Capítulo 4

MUESTREO ESTRATIFICADO

Capítulo 4. Diseños estratificados

1. Principios y objetivos: Notación
2. Diseño muestral.
3. Probabilidad de inclusión y estimación
4. Reparto (afijación) proporcional
5. Reparto óptimo o de Neyman
6. Resultados globales/locales
7. Tamaño de muestra
8. Algunas notas

1. Principios y objetivos; notación

$$V\left(\hat{\overline{Y}}\right)=\left(1-\frac{n}{N}\right)\frac{1}{n}S^2$$

Estructura de la varianza y descomposición de la varianza

Si somos capaces de:

- repartir las unidades de la población en grupos según una variable auxiliar *X* cualitativa tal que la varianza al interior de los estratos sea más pequeña que la varianza global…

- definir un estimador que sólo dependa de la varianza al interior de los estratos

entonces, se obtiene generalmente un estimador mas preciso que el estimador ASSR

$$\left(Y_{\alpha,h}-\bar{\bar{Y}}\right)=\left(Y_{\alpha,h}-\bar{Y}_h\right)+\left(\bar{Y}_h-\bar{\bar{Y}}\right)$$

Se verifica que

$$\sum_h \sum_{\alpha\in U_h}\left(Y_{\alpha,h}-\bar{\bar{Y}}\right)^2=\sum_h \sum_{\alpha\in U_h}\left(Y_{\alpha,h}-\bar{Y}_h\right)^2+\sum_h \sum_{\alpha\in U_h}\left(\bar{Y}_h-\bar{\bar{Y}}\right)^2$$

$$\sum_h \sum_{\alpha\in U_h}\left(Y_{\alpha,h}-\bar{\bar{Y}}\right)^2=\sum_h \sum_{\alpha\in U_h}\left(Y_{\alpha,h}-\bar{Y}_h\right)^2+\sum_h N_h\left(\bar{Y}_h-\bar{\bar{Y}}\right)^2$$

Variabilidad total = Variabilidad intra-grupos + Variabilidad inter-grupos

Descomposición de la varianza

$$\sigma^2=\sum_h \frac{N_h}{N}\sigma_h^2+\sum_h \frac{N_h}{N}\left(\bar{Y}_h-\bar{\bar{Y}}\right)^2$$

$$\sigma^2=\sigma^2_{INTRA}+\sigma^2_{INTER}$$

Casi-varianza

$$S^2 \cong \sum_h \frac{N_h}{N} S_h^2 + \sum_h \frac{N_h}{N} \left(\overline{Y}_h - \overline{Y}\right)^2$$

Ratio de correlación:

medida de la relación entre una variable cuantitativa y una variable categórica

$$\eta^2 = \frac{\sum_h N_h \left(\overline{Y}_h - \overline{Y}\right)^2}{\sum_h \sum_{k \in U_h} \left(y_{k,h} - \overline{Y}\right)^2} = \frac{\sigma^2_{INTER}}{\sigma^2}$$

Notación

El universo *U* esta dividido en *H* estratos: $\boldsymbol{U_h}$, $h=1,\ldots,H$

$\mathbf{U_1}$	$\mathbf{U_2}$			$\mathbf{U_h}$			$\mathbf{U_H}$

Universo	Estrato h
N	N_h
$\overline{Y}$	$\overline{Y}_h$
T	T_h
σ^2	σ_h^2
S^2	S_h^2

$N= N_1+N_2+\ldots+N_h+\ldots+N_H$

$$\overline{Y} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{Y}_h$$

$$T = T + \quad + T + \quad + T$$

2. Diseño muestral

Un diseño muestral es estratificado si:
- En cada estrato, se selecciona una muestra simple de tamaño fijo n_h
- la selección de una muestra en un estrato es independiente de la selección de las muestras en los otros estratos

(**La estratificación se puede generalizar a otros métodos de extracción**)

Se nota $\mathcal{s}_h$, la muestra aleatoria seleccionada en el estrato *h* mediante el diseño $p_h(.)$

La muestra $\mathcal{S}$ es:

$$\mathcal{S} = \bigcup_h \mathcal{S}_h$$

Un valor posible $\mathcal{s}$ de $\mathcal{S}$ se nota:

$$\mathcal{s} = \bigcup_h \mathcal{s}_h$$

Dado la independencia de las extracciones entre los estratos, la probabilidad de observar la muestra $\mathcal{s}$ es:

$$p(\mathcal{s}) = \prod_{h=1}^{H} p_h\left(\mathcal{s}_h\right)$$

Se observan *H* muestras (extracciónes independientes en cada estrato)

$$n = \sum_h n_h$$

Es posible que:

$$\frac{n_h}{n} \neq \frac{N_h}{N}$$

Muestra global	Submuestra *h* (ASSR)
n	n_h
	$f_h = \frac{n_h}{N_h}$
$\hat{\bar{Y}}$	$\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{h}}$
$\hat{T}$	$\hat{T}_h$
	$\boldsymbol{S}_{\boldsymbol{h}}^2$

3. Probabilidades de inclusión y estimación

probabilidades de inclusión

probabilidad de inclusión de la unidad α $\pi_\alpha = \frac{n_h}{N_h}, \quad k \in U_h$

probabilidad de que *a y b* pertenezcan a la muestra:

$$\pi_{\alpha\beta} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)}, \quad \alpha \; y \; \beta \in U_h$$

$$\pi_{\alpha\beta} = \frac{n_h n_l}{N_h N_l}, \quad \alpha \in U_h \; \beta \in U_l$$

Estimación del total

Estimador

$$\hat{T} = \sum_{h=1}^{H} \hat{T}_h = \sum_{h=1}^{H} N_h \hat{\bar{Y}}_h = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{y_{ih}}{\pi_{ih}}$$

Estimador sin sesgo

$\hat{T}_h$: π-estimador del total en estrato *h.*

$\pi_{i,h}=n_h/N_h$ (depende del estrato).

Varianza del estimador

$$V(\hat{T}) = V\left(\sum_{h=1}^{H} N_h \bar{y}_h\right) = \sum_{h=1}^{H} N_h^2 V(\bar{y}_h) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_h^2$$

Estimación de la varianza del estimador

$$\hat{V}(\hat{T}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} s_h^2$$

Estimador sin sesgo

Estimación de la media

Estimador

$$\hat{\bar{Y}} = \frac{\hat{T}}{N} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h$$

Estimador sin sesgo

Varianza del estimador

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} S_h^2$$

Estimación de la varianza del estimador

$$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} s_h^2$$

Estimador sin sesgo

Estimación de una proporción

Estimador

$$\hat{p} = \sum_h \frac{N_h}{N} \hat{p}_h$$

Estimador sin sesgo

Varianza del estimador

$$V(\hat{p}) = \sum_h \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{P_h(1-P_h)}{n_h}$$

Estimación de la varianza del estimador

$$\hat{V}(\hat{p}) = \sum_h \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

Aplicación numérica

Se dispone de una población de 1060 empresas. Se desea estimar el número medio de empleados por empresa. La población está compuesta de 5 estratos definidos a partir del tamaño de la empresa según el número de empleados en clase. Esta información es conocida a partir de los registros oficiales que no dan el número exacto de empleados, sino el tamaño en clase.

Se dispone de un presupuesto que permite encuestar 300 empresas. Se decide realizar un muestreo aleatorio simple en cada estrado según el reparto indicado en la tabla. Sobre cada empresa, se mide la variable *Y*:"número de empleados" y se calcula la media y la varianza de dicha variable en cada estrato. Escoger un estimador, dar su expresión. Hacer una estimación de Y por punto y por intervalo.

Estrato según tamaño empresa	N_h	n_h	$\bar{y}_h$	s_h^2
0-9	500	130	5	1.5
10-19	300	80	12	4.0
20-49	150	60	30	8.0
50-499	100	25	150	100.0
500 y más	10	5	600	2500.0
Total	1060	300		

Solución

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{h=1}^{k} N_h \bar{y}_h$$

$$\hat{V}\left[\hat{\bar{Y}}\right] = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h} = 0.055$$

$$\left[\hat{\bar{Y}} - 2\sqrt{\hat{V}\left(\hat{\bar{Y}}\right)}, \hat{\bar{Y}} + 2\sqrt{\hat{V}\left(\hat{\bar{Y}}\right)}\right] = [29.3, 30.3]$$

Nota: selección de los estratos

estratos homogéneos con respecto al tema estudiado

criterios de estratificación (variables cualitativas):
- disponibilidad
- correlacionados con el tema estudiado en la encuesta: categoría social, el nivel de instrucción, el tamaño del hogar, el tipo de hábitat

Generalmente, se emplearán:
- criterios correspondientes a una **tipología**
- criterios de **tamaño**

Una estratificación puede ser eficaz para el estudio de un fenómeno, por ejemplo la mortalidad, y serlo poco para el estudio de otros fenómenos, por ejemplo la actividad económica o los movimientos migratorios.

4. Reparto proporcional (Afijación proporcional)

El reparto proporcional consiste en utilizar la misma tasa de muestreo en todos los estratos

$$\frac{n_h}{n} = \frac{N_h}{N}$$

La expresión de la varianza del estimador en caso de muestreo estratificado:

$$V\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h} S_h^2$$

Se simplifica y se puede escribir

$$\mathbf{V}\left(\hat{\bar{Y}}\right) = (1-f)\cdot\frac{S^2_{intra}}{n}$$

y, la estimación de esta varianza se calcula mediante:

$$\hat{V}\left(\hat{\bar{Y}}\right) = (1-f)\cdot\frac{s^2_{intra}}{n}$$

Ejemplo: País dividido en dos regiones (estratos)

Estrato *h*	**Número de aldeas (N_h)**	**Población Total T_h**	**S_h**	$\overline{Y}_h$
1	**3 000**	**956 800**	**100**	**319**
2	**1 000**	**605 000**	**200**	**605**
Total	**4 000**	**1 561 800**		**390**

Estrato	Proporcional
1	60
2	20
Total	80

Ejemplo de las empresas

5. Reparto óptimo de Neyman

El reparto de Neyman consiste en minimizar la varianza del estimador, lo que conduce a

$$\begin{cases} Min\left(V\left(\hat{T}\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_h^2 \right) \\ \sum_h n_h = n \end{cases}$$

$$n_h = \frac{N_h \cdot S_h}{\sum_h N_h \cdot S_h} \cdot n$$

$$f_h = \frac{n_h}{N_h} = \frac{S_h}{\sum_h N_h \cdot S_h} \cdot n$$

Con la siguiente notación:

$$\overline{S} = \frac{1}{N} \sum_h N_h S_h$$

La expresión de la varianza del estimador en caso de muestreo estratificado:

$$V\left(\hat{\overline{Y}}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_h^2$$

se simplifica y se puede escribir:

$$V\left(\hat{\overline{Y}}\right) = \frac{1}{n}\left(\overline{S}\right)^2 - \frac{1}{N} S_{intra}^2$$

y el estimador de la varianza de la media se expresa mediante:

$$\hat{V}\left(\hat{\overline{Y}}\right) = \frac{1}{n}\left(\overline{s}\right)^2 - \frac{1}{N} s_{intra}^2$$

Ejemplo: País dividido en dos regiones (estratos)

Estrato *h*	**Número de aldeas (N_h)**	**Población Total *Y***	**S_h**	**$\overline{Y}_h$**
1	**3 000**	**956 800**	**100**	**319**
2	**1 000**	**605 000**	**200**	**605**
Total	**4 000**	**1 561 800**		**390**

Estrato	Proporcional	Neyman
1	60	48
2	20	32
Total	80	80

Ejemplo de las empresas

6. Objetivos globales/ Objetivos locales

Búsqueda de precisión a nivel de cada estrato

Cuando se desea obtener información significativa para cada estrato, habrá que dar una ventaja relativa a los estratos menos poblados, generalmente en detrimento de la precisión global.

Si se desea la misma precisión a nivel de cada estrato y si se estima que los estratos presentan la misma heterogeneidad para el carácter estudiado, se deberán tomar tamaños de muestra similares en cada uno.

Volvamos al país dividido en dos regiones…

Estrato h	Número de aldeas (N_h)	Población Total T	S_h	$\overline{Y}_h$
1	3 000	956 800	100	319
2	1 000	605 000	200	605
Total	4 000	1 561 800		390

Ahora, se desea obtener intervalos de confianza de la misma amplitud para la estimación del tamaño medio de las aldeas en cada uno de los dos estratos.

$$V(\bar{y}_1) = V(\bar{y}_2) \qquad \left(1-\frac{n_1}{N_1}\right)\frac{S_1^2}{n_1} = \left(1-\frac{n_2}{N_2}\right)\frac{S_2^2}{n_2}$$

Despreciando las tasas de muestreo para simplificar: $\frac{S_1^2}{n_1} = \frac{S_2^2}{n_2} = \frac{S_1^2 + S_2^2}{n}$

con $n = 80$, se obtiene:

$n_1 = 16$ $\qquad$ $n_2 = 64$

Aquí, para obtener una estimación precisa para el estrato 2 (que contiene menos aldeas que el estrato 1, pero con tamaños más dispersos), se debe privilegiar la asignación de unidades encuestadas hacia este estrato.

Resumen

Estrato *h*	Número de aldeas (*Nh*)	Población Total	S_h	$\overline{Y}_h$	Rep. Prop.	Rep. Neyman	Misma precisión en los dos estratos
1	3 000	956 800	100	319	60	48	16
2	1 000	605 000	200	605	20	32	64
Total	4 000	1 561 800		390			

7. Tamaño de muestra

$$\mathbf{V}\left(\hat{T}\right)=\sum_{\mathbf{h=1}}^{\mathbf{k}} N_h \frac{N_h - na_h}{na_h} S_h^2$$

Se nota: $n_h = a_h \cdot n$

Se desea: $\hat{\mathbf{V}}\left(\hat{T}\right)= V$ *fijado a priori*

$$V = \frac{1}{n}\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2 - \sum_{h=1}^{H} N_h S_h^2$$

Finalmente:

$$n = \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{V + \sum_{h=1}^{H} N_h S_h^2}$$

Problema de minimización:

$$\underset{a_h}{\boldsymbol{Min}} \; \frac{\sum_{h=1}^{H} \frac{N_h^2}{a_h} S_h^2}{V + \sum_{h=1}^{H} N_h S_h^2}$$

$$\boldsymbol{con} \quad \sum_{h=1}^{H} a_h = 1$$

$$a_h = \frac{N_h \cdot S_h}{\sum_h N_h \cdot S_h}$$

$$n^* = \frac{\left(\sum_{h=1}^{H} N_h S_h \right)^2}{V + \sum_{h=1}^{H} N_h S_h^2}$$

7. Algunas notas

1. El reparto proporcional no depende de la variable de interés *Y* (no hace falta conocimiento sobre *Y*, ni media, ni varianza, etc.). Basta conocer N_h y la variable auxiliar, a partir de la cual se forman los estratos, para toda la población.

2. EL reparto de Neyman requiere un buen conocimiento de las cuasi-varianzas S_h de Y en los estratos. Sicho conocimiento se habrá obtenido por:
   - Encuestas anteriores
   - Expertos
   - Muestreos en dos fases

3. Notar que: $$V\left(\hat{\bar{Y}}_{OPT}\right) \leq V\left(\hat{\bar{Y}}_{PROP}\right) \leq V\left(\hat{\bar{Y}}_{ASSR}\right)$$

Si las cuasi- varianzas en los estratos se conocen con una precisión suficiente