MUEST_T14. ALGUNAS TÉCNICAS ESPECIALES de MUESTREO.

MUESTREO DOBLE o BIFÁSICO.
MODELOS de CAPTURA_RECAPTURA.
MUESTREO por CUOTAS.
ESTIMACIONES en ÁREAS PEQUEÑAS.

1_ MUESTREO DOBLE o BIFÁSICO

En muchas ocasiones se consiguen mejores estimaciones de la variable en estudio X si disponemos de información anticipada de una variable auxiliar Y correlacionada con X, como es el caso del muestreo estratificado y de los estimadores de ratio y de regresión.

Cuando falta esta información sobre la variable auxiliar, puede ser relativamente económico tomar una muestra preliminar barata para medir las Y y hacer estimaciones sobre Y, para luego usarlas en las estimaciones de X, más precisas y más costosas.

El muestreo doble selecciona en una 1ª fase una muestra, relativamente grande, en la que a bajo coste pueden observarse una o varias características queridas de las unidades. En una segunda fase seleccionamos una submuestra de la primera en la que observamos la característica objeto de estimación → muestreo en dos fases, muestreo doble o muestreo bifásico.

Aspectos de notación:

1ª fase → muestra grande de tamaño n' para estudiar Y_i a bajo coste.

2ª fase → muestra de tamaño n ($n < n'$, generalmente submuestra) para observar X_i con coste mayor.

Es evidente que la conveniencia de esta técnica depende de los costes.
Si la observación de la característica en estudio X_i no tiene coste, tomaríamos una muestra del tamaño n_0 necesario para la precisión deseada de las estimaciones.

Sp. que disponemos de un presupuesto total C:
~~coste unit. 1ª muestra =~~

muestra	Coste unit.	tamaño muestral
1	c'	n'
2	c	n

Nos planteamos dos opciones:
- 1 sola muestra $\rightarrow$ $C = c_0 \cdot n_0$
- Dos fases $\rightarrow$ $C = c'n' + cn$

Igualando los costes totales, $c_0 n_0 = c'n' + cn$

$$n_0 = \frac{c'n' + cn}{c_0}$$

por lo que con la técnica en dos fases la observación de X se hace en una muestra de tamaño n, menor que n_0, con el mismo coste total $\Rightarrow$ al introducir las dos fases el tamaño de la muestra necesario es más pequeño que si hubiera una fase (muestreo aleatorio normal) y hay una pérdida en la precisión de los estimadores.

Para ver si compensa la disminución del tamaño efectivo de la muestra con el incremento de información adquirido en la 1ª fase ($\Rightarrow$ pérdida de precisión en las estim. de X) calculamos las varianzas:
- 1 fase: σ^2/n_0 (para la media)
- 2 fases: $n_0 - n = \left(\frac{c'}{c}\right)n' \Rightarrow$ cuanto menor sea $\frac{c'}{c}$ más favorable es el muestreo doble (más cerca estará n^{o} de n_0 y menor será la pérdida de precisión).

El muestreo bifásico sólo es bueno si lo que se gana en precisión con la variable Y compensa con la pérdida en precisión debido a la reducción del tamaño de la muestra.

Muestreo doble para estratificación

Supongamos que la población se estratifica en L estratos.

1ª fase: muestra aleatoria de tamaño n'

$$W_h = \frac{\text{nº elem. poblac. estrato } h}{\text{Nº total elem. poblac}} = \frac{N_h}{N}$$

$$\hat{W}_h = \frac{\text{nº elem. 1ª muestra estrato } h}{\text{nº total elem. 1ª muestra}} = \frac{n'_h}{n'} \quad \text{insesgado de } W_h$$

$\sum_{h=1}^{L} n'_h = n'$ y $\sum_{h=1}^{L} n_h = n$

2ª fase: muestra aleatoria estratificada de tamaño n

De cada estrato se elige una submuestra de tamaño $n_h \leq n'_h$ indep.

El estimador usual de la media es $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$, en muestreo doble es $\hat{\bar{X}} = \sum_{h=1}^{L} \hat{W}_h \bar{x}_h$ con $\hat{W}_h = \frac{n'_h}{n'}$ y $\bar{x}_h = \frac{x_h}{n_h}$ que es insesgado para la media.

Para calcular su varianza, se acude al teor de Nadow:

$$V(\hat{\bar{X}}) = V(E_{W'}(\hat{\bar{X}})) + E(V_{W'}(\hat{\bar{X}}))$$

la expresión de la varianza del estimador, distinta para el caso SR y CR es tal que n' (tamaño muestra 1ª fase) aparece en el denominador, de modo que cuanto mayor sea n', la pérdida de precisión por el uso de muestreo doble disminuye. (Como el coste aumenta, valorar los tamaños y la afijación óptimos en función del coste).

Muestreo doble para estimadores de ratón

El estimador usual de ratón para la media $\bar{X}$ utiliza como información conocida $\bar{Y}$ o Y.

El muestreo doble utiliza

1ª fase: muestra de tamaño n' para obtener una buena estimación de $\bar{Y}$ o de Y.

2ª fase: muestra de tamaño n para obtener $\bar{x}$ e $\bar{y}$.

El estimador de ratón para el muestreo doble es:

$$\hat{\bar{X}}_R = \frac{\bar{x}}{\bar{y}} \cdot \bar{y}' \quad , \text{ donde } \bar{y}' \equiv \text{media de la 1ª muestra}$$

El estimador será insesgado si lo es $\hat{R}$.

Se puede dar expresiones aproximadas de la varianza del estimador (varianza y covarianza para CR y cuasivarianzas y cuasicovarianzas para SR).

La varianza es igual a la varianza del estim. de ratón más una penalización por utilizar muestreo doble (cuanto mayor sea n', la ~~precisión~~ penalización es más pequeña).

* En estimadores de regresión ocurre + o − lo mismo.

2_ MODELOS de CAPTURA _ RECAPTURA

El mt. de captura y recaptura o de captura-marcado-suelta y recaptura, consiste en tomar una primera muestra, marcar las unidades capturadas (procurando que no modifique su características ni condiciones su comportamiento), y ponerlos en libertad en la zona que fueron aprehendidos. Tras un período establecido no muy largo, se toma una segunda muestra y se observan cuántos elementos están marcados.

El tamaño de la poblac. se estima por:

$$\hat{N} = \frac{n_1 \cdot n_2}{m_2}$$ (basado en la distrib. hipergeométrica)

donde $n_i \equiv$ tamaño muestra i $(i=1,2)$.

$m_2 \equiv$ número de recapturas

La probab. de obtener en una muestra de tamaño n_2, m elementos marcados, con n_1 elementos marcados en una poblac. de tamaño N (desconoc) es:

$$P(\xi = m) = \frac{\binom{n_1}{m} \cdot \binom{N-n_1}{n-m}}{\binom{N}{n_2}} = \frac{n_1!\, n_2!}{m!\,(n_1+n_2-m)!}$$

$N = n_1 + n_2 - m$

Para estimar N puede utilizarse el mt. de máxima verosimilitud, $\hat{N} = \frac{n_1 n_2}{m}$ es un estimador máx. verosímil de N que se apoya en el supuesto de que la proporción de recapturas en la segunda muestra es igual a la prop. de individuos marcados en la población: $\frac{n_1}{N} = \frac{m_2}{n_2}$

Este supuesto se basa en:

- La poblac. es cerrada en el período de estudio, N cte.
- Los individuos marcados no quedan afectados al marcarlos, y se distribuyen de manera aleatoria, con lo que la probab. de ser capturados en la 2ª muestra es constante.

Estas supuestas no son siempre plausibles, por lo que se han propuesto otros tipos de estimación para evitar sesgos basándose en supuestos sobre la población de origen. Deben mencionarse los mt. de eliminación (removal) y de capturas múltiples.

En ocasiones han de estimarse algunas totales, peso, volumen, alimento consumido, etc. lo que requiere estimadores del tipo: $\hat{N} \cdot \bar{x}$, donde ambos factores son variables aleatorias (estimadores del producto).

El mt. de captura y recaptura se aplica a las estimaciones de poblaciones móviles humanas, en cuyo caso se recomienda ponerse en contacto con los jefes tribales o superiores jerárquicos.

3_ MUESTREO por CUOTAS

El muestreo por cuotas, desarrollado en los años 30 por Cheniegton, Roper, Gallup y Gressley, ha sido adoptado por muchas organismos dedicados a realizar encuestas sobre opinión pública, estudios de mercado, etc.

No existe una definición precisa de muestreo por cuotas, es su punto débil: existen tantas variantes como aplicaciones.

Sp. que el diseño de la encuesta ha seguido los principios del muestreo probabilístico hasta llegar el momento de seleccionar las personas a entrevistar. En esta etapa se le impone al entrevistador que realice un determinado nº de entrevistas por edad, sexo, nivel económico ... y cg. otr característica sociológica o socioeconómica de interés, y se le deja libertad plena para elegir las personas que cumplan estos requisitos ⇒ sesgos que no pueden ser detectados.

Por otro lado, el desconocimiento de las probab. de selección no permite evitar los errores por ponderaciones incorrectas en el proceso de estimación y no se pueden estimar los errores debidos al muestreo.

⇒ No se puede comparar el coste del muestreo por cuotas con el coste por muestreo probabilístico.

Especialmente en encuestas de opinión, la muestra puede representar muy bien a la población para una característica y muy mal para opiniones.

Una muestra por cuotas puede proporcionar estim. muy útiles, pero la dificultad está en conocer a priori hasta qué punto lo serán.

En encuestas importantes nunca existirá muestreo probabilístico.

4_ ESTIMACIONES en ÁREAS PEQUEÑAS.

En muchas encuestas, la población se subdivide en clases para las que se requiere estimaciones separadas.

Dominio → cualquier subpoblación para la que se necesita una estim. separada, antes o después de diseñar el muestreo.

Si la necesidad de estimaciones separadas se conoce antes de hacer el diseño, se puede plantear los dominios como estratos → coste adicional, depende del nº de dominios.

Si las estimaciones a posteriori se necesitan a posteriori (por desconocimiento, economía..) las unid. muestrales pueden pertenecer al dominio o no, por lo que el nº de observaciones muestrales pertenecientes al dominio
- es una v.a.
- puede ser muy pequeño, o nulo.

Surge el problema de estimación en dominios pequeños, ya que los estimadores usuales serán menos precisos al estar basados en menos observaciones, y si no existen observaciones del dominio, no se podrán utilizar.

El término estimación en pequeñas áreas se utiliza cuando los dominios pequeños corresponden a zonas geográficas. Hay varias alternativas:

→ Estimadores directos:
- Utilizan datos únicamente de las unidades muestrales pertenecientes al dominio de estudio
- El tamaño de la muestra es una v.a.
- El estimador tiene mayor varianza.

Ejemplo: En m.a.s.s.r. sea $n_j \equiv$ nº unid. muestrales dominio (dentro).

$\bar{x}_j = \sum_{k=1}^{n_j} x_{jk} \rightarrow \hat{\bar{x}}_j = \frac{N}{n}\sum_{k=1}^{n_j} x_{jk}$. Consideramos $x'_j = \begin{cases} x_j & \text{si } i \in \text{dom } j \\ 0 & \text{si no} \end{cases}$

$X' = \sum_{i=1}^{N} x'_i = Y_j \Rightarrow N\bar{x}' = N \cdot \frac{1}{n}\sum^{N} x'_i = \bar{x}_j$

$V(\bar{x}_j) = (1-f)\cdot\frac{s'^2}{n}$, donde $s'^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x'_i - \bar{x}')^2$

→ Estimadores indirectos (sintéticos)
- Utilizan datos de unid. de otro dominio
- Gran variedad en función de los modelos que engende la pobl. finita de estudio
- Estimadores sesgados → valorar no el bueno medido de la bondad de los estimadores

Ejemplo: En m.a.s.s.r. → $X \equiv$ var. estudio, $Y \equiv$ var. correlacionada con X. $\hat{R} = \frac{\bar{y}}{\bar{y}_d}$

$X = RY$

$\hat{X}_j = \hat{R}\hat{Y}_j$

Consideramos que lo que se verifica en el modelo poblacional tb. se verifica en el dominio. El sesgo será mayor cuanto más falsa sea esta suposición.

→ Estimadores combinados:
- Media ponderada de un estim. directo (basado en el diseño) y un estim. indirecto (basado en el modelo)
- Intentan equilibrar el sesgo potencial del estimador sintético y la varianza del estimador directo.

$$\hat{Y}_d = \alpha \hat{Y}_d^{\text{directo}} + (1-\alpha)\hat{Y}_d^{\text{sintético}} \quad / \alpha \in (0,1)$$