MUEST_T18. EL MODELO de ERROR TOTAL en censos y encuestas.
FORMULACIÓN del MODELO.
ESTIMACIÓN del SESGO y de la VARIANZA de respuesta.
MEDIDA del EFECTO del ENTREVISTADOR.
SUBMUESTRAS INTERPENETRANTES
Ø – INTRODUCCIÓN: Muestreo → error → ajenos (debidos) | muestreo

1_EL MODELO de ERROR TOTAL en Censos y encuestas

En el conjunto del proceso estadístico de un censo o una encuesta intervienen diversos elementos (instrumentos y operaciones) que son fuentes potenciales de error.

Además de los sesgos que pueden producir un cuestionario mal diseñado (definición imprecisa de unidades y conceptos), también las operaciones de recogida, transcripción y grabación de los datos pueden dar lugar a que se produzcan desviaciones de los valores observados respecto de los valores reales.

Por otra parte, no debemos olvidar los efectos debidos al trabajo de los encuestadores, a la situación objetiva y subjetiva de los entrevistados, y a la interacción entre unos y otros. Tales errores se denominan "ajenos al muestreo" para distinguirlos de los producidos por la variabilidad de las muestras, "errores de muestreo".

Un censo, aunque no está sometido al error de muestreo, sufre los efectos de los errores ajenos al muestreo. El error es mayor que en una encuesta debido al mayor nº de operaciones y de personas que intervienen en el proceso, por lo que es más difícil mantener bajo control la calidad de los trabajos.

EAM ← | – Cuestionario | • mal diseñado • preguntas difíciles/dirigidas • largo
– Tipo encuesta | • total • entrevistador
– Unidades | • Motivo personales
– Entrevistadores | • mal adiestrado
| interacción
| - Codificad

El modelo de error total describe el ECM de una estimación sometida a errores de muestreo y ajenos al muestreo.

2_FORMULACION del MODELO (debido a Hansen, Hurwitz y Bershad)

De una población finita $U=\{U_1 \dots U_N\}$ estudiamos una característica X que toma los valores $\{X_1 \dots X_N\}$.

→ Añadir var. modelico.

Debido a los errores ajenos al muestreo, podemos observar un valor de la característica ≠ del verdadero.

Para cada U_i / $X_i \rightarrow$ valor verdadero
\ $Y_{it} \rightarrow$ valor observado en el instante t.

$$E_i(Y_{it}) = E(Y_{it}/_i) = Y_i \neq X_i$$

$$\Rightarrow E(Y_{it}) = E E_i(Y_{it}) = E(Y_i) = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y} \leftarrow \text{media observable}$$

Para cada observación definimos:

- Desviación de respuesta de U_i: $d_{it} = Y_{it} - Y_i$
  $E[d_{it}] = 0 \Rightarrow V(d_{it}) = E[d_{it}^2] - 0^2$
  Variante unitaria de respuesta → $\sigma_R^2 = V(d_{it}) = E\left[(Y_{it}-Y_i)^2\right]$
- Desviación de muestreo de U_i: $\delta_i = Y_i - \bar{Y}$
  $E[\delta_i] = \bar{Y} - \bar{Y} = 0 \;/\; E[\delta_i \cdot \delta_j] = 0$.
  Variante unitaria de muestreo → $V(\delta_i) = E(\delta_i^2) = \sigma_M^2$
- sesgo de respuesta de $U_i \rightarrow B_i = Y_i - X_i$
  Sesgo total de respuesta → $B = E(Y_i - X_i) = \bar{Y} - \bar{X}$
- Coef. de correlación entre las dev. respuestas:
  $$\rho(d_{it}, d_{jt}) = \frac{Cov(d_{it}, d_{jt})}{\sigma(d_{it})\,\sigma(d_{jt})} = \frac{E(d_{it}\cdot d_{jt})}{\sigma_R^2} = \rho_R \quad i \neq j.$$
- Coef. de correlación entre desviaciones de respuesta y desviaciones de muestreo:
  $$\rho(d_{it}, \delta_i) = \frac{Cov(d_{it}, \delta_i)}{\sigma(d_{it})\,\sigma(\delta_i)} = \frac{E(d_{it}\cdot\delta_i)}{\sigma_R \sigma_M} = \rho_{RM}$$

El sesgo de respuesta se debe a la inclusión en los datos de un error de carácter sistemático que sería consistente al repetir idealmente la encuesta o censo en condiciones análogas.

Error total en un censo ⊛

$$ECM(\bar{Y}_t) = E(\bar{Y}_t - \bar{X}_t)^2 = E(\bar{Y}_t - \bar{Y} + \bar{Y} - \bar{X})^2 =$$

$$= E(\bar{Y}_t - \bar{Y})^2 + E(\bar{Y} - \bar{X})^2 + 0 = E\left[\frac{1}{N}\sum_{i=1}^{N}(Y_{it} - Y_i)\right]^2 + B^2 =$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N} d_{it}\right]^2 + B^2$$

Como $E\left[\frac{1}{N}\Sigma d_{it}\right]^2 = \frac{1}{N^2} E\left[\sum_i d_{it}^2 + \sum_{i\neq j} d_{it} d_{jt}\right] =$

$$= \frac{1}{N^2}\left[\underbrace{\sum_i E(d_{it}^2)}_{N \cdot \sigma_R^2} + \underbrace{\sum_{i\neq j} E(d_{it} d_{jt})}_{N(N-1)\,\rho_R \cdot \sigma_R^2}\right] =$$

Por lo que:

$$ECM(\bar{Y}_t) = \underbrace{\underbrace{\frac{\sigma_R^2}{N}}_{\text{varianza de respuesta simple}} + \underbrace{\frac{N-1}{N}\cdot \rho_R \sigma_R^2}_{\text{correlación entre respuestas}}}_{\text{Varianza de respuesta}} + \underbrace{B^2}_{\text{sesgo}^2}$$

Cuando N es grande, la varianza ~~simple~~ de respuesta simple pierde importancia, pero no ocurre así con la componente correlacionada de la varianza total de respuesta.

El sesgo de respuesta es el promedio de los sesgos individuales y se asocia a características del procedimiento independientes de errores accidentales.

⊛ $U = \{U_1 \dots U_N\}$ (verdadero valor paramétrico, $\bar{Y}$) $\xrightarrow{\text{CENSO}} \{Y_{1t} \dots Y_{Nt}\} \longrightarrow \bar{Y}_t = \sum_{i=1}^{N} \frac{Y_{it}}{N}$.

Ahora vamos a calcular el error total asociado a $\bar{Y}_t$:

Error total en una encuesta

En una encuesta, utilizamos la media muestral de las observaciones, $\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{it}$, $E[\bar{y}_t] = \bar{Y}$

$$ECM(\bar{y}_t) = E(\bar{y}_t - \bar{X})^2 = E(\bar{y}_t - \bar{y} + \bar{y} - \bar{Y} + \bar{Y} - \bar{X})^2 =$$
$$= E(\bar{y}_t - \bar{y})^2 + E(\bar{y} - \bar{Y})^2 + E(\bar{Y} - \bar{X})^2 + 2E(\bar{y}_t - \bar{y})(\bar{y} - \bar{Y}) + 0 + 0$$

donde $E(\bar{y}_t - \bar{y})^2 \overset{\text{igual que antes}}{=} \frac{\sigma_R^2}{n} + \frac{n-1}{n}\rho_R \cdot \sigma_R^2$

$$E(\bar{y} - \bar{Y})^2 = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_{it} - \bar{Y})\right]^2 = E\left[\frac{1}{n}\sum_{i}^{n}\delta_i\right]^2 =$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} E(\delta_i^2) + \sum_{i\neq j}^{n} \underbrace{E(\delta_i \cdot \delta_j)}_{=0}\right] = \frac{1}{n^2} \cdot n\,\sigma_M^2$$

$$E(\bar{Y} - \bar{X})^2 = B^2$$

$$E(\bar{y}_t - \bar{y})(\bar{y} - \bar{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_{it} - Y_i) \cdot \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] =$$
$$= \frac{1}{n^2}E\left[\sum_{i=1}^{n} d_{it} \cdot \sum_{i=1}^{n}\delta_i\right] =$$
$$= \frac{1}{n^2}E\left[\sum_{i=1}^{n} d_{it}\delta_i + \sum_{i\neq j}^{n} d_{it}\delta_j\right] =$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} \underbrace{E(d_{it}\delta_i)}_{\rho_{RM}\sigma_R\sigma_M} + \sum_{i\neq j}^{n} \underbrace{E(d_{it}\delta_j)}_{=0}\right] =$$
$$= \rho_{RM}\sigma_R\sigma_M$$

luego $ECM(\bar{y}_t) = \underbrace{\frac{\sigma_R^2}{n} + \frac{n-1}{n}\rho_R\sigma_R^2}_{\text{varianza de respuesta}} + \frac{\sigma_M^2}{n} + \frac{2}{n}\rho_{RM}\sigma_R\sigma_M + B^2$

- $\frac{\sigma_R^2}{n}$: var. resp. simple
- $\frac{n-1}{n}\rho_R\sigma_R^2$: correlac. de respuesta
- $\frac{\sigma_M^2}{n}$: varianza muestral
- $\frac{2}{n}\rho_{RM}\sigma_R\sigma_M$: correlación entre la respuesta y el muestreo
- B^2: sesgo^2

El error cuadrático medio en las encuestas es igual al error cometido en el censo más la variabilidad debida al muestreo y la correlación entre el muestreo y la respuesta, que al aumentar el tamaño de la muestra se anulan.

De los restantes términos, el sesgo es siempre constante y la correlación de respuesta permanece constante o incluso aumenta.

$ECM(\bar{y}_t) = \text{Variabilidad total} + \text{sesgo}^2$

Variabilidad total = Variabilidad total de respuesta (var. simple de respuesta + comp. correlacionada de la var. total respuesta)

+ Variabilidad del muestreo

+ Interacción entre desviaciones de respuesta y desviac. de muestreo

---

En el caso del censo ($n = N$), desaparecen la componente debida al muestreo y la interacción

? BUSCAR

3_ ESTIMACIÓN del SESGO y de la VARIANZA TOTAL de RESPUESTA.

Supongamos que se selecciona una submuestra aleatoria de entre todos los agentes que han participado en la encuesta o censo, y se repiten sus observaciones por agentes independientes con adiestramiento similar (En instante t' y tamaño m).

Utilizamos como estimador de la varianza total de respuesta

$$\hat{V}TR = \frac{1}{2}J = \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m}(\underbrace{Y_{it}}_{1^a\text{ vuelta}} - \underbrace{Y_{it'}}_{2^a\text{ vuelta}})\right]^2$$

$$E[\tfrac{1}{2}J] = \frac{1}{2}\cdot\frac{1}{m^2}E\left[\sum_{i=1}^{m}(d_{it}-d_{it'})\right]^2 = \frac{1}{2m^2}E\left[\sum_{i=1}^{m}(d_{it}-d_{it'})^2 + \sum_{j\neq i}^{m}(d_{it}-d_{it'})(d_{jt}-d_{jt'})\right] = \frac{1}{2m^2}\left[\sum_{i=1}^{m}\underbrace{E(d_{it}-d_{it'})^2}_{2\sigma_R^2} + \sum_{i\neq j}^{m}\underbrace{E(d_{it}-d_{it'})(d_{jt}-d_{jt'})}_{2\rho_R\sigma_R^2}\right] = \frac{\sigma_R^2}{m} + \frac{m-1}{m}\rho_R\sigma_R^2$$

Cuando se utiliza el método de la entrevista se ha encontrado que la componente correlacionada es en general mucho mayor que la varianza de respuesta simple, excepto en caracteres como edad, sexo y estado civil.

Como estimador de la varianza de respuesta simple

$$\hat{V}RS = \frac{1}{2}G = \frac{1}{2}\left[\frac{1}{m}\sum_{i=1}^{m}(Y_{it}-Y_{it'})^2\right]$$

$$E[\tfrac{1}{2}G] = \frac{1}{2m}\sum_{i=1}^{m}E(d_{it}-d_{it'})^2 = \frac{1}{2m}m\,2\sigma_R^2 = \sigma_R^2$$

Para garantizar la independencia entre Y_{it} e $Y_{it'}$, se recomienda que el 2º entrevistador no conozca los datos de la 1ª entrevista, ya que el "factor memoria" puede dar lugar a $COV(d_{it}, d_{it'}) > 0 \Rightarrow Y_{it}$ DEPENDIENTE $Y_{it'}$, por

lo que el estimador por defecto de σ_R^2 sería $\frac{1}{2}\sigma$, que tb. ocurre si uno de los entrevistadores está mejor adiestrado.

~~Hansen, Hurwitz y Pritzker (1964)~~

Cuando se trata de una variable cualitativa binaria:

Entr. repetida \ Entrevist. original	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	m

$\hat{VRS} = \hat{\sigma}_R^2 = \frac{b+c}{2m}$ → estimador de la var. respuesta simple cuando las entrevistas son indep

$\hat{B} = \frac{b-c}{m} \cdot 100$ → estimador del sesgo cuando la 2ª observación se considera superior a la 1ª.

↑ tasa de dif. neta en %.

$\frac{b-c}{a+b} \cdot 100$ → índice de cambio neto

$\frac{b+c}{m} \cdot 100$ → tasa de dif. bruta

$\frac{b+c}{a+b} 100$ → índice de cambio bruto.

Estimador	Valor esperado
$T_1 = \frac{c+b}{2m^2}$	σ_R^2/m
$T_2 = \frac{(c-b)^2}{2m^2}$	$\sigma_R^2/m\,(1+(m-1)\rho_R)$
$T_3 = \frac{(a+c)(b+d)}{(m-1)m^2}$	$\frac{\sigma_R^2+\sigma_M^2}{m}$

T_1 estima VRS

$T_2 - T_1$ estima comp. correl.

⇒

$T_3 - T_1$ estima comp. sesgo

$\frac{T_2 - T_1}{T_1(m-1)}$ estima correlación entre desviaciones

4_MEDIDA del EFECTO del ENTREVISTADOR

Un censo o encuesta efectuado con entrevistadores disminuye la no respuesta pq. el entrevistador puede ayudar y aumenta la calidad pq. el entrevistador puede añadir preguntas dependiendo de las respuestas.

Pero puede ocurrir que los entrevistadores ejerzan efectos distintos sobre las respuestas, debidos a su conducta individual y al diferente grado de adiestramiento. Los cursos de formación tratan de conseguir uniformidad de criterio de los entrevistadores, pero el efecto del entrevistador en la práctica es inevitable.

Recordemos que

$$VTR = VRS + \text{Correlac. entre desviaciones}$$

Llamamos $Y_{iu} \equiv$ respuesta de la unidad u_i al entrevistador h.

$d_{ih} = Y_{ih} - X_i \rightarrow$ desviación de respuesta debida al entrevistador h.

$$E[d_{ih}] = E[Y_{ih}] - E[X_i] = E_h E_i[Y_{ih}] - E_h E_i(X_i) =$$
$$= E_h[\bar{Y}_h] - E_h(\bar{X}) = \bar{Y} - \bar{X} = B \leftarrow \text{sesgo.}$$

($\bar{Y}_h$: media poblac. del entrev. h; $\bar{Y}$: media poblac. de todos los entrev.)

$$\sigma_Y^2 = \sigma_W^2 + \sigma_b^2$$

$\sigma_b^2 \rightarrow$ varianza entre entrevistadores, medida del efecto del entrevistador

Veamos los ≠ casos que pueden presentarse:

1) Todos los entrevistadores recogen datos verdaderos:

$Y_{ih} = X_i \ , \forall h \Rightarrow \nexists$ desv. respuesta $\Rightarrow$ | $\nexists$ sesgo, $\nexists$ efecto entrevist.

(Situación ideal)

$\Rightarrow \bar{Y} = \bar{X}$

$\Rightarrow \sigma^2_Y = \sigma^2_X$.

2) Los errores de respuesta se compensan dentro de cada entrevistador

$d_{ih} \neq 0$, pero $E[d_{ih}] = 0 \ , \forall h$.

$\Rightarrow \bar{Y}_h = \bar{Y} = \bar{X} \Rightarrow \nexists$ sesgo

$\nexists$ efecto entrevistador

$\Rightarrow \sigma^2_Y = V(d_{ih}) + \sigma^2_x + 2\,cov(d_{ih}, d_x)$.

No existe sesgo ni efecto del entrevistador, pero la var. del modelo aumenta al aumentar la variabilidad de respuesta.

3) Todos los entrevistadores recogen datos cuyos promedios tienen el mismo sesgo.

$d_{ih} \neq 0$

$\bar{Y}_h = \bar{Y} \Rightarrow \sigma^2_b = 0 \Rightarrow \nexists$ efecto entrevistador

$\bar{Y}_h - \bar{X} = B = cte \Rightarrow \exists$ sesgo

$\Rightarrow \sigma^2_Y = V(d_{ih}) + \sigma^2_x + 2\,cov(d_{ih}, d_x) + B^2$

4) El valor promedio obtenido por uno o más entrevistadores es distinto a los demás.

$\sigma^2_b > 0 \Rightarrow$ | $\exists$ efecto entrevistador |

El sesgo puede existir o no, dependiendo de que el promedio $E[Y_h] = \bar{Y}$ sea igual o distinto a $\bar{X}$.

En general,

$$\sigma^2_Y = V(d_{ih}) + \sigma^2_x + 2\,cov(d_{ih}, d_x) + B^2$$

5_ SUBMUESTRAS INTERPENETRANTES

Supongamos que una muestra aleatoria de tamaño n se divide aleatoriamente en K submuestras de tamaño m ($n = Km$), que se asignan a K agentes independientes (≠ correlación entre sus observaciones).

Podemos distinguir dos fuentes de variación: entre agentes ó submuestras y dentro de las submuestras:

$$SCT = SCD + SCE$$

$$\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{m}(\bar{Y}_i-\bar{Y})^2$$

que se puede resumir en el cuadro ANOVA:

Fuente	G.L.	Suma cuadr	Cuadr. medio
Entre subm	$K-1$	$\sum_{i=1}^{K}\sum_{j=1}^{m}(\bar{Y}_i-\bar{Y})^2$	$\hat{S}_b^2 = \frac{SCE}{K-1}$
Dentro subm	$K(m-1)$	$\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_i)^2$	$\hat{S}_W^2 = \frac{SCD}{K(m-1)}$
Total	$\underbrace{n-1}_{Km}$	$\sum_{i=1}^{K}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2$	$\hat{S}^2 = \frac{SCT}{\underset{Km}{(n)}-1}$

H_0: ≠ efecto del entrevistador $\Rightarrow \sigma_Y^2 = \hat{S}^2$

$\left.\begin{matrix} H_0: \sigma_b^2 = 0 \\ H_1: \sigma_b^2 \neq 0 \end{matrix}\right\}$

Estadístico: $\dfrac{\hat{S}_b^2}{\hat{S}_W^2} \longrightarrow F_{K-1,\,K(m-1)}$

Decisión: Si $\dfrac{\hat{S}_b^2}{\hat{S}_W^2} > F_{K-1,\,K(m-1),\,\alpha}$ rechazo $H_0 \Rightarrow$

$\Rightarrow \sigma_Y^2 = \sigma_b^2 + \sigma_W^2$

$E[\hat{S}_W^2] = \sigma_W^2$

$E[\hat{S}_b^2] = \sigma_W^2 + \sigma_b^2$

} luego $\hat{\sigma}_b^2 = \hat{S}_b^2 - \hat{S}_W^2$