MUEST_T16. ERRORES AJENOS AL MUESTREO I: MARCOS IMPERFECTOS. EL PROBLEMA de las UNIDADES VACÍAS. ESTIMACIÓN del TOTAL y de la MEDIA. CÁLCULO de la VARIANZA y COMPARACIÓN con la VARIANZA del MARCO DEPURADO.

---

1. ERRORES AJENOS AL MUESTREO

En las encuestas por muestreo puede definirse el "error" de una determinada estimación como la diferencia entre el valor observado $\hat{\theta}$ y el valor desconocido de la característica poblacional θ que tratamos de estimar.

$$\text{error} = |\hat{\theta} - \theta|$$

Los errores se deben a causas diversas, pudiendo clasificarse en errores de carácter aleatorio y errores de carácter sistemático o sesgos.

Pueden originarse errores en los resultados de una muestra particular debido a los respondientes, a los encuestadores, codificadores, etc, así como a la posible interdependencia entre ellos.

Los errores de carácter aleatorio y los errores de carácter sistemático tienen, en general, distintas fuentes, efectos y métodos de medida. La reducción de los errores aleatorios requiere hacer "más de algo" (aumentar el tamaño muestral), mientras que la reducción de los errores sistemáticos requiere hacer "algo más" (supervisar, controlar...)

Otra clasificación distingue entre errores de muestreo (originados por la variabilidad de los valores obtenidos en el muestreo, de carácter aleatorio) y los errores ajenos al muestreo (que se producen por causas no proba-

bilísticos, ou errores no aleatorios).

Mientras los errores de muestreo decrecen al aumentar el tamaño de la muestra, los errores ajenos al muestreo suelen crecer al aumentar el tamaño de la investigación, o en el caso no decrecen.

Los errores de muestreo se estiman con los datos de la muestra, los errores ajenos al muestreo suelen requerir para su estimación datos extramuestrales.

- Errores
  - de muestreo (aleatorios) → $E(\hat{\theta}-\theta)^2$ → Exactitud; $E(\hat{\theta}-E(\hat{\theta}))^2$ → Varianza; $\sigma(\hat{\theta})$ → Error de muestreo
  - ajenos al muestreo (sistemáticos) → Errores de cobertura; Errores de respuesta; Falta de respuesta

- Errores de cobertura ≡ marco imperfecto
  El marco no se corresponde con la poblac. objetivo y subestima o sobreestima la característica.

- Errores de respuesta ≡ debidos al proceso de recogida de la información (la encuesta) y tratam. distintas respuestas, posibles influencias de entrevistadores, codificadores, etc.

- Falta de respuesta ≡ imposibilidad de obtener toda la información por falta de unidades respondientes, cuestionarios incompletos...

ARTICULO 50

1. LAS INFRACCIONES SE CLASIFICAN EN MUY GRAVES, GRAVES Y LEVES.

2. SON INFRACCIONES MUY GRAVES:

A) EL INCUMPLIMIENTO DEL DEBER DEL SECRETO ESTADISTICO.

B) LA UTILIZACION PARA FINALIDADES DISTINTAS DE LAS ESTADISTICAS DE LOS DATOS PERSONALES OBTENIDOS DIRECTAMENTE DE LOS INFORMANTES POR LOS SERVICIOS ESTADISTICOS.

C) EL SUMINISTRO DE DATOS FALSOS A LOS SERVICIOS ESTADISTICOS COMPETENTES.

D) LA RESISTENCIA NOTORIA, HABITUAL O CON ALEGACION DE EXCUSAS FALSAS EN EL ENVIO DE LOS DATOS REQUERIDOS, CUANDO HUBIERE OBLIGACION DE SUMINISTRARLOS.

3 graves en 1 año

E) LA COMISION DE UNA INFRACCION GRAVE CUANDO EL INFRACTOR HUBIERE SIDO SANCIONADO POR OTRAS DOS GRAVES DENTRO DEL PERIODO DE UN AÑO.

3. SON INFRACCIONES GRAVES:

A) LA NO REMISION O EL RETRASO EN EL ENVIO DE LOS DATOS REQUERIDOS CUANDO SE PRODUJESE GRAVE PERJUICIO PARA EL SERVICIO, Y HUBIERE OBLIGACION DE SUMINISTRARLOS.

B) EL ENVIO DE DATOS INCOMPLETOS O INEXACTOS CUANDO SE PRODUJESE GRAVE PERJUICIO PARA EL SERVICIO, Y HUBIERE OBLIGACION DE SUMINISTRARLOS.

3 leves en 1 año

C) LA COMISION DE UNA INFRACCION LEVE CUANDO EL INFRACTOR HUBIERA SIDO SANCIONADO POR OTRAS DOS LEVES DENTRO DEL PERIODO DE UN AÑO.

4. SON INFRACCIONES LEVES:

A) LA REMISION O EL RETRASO EN EL ENVIO DE DATOS CUANDO NO HUBIERE CAUSADO PERJUICIO GRAVE PARA EL SERVICIO, Y HUBIERE OBLIGACION DE SUMINISTRARLOS.

B) EL ENVIO DE DATOS INCOMPLETOS O INEXACTOS CUANDO NO HUBIERE CAUSADO PERJUICIO GRAVE PARA EL SERVICIO, Y HUBIERE OBLIGACION DE SUMINISTRARLOS.

ARTICULO 51 Sanciones: de 10000 a 5.000.000 ptas

1. LAS INFRACCIONES MUY GRAVES SERAN SANCIONADAS CON MULTAS DE 500.001 A 5.000.000 DE PESETAS.

2. LAS INFRACCIONES GRAVES SERAN SANCIONADAS CON MULTAS DE 50.001 A 500.000 PESETAS.

10.000 – 50.000
50.001 – 500.000
500.001 – 5.000.000

3. LAS INFRACCIONES LEVES SE SANCIONARAN CON MULTAS DE 10.000 A 50.000 PESETAS.

4. LA CUANTIA DE LAS SANCIONES ESTABLECIDAS EN LOS APARTADOS ANTERIORES SE GRADUARA ATENDIENDO, EN CADA CASO, A LA PROPIA GRAVEDAD DE LA INFRACCION, A LA NATURALEZA DE LOS DAÑOS Y PERJUICIOS CAUSADOS Y A LA CONDUCTA ANTERIOR DE LOS INFRACTORES.

ARTICULO 52

2_FALTA de RESPUESTA y sus EFECTOS

Prácticamente en todos los casos (muestreo exhaustivo) y ~~casi~~ todas las encuestas (por muestreo), en las que se utilizan cuestionarios y entrevistadores para la recogida de datos se produce el problema de la falta de respuesta ⇒ falta información de una parte de la población o de la muestra seleccionada.

Falta de respuesta total → una unidad respondiente no ha contestado ninguna pregunta del cuest.

Falta de respuesta parcial → la unidad ha dejado alguna pregunta sin responder.

Las posibles causas del problema de la no respuesta son:

(↗) MIRAR ATRÁS

1_ Imposibilidad de identificar la unidad sobre el terreno o de acceder a la misma (zonas difíciles, inclemencias tiempo...)

2_ Ausencia temporal del entrevistado

3_ Incapacidad para contestar por parte del entrevistado

↳ Contribuye el hecho de que el respondiente deba ser una persona específica, al no ser válida la respuesta de otra persona del hogar, p. ej.

4_ Negativa a cooperar en la encuesta por parte del entrevistado

↳ Por razones personales, subjetivas

↳ Si las preguntas son incómodas (íntimas, económicas)

↳ Por temas de cumplimentación y reenvío en cuest. postales

→ Destacar la legislación estadística, que obliga a cumplimentar los cuestionarios del INE. (ATRÁS)

5_ Pérdida de información

↳ En el proceso de recogida, codificación...

↳ Falta de conocimiento del entrevistador

adiestramiento, motivación...

EFECTOS:

Población: N $\begin{cases} N_1 & \text{, que contestan} \\ N_2 & \text{, que no contestan} \longrightarrow W = \frac{N_2}{N} \end{cases}$ } estrato

$W \equiv$ proporción de los no respondientes

los parámetros poblacionales quedan:

$$X = \sum_{i=1}^{N} X_i = \not{\Phi} X_1 + X_2$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = (1-W)\bar{X}_1 + W\bar{X}_2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2 = \sum_{h=1}^{2}\frac{N_h}{N}\sigma_i^2 + \sum_{h=1}^{2}\frac{N_h}{N}(\bar{X}_h-\bar{X})^2$$

(dentro estrato) (entre estr. con total)

La falta de respuesta produce:

1_ Disminución del tamaño de la muestra $\Rightarrow$ disminuye la precisión $n \longrightarrow n_1$

2_ Aparece un sesgo independiente del tamaño muestral

El primer problema se puede solucionar eligiendo un tamaño muestral suficientemente grande, para que n_1 tb. lo sea.

El segundo problema sí que es importante, porque:

$\bar{x} \longrightarrow \bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ con $E[\bar{x}_1] = \bar{X}_1 \Rightarrow$ sesgado para $\bar{X}$

cuyo sesgo es

$$B(\bar{x}_1) = E(\bar{x}_1) - \bar{X} = \bar{X}_1 - \bar{X} = \bar{X}_1 - (1-W)\bar{X}_1 - W\bar{X}_2 = W(\bar{X}_1 - \bar{X}_2)$$

$\Rightarrow$ el sesgo es proporcional al nº de unid. no respondientes.

$\bar{x}$ / $B(\bar{x})$ $\begin{cases} \cdot \text{ indep. de } n \\ \cdot \text{ proporcional a } N_2 \end{cases}$

En el caso del total poblacional, si consideramos $\hat{X}_1 = N\bar{x}_1$, también es sesgado pues $E[\hat{X}_1] = N\bar{X}_1$

$$B(\hat{X}_1) = NW(\bar{X}_1 - \bar{X}_2)$$

$\hat{X}_2 = \frac{N}{n}\sum_{i=1}^{n_1} X_i$ tb. es sesgado.

La cuasivariauza muestral tb. es sesgado para la variauza poblacional.

$$\hat{X}_1 \equiv N\bar{x}_1 = \frac{N}{n_1}\sum_{i=1}^{n_1} X_i \quad \text{con} \quad E[\hat{X}_1] = N\cdot\bar{X}_1$$
$$B(\hat{X}_1) = N(\bar{X}_1 - \bar{X}_\cdot)$$

$$\hat{X}_2 = \frac{N}{n}\sum_{i=1}^{n_1} X_i \quad \text{con} \quad E[\hat{X}_2] = \frac{N}{n}\cdot E\left[\sum_{i=1}^{n_1} X_i\right] = \frac{N}{n}\, n_1 \cdot \bar{X}_1$$
$$B(\hat{X}_2) = E[\hat{X}_2] - X$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{x}_1)^2 \quad \text{con} \quad E[s_1^2] = \sigma_1^2 \neq \sigma^2$$
$$B(s_1^2) = \sigma_1^2 - \sigma^2 = W(\sigma_1^2 - \sigma_2^2).$$

Parámetro	Estimador	Sesgo
$X = X_1 + X_2$	$\hat{X}_1 = N\cdot\bar{x}_1$	$N\cdot W_2(\bar{X}_1 - \bar{X}_2)$
	$\hat{X}_2 = N\cdot\frac{\sum X_i}{n}$	$\frac{N}{n}\cdot n_1\cdot\bar{X}_1 - X$
$\bar{X} = W_1\bar{X}_1 + W_2\bar{X}_2$	$\bar{x}_1 = \frac{\sum X_i}{n_1}$	$W_2(\bar{X}_1 - \bar{X}_2)$
$\sigma^2 = \sum_{h=1}^{2}\frac{N_h}{N}\sigma_h^2 + \sum_{h=1}^{2}\frac{N_h}{N}(\bar{X}_h - \bar{X})^2$	$\hat{S}_1^2 = \frac{1}{n_1 - 1}\sum(X_i - \bar{x}_1)^2$	$W_2(\sigma_1^2 - \sigma_2^2)$

Modelos de Deming (1952)

Procedimiento para aminorar la no respuesta, especialmente producida por ausencia de la persona específica seleccionada para la encuesta, realizando visitas sucesivas.

Para ello se fija un nº mínimo de "revisitas" que deben hacerse a cada unidad antes de abandonarlo como "contacto imposible".

Este procedimiento encarece el coste de la encuesta y puede retrasar considerablemente los resultados finales → planificar previamente la duración de la recogida de datos.

Deming desarrolló el siguiente modelo:
se divide la población en L clases de acuerdo con la probabilidad de que el encuestado se encuentre en casa.

w_{ij} = probab. de que un encuestado de la clase j sea entrevistado en la visita i

P_j = proporción de la población de la clase j.

$\bar{X}_j$, G^2_j } media y variance pobl. de la clase j.

Sp $w_{ij} > 0 \quad \forall\, j = 1 \dots L$.

Después de i visitas la composición de la muestra está formada por L clases, con elementos del estrato j que han respondido en los i primeros intentos, más una clase que no ha respondido en estos i primeros intentos.

Si n_0 es el tamaño inicial de la muestra, n_j es una multinomial.
Después de i visitas, el nº de respuestas n_i es una binomial:

$$n_i \longrightarrow B\left(n_0, \sum_{j=1}^{L} w_{ij} P_j\right)$$

$$E[n_i] = n_0 \cdot \sum_{j=1}^{L} w_{ij} P_j$$

El total muestral se puede escribir como

$$X = X_1 + X_2$$

donde $X_2 \equiv$ unidades sin respuesta en la primera vuelta.
Si llamamos $f_{21} \equiv$ fracción de muestreo en el estrato 2 en segunda vuelta, $f_{21} = \frac{n_{21}}{n_2}$, un estimador insesgado de X_2 es: $\hat{\hat{X}}_2 = \frac{n_2}{n_{21}} x_{21} = n_2 \cdot \bar{x}_{21}$ $\left(\bar{x}_{21} = \sum \frac{x_i'}{n_{21}}\right)$.

$$\hat{X} = \hat{X}_1 + \hat{X}_2 = N\left(\frac{1}{n} x_1 + \frac{1}{n} \hat{\hat{X}}_2\right) = \frac{N}{n}\left(x_1 + \hat{\hat{X}}_2\right)$$

$$E[\hat{X}] = E\, E_{\underbrace{n_1 n_2}_{\text{fijos}}}(\hat{X}) = E\left[\frac{N}{n}\left(x_1 + n_2 \underbrace{E_{n_1 n_2}(\bar{x}_{21})}_{\bar{x}_2}\right)\right] = E\left(\frac{N}{n} X\right) = E[N\bar{x}] = X$$

luego $\hat{X}$ es insesgado para X.

$$V(\hat{X}) = E\, V_{n_1 n_2}(\hat{X}) + V\, E_{n_1 n_2}(\hat{X})$$

donde $V_{n_1 n_2}(\hat{X}) = V_{n_1 n_2}\left[\frac{N}{n}\left(x_1 + \hat{\hat{X}}_2\right)\right] = \left(\frac{N}{n}\right)^2 V_{n_1 n_2}\left(\hat{\hat{X}}_2\right) = \left(\frac{N}{n}\right)^2 n_2^2\, V_{n_1 n_2}(\bar{x}_{21}) =$

$$= \frac{1}{f^2} \cdot n_2^2 \cdot \frac{n_2 - n_{21}}{n_2} \cdot \frac{S_{21}^2}{n_{21}} = \frac{1 - f_{21}}{f^2 \cdot f_{21}} \cdot n_2 S_{21}^2$$

$$E\, V_{n_1 n_2}(\hat{X}) = \frac{1 - f_{21}}{f^2 \cdot f_{21}} E[n_2 S_{21}^2] = \frac{1 - f_{21}}{f^2 \cdot f_{21}} n S_2^2 \underbrace{E\left[\frac{n_2}{n}\right]}_{N_2/N} =$$

$$= \frac{1 - f_{21}}{f \cdot f_{21}} N_2 S_2^2$$

$$E_{n_1 n_2}(\hat{X}) = \frac{N}{n} X$$

$$V\, E_{n_1 n_2}(\hat{X}) = V\left(\frac{N}{n} x\right) = N^2 V(\bar{X}) = N^2 \cdot \frac{1 - f}{n} \cdot S^2$$

Entonces: $V(\hat{X}) = \frac{1 - f_{21}}{f \cdot f_{21}} N_2 S_2^2 + N^2 \cdot \frac{1 - f}{n} S^2$

donde $S^2 \equiv$ cuasiv. poblacional y $S_2^2 \equiv$ cuasiv. pobl. estrato 2.

que puede expresarse como: $V(\hat{\bar{X}})$ es la suma de la varianza del estimador del total si no hubiera falta de respuesta más el incremento de varianza debido al submuestreo de los no respondientes.

Para estimar la media poblacional $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

$$\hat{\bar{X}} = \frac{1}{n}(X_1 + \hat{\bar{X}}_2)$$

$$V(\hat{\bar{X}}) = \frac{1-f}{n} S^2 + \frac{1-f_{21}}{nNf_{21}} N_2 S_2^2$$

Para determinar el tamaño de la muestra, consideramos la función de coste total:

$$C = C_0 n + C_1 n_1 + C_2 n_{21}$$

Como n_2 es una cantidad aleatoria, utilizamos el coste esperado:

$$C = n\left(C_0 + C_1 \frac{n_1}{n} + C_2 \frac{n_{21}}{n}\right) = n\left(C_0 + C_1 \frac{n_1/n}{1-W} + C_2 \cdot f_{21} \frac{n_2}{n}\right)$$

$$\Rightarrow E(C) = n\left(C_0 + C_1(1-W) + C_2 f_{21} W\right)$$

Utilizando multiplicadores de Lagrange para optimizar el tamaño muestral para obtener una varianza dada, V:

$$n = \underbrace{\frac{N^2 S^2}{V + NS^2}}_{\text{Tamaño muestral necesario para estimar el total con un error } E=\sqrt{V} \text{ si no hubiera falta de respuesta}} + \left.\frac{NN_2 \frac{1-f_{21}}{f_{21}} S_2^2}{V + NS^2}\right\} \rightarrow \text{Penalización debida a la falta de respuesta}$$

3_TRATAMIENTO de la FALTA de RESPUESTA

Una primera idea consiste en seguir seleccionando unidades aleatoriamente hasta conseguir el tamaño muestral requerido. Conseguiríamos disminuir la varianza del estimador, pero no el sesgo, que permanecería constante ya que todas las unidades de la muestra corresponderían al estrato de los que sí responden.

Si se dispone de información directa del estrato de no respondientes que justifique la hipótesis $\bar{X}_1 \simeq \bar{X}_2$, el estimador insesgado para el total sería:

$\hat{X}_3 = \frac{N}{n} \sum_{i=1}^{n_1} x_i$, sólo insesgado si $\bar{X}_1 \simeq \bar{X}_2$

Si esta hipótesis no es sostenible, estaríamos en la situación de $\hat{X}_1$, que es sesgado, aunque su varianza el menor.

Otras técnicas están basadas en modelos, aplicables a situaciones particulares en las que se produce la falta de respuesta.

Mt. de Hansen y Hurwitz

Método de aplicación general, que inicialmente fue diseñado para encuestas que se enviaban por correo.

Sea $n \equiv$ tamaño muestral, con (por m.a.s.) $\begin{cases} n_1 \text{ unid. que sí responden} \\ n_2 \text{ unid. que NO responden} \end{cases}$

De las n_2 unid. que no responden, seleccionamos una muestra aleatoria de tamaño $n_{21} \leq n_2$ y se envían entrevistadores para conseguir su respuesta en 2ª vuelta.

Para un n_i fijo, el nº de entrevistas obtenidas en cada una de las clases es una multinomial condicionada, $E(\bar{x}_i/n_i) = \bar{X}_i$, que no depende de $n_i \Rightarrow$ la media de la muestra después de i visitas, independientemente del nº de ~~visitas~~ respuestas, es también $\bar{x}_i$.

$B(\bar{x}_i) = \bar{X}_i - \bar{X}$

Variable

Modelo de respuesta aleatorizada. Modelo de Warner (1965)

Es difícil contar con la colaboración veraz de los encuestados frente a preguntas íntimas. Se puede dejar de contestar o bien dar deliberadamente respuestas falsas.

El individuo selecciona una de dos preguntas aleatoriamente, y contesta SÍ o NO. El entrevistador desconoce la pregunta seleccionada $\longrightarrow$ Queda garantizado el secreto.

$\pi_A \rightarrow$ proporción desconocida de personas que contestan SÍ a una pregunta íntima

$P \rightarrow$ probabilidad conocida de que sea elegida la pr. íntima

$\pi_y \rightarrow$ proporción conocida de respuestas SÍ a una pr. intrascendente.

$n \rightarrow$ tamaño de la muestra

$$P(\text{Sí}) = \lambda = P \cdot \pi_A + (1-P)\pi_y \Rightarrow \pi_A = \frac{\lambda - (1-P)\pi_y}{P}$$

estimando λ con $\hat{\lambda}$:

$$\hat{\pi}_A = \frac{\hat{\lambda} - (1-P)\pi_y}{P}$$

$$V(\hat{\pi}_A) = V\left(\frac{\hat{\lambda}}{P}\right) = \frac{\lambda(1-\lambda)}{P^2 \cdot n}$$

4_IMPUTACIÓN. TÉCNICAS de REPONDERACIÓN

Con el propósito de disminuir el posible sesgo introducido por la falta de respuesta se suelen utilizar varios tipos de ajuste.

- Ajuste sobre el terreno: el entrevistador recibe el listado de las unidades muestrales y un listado de suplentes.
  ↳ No reduce el sesgo de los que no responden.

- Ponderaciones
  Si se dispone información suplementaria sobre la proporción de unidades para ciertas clases de la población, se puede estratificar la muestra a posteriori y utilizar las mencionadas proporciones como ponderaciones.
  Si no existe información suplementaria, se pueden utilizar como pesos las inversas de las tasas de respuesta.

- Imputación
  Cuando un cuestionario no está completo, se puede realizar una imputación para los datos que faltan basada en la posible correlación entre el dato omitido y el resto de los datos disponibles.
  Fichero caliente:
  1_Se establecen una serie de caracteres que se suponen correlacionados con el que queremos imputar.
  2_Se introducen en el ordenador unos valores iniciales, fichero frío, obtenidos de encuestas anteriores.
  3_Si en la primera ficha de la encuesta falta el dato, se imputa al fichero caliente. Si está el dato, se imputa al fichero frío.

  y así sucesivamente

Se ha demostrado que el procedimiento "fichero caliente" produce un incremento de la varianza del estimador que no se reflejan en los métodos de estimación disponibles hasta ahora.