no existe una teoría asintótica general. Existen resultados particulares para los planes simples (Madow, 1948) y para algunos planos con probabilidades desiguales (Rosen, 1972a, 1972b). Vamos a suponer que los $\widehat{Y}_j$ verifican las condiciones siguientes :

1. Los $\widehat{Y}_j$ son lineales homogéneos, es decir que pueden ser escritos de la manera siguiente

$$\widehat{Y}_j = \sum_{k\in S} w_k(S) y_{kj}, j = 1,...,p, \tag{9.2}$$

   donde $y_{kj}$ es el valor tomado por la $j$-isima variable sobre la unidad $k$. El caso más simple viene dado por el estimador de Horvitz-Thompson donde $w_k(S) = 1/\pi_k$.

2. $$\frac{\widehat{Y}_j - Y_j}{N} = O_p\left(\frac{1}{\sqrt{n}}\right), j = 1,...,p.$$

3. Tenemos un estimador de la varianza de cada uno de los $\widehat{Y}_j$ que se nota por $\widehat{Var}\left(\widehat{Y}_j\right)$.

4. Las $\widehat{Var}\left(\widehat{Y}_j\right)^{-1/2}\left(\widehat{Y}_j - Y_j\right)$ tienen una distribución normal centrada reducida.

Estas cuatro hipótesis son bastante simple y son verificadas para los planes simples y los planes estratificados (si el numero de estratos es $n$) y para los planes con conglomerados (si el numero de conglomerados crece con $n$).

**Definición 10** *La variable*

$$v_k = \sum_{j=1}^{p} y_{kj} \left. \frac{\partial f(a_1, ..., a_p)}{\partial a_j} \right|_{a_1=Y_1,...,a_p=Y_p}, k \in U, \tag{9.3}$$

*es llamada la variable linearizada de* $\theta = f(Y_1, ..., Y_p)$.

**Teorema 14** *Sea* $v_k, k \in U$, *la variable linealizsada de una función de interés* $\theta$ *de grado* $\alpha$ *estimada por* $\widehat{\theta}$, *sobre los dos primeras condiciones, entonces*

$$N^{-\alpha}\widehat{\theta} = N^{-\alpha}\theta + N^{-\alpha}\left(\widehat{V} - V\right) + O_p\left(\frac{1}{n}\right),$$

*donde*

$$V = \sum_{k \in U} v_k,$$
$$\widehat{V} = \sum_{k \in S} w_k(S) v_k,$$

*y los* $w_k(S)$ *están definidos de la misma manera que en (9.2).*

**Demostración**
Si se nota $\widehat{\overline{Y}}_j = \widehat{Y}_j/N$, tenemos

$$N^{-\alpha}\widehat{\theta} = N^{-\alpha} f(\widehat{Y}_1, ..., \widehat{Y}_j, ..., \widehat{Y}_p) = N^{-\alpha} f(N\widehat{\overline{Y}}_1, ..., N\widehat{\overline{Y}}_j, ..., N\widehat{\overline{Y}}_p).$$

La condición 2 implica que $\widehat{\overline{Y}}_j = \mathbf{y}_j + O_p(n^{-1/2})$ y con el teorema 13, tenemos

$$\begin{aligned}
N^{-\alpha}\widehat{\theta} &= N^{-\alpha} f(N\widehat{\overline{Y}}_1, ..., N\widehat{\overline{Y}}_j, ..., N\widehat{\overline{Y}}_p) \\
&= N^{-\alpha} f(N\mathbf{y}_1, ..., N\mathbf{y}_j, ..., N\mathbf{y}_p) \\
&\quad + N^{-\alpha} \sum_{j=1}^{p} \left(\widehat{\overline{Y}}_j - \mathbf{y}_j\right) \left. \frac{\partial f(Na_1, ..., Na_p)}{\partial a_j} \right|_{a_1=\mathbf{y}_1,...,a_p=\mathbf{y}_p} \\
&\quad + O_p\left(\frac{1}{n}\right) \\
&= N^{-\alpha}\theta \\
&\quad + N^{-\alpha} \sum_{j=1}^{p} \left(\widehat{Y}_j - Y_j\right) \left. \frac{\partial f(a_1, ..., a_p)}{\partial a_j} \right|_{a_1=Y_1,...,a_p=Y_p} \\
&\quad + O_p\left(\frac{1}{n}\right) \\
&= N^{-\alpha}\theta + N^{-\alpha}\left(\widehat{V} - V\right) + O_p\left(\frac{1}{n}\right).
\end{aligned}$$

□

Observar que $N^{-\alpha}\left(\widehat{V} - V\right) = O_p(n^{-1/2})$. La varianza del estimador de la función de interés puede ser aproximada simplemente. En efecto,

$$\begin{aligned}
&Var\left[N^{-\alpha}\widehat{\theta}\right] \\
&= Var\left[N^{-\alpha}\theta + N^{-\alpha}(\widehat{V} - V) + O_p\left(\frac{1}{n}\right)\right] \\
&= Var\left[N^{-\alpha}\widehat{V}\right] + 2E\left[N^{-\alpha}(\widehat{V} - V) \times O_p\left(\frac{1}{n}\right)\right] + E\left[O_p\left(\frac{1}{n}\right)^2\right] \\
&= Var\left[\frac{\widehat{V}}{N^{\alpha}}\right] + EO_p\left(\frac{1}{n^{3/2}}\right).
\end{aligned}$$

Considerando que $EO_p\left(n^{-3/2}\right)$ es despreciable, se puede construir una aproximación de la varianza

$$AVar[\widehat{\theta}] = Var\left[\widehat{V}\right].$$

## 9.3. Estimación de la varianza

Para estimar la varianza, no se puede usar directamente los $v_k$, porque los $v_k$ dependen de los totales de la población $Y_j$ quienes son desconocidos. Se aproximan los $v_k$ combinado los totales desconocidos por los estimadores, y $\widehat{v}_k$ es la aproximación de la variable linealizsada. Deville (1999) ha probado que si el número de totales a estimar en $v_k$ no crece con $n$, entonces la aproximación de la varianza lograda con los $\widehat{v}_k$ es válida para grandes tamaños de muestra.

Al final, para estimar la varianza de $\widehat{\theta}$, se usa un estimador de la varianza. Si los $\widehat{Y}_j$ son estimadores de Horvitz-Thompson, se puede usar de manera general el estimador de la varianza de Horvitz-Thompson :

$$\widehat{Var}\left[\widehat{\theta}\right] = \sum_{k\in S} \frac{\widehat{v}_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}} \frac{\widehat{v}_k\widehat{v}_\ell}{\pi_k\pi_\ell}\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}.$$

**Ejemplo 10.** El problema más clásico consiste en estimar la razón $R = Y/X$ y la varianza en un plan simple. Primero, se define $f(a_1, a_2) = a_1/a_2$ y entonces

$$R = f(Y, X).$$

El estimador viene directamente dado par

$$\widehat{R} = f(\widehat{Y}, \widehat{X}) = \frac{\widehat{Y}}{\widehat{X}}.$$

Luego, se calculan las derivadas parciales

$$\begin{aligned} \left.\frac{\partial f(a_1,a_2)}{\partial a_1}\right|_{a_1=Y,a_2=X} &= \frac{1}{X} \\ \left.\frac{\partial f(a_1,a_2)}{\partial a_2}\right|_{a_1=Y,a_2=X} &= -\frac{Y}{X^2} \end{aligned}$$

y por (9.3), se obtiene

$$v_k = \frac{y_k}{X} - \frac{Y}{X^2}x_k = \frac{1}{X}\left(y_k - Rx_k\right). \tag{9.4}$$

La varianza aproximada se escribe

$$AVar\left(\widehat{R}\right) = N\frac{N-n}{n}S_v^2,$$

donde

$$S_v^2 = \frac{1}{N-1}\sum_{k\in U}\left(v_k - \frac{V}{N}\right)^2 = \frac{1}{X^2}\left(S_y^2 - 2RS_{xy} + R^2S_x^2\right).$$

Para estimar la varianza de $\widehat{R}$, se empieza a estimar los $v_k$ por

$$\widehat{v}_k = \frac{1}{\widehat{X}}\left(y_k - \widehat{R}x_k\right),$$

y, como

$$\widehat{V} = \frac{N}{n}\sum_{k\in S}\widehat{v}_k = 0,$$

se obtiene el estimador de la varianza

$$\widehat{VAR}\left(\widehat{R}\right) = N\frac{N-n}{n}\frac{1}{n-1}\sum_{k\in S}\widehat{v}_k^2 = N\frac{N-n}{n}\frac{1}{\widehat{X}^2}\left(s_y^2 - 2\widehat{R}s_{xy} + \widehat{R}^2s_x^2\right).$$

**Ejemplo 11.** En un plan complejo con probabilidades de inclusión de segundo orden positivas, se quiere estimar la varianza del vector de coeficientes de regresión

$$\widehat{\mathbf{b}} = \left( \sum_{k \in S} \frac{c_k \mathbf{x}_k \mathbf{x}_k'}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{c_k \mathbf{x}_k y_k}{\pi_k}.$$

La función de interés a estimar es

$$\mathbf{b} = \left( \sum_{k \in U} c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in U} c_k \mathbf{x}_k y_k.$$

Si se nota por

$$\mathbf{T} = \sum_{k \in U} c_k \mathbf{x}_k \mathbf{x}_k',$$

el vector de las variables linealizsadas es egual a

$$\begin{aligned} \mathbf{v}_k &= \mathbf{T}^{-1} c_k \mathbf{x}_k y_k - \mathbf{T}^{-1} c_k \mathbf{x}_k \mathbf{x}_k' \mathbf{T}^{-1} \sum_{k \in U} c_k \mathbf{x}_k y_k \\ &= \mathbf{T}^{-1} \mathbf{x}_k c_k \left( y_k - \mathbf{x}_k' \mathbf{b} \right). \end{aligned}$$

Si se nota por $e_k = y_k - \mathbf{x}_k' \mathbf{b}$, tenemos

$$\mathbf{v}_k = \mathbf{T}^{-1} \mathbf{x}_k c_k e_k. \tag{9.5}$$

Al final, se estima $\mathbf{v}_k$ por

$$\widehat{\mathbf{v}}_k = \widehat{\mathbf{T}}^{-1} \mathbf{x}_k c_k \widehat{e}_k,$$

donde

$$\widehat{\mathbf{T}} = \sum_{k \in S} \frac{c_k \mathbf{x}_k \mathbf{x}_k'}{\pi_k},$$

y

$$\widehat{e}_k = y_k - \mathbf{x}_k' \widehat{\mathbf{b}}.$$

## 9.4. Linealización por etapas

## 9.5. Descomposición en etapas de la linealización

La técnica de linealización puede ser aplicada por etapas. Suponemos que $\theta = f(Y_1, ..., Y_j, ..., Y_p, \lambda)$ donde $\lambda$ es también una función de totales de la qual conocemos la variable linealizada $u_k$, entonces es fácil demostrar que la linealizada de $\theta$ puede escribirse de la forma

$$v_k = \sum_{j=1}^{p} y_{kj} \left. \frac{\partial f(a_1, ..., a_p, \lambda)}{\partial a_j} \right|_{a_1 = Y_1, ..., a_p = Y_p} + u_k \left. \frac{\partial f(Y_1, ..., Y_p, z)}{\partial z} \right|_{z = \lambda}.$$

**Ejemplo 12.** Para un plan con probabilidades de orden 1 y 2 conocidas, queremos calcular la varianza del cuadrado del estimador de razón de Hájek dado por

$$\widehat{\overline{Y}}_H = \left( \sum_{k \in S} \frac{1}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Se observa que la linealizada para la media $\mathbf{y} = Y/N$ se deduce de la linealizada de un razón (9.4) :

$$u_k = \frac{1}{N} \left( y_k - \mathbf{y} \right).$$

Aplicando el método de linealización por etapas, la linealizada de $\mathbf{y}^2$ es

$$v_k = 2\mathbf{y}u_k = \frac{2\mathbf{y}}{N}\left(y_k - \mathbf{y}\right).$$

Se estima $v_k$ por

$$\widehat{v}_k = \frac{2\widehat{\overline{Y}}_H}{\widehat{N}}\left(y_k - \widehat{\overline{Y}}_H\right).$$

## 9.6. Linealización del estimador de regresión

El estimador de regresión se definió en (8.2) :

$$\widehat{Y}_{greg} = \widehat{Y}_\pi + (\mathbf{t}_x - \widehat{\mathbf{t}}_{x\pi})'\widehat{\mathbf{b}}.$$

Podemos escribirlo de la forma

$$\widehat{Y}_{greg} = f(\widehat{Y}_\pi, \widehat{\mathbf{t}}_{x\pi}, \widehat{\mathbf{b}}).$$

Este estimador depende de dos totales $\widehat{Y}_\pi$, $\widehat{\mathbf{t}}_{x\pi}$, y de $\widehat{\mathbf{b}}$ del cual conocemos la linealizada (9.5). Si se usa la técnica de linealisación por etapas $f(Y, \mathbf{t}_x, \mathbf{b})$, se obtiene

$$\begin{aligned} u_k &= y_k - \mathbf{x}_k'\mathbf{b} + (\mathbf{t}_x - \mathbf{t}_x)'v_k \\ &= y_k - \mathbf{x}_k'\mathbf{b} \\ &= e_k, \end{aligned} \tag{9.6}$$

donde $v_k$ es la linealizada de $\mathbf{b}$ que no interviene en la linealizada de $u_k$. El estimador de la linealizada es entonces

$$\widehat{u}_k = \widehat{e}_k = y_k - \mathbf{x}_k'\widehat{\mathbf{b}}.$$

La varianza puede ser estimada simplemente mediante el residuo de la regresión.

# Capítulo 10

# Referencias


Ardilly, P. (1994), *Les Techniques de Sondage*, Paris, Technip.

Ardilly, P. (1991), Echantillonnage représentatif optimum à probabilités inégales, *Annales d'Economie y de Statistique,* **23**, 91-113.

Arora, H.R. y Brackstone, G.J. (1977), An investigation of the properties of raking ratio estimator : I with simple random sampling, *Survey Methodology*, **3**, 62-83.

Basu, D. (1958), On sampling with and without replacement, *Sankhyā*, **20**, 287-294.

Basu, D. (1964), Recovery of ancillary information, *Sankhyā*, **A26**, 3-16.

Basu, D. (1969), Role of the sufficiency and likelihood principles in sample survey theory, *Sankhyā*, **A31**, 441-454.

Basu, D. (1971), An essay on the logical foundations of survey sampling,*in* Godambe, V.P. y Sprott, D.A. Éds., *Foundations of Statistical Inference*, Toronto, Holt, Rinehart and Winston, pp. 203-233.

Basu, D. y Ghosh, J.K. (1967), Sufficient statistics in sampling from a finite universe, *Proceedings of the 36th Session of International Statistical Institute*, 850-859.

Bebbington, A.C. (1975), A simple method of drawing a sample without replacement, *Applied Statistics*, **24**, 136.

Bellhouse D.R. (1988), A brief history of random sampling methods,*in* Krishnaiah, P.R. y Rao, C.R. Éds., *Handbook of Statistics,* Vol 6 (Sampling), New York, Elsevier Science Publishers, pp. 1-14.

Berger, Y. (1996), Asymptotic variance for sequential sampling without replacement with unequal probabilities, Variance asymptotique pour un plan séquentiel sans remise à probabilités inégales, *Survey Methodology, Techniques d'enquête*, **22**,167-173.

Berger, Y. (1998a), *Comportements asymptotiques des plans de sondage à probabilités inégales pour un modèle de population fixe,* Ph.D., Université Libre de Bruxelles.

Berger, Y. (1998b), Variance estimation using list sequential scheme for unequal probability sampling, *Journal of Official Statistics*, **14**, 315-323.

Berger, Y. (1998c), Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator, *Journal of Statistical Planning and Inference*, **74**, 149-168.

Berger, Y., El Haj Tirari, M., Tillé, Y. (2000), *Optimal generalised regression estimation under complex sampling designs*, Document de travail, Rennes, CREST-ENSAI.

Bethlehem, J.G. y Keller J.W. (1987), Lineal weighting of sample survey data, *Journal of Official Statistics*, **3**, 141-153.

Bethlehem, J.G. y Schuerhoff, M.H. (1984), Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities, *Biometrika*, **71**, 642-644.

Binder, D.A. y Patak, Z. (1994), Use of estimating functions for estimation from complex surveys, *Journal of the American Statistical Association,* **89,** 1035-1043.

Binder, D.A. y Theberge, A. (1988), Estimating the variance of raking-ratio estimators, *Canadian Journal of Statistics*, **16** supplement, 47-55.

Brackstone, G.J. y Rao, J.N.K. (1979), An investigation of raking ratio estimators, *Sankhyā*, **C41**, 97-114.

Brewer, K.R.W. (1963), Ratio estimation in finite populations : some results deductible from the assumption of an underlying stochastic process, *Australian Journal of Statistics*, **5**, 93-105.

Brewer, K.R.W. (1975), A simple procedure for $\pi$pswor, *Australian Journal of Statistics*, **17**, 166-172.

Brewer, K.R.W. y Hanif, M. (1983), *Sampling with Unequal Probabilities*, New York, Springer-Verlag.

Bülher, W. y Deutler, T. (1975), Optimal stratification and grouping by dynamic programming, *Metrika*, **22**, 161-175.

Caron, N. (1996), *Les principales techniques de correction de la non-réponse, y les modèles associés*, Document de

travail n°9604, Méthodologie statistique, INSEE.

Caron, N. (1999), Le logiciel POULPE aspects méthodologiques, Actes des Journées de Méthodologie statistique, des 17 y 18 mars 1998, INSEE Méthodes **84-85-86**, pp. 173-200.

Cassel, C.-M., Särndal, C.-E. y Wretman, J.H. (1976), Some results on generalized difference estimation and generalized regression estimation for finite population, *Biometrika*, **63**, 615-620.

Cassel, C.-M., Särndal, C.-E. y Wretman, J.H. (1993), *Foundations of Inference in Survey Sampling*, New York, Wiley.

Causey, B.D. (1972), Sensitivity of raked contingency table totals to change in problem conditions, *Annals of Mathematical Statistics*, **43**, 656-658.

Chao, M.T. (1982), A general purpose unequal probability sampling plan, *Biometrika*, **69**, 653-656.

Chaudhuri, A. (1988), Optimality of sampling strategies,*in* Krishnaiah, P.R. y Rao, C.R. Éds., *Handbook of Statistics, Vol 6 (Sampling)*, New York, Elsevier Science Publishers, pp. 47-96.

Chen, X.-H., Dempster, A.P., y Liu, S.L. (1994), Weighted finite population sampling to maximize entropy, *Biometrika*, **81**, 457-469.

Cochran, W.G. (1939), The use of the analysis of variance in enumeration by sampling, *Journal of the American Statistical Association*, **24**, 492-510.

Cochran, W.G. (1942), Sampling theory when the sampling units are of unequal sizes, *Journal of the American Statistical Association*, **37**, 199-212.

Cochran, W.G. (1946), Relative accuracy of systematic and stratified random samples for a certain class of population, *Annals of Mathematical Statistics*, **17**, 164-177.

Cochran, W.G. (1961), Comparison of methods for determining stratum boundaries, *Proceedings of the International Statistical Institute*, **38**, 245-358.

Cochran, W.G. (1977), *Sampling Techniques*, 3ème édition, New York, Wiley.

Connor, W.S. (1966), An exact formula for the probability that specified sampling units will occur in a sample drawn with unequal probabilities and without replacement, *Journal of the American Statistical Association*, **61**, 384-490.

Cornfield, J. (1944), On samples from finite populations, *Journal of the American Statistical Association*, **39**, 236-239.

Deming, W.E. (1950), *Some Theory of Sampling*, New York, Dover Publications.

Deming, W.E. (1948), *Statistical Adjustment of Data*, New York, Wiley.

Deming, W.E. (1960), *Sample Design in Business Research*, New York, Wiley.

Deming, W.E. y Stephan, F.F. (1940), On a least square adjustment of sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, **11**, 427-444.

Deville, J.-C. (sans date), *Cours de Sondage, Chapitre III : Les Outils de Base*, Polycopié, Paris, ENSAE.

Deville, J.-C. (1988), Estimation linéaire y redressement sur informations auxiliaires d'enquêtes par sondage,*in* Monfort, A. y Laffond, J.J. Éds., *Essais en l'honneur d'Edmond Malinvaud*, Paris, Economica, pp. 915-929.

Deville, J.-C., (1992), Constrained samples, conditional inference, weighting : three aspects of the utilisation of auxiliary information, *Proceedings of the Workshop Auxiliary Information in Surveys*, Örebro (Suède).

Deville, J.-C. (1998a), *Une nouvelle (encore une!) méthode de tirage à probabilités inégales*, Document de travail n°9804, Méthodologie statistique, INSEE.

Deville, J.-C. (1998b), La correction de la non-réponse par calage ou par échantillonnage équilibré,*in Recueil de la Section des méthodes d'enquêtes des communications présentées au 26ème congrès de la Société Statistique du Canada*, Sherbrooke, pp.103-110.

Deville, J.-C. (1999), Estimation de variance pour des statistiques y des estimateurs complexes : techniques de résidus y de linéarisation, Variance estimation for complex statistics ans estimators : linealization and residual techniques, *Techniques d'enquête, Survey methodology*, **25,** 219-230 (fr.), 193-204 (angl.).

Deville, J.-C. (2000a), Note sur l'algorithme de Chen, Dempster y Liu, *Note manuscrite, CREST-ENSAI.*

Deville, J.-C. (2000b), Generalized calibration and application to weighting for non-response, *Communication invitée,* Utrecht, COMPSTAT.

Deville, J.-C. y Dupont, F. (1993), Non-réponse : principes y méthodes,*in Actes des Journées de Méthodologie statistique des 15 y 16 décémbre 1993, INSEE Méthodes n° 56-57-58*, Paris, INSEE, pp. 53-70.

Deville, J.-C. y Grosbras, J.-M. (1987), Algorithmes de tirage,*in* Droesbeke, J.-J., Fichet, B. y Tassi, P. Éds., *Les Sondages*, Paris, Economica, pp. 209-233.

Deville, J.-C., Grosbras, J.-M., y Roth N. (1988), Efficient sampling algorithms and balanced sample, *COMPSTAT, Proceeding in computational statistics*, Physica Verlag, pp. 255-266.

Deville, J.-C., y Särndal, C.-E. (1990), *Estimateur par calage y technique de ratissage généralisé dans les enquêtes par sondage*, Document de travail, Paris, INSEE.

Deville, J.-C., y Särndal, C.-E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical*

*Association*, **87**, 376-382.

Deville, J.-C., Särndal, C.-E. y Sautory, O. (1993), Generalized Raking procedure in survey sampling, *Journal of the American Statistical Association*, **88**, 1013-1020.

Deville, J.-C., y Tillé, Y. (1998), Unequal probability sampling without replacement through a splitting method, *Biometrika*, **85**, 89-101.

Deville, J.-C., y Tillé, Y. (2000), *Balanced sampling by means of the cube method*, Document de travail, Rennes, CREST-ENSAI.

Deroo, M. y Dussaix, A.-M. (1980), *Pratique y analyse des enquêtes par sondage*, Paris, P.U.F.

Durbin, J. (1953), Some results in sampling when the units are selected with unequal probabilities, *Journal of the American Statistical Association*, **61**, 384-490.

Dussaix A.-M. (1987), Modèles de surpopulation,*in* Droesbeke, J.-J., Fichet, B. y Tassi, P. Éds., *Les Sondages*, Paris, Economica, pp. 66-88.

Dussaix A.-M. y Grosbras, J.-M. (1992), *Exercices de sondages*, Paris, Economica.

Fan C.T., Muller, M.E. y Rezucha I. (1962), Development of sampling plans by using sequential (item by item) selection techniques and digital computer, *Journal of the American Statistical Association*, **57**, 387-402.

Fienberg, S.E. (1970), An iterative procedure for estimation in contingency tables, *Annals of Mathematical Statistics*, **41**, 907-917.

Frielander, D., (1961), A technique for estimating a contingency table, given the marginal totals and some supplementary data, *Journal of the Royal Statistical Society*, **A124**, 412-420.

Fuller, W.A., y Isaki, C.T. (1981), Survey design under superpopulation models, *inCurrents topics in survey sampling*, Eds Krewski, D., Platek, R., Rao, J.N.K., y Singh, M.P., New York, Academic Press, 196-226.

Fuller, W.A. (1976), *Introduction to Statistical Time Series*, New York, Wiley.

Gabler, S. (1984), On unequal probability sampling : sufficient conditions for the superiority of sampling without replacement, *Biometrika*, **71**, 171-175.

Gabler, S. (1990), *Minimax solutions in sampling from finite populations*, Lecture Notes in Statistics, 64, Berlin, Spinger Verlag.

Ghiglione, R, y Matalon, B. (1991), *Les enquêtes sociologiques : théorie y pratique*, Paris, Armand Colin.

Hájek, J. (1960), Limiting distributions in simple random sampling from finite population, *Matematikai Kutató Intézeténk közleményei (Publication of the Mathematical Institute of the Hungarian Academy of Sciences)*, **A5**, 361-374.

Hájek, J. (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population, *Annals of Mathematical Statistics*, **35**, 1491-1523.

Hájek, J. (1971), Comment on a paper of D. Basu,*in* Godambe, V.P. y Sprott, D.A. Éds., *Foundations of Statistical Inference*, Toronto, Holt, Rinehart y Winston, p.236.

Hájek, J. (1981), *Sampling in Finite Population*, New York, Marcel Dekker.

Hanif, M. y Brewer, K.R.W. (1980), Sampling with unequal probabilities without replacement : a review, *International Statistical Review*, **48**, 317-335.

Hansen, M.H., Dalenius, T.D. y Tepping B.J. (1985), The development of sample survey in finite population,*in* Atkinson, A. y Fienberg, S. Éds., *A Celebration of Statistics*, The ISI Centenary Volume, Springer-Verlag. pp. 327-353.

Hansen, M.H., Hurwitz, W.N. (1943), On the theory of sampling from finite populations, *Annals of Mathematical Statistics*, **14**, 333-362.

Hansen, M.H., Hurwitz, W.N. (1949), On the determination of the optimum probabilities in sampling, *Annals of Mathematical Statistics*, **20**, 426-432.

Hansen, M.H., Hurwitz, W.N. y Madow, W.G. (1953a réimprimé en 1993), *Sample Survey Methods and Theory, I*, New York, Wiley.

Hansen, M.H., Hurwitz, W.N. y Madow, W.G. (1953b réimprimé en 1993), *Sample Survey Methods and Theory, II*, New York, Wiley.

Hansen, M.H. y Madow, W.G. (1974), Some important events in the historical development of sample survey,*in* Owen, D., Éd. *On the History of Statistics and Probability*, New York, Marcel Dekker.

Hansen, M.H., Madow, W.G. y Tepping B.J. (1983), An evaluation of model dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, **78**, 776-793, Comments and rejoinder, 794-807.

Hanurav, T.V. (1962a), Some sampling schemes in probability sampling, *Sankhyā*, **A24**, 421-428.

Hanurav, T.V. (1962b), On Horvitz and Thompson estimator, *Sankhyā*, **A24**, 429-436.

Hanurav, T.V. (1965), *Optimum Sampling Strategies and some Related Problems*, Thèse de doctorat, Indian Statistical Institute.

Hanurav, T.V. (1966), Some aspects of unified sampling theory, *Sankhyā*, **A28**, 175-204.

Hanurav, T.V. (1967), Optimum utilization of auxiliary information : IIPS sampling of two units from a stratum, *Journal of the Royal Statistical Society*, **B29**, 374-391.

Hanurav, T.V. (1968), Hyper-admissibility and optimum estimators for sampling finite population, *Annals of Mathematical Statistics*, **39**, 621-642.

Hartley, H.O. y Rao, J.N.K. (1962), Sampling with unequal probabilities and without replacement, *Annals of Mathematical Statistics*, **33**, 350-374.

Hartley, H.O. y Rao, J.N.K (1968), A new estimation theory for sample survey, *Biometrika*, **55**, 547-557.

Hedayat, A.S., Majumdar, Dibyen (1995), Generating desirable sampling plans by the technique of trade-off in experimental design, *Journal of Statistical Planning and Inference,* **44**, 237-247.

Hedayat, A.S., y Sinha, B.K. (1991), *Finite Population Sampling*, New York, Wiley.

Hedayat, A.S., Bing-Ying Lin y Stufken, J. (1989), The construction of Π*PS* sampling designs through a method of emptying boxes, *Annals of Statistics*, **4**, 1886-1905.

Holt, D. y Smith, T.M.F. (1979), Poststratification, *Journal of the Royal Statistical Society*, **A142**, Part 1, 33-46.

Horvitz, D.G. y Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663-685.

Ireland, C.T. y Kullback, S. (1968), Contingency tables with given marginals, *Biometrika*, **55**, 179-188.

Isaki, C.T. y Fuller, W.A. (1982), Survey design under a regression population model, *Journal of the American Statistical Association*, **77**, 89-96.

Jagers, P. (1986), Poststratification against bias in sampling, *International Statistical Review*, **54**, 159-167.

Jagers, P., Odén, A. y Trulsson, L. (1985), Poststratification and ratio estimation : usages of auxiliary information in survey sampling and opinion polls, *International Statistical Review*, **53**, 221-238.

Jessen, R.J. (1978), *Statistical Survey Techniques*, New York, Wiley.

Johnson, N.L. y Smith, H. Éds. (1969), *New Developments in Survey Sampling*, New York, Wiley.

Johnson, N.L, Kotz, S. y Kemp, A.W. (1992), *Univariate Discrete Distributions*, New York, Wiley.

Keverberg, Baron de, (1827), Notes,*in* Quetelet, A., *Nouveaux Mémoires de l'Académie royale des Sciences y Belles Lettres de Bruxelles*, **4**, 175-192.

Kiaer, A. (1896), Observations y expériences concernant des dénombrements représentatifs, *Bulletin de l'Institut International de Statistique*, Berne, **9**, livre 2, 176-183.

Kish, L. (1965), *Survey Sampling*, New York, Wiley.

Konijn, H.S. (1973), *Statistical Theory of Sample Survey Design and Analysis*, North-Holland, Amsterdam.

Konijn, H.S. (1981), Biases, variances and covariances of raking ratio estimators for marginal and cell totals and averages of observed characteristics, *Metrika*, **28**, 109-121.

Lanke, J. (1973), On the UMV-estimators in survey sampling, *Metrika*, **20**, 196-202.

Lanke, J. (1975), *Some contributions to the theory of survey sampling*, Lund, AV-Centralen.

Lavallée, P. y Hidiroglou, M.A. (1987), On the stratification of skewed populations, Sur la stratification de populations asymétriques, *Survey Methodology, Techniques d'enquête*, **14**, 33-43.

Laplace, S.P., (1847), *Théorie analytique des probabilités*, Paris, Imprimerie royale.

McLeod, A.I. y Bellhouse, D.R. (1983), A convenient algorithm for drawing a simple random sampling, *Applied Statistics*, **32**, 182-184.

Madow, W.G. (1948), On the limiting distribution based on samples from finite universes, *Annals of Mathematical Statistics,* **19,** 535-545.

Madow, W.G. (1949), On the theory of systematic sampling, II, *Annals of Mathematical Statistics*, **20**, 333-354.

Marcus, M. et Minc, H. (1964), *A survey of matrix theory and matrix inequalities.* Boston : Allyn and Bacon.

Montanari, G.E. (1987), Post sampling efficient QR-prediction in large sample survey, *International Statistical Review*, **55**, 191-202.

Narain, R.D. (1951), On sampling without replacement with varying probabilities, *Journal of Indian Society for Agricultural Statistics*, **3**, 169-174.

Neyman, J., (1934), On the two different aspects of representative method : the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, **97**, 558-606.

Owen, D.B., Cochran, W.G. (1976), On the history of statistics and probability, *Proceedings of a Symposium on the American Mathematical Heritage, to celebrate the bicentennial of the United States of America, held at Southern Methodist University*, New York, M. Dekker.

Raj, D. (1968), *Sampling Theory*, New York, McGraw-Hill.

Raj, D. y Khamis, S.D. (1958), Some remarks on sampling with replacement, *Annals of Mathematical Statistics*, **29**, 550-557.

Rao, T.J. (1971), Π-ps sampling designs and the H.T. estimator, *Journal of the American Statistical Association*, **66**, 872-875.

Rao, C.R. (1971), Some aspects of statistical inference in problems of sampling from finite population, *in* Godambe, V.P. y Sprott, D.A. Éds., *Foundations of Statistical Inference*, Toronto, Montréal.

Rao, J.N.K. (1975), On the foundations of survey sampling, *in* Shrivastava, J.N. Éds., *A Survey of Statistical Design and Lineal Models*, La Haye, North Holland, pp. 489-505.

Rao, J. N. K. (1985), Conditional inference in survey sampling, Inférence conditionelle dans les enquêtes par sondage, *Survey Methodology, Techniques d'enquête*, **11**, 15-31.

Rao, J. N. K. (1994), Estimating totals and distribution functions using auxiliary information at the estimation stage, *Journal of Official Statistics*, **10**, 153-165.

Rao, J. N. K. (1997), Development in sample survey theory : an appraisal, *Canadian Journal of Statistics*, **25**, 1-21.

Rao, J.N.K., Hartley, H.O. y Cochran, W.G. (1962), On a simple procedure of unequal probability sampling without replacement, *Journal of the Royal Statistical Society*, **B24**, 482-491.

Rosen (1972a), Asymptotic theory for successive sampling I, *Annals of Mathematical Statistics*, **43**, 373-397.

Rosen (1972b), Asymptotic theory for successive sampling II, *Annals of Mathematical Statistics*, **43**, 748-776.

Royall, R., (1968), An old approach to finite population sampling theory, *Journal of the American Statistical Association*, **63**, 1269-1279.

Royall, R., (1970), On finite population sampling theory under certain lineal regression models, *Biometrika*, **57**, 377-387.

Royall, R., (1971), Lineal regression models in finite population sampling theory, *in* Godambe, V.P. y Sprott, D.A. Éds., *Foundations of Statistical Inference*, Toronto, Montréal.

Royall, R., (1976), The lineal least squares prediction approach to two-stage sampling, *Journal of the American Statistical Association*, **71**, 657-664.

Royall, R. y Cumberland, W.G. (1981), An empirical study of the ratio estimator and its variance, *Journal of the American Statistical Association*, **76**, 66-77.

Royall, R. y Eberhardt, K.R. (1975), Variance estimates for the ratio estimator, *Sankhyā*, **C37**, 43-52.

Royall, R. y Herson, J. (1973a), Robust estimation in finite populations I, *Journal of the American Statistical Association*, **68**, 880-889.

Royall, R. y Herson, J. (1973b), Robust estimation in finite populations II : stratification on a size variable, *Journal of the American Statistical Association*, **68**, 891-893.

Särndal, C.-E. (1980), On $\pi$-inverse weighting versus best lineal unbiased weighting in probability sampling, *Biometrika*, **67**, 639-650.

Särndal, C.-E (1982), Implication of survey design for generalized regression estimation of lineal functions, *Journal of Statistical Planning and Inference*, **7**, 155-170.

Särndal, C.-E. (1984), *Inférence statistique y analyse des données sous des plans d'échantillonnage complexes*, Montréal, Presses de l'Université de Montréal.

Särndal, C.-E. (1984), Design-Consistent versus Model dependent estimation for small domains, *Journal of the American Statistical Association*, **68**, 880-889.

Särndal, C.-E. y Swensson, B. (1987), A general view of estimation for two phases of selection with applications to two-phase sampling and non-response, *International Statistical Review*, **55**, 279-294.

Särndal, C.-E., Swensson, B. y Wretman, J.H. (1989), The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika,* **76,** 527-537.

Särndal, C.-E., Swensson, B. y Wretman, J.H. (1992), *Model Assisted Survey Sampling*, New York, Springer Verlag.

Särndal, C.-E. y Wright, R.L. (1984), Cosmetic form of estimators in survey sampling, *Scandinavian Journal of Statistics*, **11**, 146-156.

Scott, A.J. (1975a), On admissibility and uniform admissibility in the finite sampling, *Annals of Statistics*, **2**, 489-491.

Scott, A.J. (1975b), Some comments on the problem of randomization in surveys, *Proceedings of the 40th Session of the International Statistical Institute*, Varsovie.

Sen, A.R. (1953), On the estimate of the variance in sampling with varying probabilities, *Journal of Indian Society for Agricultural Statistics*, **5**, 119-127.

Sengupta, S. (1989), On Chao's unequal probability sampling plan, *Biometrika*, **76**, 192-196.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* New York, Wiley.

Sheynin O.B. (1986), Adolphe Quetelet as a statistician, *Archive for History of Exact Science*, **36**, 282-325.

Sinha, B.K. (1973), On sampling schemes to realize preassigned sets of inclusion probabilities of first two orders, *Bulletin of the Calcutta Statistical Association*, **22**, 89-110.

Skinner, C.J. (1991), On the efficiency of raking estimation for multiple frame surveys, *Journal of the American Statistical Association*, **86**, 779-784.

Smith, T.M.F. (1976), The foundations of survey sampling : a review, *Journal of the Royal Statistical Society*, **A139**, 183-204.

Stephan, F. (1942), An iterative method of adjusting sample frequency data tables when expected marginal totals are known, *Annals of Mathematical Statistics*, **13**, 166-178.

Stephan, F. (1945), The expected value and variance of the reciprocal and other negative powers of a positive Bernoullan variate, *Annals of Mathematical Statistics*, **16**, 50-61.

Stephan, F. (1948), History of the uses of modern sampling procedures, *Journal of the American Statistical Association*, **43**, 12-49.

Sukhatme, P.V, Sukhatme, B.V. (1970), *Sampling Theory of Surveys with Applications*, 2ème édition, London, Asia Publishing House.

Sunter, A. (1977), List sequential sampling with equal or unequal probabilities without replacement, *Applied Statistics*, **26**, 261-268.

Sunter, A. (1986), Solutions to the problem of unequal probability sampling without replacement, *International Statistical Review*, **54**, 33-50.

Thionet, P. (1953). *La théorie des sondages.* Etudes théoriquew 5, INSEE, Paris, Imprimerie nationale.

Thionet, P. (1959), L'ajustement des résultats des sondages sur ceux des dénombrements, *Revue de l'Institut International de Statistique*, **27**, 8-25.

Thionet, P. (1976), Construction y reconstruction de tableaux statistiques, *Annales de l'INSEE*, **22-23**, 5-27.

Thompson, S.K. (1992), *Sampling*, New York, Wiley.

Tillé, Y. (1996a), An elimination procedure of unequal probability sampling without replacement, *Biometrika*, **83**, 238-241.

Tillé, Y. (1996b), Some remarks on unequal probability sampling designs without replacement, *Annales d' Économie y de Statistique*, **44**, 177-189.

Tillé, Y. (1996c), A moving stratification algorithm, Un algorithme de stratification mobile, *Survey Methodology, Techniques d'enquête*, **22**, 1, 85-94.

Tillé, Y. (1998), Estimation in surveys using conditional inclusion probabilities : simple random sampling, *International Statistical Review*, **66**, 303-322.

Tillé, Y. (1999), Estimation in surveys using conditional inclusion probabilities : complex design, Estimation dans les enquêtes par sondage en utilisant des probabilités d'inclusion conditionnelles : plans complexes, *Survey Methodology, Techniques d'enquête*, **25**, 57-66.

Tillé, Y., Newman, J.A. y Healy, S.D. (1996), New tests for departures from random behavior in spatial memory experiments, *Animal Learning and Behavior*, **24**, 327-340.

Tillé, Y., (2001), *Théorie des sondages, : échantillonnage et estimation en populations finies*, Paris, Dunod.

Tschuprow, A. (1923), On the mathematical expectation of the moments of frequency distributions in the case of correlated observation, *Metron*, **3**, 461-493, 646-680.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, New York, Springer-Verlag.

Woodruff, R.S. (1971), A simple method for approximating de variance of a complicated estimate, *Journal of the American Statistical Association,* **66,** 411-414.

Wynn, H.P. (1977), Convex sets of finite population plans, *Annals of Statistics*, **5**, 414-418.

Wright, R.L. (1983), Finite population sampling with multivariate auxiliary information, *Journal of the American Statistical Association*, **78**, 879-884.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*, London, Griffin.

Yates, F. y Grundy, P.M. (1953), Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society*, **B15**, 235-261.