2 _ MARCOS IMPERFECTOS

Definiciones previas:

Población objetivo → Colección de elementos sobre los que se desea hacer alguna inferencia.

Población investigada → Población que realmente es objeto de estudio.

marco → listado material de unidades de muestreo, de la que se selecciona la muestra.

Lo ideal sería disponer de un marco tal que la lista de unidades muestrales que lo componen coincida con la población objetivo. Pero en la práctica el marco contiene impurezas debidas a desactualizaciones, errores, omisiones, y otras causas que hacen que el marco no coincida con la población objetivo, lo que no impide que el marco sea la contrapartida en el mundo real de la poblac. objetivo.

De todas formas, la separación entre el marco y la poblac. objetivo ha de ser lo suficientemente pequeña como para permitir que se hagan inferencias sobre la población basándose en una muestra obtenida en el marco.

En la práctica, los marcos imperfectos dan lugar a la aparición de sesgos y a la alteración de las varianzas de los estimadores usuales.

Los errores de cobertura son difíciles de estimar, y requieren investigaciones especiales o la utilización de fuentes externas a la encuesta.

Estos errores pueden estimarse mediante el método de reenumeración (*) ó mediante el mt. de las principales componentes demográficas (**).

ATRÁS →

3_ EL PROBLEMA DE LAS UNIDADES VACÍAS

La imperfección del marco suele tener como origen la existencia de duplicaciones de algunas unidades, omisiones de otras y la presencia de unidades extrañas y vacías.

En la práctica, mantener un listado actualizado es imposible.

Unidad vacía → unidad de muestreo erróneamente incluida en el marco y que no pertenece a la pobl objetivo (aunque está relacionada con ella).

(viviendas vacías en encuesta poblac.)

Unidad extraña → unidad que aparece en el marco y de ninguna manera debería estar en él (porque no guarda relación alguna con la pob. objetivo).

(explot. agrícola que no produce leche en un estudio de leche)

IR a (*)

Si eliminamos del marco las unidades erróneamente incluidas en él (unidades extrañas, unidades vacías y duplicaciones) y le añadimos las omisiones, obtenemos la población objetivo. Este proceso se conoce como depuración de marcos imperfectos.

Si eliminamos del marco las unidades inaccesibles, unidades que no colaboran ni responden, unidades ausentes, etc. obtenemos la población investigada.

Marco en sentido ~~estricto~~ restringido → sólo las unidades de las que se va a extraer la muestra.

Marco en sentido amplio → restringido + unid. complementarias (var. auxiliares para estratificar ó para razón, encuestas anteriores, pilotos ...

La presencia de unidades vacías y extrañas se produce principalmente en dos tipos de situaciones (conceptualmente equivalentes):

- la lista, por no estar actualizada, incluye unidades que han dejado de pertenecer al colectivo que se desea muestrear.
- la población que se desea muestrear es una subpoblación de la cubierta por el marco.

Las unidades vacías y unidades extrañas son equivalentes en cuanto al problema metodológico que plantea su presencia en el marco de muestreo.

Para solucionar el problema de los marcos imperfectos, pueden adoptarse varias soluciones cuya puesta en práctica depende de los recursos disponibles:

- Depuración directa del marco: Eliminar del marco las unidades vacías o extrañas, conociendo cuántas unidades contiene el marco depurado. No siempre es posible, por falta de información, tiempo, dinero, ...
- Sustitución de las unidades vacías de la muestra; Se selecciona la muestra en el marco disponible no depurado, sustituyendo las vacías por otras aleatoriamente seleccionadas hasta completar el tamaño. Aparece sesgo en las estimaciones.
- Utilización de la información disponible acerca del nº de unidades vacías: Se conoce cuántas unid. vacías hay, pero no cuáles son.

4 y 5. ESTIMACIÓN del TOTAL y de la MEDIA. CÁLCULO de la VARIANZA y COMPARACIÓN con la VAR. del MARCO DEPURADO.

Definiciones previas:

$N' =$ nº unidades no vacías $\begin{cases} M = \{A_1 \dots A_N\} \text{ marco disponible} \\ M' = \{A_1 \dots A_{N'}\} \text{ marco depurado} \end{cases}$

W = proporción de unidades vacías en el marco disponible

$$W = \frac{N-N'}{N} = 1 - \frac{N'}{N} \quad , \quad N' \leq N$$

Si X_i es la característica que pretendemos estudiar sobre la unidad A_i :

$$X_i = \begin{cases} X_i & \text{si } U_i \text{ es no vacía} \\ 0 & \text{si } U_i \text{ es vacía} \end{cases}$$

por lo que la contribución al total de las unid. vacías es nula $\longrightarrow X = X'$

$$X = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N'} X_i + \sum^{N-N'} 0 = X'$$

$\longrightarrow$ el total del marco dispon. coincide con el total del marco depurado. (los totales)

$$\sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N'} X_i^2 + \sum^{N-N'} 0 = \sum_{i=1}^{N'} X_i^2$$

$$\sum_{i \neq j}^{N} X_i X_j = \sum_{i \neq j}^{N'} X_i X_j .$$

Pero las medias no coinciden (num. iguales, denom. $\neq$).

marco disponible $\rightarrow \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = \frac{1}{N} \sum_{i=1}^{N'} X_i = \frac{1}{N} \cdot \frac{N'}{N'} \sum_{i=1}^{N'} X_i =$

$$= \frac{N'}{N} \cdot \frac{1}{N'} \sum_{i=1}^{N'} X_i = \frac{N'}{N} \bar{X}'$$ $\leftarrow$ marco depurado

$$\bar{X} = (1-W) \cdot \bar{X}'$$

Las varianzas en ambos marcos tampoco son iguales:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\bar X)^2 = \frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar X^2 = \frac{1}{N}\sum_{i=1}^{N'}X_i^2 - \bar X^2 =$$

$$= \frac{N'}{N}\cdot\frac{1}{N'}\sum_{i=1}^{N'}X_i^2 - \left[(1-W)\bar X'\right]^2 = (1-W)\cdot\frac{1}{N'}\sum_{i=1}^{N'}X_i^2 - (1-W)^2\bar X'^2 =$$

$$= (1-W)\left[\frac{1}{N'}\sum_{i=1}^{N'}X_i^2 - (1-W)\bar X'^2\right] = (1-W)\left[\frac{1}{N'}\sum_{i=1}^{N'}X_i^2 - \bar X'^2 + W\bar X'^2\right] =$$

$$= (1-W)\,\sigma'^2 + W(1-W)\,\bar X'^2$$

$$N\sigma^2 = \sum_{i=1}^{N}X_i^2 - \frac{X^2}{N} = \underbrace{\sum_{i=1}^{N'}X_i^2 - \frac{\overset{X'^2}{\overset{\|}{X^2}}}{N'}} + \frac{X^2}{N'} - \frac{X^2}{N} =$$

$$= N'\sigma'^2 + X^2\underbrace{\left(\frac{1}{N'}-\frac{1}{N}\right)}_{\frac{N-N'}{NN'} = W/N'} = N'\sigma'^2 + X^2\cdot\frac{W}{N'}$$

Para tamaños poblacionales sufic. grandes $\begin{cases} N \approx N-1 \\ N' \approx N'-1 \end{cases}$, las varianzas equivalen a las cuasivarianzas:

$$\longrightarrow S^2 = (1-W)\,S'^2 + W(1-W)\,\bar X'^2$$

ó bien

$$S^2 = (1-W)\,S'^2 + \frac{W}{1-W}\,\bar X^2$$

Definimos una var. auxiliar dicotómica que indica si la unidad Ai es no vacía:

$$T_i = \begin{cases} 1 & \text{si } A_i \text{ es no vacía} \\ 0 & \text{si } A_i \text{ es vacía} \end{cases}, \quad \text{donde } COV(X_i,T_i) = W\bar X$$

$$COV(X_i,T_i) = E(X_i,T_i) - E(X_i)E(T_i) = \underbrace{\frac{1}{N}\sum_{i=1}^{N}X_iT_i} - \left(\frac{1}{N}\sum_{i=1}^{N}X_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}T_i\right) =$$

$$= \frac{X}{N} - \frac{X}{N}(1-W) = \bar X\cdot(1-1+W) = \left\{\begin{array}{c} \frac{1}{N}\sum^{N}X_i = \frac{1}{N}\sum^{N'}X_i \\ \frac{X}{N} = \frac{X'}{N} \end{array}\right\} \quad \downarrow \frac{X}{N} \qquad \downarrow \frac{N'}{N} = 1-W$$

$$= W\bar X$$

Estimación del total en el marco depurado:

Tamaño poblacional $\to N'$
" muestral $\to n$
} elementos en M'

En m.a.s.s.r. (p.i.), los estimadores insesgados del total y su var:

$$\hat{X} = N'\bar{x} = N' \cdot \frac{x}{n} = \frac{N'}{n} \cdot x$$

$$V(\hat{X}) = N'^2 \cdot (1-f') \cdot \frac{S'^2}{n} = N'^2 \underbrace{\left(1 - \frac{n}{N'}\right)}_{\frac{N'-n}{N'}} \cdot \frac{S'^2}{n} = \frac{N'^{\cancel{2}}(N'-n)S'^2}{\cancel{N'} \cdot n}$$

Para la media: $e_i = \begin{cases} 1 & \text{si } \in \text{muestra} \\ 0 & \text{si } \notin \end{cases} \longrightarrow E[e_i] = n/N'$

$$\hat{X}' = \bar{x} = \frac{x}{n} \qquad E[x/n] = E\left[\sum_{i=1}^{n} \frac{x_i}{n}\right] = E\left[\sum_{i=1}^{N'} \frac{x_i}{n} e_i\right] = \sum_{i=1}^{N'} \frac{x_i}{n} E[e_i] =$$

$$\underset{\substack{\uparrow \\ \text{SR y p.i.}}}{=} \frac{1}{\cancel{n}} \cdot \frac{\cancel{n}}{N'} \sum_{i=1}^{N'} x_i = \frac{X}{N'} = \bar{X}' \quad \text{(insesgado)}$$

$$V(\hat{X}') = (1-f') \cdot \frac{S'^2}{n} = \frac{(N'-n)S'^2}{N' \cdot n}$$

Estimación del total en el marco no depurado:

a) Se desconoce N' y no se sustituyen las unid. vacías muestrales

Tamaño poblacional $\to N$
" muestral $\to n$
} elementos en M

En m.a.s.s.r., el estimador insesgado del total y su varianza:

$$\hat{X}_1 = N\bar{x} = N \cdot \frac{x}{n}$$

$$V(\hat{X}_1) = N^2(1-f) \cdot \frac{S^2}{n} = \frac{N^2(N-n)S^2}{Nn} = \frac{N(N-n)}{n} S^2$$

$$E[\hat{X}_1] = E\left[N \cdot \frac{x}{n}\right] = \frac{N}{n} E\left[\sum_{i=1}^{n} x_i\right] = \frac{N}{n} E\left[\sum_{i=1}^{N} x_i e_i\right] = \frac{N}{n} \sum_{i=1}^{N} x_i E[e_i] =$$

$$= \frac{N}{n} \cdot \frac{n}{N} \sum_{i=1}^{N} x_i = X \; (= X')$$

$$V(\hat{X}_1) = \frac{N(N-n)}{n} S^2 > V(\hat{X})$$

Lo vemos:

Sp. $N \approx N-1$ y $N' \approx N'-1$

$$\frac{S^2}{S'^2} = \frac{(1-W)S'^2 + W(1-W)\bar{X}'^2}{S'^2} = 1-W + W(1-W)\cdot\frac{\bar{X}'^2}{S'^2} \Rightarrow$$

$$\frac{V(\hat{X}_1)}{V(\hat{X})} = \frac{N(N-n)S^2}{N'(N'-n)S'^2} = \frac{N(N-n)}{N'(N'-n)}\left[1-W+W(1-W)\frac{\bar{X}'^2}{S'^2}\right] =$$

$$= \frac{\cancel{N}(N-n)}{\cancel{N'}(N'-n)}\left[\frac{N'}{N} + \frac{N'}{N}W\cdot\frac{\bar{X}'^2}{S'^2}\right] = \underbrace{\frac{N-n}{N'-n}}_{>1}\underbrace{\left[1+W\frac{\bar{X}'^2}{S'^2}\right]}_{>1} > 1.$$

Por lo que, cuando hay unidades vacías se verifica que la varianza del estimador del total en el marco no depurado es mayor que la varianza del estimador del total en el marco depurado. Será tanto mayor cuanto mayor sea W, proporción de unidades vacías, y menor sea la cuasivarianza relativa $\frac{S'^2}{\bar{X}'^2}$ de la característica estudiada en las unidades no vacías.

Para la media:

$$\hat{\bar{X}}_1 = \bar{x} = \frac{x}{n} = \sum_{i=1}^{n}\frac{X_i}{n}$$ estim. insesgado de $\bar{X}$

$$V(\hat{\bar{X}}_1) = (1-f)\cdot\frac{S^2}{n} = \frac{N-n}{N}\cdot\frac{S^2}{n}$$

si la comparamos con $V(\hat{\bar{X}}) = (1-f')\cdot\frac{S'^2}{n} = \frac{N'-n}{N'}\cdot\frac{S'^2}{n}$

$$\frac{V(\hat{\bar{X}}_1)}{V(\hat{\bar{X}})} = \frac{(N-n)/N\cancel{n}}{(N'-n)/N'\cancel{n}}\cdot\frac{S^2}{S'^2} = \frac{N'(N-n)}{N(N'-n)}\left[\frac{N'}{N} + \frac{N'}{N}W\cdot\frac{\bar{X}'^2}{S'^2}\right] =$$

$$= \underbrace{\frac{N'^2}{N^2}}_{<1}\cdot\underbrace{\frac{N-n}{N'-n}}_{>1}\cdot\underbrace{\left[1+W\cdot\frac{\bar{X}'^2}{S'^2}\right]}_{>1} \quad ??$$

b) Se desconoce N' y se sustituyen aleatoriamente las unidades vacías de la muestra hasta seleccionar n unid. no vacías

$\hat{X}_2 = \frac{N}{n} x$ tiene un sesgo positivo de valor $B_2 = \frac{W}{1-W} X'$

$$V(\hat{X}_2) = \frac{N^2 (N'-n) S'^2}{N' n} > V(\hat{X})$$

Nótese que la sustitución aleatoria de las unid. vacías por unid. no vacías equivale a trabajar en el marco depurado, por lo que $E[e_i] = \frac{n}{N'}$

$$E[\hat{X}_2] = \frac{N}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{N}{n} \sum_{i=1}^{N} X_i \underbrace{E[e_i]}_{n/N'} = \frac{N}{\cancel{n}} \cdot \frac{\cancel{n}}{N'} X' = \frac{1}{1-W} X'$$

por lo que el sesgo es

$$E[\hat{X}_2] = X' + B_2(\hat{X}_2) = \frac{1-W+W}{1-W} X' \Rightarrow B_2(\hat{X}_2) = \frac{W}{1-W} X'$$

$$V(\hat{X}_2) = N^2 V(x) = N^2 \cdot \frac{(N'-n)}{N'} \cdot \frac{S'^2}{n} = N^2 (1-f') \cdot \frac{S'^2}{n}$$

comparándola con $V(\hat{X})$.

$$\frac{V(\hat{X}_2)}{V(\hat{X})} = \frac{\frac{N^2 \cancel{(N'-n)}}{N'} \cdot \frac{\cancel{S'^2}}{\cancel{n}}}{N' \frac{\cancel{(N'-n)}}{\cancel{n}} \cdot \cancel{S'^2}} = \frac{N^2}{N'^2} = \frac{1}{(1-W)^2} > 1$$

Por lo que cuando existen unidades vacías y se sustituyen la varianza del estimador es mayor que la varianza del estimador trabajando en el marco depurado.

Además,

$$ECM(\hat{X}_2) = \frac{1}{(1-W)^2} \left[V(\hat{X}) + W^2 X'^2 \right].$$

Para la media:

$\hat{\bar{X}}_2 = \bar{x} = \frac{x}{n} \to E[\bar{x}] = \frac{X'}{N'} = \bar{X}'$ ⇒ insesgado

$$V(\hat{\bar{X}}_2) = V(\bar{x}) = (1-f') \cdot \frac{S'^2}{n}$$

c) Se conoce N' y no se sustituyen las unid. vacías muestrales

Tamaño muestral $= n$

de las que hay n' unid. muestrales no vacías ↳ trabajamos en M.

$\hat{X}_3 = N' \cdot \bar{x}' = \frac{N'}{n'} \cdot x$, estim. insesgado de X

APROX.

$V(\hat{X}_3) \cong \frac{N'(N-n)}{n} S'^2 =$

y se cumple $V(\hat{X}) < V(\hat{X}_3) < V(\hat{X}_1)$

$E[\hat{X}_3] = N' \cdot E[\bar{x}'] = N' \cdot E[\bar{x}'] = N'\bar{X}' = X'$

Para calcular la varianza se puede calcular de manera análoga al caso de estimadores de razón, porque n' es v.a.

$V(\hat{X}_3) > V(\hat{X})$ inmediata : si hay unidades vacías $N > N'$

$V(\hat{X}_3) < V(\hat{X}_1)$ porque $N'S'^2 < NS^2$

d) Se conoce N' y se sustituyen aleatoriamente las unidades vacías que aparezcan en la muestra hasta conseguir n no vacías

Tamaño poblacional $\to N'$

Tamaño muestral $\to n$

$\hat{X}_4 = N' \cdot \bar{x} = N' \cdot \frac{x}{n} = \frac{N'}{n} x$, insesgado

$V(\hat{X}_4) = N'^2 \cdot (1-f') \cdot \frac{S'^2}{n} = \frac{N'^2(N'-n) S'^2}{N' \cdot n} = \frac{N'(N'-n)S'^2}{n}$

$V(\hat{X}_4) = V(\hat{X}_0)$

Hay que tener en cuenta que $\frac{N'}{n}$ es uno de conocido, y que la sustitución aleatoria de las unid. vacías equivale a trabajar en el marco depurado

Conclusiones:

1_ Si se conoce el nº de unid. vacías, el problema puede resuelto aplicando $\hat{X}_4$, que es insesgado y tiene la misma varianza que si se trabaja en el marco depurado. Si no se sustituyen las unid. vacías, $\hat{X}_3$ tiene mayor varianza.

2_ Si no se conoce el nº de unid. vacías, es mejor no sustituir las que aparezcan en la muestra, porque $\hat{X}_1$ es insesgado y $\hat{X}_2$ es sesgado.

3_ En el caso de la media poblacional, $\bar{X}' = \frac{X'}{N'}$, el conocimiento de unidades vacías es irrelevante, ya que tanto $\hat{\bar{X}}_3 = \frac{x}{n'}$ y $\hat{\bar{X}}_4 = \frac{x}{n}$ son insesgado. Se prefiere 4 a 3, por tener menos varianza.

RESUMEN:

Marco	N' n'	Cambio unid. vacías	Estimador	Nº unid. No vacías	Varianza	Sesgo
~~No~~ Depurado			$\hat{X} = \frac{N'}{n} x$	n	$N'(N'-n)\frac{S'^2}{n}$	0
NO Depurado	NO	NO	$\hat{X}_1 = \frac{N}{n} x$	n'	$N(N-n)\frac{S^2}{n}$	0
		SÍ	$\hat{X}_2 = \frac{N}{n} x$	n	$N^2 \frac{(N'-n)}{N'} \frac{S'^2}{n}$	$\frac{W}{1-W} X'$
	SÍ	NO	$\hat{X}_3 = \frac{N'}{n'} x$	n'	$N'(N-n)\frac{S'^2}{n}$	0
		SÍ	$\hat{X}_4 = \frac{N'}{n} x$	n	$N'(N'-n) \cdot \frac{S'^2}{n'}$	0