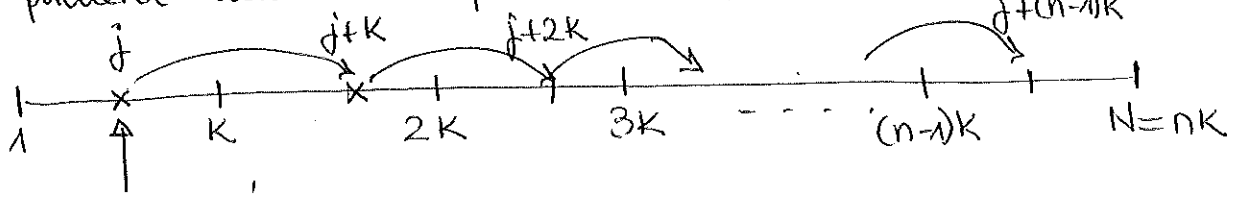
MUEST_T10. MUESTREO SISTEMÁTICO.
ESTIMADORES Y VARIANZAS.
RELACIÓN CON EL MUESTREO ESTRATIFICADO.
RELACIÓN CON EL MUESTREO de CONGLOMERADOS.
RELACIÓN CON EL MAS.
PROBLEMÁTICA de la ESTIM. de VARIANZAS.

1_MUESTREO SISTEMÁTICO.

A partir de una población de tamaño N de la que se quiere extraer una muestra de tamaño n, el muestreo sistemático divide la población en n zonas (tantas como el tamaño muestral) de igual tamaño K, $N = nK$ ¿qué pasa si $N \neq n$? pág. sgte

Se elige al azar una unidad de la primera zona, y las n-1 unidades restantes se eligen tomando de cada zona la unidad que ocupa el mismo lugar en su zona que la primera unidad elegida en la primera zona.



arranque ó semilla aleatoria ← normalmente se elige por mas, con prob $= \frac{1}{K}$.

Tal y como hemos definido el método de selección sistemática, al hacer variar j desde 1 hasta K se obtienen las posibles muestras sistemáticas, por lo que el espacio muestral estará formado por las K muestras posibles:

muestra 1 $\rightarrow \{u_1, u_{1+K}, u_{1+2K} \dots u_{1+(n-1)K}\}$

– – – – – –

muestra j $\rightarrow \{u_j, u_{j+K}, u_{j+2K} \dots u_{j+(n-1)K}\}$

– – – –

muestra K $\rightarrow \{u_K, u_{K+K}, u_{K+2K} \dots u_{K+(n-1)K}\}$

todas ellas con la misma probab. de ser seleccionadas

$$P(M) = \frac{1}{K} = \frac{n}{N}$$

zonas	elementos				
1	u_{11}	. .	u_{1j}	. .	u_{1K}
⋮					
n	u_{n1}	. .	u_{nj}	. .	u_{nK}

$N = nK$

Muestra = todos los elem. que ocupan la posic. j.

Además, como es imposible que una unidad de la población aparezca más de una vez en una muestra, es un muestreo SIN reposición, donde la probab. de una unidad de pertenecer a la muestra es la misma que la probab. de la muestra:

$$\pi_i = P(u_i \in M) = \frac{1}{K} = \frac{n}{N}$$

Si el tamaño poblacional no es múltiplo del tamaño muestral, $N = Kn + r$, entonces se obtienen r muestras sistemáticas de tamaño $(n+1)$ y $K-r$ muestras sistemáticas de tamaño n, aunque si el tamaño muestral n es sufic. grande, un aumento en una unidad no suele producir perturbaciones en la distrib. de los estimadores.

Existen algunas variedades en la forma de selección de una muestra sistemática.

- Muestreo estrictamente sistemático (ó muestreo sistemático centrado) → seleccionar las unidades que ocupan el punto medio de cada zona.
  Ya no es probabilístico, sino intencional, aunque sus resultados no difieren mucho.

En cuanto a la disposición de las unidades poblacionales en una lista y su numeración, puede ser:

- aleatoria (equivale a mas)
- por orden alfabético
- "serpentina", conveniente en la selección de bloques, manzanas o edificios.

El desarrollo teórico del muestreo sistemático aleatorio se debe a Madow y Madow (1944) y a Cochran (1946)

VENTAJAS:

- Extiende la muestra a toda la población (ninguna sucesión grande de la muestra queda sin representar)
- Recoge el posible efecto de estratificación (si las unidades están ordenadas en la lista de forma parecida)
- Permite la consideración de conglomerados en la población.
- Fácil de aplicar y comprobar
- No presenta problemas de cálculo
- No precisa distinción entre muestreo SIN y CON reposición
- El error de muestreo suele ser menor que el de mas o incluso que muestreo estratificado.

INCONVENIENTES:

- Existe la posibilidad de aumento de la varianza si existe periodicidad en la población
- Tiene un problema teórico en la estimación de varianzas
- No hay independencia entre las unidades seleccionadas en las $\neq$s tomas con la primera toma
- En general sólo hay selección aleatoria en la primera unidad de la muestra

2 - ESTIMADORES y VARIANZAS.

Dado que el muestreo sistemático es sin reposición, utilizaremos el estimador lineal insesgado de Horvitz y Thompson:

$$\hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i} \quad \text{estim. insesgado de } \theta = \sum_{i=1}^{N} Y_i$$

En muestreo sistemático, $\pi_i = \frac{1}{K} = \frac{n}{N}$

llamamos u_{ij} $\begin{cases} \text{unidad } j\text{-ésima de la} & (j=1 \ldots K) \\ \text{zona } i\text{-ésima} & (i=1 \ldots n) \end{cases}$

por lo que $$\hat{\theta} = \sum_{i=1}^{n} \sum_{j=1}^{1} \frac{Y_{ij}}{\frac{1}{K}} \quad \text{estim. } \underline{\text{insesgado}} \text{ de } \theta = \sum_{i=1}^{n} \sum_{j=1}^{K} Y_{ij}$$

(*) mirar atrás

que se puede particularizar para los parámetros poblacionales más conocidos:

Total poblacional: $\theta = X = \sum_{i=1}^{n} \sum_{j=1}^{K} X_{ij} \longrightarrow Y_{ij} = X_{ij}$

$$\hookrightarrow \hat{X} = \sum_{i=1}^{n} \sum_{j=1}^{1} \frac{X_{ij}}{\frac{1}{K}} = \sum_{i=1}^{n} \underset{N/n}{\underbrace{K}} X_{ij} = N \cdot \frac{1}{n} \sum_{i=1}^{n} X_{ij} = \boxed{N \overline{x}_j}$$

Media poblacional: $\theta = \overline{X} = \sum_{i=1}^{n} \sum_{j=1}^{K} \frac{X_{ij}}{\underset{Kn}{\underbrace{N}}} \longrightarrow Y_{ij} = \frac{X_{ij}}{N}$

$$\hookrightarrow \hat{\overline{X}} = \sum_{i=1}^{n} \sum_{j=1}^{1} \frac{X_{ij}/N}{\frac{1}{K}} = \sum_{i=1}^{n} \sum_{j=1}^{1} \frac{X_{ij}}{n} = \boxed{\overline{x}_j}$$

Proporción $\longrightarrow$ $\hat{P} = \sum_{i=1}^{n} \sum_{j=1}^{1} \frac{A_{ij}/nK}{1/K} = \frac{1}{n} \sum_{i=1}^{n} A_{ij} = \hat{P}_j$

Total de clase $\longrightarrow$ $\hat{A} = \ldots\ldots = N\hat{P}_j$

Por lo que vemos que un estimador lineal insesgado de la media poblacional es la media sistemática, etc...

$\hat{\Theta}$ es insesgado de Θ porque:

$$E[\hat{\Theta}] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{1}\frac{y_{ij}}{1/K}\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{y_{ij}}{1/K}\, e_{ij}\right] = \sum_{i=1}^{n}\sum_{j=1}^{K}\frac{y_{ij}}{1/K}\, E[e_{ij}] =$$

donde $e_{ij} = \begin{cases} 1 & \text{si } u_{ij} \in \text{muestra con probab. } 1/K \\ 0 & \text{en otro caso} \end{cases}$ $\quad e_{ij} \to \text{Bern}\left(\frac{1}{K}\right)$

$$= \sum_{i=1}^{n}\sum_{j=1}^{K}\frac{y_{ij}}{1/K}\cdot 1/K = \Theta.$$

VARIANZAS:

Antes de hallar las expresiones de las varianzas para los estim. de los parámetros poblacionales, descomponemos la suma de cuadrados:

$$SCT = SCD + SCE$$

$$\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{x}_j+\bar{x}_j-\bar{X})^2 = \sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{x}_j)^2 +$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{K}(\bar{x}_j-\bar{X})^2 + 2\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{x}_j)(\bar{x}_j-\bar{X})}_{=0} \Rightarrow$$

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X})^2}_{\text{total}} = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{x}_j)^2}_{\text{dentro de las muestras}} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{K}(\bar{x}_j-\bar{X})^2}_{\text{entre las muestras}}$$

Si definimos:

$$S_{bs}^2 = \frac{1}{K-1}\sum_{i=1}^{n}\sum_{j=1}^{K}(\bar{x}_j-\bar{X})^2 \leftarrow \text{cuasivarianza intermuestral}$$

$$S_{WS}^2 = \frac{1}{N-K}\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{x}_j)^2 \leftarrow \text{cuasivarianza intramuestral}$$

podemos escribir la descomp. de sumas de cuadrados:

$$(N-1)S^2 = (\overset{K(n-1)}{N-K})S_{WS}^2 + (K-1)S_{bs}^2$$

Por lo que podemos expresar la varianza de los estimadores:

$$V(\hat{X}) = V(\bar{x}_j) \overset{\text{Por def. de varianza}}{=} \frac{(1-f)}{n}\cdot S_{bs}^2 \overset{f=\frac{n}{N}=\frac{1}{K}}{=} \frac{K-1}{Kn}\cdot S_{bs}^2 = \frac{K-1}{N}S_{bs}^2 =$$

$$= \frac{(N-1)S^2-(N-K)S_{WS}^2}{N} = \frac{N-1}{N}S^2 - \underset{\left(1-\frac{1}{n}\right)}{\frac{N-K}{N}}S_{WS}^2 =$$

$$= \sigma^2 - \frac{n-1}{n}S_{WS}^2$$

MIRAR ATRÁS

n var. aleat. discretas
i' equiprobables

$\ast$ y / n

$$(\ast)\ V(\bar{x}_j) = E\Big[\bar{x}_j - \underbrace{E[\bar{x}_j]}_{=\bar{X}}\Big]^2 = \frac{1}{K}\sum_{j=1}^{K}(\bar{x}_j-\bar{X})^2 = \frac{1}{nK}\sum_{i=1}^{n}\sum_{j=1}^{K}(\bar{x}_j-\bar{X})^2 =$$

$$= \frac{1}{nK}(K-1)\,S^2_{bs}$$

$$= \frac{1}{n}\left(\frac{K-1}{K}\right)S^2_{bs} = \frac{1}{n}(1-f)\cdot S^2_{bs} = \left(1-\frac{1}{K}\right)\frac{S^2_{bs}}{n}$$

Relacionándolo con las otras cuasiV/var.

$$\frac{(K-1)}{N}S^2_{bs} = \frac{1}{N}\left[(N-1)S^2 - (N-K)S^2_{WS}\right] =$$

$$= \sigma^2 - \frac{N-K}{N}S^2_{WS}$$

$$\frac{n\cancel{K}-\cancel{K}}{n\cancel{K}}$$

$$= \sigma^2 - \frac{n-1}{n}S^2_{WS} = \frac{N-1}{N}S^2 - \frac{n-1}{n}S^2_{WS}$$

A partir de esta expresión se pueden obtener las varianzas de los demás estimadores:

$$V(\hat{X}) = V(N\bar{x}_j) = N^2 V(\bar{x}_j) = N^2 \cdot \frac{1-f}{n} \cdot S_b^2 = N\left[(N-1)S^2 - (N-K)S_{WS}^2\right]$$

En el caso de características dicotómicas:

$$S_{bs}^2 = \frac{1}{K-1}\sum_{i=1}^{n}\sum_{j=1}^{K}(\bar{x}_j - \bar{X})^2 = \frac{n}{K-1}\sum_{j=1}^{K}(\bar{x}_j - \bar{X})^2 =$$
$$= \frac{n}{K-1}\sum_{j=1}^{K}(P_j - P)^2$$

por lo que:

$$V(\hat{P}_j) = \frac{1-f}{n} \cdot S_{bs}^2 = \frac{K-1}{Kn} \cdot \frac{n}{K-1}\sum_{j=1}^{K}(P_j - P)^2 = \frac{1}{K}\sum^{K}(\hat{P}_j - P)^2$$

$$V(\hat{A}) = V(N\hat{P}_j) = N^2 V(\hat{P}_j) = Nn\sum_{j=1}^{K}(\hat{P}_j - P)^2$$

sí

~~Aplicando la igualdad~~

Se observa que la varianza de los estimadores es mayor cuanto mayor sea la cuasivarianza intermuestral $S_{bs}^2 \Rightarrow$ conviene que haya homogeneidad entre las muestras y que todas las muestras posibles sean lo + parecidas entre sí.

Utilizando la expresión $V(\bar{x}_j) = \sigma^2 - \frac{n-1}{n} S_{WS}^2$, se observa que las varianzas de los estimadores son menores cuanto mayor sea la varianza intramuestral $\Rightarrow$ conviene que haya heterogeneidad $\boxed{\text{dentro de las muestras}}$ $\Rightarrow$ no dentro de las zonas

Si
$$V[\hat{P}] = \sigma^2 - \frac{n-1}{n} S_{WS}^2 = PQ - \frac{n-1}{n} \cdot \frac{n}{N-K}\sum \hat{P}_j(1-\hat{P}_j) = PQ - \frac{1}{K}\sum_j(\hat{P}_j\hat{Q}_j)$$

$$V(\hat{A}) = N^2 \cdot V(\hat{P})$$

3 _ RELACIÓN CON EL MUESTREO ESTRATIFICADO

En el muestreo sistemático puede considerarse cada zona de k elementos consecutivos a partir del primero como un estrato → se divide la población en n estratos.
La muestra ~~estratif~~ sistemática equivale entonces a una muestra estratificada con una unidad por estrato.

Hay que tener en cuenta que en el m.e. la selección se hace de manera independiente en cada estrato, y en el m.s. no hay aleatoriedad de selección.

En el m.e. los estratos han de ser heterogéneos entre sí y homogéneos dentro, mientras que en m.s. ~~las zonas debe ser~~ conviene que haya homogeneidad entre las muestras y heterogeneidad dentro de las muestras, lo que se traduce a heterogeneidad entre las zonas y homogeneidad dentro de cada zona.

Considerando cada zona como un estrato de k unidades, en el muestreo estratificado se cumple:

$$SCT = SCD + SCE$$

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X})^2}_{\text{variabilidad total}} = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2}_{\text{var. dentro de cada estrato}} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{X}_i-\bar{X})^2}_{\text{var. entre los estratos}}$$

que puede escribirse en términos de cuasivarianzas como:

$$(N-1)\,S^2 = (N-n)\,S^2_{W\,st} + (n-1)\,S^2_{b\,st}$$

donde:

$$S^2_{W\,st} = \sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_i)^2/(N-n) \rightarrow \text{cuasiv. intraestratal}$$

$$S^2_{b\,st} = \sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{X}_i-\bar{X})^2/(n-1) \rightarrow \text{cuasiv. interestratal}$$

Como una muestra sistemática (columna) equivale a la estratif. con una unidad por estrato podemos utilizar las fórmulas de estimaciones del muestreo estratificado: (*)

$$V(\hat{X}) = \underline{V(\bar{x}_{st})} = \sum_{h=1}^{L} W_h^2 V(\bar{x}_h) = \sum_{i=1}^{n} W_i^2 V(\bar{x}_i) = \sum_{i=1}^{n} \frac{1}{n^2}(1-f_i)\frac{S_i^2}{n_i} =$$

$$= \sum_{i=1}^{n} \frac{1}{n^2}\left(1-\frac{1}{K}\right)\cdot S_i^2 = \frac{1}{n^2}\left(1-\frac{1}{K}\right)\sum_{i=1}^{n} S_i^2 =$$

$$= \frac{1}{n^2}\left(1-\frac{1}{K}\right)\underbrace{\sum_{i=1}^{n}\frac{1}{K-1}\sum_{j=1}^{K}(X_{ij}-\bar{X}_i)^2}_{(N-n)S^2_{W\,st}} =$$

$$= \frac{1}{\underset{\underset{\frac{N}{n}}{\shortparallel}}{n^2 K}}(N-n)S^2_{W\,st} = \frac{N-n}{Nn}S^2_{W\,st} = \boxed{(1-f)\cdot\frac{S^2_{W\,st}}{n}}$$

coincide con m.s.

Falla la premisa (*)

Pero a esta expresión hemos llegado considerando que las unidades muestrales se han seleccionado de manera indep.

Como en muestreo sistemático no se da la independencia, intentamos dar una expresión de la varianza sin suposición de independencia, a partir del coef. de correlación $\rho_{\omega st}$.

$\rho_{\omega st}$ ≡ coef. de correlación lineal entre las desviaciones respecto de las medias de cada estrato de todos los pares de valores muestrales → COEF. CORRELACIÓN ENTRE ESTRATOS

omega

$$\rho_{\omega st} = \frac{cov(X_{ij}, X_{zj})}{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X}_i)^2}_{\frac{(N-n)S^2_{W\,st}}{\underset{\underset{K-1}{\shortparallel}}{n}}}} = \frac{\sum_{i\neq z}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X}_i)(X_{zj}-\bar{X}_z)}{\underbrace{n(n-1)}_{\text{combinac. posibles}}\cdot(K-1)S^2_{W\,st}}$$

σ_i^2

⟹ MIRAR ATRÁS

(*) $W_h = \frac{N_h}{N} = \frac{K}{N} = \frac{K}{nK} = \frac{1}{n}$

$f_h = f_i = \frac{n}{N} = \frac{1}{K}$

$n_i = 1$ (1 unidad por estrato)

$S_i^2 = \frac{1}{K-1}\sum_{j=1}^{K}(X_{ij}-\bar{X}_i)^2$

Para cada j, $i \neq z$.

$$\rho_{iz} = \frac{cov(X_{ij}, X_{zj})}{\sigma(X_{ij})\sigma(X_{zj})} = \frac{E\left[(X_{ij}-\overline{X}_i)\cdot(X_{zj}-\overline{X}_z)\right]}{\frac{1}{nK}\sum_{j=1}^{K}\sum_{i=1}^{n}(X_{ij}-\overline{X}_i)^2} =$$ para todo $\downarrow$

$$= \frac{\frac{1}{Kn(n-1)}\sum_{j=1}^{K}\sum_{i\neq z}^{n}(X_{ij}-\overline{X}_i)(X_{zj}-\overline{X}_z)}{\frac{1}{N}\underbrace{\sum_{j=1}^{K}\sum_{i=1}^{n}(X_{ij}-\overline{X}_i)^2}_{(N-n)S^2_{W\,st}}} = \text{\_\_\_\_\_\_\_\_\_\_}$$

$Kn-n$

$n(K-1)$

(no hace falta).

$$\rho_{w\,st} = \frac{cov(X_{ij}, X_{zj})}{\sigma(X_{ij})\sigma(X_{zj})} = \frac{cov(X_{ij}, X_{zj})}{\frac{1}{N}\underbrace{\sum_{j=1}^{K}\sum_{i=1}^{n}(X_{ij}-\overline{X}_i)^2}_{(N-n)S^2_{W\,st}}} = \frac{cov(X_{ij}, X_{zj})}{\underbrace{\frac{N-n}{N}}_{1-f}\cdot S^2_{W\,st}}$$

de donde se deduce:

$$cov(X_{ij}, X_{zj}) = (1-f)\cdot\rho_{w\,st}\cdot S^2_{W\,st}$$

por lo que:

$$V(\overline{x}_j) = V\left(\frac{1}{n}\sum_{i=1}^{n}X_{ij}\right) = \frac{1}{n^2}\sum_{i=1}^{n}V(X_{ij}) + \frac{1}{n^2}\cdot\sum_{i\neq z}cov(X_{ij}, X_{zj})$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}V(\overline{x}_i) + \frac{1}{n^{\cancel{2}}}\cancel{n}(n-1)\cdot(1-f)\cdot\rho_{w\,st}\cdot S^2_{W\,st} =$$

$$= \sum W_i^2\cdot V(\overline{x}_i) +$$

$$= V(\overline{X}_{st}) +$$

$$= \frac{1-f}{n}\cdot S^2_{W\,st} + \frac{1-f}{n}\cdot S^2_{W\,st}\,(n-1)\rho_{w\,st} =$$

$$= \frac{1-f}{n}S^2_{W\,st}\left[1+(n-1)\rho_{w\,st}\right]$$

$$W_i = \frac{N_i}{N} = \frac{K}{N} = \frac{1}{n}$$

de donde se deduce: $\operatorname{cov}(X_{ij}, X_{zj}) = \rho_{w\,st}\,(1-f)\cdot S^2_{w\,st}$

y $V(\hat{\bar{X}}) = V(\bar{x}_j) = \dots = (1-f)\cdot \dfrac{S^2_{w\,st}}{n}\left(1+(n-1)\rho_{w\,st}\right)$

Teniendo en cuenta que $V(\bar{x}_{st}) = (1-f)\cdot \dfrac{S^2_{w\,st}}{n}$, podemos escribir:

$$V(\hat{\bar{X}}) = V(\bar{x}_{st})\cdot\left[1+(n-1)\rho_{w\,st}\right]$$

Interpretación:

1) Precisión máxima $(V(\bar{x}_j)=0) \Leftrightarrow (n-1)\rho_{w\,st} = -1$
$\Leftrightarrow \rho_{w\,st} = -\dfrac{1}{n-1}$

2) Precisión mínima $(V(\bar{x}_j)\ MAX) \Leftrightarrow \rho_{w\,st} = 1$

3) $\rho_{w\,st} = 0 \Rightarrow V(\hat{\bar{X}}) = V(\bar{x}_{st}) \Rightarrow$ el muestreo sistemático coincide en precisión con el muestreo estratificado con selección indep. en cada estrato.

4) $-\dfrac{1}{n-1} < \rho_{w\,st} < 0 \Rightarrow$ Sistemático mejor que estratificado

5) $0 < \rho_{w\,st} < 1 \Rightarrow$ Estratificado mejor que sistemático

Así, se puede interpretar $\rho_{w\,st}$ como una medida de la falta de aleatoriedad en la selección de unidades para la muestra en las distintas zonas sistemáticas (zonas o estratos)

4_ RELACIÓN con el MUESTREO de CONGLOMERADOS

La muestra sistemática puede considerarse como una muestra de conglomerados monoetápica de tamaño 1 ⟶ la población está formada por K conglomerados de tamaño n (columnas), se elige aleatoriamente 1 de ellos, con probab. $\frac{1}{K}$, y se observan todas sus unidades elementales.

Al seleccionar el conglomerado j-ésimo, se calcula su media poblacional $\overline{X}_j$, que será la estimación de la media global.

$$\overline{X} = \frac{1}{K} \cdot \sum_{j=1}^{K} \overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{K} \sum_{j=1}^{K} X_{ij}$$

$$\hat{\overline{X}} = \overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

$$V(\overline{\overline{x}}) = V(\overline{X}_j) = (1-f) \cdot \frac{S^2_{\overline{x}}}{1} = \frac{K-1}{K} \cdot \frac{1}{K-1} \sum_{j=1}^{K} (\overline{X}_j - \overline{X})^2 =$$

$$= (1-f) \cdot \frac{S^2_{bs}}{n}$$

donde $S^2_{bs} = \frac{1}{K-1} \sum_{j=1}^{K} \underbrace{\sum_{i=1}^{n}}_{= n} (\overline{X}_j - \overline{X})^2$ ⟶ cuasiV. entre conglom.

Por tanto, las expresiones de los estimadores y de sus varianzas son idénticas, debido a que los dos tipos de muestreo son equivalentes.

5_RELACIÓN con el MUESTREO ALEATORIO SIMPLE

Para una m.a.s.(n)

$$V(\bar{x}) = (1-f)\cdot\frac{S^2}{n} \underset{f=1/K}{=} \frac{K-1}{K}\cdot\frac{S^2}{n}$$

Para el muestreo sistemático

$$V(\bar{x}_j) = (1-f)\cdot\frac{S^2_{bs}}{n} = \frac{K-1}{K}\cdot\frac{S^2_{bs}}{n}$$

Por lo que comparar el muestreo sistemático con el m.a.s. se reduce a comparar la cuasivarianza poblacional S^2 con la cuasivarianza ~~intra~~ inter muestral S^2_b.

- $S^2_{bs} = 0 < S^2 \Rightarrow V(\bar{x}_j) = 0 \rightarrow$ caso ideal para utilizar el muestreo sistemático
- $S^2_{bs} < S^2 \Rightarrow$ muestreo sistemático más preciso
- $S^2_{bs} = S^2 \Rightarrow$ muestreo sistemático igual que mas
- $S^2_{bs} > S^2 \Rightarrow$ muestreo sistemático menos preciso que S^2.

El muestreo sistemático será más preciso cuanto más homogéneas sean las ~~zonas~~ muestras en que se divide la población, es decir cuando la heterogeneidad esté dentro de la muestra. La igualdad se da cuando la disposición de los elementos sea aleatoria, que se consigue simplemente numerándolos aleatoriamente antes de realizar las zonas sistemáticas para la selección de la muestra.

Sin embargo, si en la ordenación de los elementos poblac. existe cierta periodicidad, la representatividad de la muestra sistemática disminuye y es preferible mas.

Lo que hemos comentado anteriormente utilizando S^2_{bs}, lo podemos hacer también con la cuasivarianza intramuestral S^2_{ws}, recordando la relación:

$$(N-1)S^2 = K(n-1)S^2_{ws} + (K-1)S^2_{bs}$$

$$V(\bar{x}_j) = \frac{K-1}{K}\cdot\frac{S^2_b}{n} = \frac{1}{Kn}\left[(N-1)S^2 - K(n-1)S^2_{ws}\right] =$$

$$\cancel{= \frac{N-1}{KnN}S^2 - \frac{K(n-1)}{Kn}S^2_{ws}} = \frac{N-1}{N}\cdot\frac{n}{1-f}V_{MAS}(\bar{x}) - \frac{n-1}{n}S^2_{ws}$$

$$= \cancel{\tfrac{1}{K}}\cdot\frac{(Kn-1)}{K-1}\left(\frac{(K+1)}{K}\cdot\frac{S^2}{n}\right) - \frac{K(n-1)}{Kn}S^2_{ws}$$

$$= \frac{Kn-1}{K-1}\cdot V_{MAS}(\bar{x}) - \frac{n-1}{n}S^2_{ws}$$

También se puede hacer la comparación utilizando el coeficiente de correlación intramuestral, ρ_w, que mide la interrelación entre las unidades dentro de las muestras.

Interesa que ρ_w sea muy pequeño, porque en el muestreo sistemático interesa la heterogeneidad intramuestral con la finalidad de que una única muestra sistemática represente lo mejor posible a toda la población.

$$\rho_w = \frac{cov(X_{ij}, X_{zj})}{\sigma(X_{ij})\sigma(X_{zj})} = \ldots = \frac{\sum_{i\neq z}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X})(X_{zj}-\bar{X})}{Kn(n-1)\sigma^2}$$

donde $\sigma^2 = \frac{1}{nK}\sum_{i=1}^{n}\sum_{j=1}^{K}(X_{ij}-\bar{X})^2$

Se puede expresar la varianza del estimador de la media poblacional en función del coeficiente de correlación intramuestral:

$$V(\bar{x}_j) = \frac{1}{K}\sum_{j=1}^{K}(\bar{x}_j - \bar{X})^2 = \dots = \frac{N-1}{N} \cdot \frac{S^2}{n}\left[1+(n-1)\rho_W\right]$$

- Precisión máxima ($V(\bar{x}_j)=0$) si $(n-1)\rho_W = -1$
$\Rightarrow \rho_W = -\frac{1}{n-1}$

- Precisión mínima ($V(\bar{x}_j)$ MÁX) si $\rho_W = +1$

- si $\rho_W = 0 \Rightarrow V(\bar{x}_j) = \frac{\sigma^2}{n}$ → el muestreo sistemático coincide con el m.a.s. CON reposición.

- $-\frac{1}{n-1} < \rho_W < 0 \Rightarrow$ muestreo sistemático +preciso que MAS.

- $0 < \rho_W < 1 \Rightarrow$ muestreo sistemático menos preciso que MAS.

También se puede llegar a aproximaciones de la varianza del estimador de la media poblacional a partir de los valores que tomen

$\rho_w \to$ coef. correlación intramuestral (intracongl.)

$\rho_{w\,st} \to$ coef. correlación entre las medias de los estratos

a) $\rho_w \sim 0 \Rightarrow$ puede suponerse la poblac. aleatoria.

MS $\sim$ MAS

$$\hat{V}(\bar{x}) \simeq (1-f) \cdot \frac{\hat{S}^2}{n} \quad / \quad \hat{S}^2 \equiv \text{cuasiv. de la muestra sistemática.}$$

b) $\rho_{w\,st} \sim 0 \Rightarrow$ puede utilizarse muestreo estratificado.

En la práctica, se transforman las n zonas con K unid. en $\frac{n}{2}$ zonas con $2K$ unid, para tener dos unidades muestrales por zona.

$$\hat{V}(\bar{x}_{st}) \simeq \frac{1-f}{n^2} \sum_{h=1}^{n/2} (x_{h1} - x_{h2})^2$$

c) Ni ρ_w ni $\rho_{w\,st}$ están próximos a 0.

En este caso hay que utilizar alguno de los mt. especiales, como el mt. de las muestras interpenetrantes, que se utiliza cuando tenemos un conjunto de muestras, elegidas con el mismo esquema de muestreo (indep. o no), t.q. una de las muestras proporcione una estimación válida del parámetro con el mismo error de muestreo.

En muestreo sistemático, este mt. elije muestras con cranques aleatorios y se forma un estimador combinado de la media poblacional basado en las medias de las muestras:

$$\bar{x}_c = \frac{1}{t} \sum \bar{x}_t \quad / \quad \hat{V}(\bar{x}_c) = \frac{1}{t(t-1)} \sum (\bar{x}_j^2 - \bar{x}_c^2)$$

Al aumentar el n.º de arranques, se aproxima a MAS

6_ PROBLEMÁTICA de la ESTIMACIÓN de VARIANZAS

las varianzas de los estimadores están expresadas en función de la cuasivarianza entre muestras S^2_{bs}.

Hemos visto que el muestreo sistemático era comparable al muestreo de conglomerados ~~donde~~ monoetápico, con conglomerados igual a las muestras (columnas).

En MCM, el estimador insesgado de la cuasivarianza entre conglomerados es la cuasivarianza muestral entre conglomerados $\hat{S}^2_{bs}$:

$$V(\bar{x}) = (1-f) \cdot \frac{S^2_{bs}}{n} \longrightarrow \hat{V}(\bar{x}) = (1-f) \cdot \frac{\hat{S}^2_{bs}}{n} \quad n=1$$

El problema está en que para calcular la cuasivarianza muestral entre conglomerados es necesario seleccionar más de un conglomerado para que forme parte de la muestra, lo que resulta imposible en muestreo sistemático.

Madow y Madow (1944) probaron que si la ordenación de las unidades poblacionales es aleatoria, el muestreo sistemático es, en promedio, equivalente al mas, de modo que:

$$E[V_s] = V_{mas}$$

donde $V_s \rightarrow$ promedio de las $N!$ posibles permutaciones de la población.

~~de~~ $V_{mas} \rightarrow$ es la misma para toda la poblac.

Por lo que para poblaciones ordenadas aleatoriamente y de tamaño N sufic. grande el m.s. $\sim$ m.a.s.

$$V(\bar{X}_j) \sim (1-f) \frac{S^2}{n} \quad / \quad f = \frac{n}{N} = \frac{n}{nk} = \frac{1}{k}$$

$$\hat{V}(\bar{X}_j) \simeq (1-f) \cdot \frac{\hat{S}^2}{n} \quad / \quad \hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X})^2 \simeq \hat{S}^2_j.$$