TEMA 6 MUESTREO POR CONGLOMERADOS MONOETÁPICO

Contenido

1- Definición. Aplicación. Selección de una muestra por Conglomerados. Etapas. Notación.
2- Muestreo monoetápico con conglomerados de igual tamaño. Estimación de la media, el total y la proporción. Coeficiente de correlación intra-conglomerados. Descomposición de la varianza. Elección del tamaño del conglomerado.
3- Muestreo monoetápico con conglomerados de tamaño desiguales. Estimadores insesgados y de razón para la media y el total. Tamaño de muestra. Estimación de la proporción y tamaño de muestra.
4- Muestreo por conglomerados con probabilidad proporcional al tamaño y con restitución. Método de Hansen y Hurwitz y de Lahin de selección. Estimadores ppt del total y de la media. Exavtitud relativa de los 3 estimadores.
5- Muestreo con probabilidades diferentes de selección y sin restitución. Estimador de Horvitz-Thompson.
6- Muestreo estratificados de conglomerados desiguales.

Muestreo por conglomerados

Definición:

El muestreo por conglomerados es un muestreo aleatorio donde cada unidad de muestreo (conglomerado) comprende a varias unidades elementales.

El muestreo por conglomerados es en muchos casos, un diseño efectivo para obtener la información deseada reduciendo los costos. El diseño por conglomerados no requiere de marco muestral completo de las unidades elementales.

El muestreo por conglomerados es diferente al estratificado, donde todos los estratos tienen representación en la muestra y cuyo objetivo es reducir la varianza de los estimadores.

Muestreo por conglomerados

Se aplica conglomerados porque:

i. No se dispone de marco muestral de las unidades últimas pero si de conglomerados y el costo de construir un marco sobrepasa los del estudio.
ii. Se minimizan costos al limitar los traslados entre conglomerados
iii. Es difícil fijar con acuracidad los límites de las unidades últimas.
iv. Consideraciones de: los objetivos de estudio, estructura de la población o administrativas definen la necesidad de conglomerados.

A diferencia del estratificado en el conglomerado la varianza del estimador se hace pequeña al hacer cada conglomerado heterogéneo dentro de sí y semejantes entre si.

Muestreo por conglomerados

Cómo seleccionar una muestra por conglomerados.

1. Definir el conglomerado.
   Tamaño: igual o diferente
   Tamaño apropiado: estructura de la población, costos, variabilidad del estimador e información disponible.
2. Formar el Marco Muestral (directorio de conglomerados)
3. Selección aleatoria de muestra de conglomerados.
4. Encuesta u observación (Etapas)
   - Monoetápico: Se observan todas las unidades de los conglomerados de la muestra.
   - Bietápico: Se seleccionan muestras aleatorias dentro de los conglomerados seleccionados en la primera etapa.
   - Polietápico: Se seleccionan conglomerados que a su vez están formados por conglomerados, donde a su vez se muestrea y así sucesivamente.

Muestreo por conglomerados

Notación.

El estudio del diseño por conglomerados requiere de una notación un poco mas compleja (un subíndice por etapa)

Población – P	Muestra – m
N = núm de conglomerados en P	n = núm de conglomerados en m
M_i = número de unidades en el conglomerado i	m_i = número de unidades del conglomerado i en la muestra
$M_o = \sum^N M_i$ = número total de unidades en la población	
$\overline{M} = M_o / N$ = tamaño medio del conglomerado	
$Y_{ij} \equiv y_{ij}$ observación j-ésima del i-ésimo conglomerado	

Muestreo por conglomerados

Notación.

Población – P	**Muestra – m**
Y_i Total del conglomerado i $Y_i = \sum^{N} Y_{ij}$	y_i Total muestral del conglomerado i $y_i = \sum^{N} y_{ij}$
$\overline{Y}_i = Y_i / M_i$ Media del conglomerado i	$\bar{y}_i = y_i / m_i$ Media muestral del conglomerado i
$\overline{\overline{Y}}_i = \sum \sum Y_{ij} / M_o$ Media Poblacional	$\overline{\overline{y}}_i$ Media muestral
$Y = \sum \sum Y_{ij}$ Total Poblacional	$\hat{Y}$ Total estimado
$\overline{Y} = \sum \sum Y_i / N$ Media del total por conglomerado	$\bar{\bar{y}}$ Media muestral del total por conglomerado

Muestreo por conglomerados

Estudiaremos ahora diferentes casos del muestreo por **conglomerados monoetápico**

Si el muestreo es monoetápico, observamos todas las unidades últimasde los n conglomerados seleccionados y mi = Mi, yi = Yi, $\bar{y}_i = \bar{Y}_i$. Distinguiendo dos casos: cuando los conglomerados son de igual tamaño y cuando son de diferente tamaño.

<u>Monoetápico con conglomerados de igual tamaño.</u>

$M_i = \bar{M}$ para todo i (todos los M_i son iguales)

Estimación de $\bar{\bar{Y}}_i$

$$\hat{\bar{\bar{Y}}} = \bar{\bar{y}} = \frac{\sum^{n}\sum^{\bar{M}} y_{ij}}{n\bar{M}} = \frac{\sum^{n} Y_{i,}}{n\bar{M}} = \frac{\bar{y}}{\bar{M}} = \bar{\bar{y}} = \frac{\sum \bar{Y}_i}{n}$$

Muestreo por conglomerados

En el monoetápico de igual tamaño, la varianza de $\bar{y}$ de la media muestral del total por conglomerado, es semejante a la varianza de la media muestral en el aleatorio simple

$$V\left(\bar{y}\right)=\frac{N-n}{Nn}\frac{\sum\left(Y_i-\bar{\bar{Y}}\right)^2}{N-1}$$

como $$\bar{\bar{y}}=\frac{\sum^{n}\sum^{M_i}y_{ij}}{n\bar{\bar{M}}}=\frac{\sum^{n}Y_i}{n\bar{\bar{M}}}=\frac{\bar{y}}{\bar{\bar{M}}}$$

luego $$V\left(\bar{\bar{y}}\right)=\frac{1}{\bar{\bar{M}}^2}V\left(\bar{y}\right)=\frac{1-f}{\bar{\bar{M}}^2 n}\frac{\sum^{N}\left(Y_i-\bar{\bar{Y}}\right)^2}{N-1}=\frac{1-f}{n}\frac{\sum^{N}\left(\bar{\bar{Y}}_i-\bar{\bar{\bar{Y}}}\right)^2}{N-1}$$

y por igual razón que en el m.a.s.

$$\hat{V}\left(\bar{\bar{y}}\right)=\frac{N-n}{Nn}\frac{\sum^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n-1}$$

Muestreo por conglomerados

Estimación del total

El total poblacional en el conglomerado monoetápico de igual tamaño $\mathrm{Y} = \sum^{N} \sum^{\overline{M}} y_{ij} = \sum Y_i = N\overline{Y} = N\overline{\overline{M}}\,\overline{\overline{Y}}$

su estimador es $\hat{\mathrm{Y}} = N\overline{\overline{M}}\,\overline{\overline{y}} = N\overline{y}$ con varianza

$$V(\hat{Y}) = N^2 V\left(\overline{y}\right) = N^2 \overline{M}^2 \frac{N-n}{Nn} \frac{\sum^{N}\left(\overline{Y}_i - \overline{\overline{Y}}\right)^2}{N-1}$$

y su estimador insesgado de $V(\hat{Y})$ es

$$\hat{V}(\hat{Y}) = N^2 \overline{M}^2 \frac{N-n}{Nn} \frac{\sum^{n}\left(\overline{y}_i - \overline{\overline{y}}\right)^2}{n-1} = \frac{N^2(N-n)}{Nn} \frac{\sum^{n}\left(y_i - \overline{y}\right)^2}{n-1}$$

Estimación de la proporción

Basándose en lo visto para la media proponga un estimador para la proporción, determine la varianza y la varianza estimada.

Muestreo por conglomerados

Coeficiente de correlación intra-conglomerados.

Definido por: $\rho = \frac{E\left[\left(Y_{ij} - \bar{\bar{Y}}\right)\left(Y_{il} - \bar{\bar{Y}}\right)\right]}{E\left[Y_{ij} - \bar{\bar{Y}}\right]^2}$ el numerador esta formado por $N\overline{M}(\overline{M} - 1)/2$ pares de unidades, así:

$$\rho = \frac{2\sum\sum\left(y_{ij} - \bar{\bar{Y}}\right)\left(y_{il} - \bar{\bar{Y}}\right)/N\overline{M}(\overline{M} - 1)}{\sigma^2} = \frac{2\sum\sum\left(y_{ij} - \bar{\bar{Y}}\right)\left(y_{il} - \bar{\bar{Y}}\right)/(\overline{M} -}{s^2(N\overline{M} - 1)}$$

así $2\sum\sum\left(y_{ij} - \bar{\bar{Y}}\right)\left(y_{il} - \bar{\bar{Y}}\right) = (\overline{M} - 1)(N\overline{M} - 1)s^2\rho$

Al expresar la varianza de $\bar{\bar{y}}$ en función del coeficiente de correlación y aproximar $N\overline{M} - 1 = N\overline{M}$ y $N - 1 = N$ se obtiene

$$V(\bar{\bar{y}}) \cong \frac{1 - f}{\overline{M}} \frac{s^2}{n}\left(1 + \left(\overline{M} - 1\right)\rho\right)$$

Muestreo por conglomerados

Coeficiente de correlación intra-conglomerados (cont.).

Esta expresión va a permitir hacer comparaciones entre el muestreo aleatorio simple y el muestreo por conglomerados.
Sean n_a y n_c los tamaños de la muestra en la misma población para el m.a.s. y el conglomerado

$$V(a) = (1-f)\frac{s^2}{n_a} \qquad \text{y} \qquad V(c) \cong (1-f)\frac{s^2}{n_c \overline{M}}\left(1+\left(\overline{M}-1\right)\rho\right)$$

Si la precisión en ambos diseños es igual

$$V(a) = V(c) \rightarrow n_c = n_a\left(1+\left(\overline{M}-1\right)\rho\right)$$

Luego

$\left(1+\left(\overline{M}-1\right)\rho\right)$ esta expresión la denomina Kish “efecto de diseño”

Muestreo por conglomerados

Coeficiente de correlación intra-conglomerados (cont.).

1. por el hay que multiplicar n_a para obtener n_c
2. ρ decrece mientras aumenta $\overline{M}$, pero su tasa de decrecimiento suele ser inferior a la del crecimiento de $\overline{M}$
3. El término $\left(\overline{M}-1\right)\rho$ expresa el aumento de la varianza debido a la selección de n conglomerados de tamaño $\overline{M}$ en lugar de $n\overline{M}$ unidades en el m.a.s.
4. De $V\left(\bar{\bar{y}}\right)=\frac{1-f}{\overline{M}}\frac{s^2}{n}\left(1+\left(\overline{M}-1\right)\rho\right)$

   Para $\rho>0$ existe un incremento en $V\left(\bar{\bar{y}}\right)$ para el muestreo por conglomerados en relación al m.a.s. de tamaño $n\overline{M}$, y el caso mas favorable al conglomerado es cuando $\rho=-1/(\overline{M}-1)$ que la varianza es nula. En el caso $\rho=0$ ambos métodos proporcionan igual precisión.

Muestreo por conglomerados

Descomposición de la varianza.

Es necesario determinar la variación entre y dentro de los conglomerados por ser la población finita se puede establecer el ANAVA para la muestra y para la población

$$\sum\sum\left(y_{ij}-\bar{\bar{Y}}\right)^2=\sum\sum\left(y_{ij}-\bar{Y}_i\right)^2+\sum\sum\left(\bar{Y}_i-\bar{\bar{Y}}\right)^2$$

$$S^2=\frac{\sum\sum\left(y_{ij}-\bar{\bar{Y}}\right)^2}{\sum M_i-1}=\frac{\sum\sum\left(y_{ij}-\bar{\bar{Y}}\right)^2}{N\bar{M}-1}$$

cuasivarianza poblacional

$$S_w^2=\frac{\sum\sum\left(y_{ij}-\bar{Y}_i\right)^2}{N(\bar{M}-1)}$$

cuasivarianza dentro de los conglomerados

$$S_b^2=\frac{\sum\sum\left(\bar{Y}_i-\bar{\bar{Y}}\right)^2}{N-1}$$

cuasivarianza entre los conglomerados

Muestreo por conglomerados

así

$$(N\overline{M}-1)S^2=(N-1)S_b^2+N(\overline{M}-1)S_w^2$$

$$S^2=\frac{(N-1)}{(N\overline{M}-1)}S_b^2+\frac{N(\overline{M}-1)}{(N\overline{M}-1)}S_w^2$$

$$S_b^2=\frac{(N\overline{M}-1)}{(N-1)}S^2-\frac{N(\overline{M}-1)}{(N-1)}S_w^2$$

$$S_w^2=\frac{(N\overline{M}-1)}{N(\overline{M}-1)}S^2-\frac{(N-1)}{N(\overline{M}-1)}S_b^2$$

Muestreo por conglomerados

Análisis de Varianza

Población

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios
Conglomerados	$N-1$	$\sum\sum\left(\overline{Y}_i-\overline{\overline{Y}}\right)^2$	S_b^2
Elementos	$N(\overline{M}-1)$	$\sum\sum\left(Y_{ij}-\overline{Y}_i\right)^2$	S_w^2
Total	$N\overline{M}-1$	$\sum\sum\left(Y_{ij}-\overline{\overline{Y}}\right)^2$	S^2

Muestra

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios
Conglomerados	$n-1$	$\sum\sum\left(\overline{y}_i-\overline{\overline{y}}\right)^2$	s_b^2
Elementos	$n(\overline{M}-1)$	$\sum\sum\left(y_{ij}-\overline{y}_i\right)^2$	s_w^2
Total	$n\overline{M}-1$	$\sum\sum\left(y_{ij}-\overline{\overline{y}}\right)^2$	s^2

Muestreo por conglomerados (ejemplos)

Población	Variables	Elementos	Conglomerados o unidades de muestreo
Ciudad A	Característica de la vivienda	Viviendas	Manzanas
Ciudad B	Compras de ropa	Personas	Viviendas
Aeropuerto	Información acerca de viajes	Pasajeros que salen	Vuelos
Escuela	Notas	Estudiantes	Salones
Gente de pueblo	Actitudes sociales	Adultos	Pueblos
Tránsito anual en puente	Origen y destino	Vehículos	Intervalos de 40 minutos
Archivo de propiedad de terrenos en ciudad	Información sobre impuestos	Propiedades de terreno	Páginas de registro (o libros)
Granja	Características de las naranjas	Naranjas	Arboles

Muestreo por conglomerados

Elección del tamaño del conglomerado

En el muestreo por conglomerados, con conglomerados de igual tamaño es importante determinar el tamaño apropiado del conglomerado ($\overline{M}$).
El tamaño depende entre otros de los siguientes factores: tipo y estructura de la población, posibilidad de cambiar la estructura de agrupamiento, información disponible de la población, variabilidad de la población y de los conglomerados y la estructura de costos.
La bibliografía presenta diversas metodologías para determinar el tamaño optimo de los conglomerados, por ejemplo tres métodos (cochran)

1. Si se dispone de información poblacional para diferentes tamaños de conglomerados.
2. Si la comparación de la precisión se hace a partir de datos muéstrales.
3. Hipótesis de la existencia de una ley que regula el comportamiento dentro de los conglomerados S_w^2 y se relaciona con el tamaño del conglomerado.

Muestreo por conglomerados

Tamaño del conglomerado, en base a:

1. *Información poblacional para diferentes tamaños*

   Un principio general para seleccionar el tamaño del conglomerado es el criterio de menor varianza para un costo dado, o equivalente, el menor costo para una varianza prefijada.

   Este criterio se basa en que la precisión relativa es proporcional a $M_u^2 / C_u S_u^2$, donde C_u es el costo de encuesta por unidad, M_u es el tamaño relativo de la unidad, S_u^2 varianza entre los totales de unidades, por lo cual disponemos de un criterio para seleccionar el tamaño de conglomerado adecuado.

   Cuando hay mas de una característica a considerar se requiere tomar decisiones que estudien las diferentes alternativas.

Muestreo por conglomerados

Tamaño del conglomerado, en base a:

2. *Precisiones en base de datos muéstrales*
   Para una encuesta con unidades de tamaño M, si se registran los datos para cada una de las M unidades menores, se puede hacer comparaciones entre las precisiones de los diferente tamaño de conglomerados, un instrumento de utilidad en este método es el análisis de varianza acompañado de un análisis de costo.
3. *Funciones de varianza*
   En este enfoque se considera M como una variable continua y allí encontrar el optimo. Este método también utiliza el análisis de varianza para predecir S_b^2 y S_w^2 relacionando $S_w^2 = AM^g$ y ajustando por $\log(S_w^2) = \log(A) + g * \log(M)$, necesitando al menos tres valores de S_w^2 y M para estimar A y g, y apreciar la linealidad del ajuste.

Muestreo por conglomerados

Muestreo por conglomerados monoetápico de tamaños desiguales.

En la mayoría de las aplicaciones los conglomerados son de tamaño diferente (poblaciones naturales)

Estimación del total poblacional: $Y = \sum^{N} \sum^{M_i} y_{ij}$

Dos estimadores diferentes de Y

Estimación insesgada

Un estimador insesgado de Y en el muestreo por conglomerados monoetápico es: $\hat{Y} = \frac{N}{n} \sum^{n} y_i$

donde y_i es el total del conglomerado i-ésimo, $y_i = \sum_{j=}^{M_i} y_{ij}$

También $\hat{Y} = \frac{N}{n} \sum^{n} y_i = N \overline{y}$ donde $\overline{y}$ es la media muestral del total por conglomerado.

Muestreo por conglomerados

Sabemos que en el m.a.s. $V(\bar{y}) = \frac{1-f}{n}\frac{\sum(Y_i - \bar{Y})^2}{N-1}$ (note que $y_i = Y_i$)

Así $V(\hat{Y}) = N^2 V(\bar{y}) = \frac{N^2(1-f)}{n}\frac{\sum(Y_i - \bar{Y})^2}{N-1}$

$\bar{Y}$ es la media poblacional del total por conglomerado

A pesar de ser $\hat{Y}$ un estimador insesgado puede ser poco preciso, debido a que no toma en cuenta las ponderaciones M_i, fundamentalmente cuando los $\bar{y}_i$ (media del conglomerado i) varían poco y los M_i varian considerablemente, y en este caso los $y_i = M_i \bar{y}_i$ varian considerablemente y la varianza $V(\hat{Y})$ es grande.

Note que en $\hat{Y} = \frac{N}{n}\sum^n y_i$ cada y_i es ponderado por el mismo peso. Una forma de corregir esta impresión es tomar en cuenta los valores M_i