Maestría en Generación de Información Estadística

Teoría y Técnicas de Muestreo

Augusto E. Hoszowski

UNTREF

2024

Tabla de Contenidos

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Bibliografía

- Muestreo: Diseño y Análisis, Sharon Lohr (2000) . Thomson
- Teoría de Muestreo, Yves Tillé (en castellano, 2005)
- Sampling: Design and Analysis, Sharon Lohr (2021, 3Ed) . Thomson
- Model Assisted Survey Sampling, C Sarndal, B Swenson, J Wretman (1992). Springer
- Practical Tools for Designing and Weighting Survey Samples (2ºEd), R Valliant et.al. Springer
- Complex Surveys: A Guide to Analysis Using R. T. Lumley. Wiley

Muestreo Polietápico

En el muestreo por conglomerado es usual que se subseleccionen unidades dentro de los conglomerados
$\Rightarrow$ Muestreo polietapico

Radios censales $\Rightarrow$ Viviendas
Radios censales $\Rightarrow$ Manzanas $\Rightarrow$ Viviendas
Colegio $\Rightarrow$ Aulas
Radios censales $\Rightarrow$ Viviendas $\Rightarrow$ Hogar $\Rightarrow$ Persona
Región (NEA, NOA, ...) $\Rightarrow$ Provincia $\Rightarrow$ Radio Censal (??)

Muestreo Polietápico

En el muestreo por conglomerado es usual que se subseleccionen unidades dentro de los conglomerados
$\Rightarrow$ Muestreo polietapico

Radios censales $\Rightarrow$ Viviendas
Radios censales $\Rightarrow$ Manzanas $\Rightarrow$ Viviendas
Colegio $\Rightarrow$ Aulas
Radios censales $\Rightarrow$ Viviendas $\Rightarrow$ Hogar $\Rightarrow$ Persona
Región (NEA, NOA, ...) $\Rightarrow$ Provincia $\Rightarrow$ Radio Censal (??)

Muestreo Polietápico

En el muestreo por conglomerado es usual que se subseleccionen unidades dentro de los conglomerados
$\Rightarrow$ Muestreo polietapico

Radios censales $\Rightarrow$ Viviendas
Radios censales $\Rightarrow$ Manzanas $\Rightarrow$ Viviendas
Colegio $\Rightarrow$ Aulas
Radios censales $\Rightarrow$ Viviendas $\Rightarrow$ Hogar $\Rightarrow$ Persona
Región (NEA, NOA, ...) $\Rightarrow$ Provincia $\Rightarrow$ Radio Censal (??)

Muestreo Polietápico

En el muestreo por conglomerado es usual que se subseleccionen unidades dentro de los conglomerados
$\Rightarrow$ Muestreo polietapico

Radios censales $\Rightarrow$ Viviendas
Radios censales $\Rightarrow$ Manzanas $\Rightarrow$ Viviendas
Colegio $\Rightarrow$ Aulas
Radios censales $\Rightarrow$ Viviendas $\Rightarrow$ Hogar $\Rightarrow$ Persona
Región (NEA, NOA, ...) $\Rightarrow$ Provincia $\Rightarrow$ Radio Censal (??)

Muestreo Polietápico

En el muestreo por conglomerado es usual que se subseleccionen unidades dentro de los conglomerados
$\Rightarrow$ Muestreo polietapico

Radios censales $\Rightarrow$ Viviendas
Radios censales $\Rightarrow$ Manzanas $\Rightarrow$ Viviendas
Colegio $\Rightarrow$ Aulas
Radios censales $\Rightarrow$ Viviendas $\Rightarrow$ Hogar $\Rightarrow$ Persona
Región (NEA, NOA, ...) $\Rightarrow$ Provincia $\Rightarrow$ Radio Censal (??)

Muestreo por conglomerado bietápico

1. El universo se divide en N_I subuniversos, llamados Unidades de Primera Etapa (UPE)
2. Mediante algún diseño muestral se selecciona una muestra s_I de UPE's
3. Dentro de cada UPE i seleccionada ($i \in s_I$) se selecciona una muestra s_i de elementos de U_i, en forma independiente de UPE a UPE

El factor de expansión final de cada elemento es el producto de los factores de expansión de cada una de las dos etapas

Muestreo por conglomerado bietápico

1. El universo se divide en N_I subuniversos, llamados Unidades de Primera Etapa (UPE)
2. Mediante algún diseño muestral se selecciona una muestra s_I de UPE's
3. Dentro de cada UPE i seleccionada ($i \in s_I$) se selecciona una muestra s_i de elementos de U_i, en forma independiente de UPE a UPE

El factor de expansión final de cada elemento es el producto de los factores de expansión de cada una de las dos etapas

Muestreo por conglomerado bietápico

1. El universo se divide en N_I subuniversos, llamados Unidades de Primera Etapa (UPE)
2. Mediante algún diseño muestral se selecciona una muestra s_I de UPE's
3. Dentro de cada UPE i seleccionada ($i \in s_I$) se selecciona una muestra s_i de elementos de U_i, en forma independiente de UPE a UPE

El factor de expansión final de cada elemento es el producto de los factores de expansión de cada una de las dos etapas

Muestreo por conglomerado bietápico

1. El universo se divide en N_I subuniversos, llamados Unidades de Primera Etapa (UPE)
2. Mediante algún diseño muestral se selecciona una muestra s_I de UPE's
3. Dentro de cada UPE i seleccionada ($i \in s_I$) se selecciona una muestra s_i de elementos de U_i, en forma independiente de UPE a UPE

El factor de expansión final de cada elemento es el producto de los factores de expansión de cada una de las dos etapas

Muestreo por conglomerado bietápico

1. El universo se divide en N_I subuniversos, llamados Unidades de Primera Etapa (UPE)
2. Mediante algún diseño muestral se selecciona una muestra s_I de UPE's
3. Dentro de cada UPE i seleccionada ($i \in s_I$) se selecciona una muestra s_i de elementos de U_i, en forma independiente de UPE a UPE

El factor de expansión final de cada elemento es el producto de los factores de expansión de cada una de las dos etapas

Muestreo polietápico y muestreo en dos fases

Muestreo polietápico

'En el muestreo polietápico el diseño de muestra en cada conglomerado es independiente de la muestra de conglomerados seleccionados'

$$\neq$$

Muestreo en dos fases

$$U \Rightarrow s \Rightarrow s'$$

Muestreo polietápico y muestreo en dos fases

Supongamos un colegio con tres aulas de sexto grado. En cada aula hay al menos un varón y una mujer

Muestreo bi-etápico
Seleccionamos una MAS de dos aulas. En cada aula seleccionamos una muestra aleatoria estratificada por sexo: un MAS de un varón y una MAS de una mujer $\Rightarrow$

Muestreo en dos fases
Seleccionamos una MAS de dos aulas. Del conjunto de alumnos seleccionados, seleccionamos una muestra aleatoria estratificada por sexo: un MAS de dos varones y una MAS de dos mujeres $\Rightarrow$

En el muestreo en dos fases un factor de expasión se puede construir multiplicando los factores de expansión de cada fase

Muestreo polietápico y muestreo en dos fases

Ejemplo

Aula	Alumno	Sexo
A	1	V
	2	V
	3	M
B	4	V
	5	M
	6	M
	7	M
C	8	V
	9	M
	10	M

Hallar los factores de expansión finales en cada caso

A:

Primera etapa: $\{A, B\}$; Segunda etapa: $\{1, 4, 3, 5, \}$
Primera fase: $\{A, B\}$; Segunda fase: $\{1, 4, 3, 5\}$

B: Primera etapa: $\{A, C\}$; Segunda etapa: $\{1, 3, 8, 9\}$
Primera fase: $\{A, C\}$; Segunda fase: $\{1, 3, 8, 9\}$

Muestreo polietápico y muestreo en dos fases

Ejemplo

Aula	Alumno	Sexo
A	1	V
	2	V
	3	M
B	4	V
	5	M
	6	M
	7	M
C	8	V
	9	M
	10	M

Hallar los factores de expansión finales en cada caso

A:

Primera etapa: $\{A, B\}$; Segunda etapa: $\{1, 4, 3, 5, \}$
Primera fase: $\{A, B\}$; Segunda fase: $\{1, 4, 3, 5\}$

B: Primera etapa: $\{A, C\}$; Segunda etapa: $\{1, 3, 8, 9\}$
Primera fase: $\{A, C\}$; Segunda fase: $\{1, 3, 8, 9\}$

Muestreo polietápico y muestreo en dos fases

Ejemplo

Aula	Alumno	Sexo
A	1	V
	2	V
	3	M
B	4	V
	5	M
	6	M
	7	M
C	8	V
	9	M
	10	M

Hallar los factores de expansión finales en cada caso

A:

Primera etapa: $\{A, B\}$; Segunda etapa: $\{1, 4, 3, 5, \}$

Primera fase: $\{A, B\}$; Segunda fase: $\{1, 4, 3, 5\}$

B: Primera etapa: $\{A, C\}$; Segunda etapa: $\{1, 3, 8, 9\}$

Primera fase: $\{A, C\}$; Segunda fase: $\{1, 3, 8, 9\}$

Muestreo bi etápico MAS / MAS

N conglomerados de igual tamaño M

Al ser los conglomerados de igual tamaño y MAS en ambas etapas un estimador insesgado de $\bar{Y}$ es:

$$\hat{\bar{t}}_{\pi y} = \frac{1}{n} \cdot \sum_{i \in s_I} \bar{Y}_i$$

siendo Y_i la media en el conglomerado i. Y

$$V(\hat{\bar{t}}_{\pi y}) = (1 - n/N) \cdot \frac{S_1^2}{n} + (1 - m/M) \cdot \frac{S_2^2}{nm}$$

$$S_1^2 = \frac{1}{N-1} \cdot \sum_{i \in s} \left(\bar{Y}_i - \bar{\bar{Y}} \right)^2 \; ; \quad S_{2i}^2 = \frac{1}{M-1} \cdot \sum_j \left(Y_{ij} - \bar{Y}_i \right)^2$$

$$S_2^2 = \frac{1}{N} \cdot \sum_i S_{2i}^2$$

Nota: En el caso de un diseño polietápico estratificado en cada etapa y con

Muestreo bi−etápico

Caso general

N conglomerados de tamaño N_i. Seleccionamos una muestra aleatoria s_I de n conglomerados, mediante un diseño arbitrario. En cada conglomerado de s_I seleccionamos una muestra aleatoria de unidades finales, con un diseño que no dependa de la muestra de conglomerados seleccionada
Un estimador insesgado del total Y es:

$$\hat{t}_{\pi y} = \sum_{i \in s_I} \frac{\hat{t}_{yi}}{\pi_i}$$

siendo $\hat{t}_{yi}$ un estimador del total de Y en el conglomerado i
- La varianza de $\hat{t}_{\pi y}$ tiene dos componentes:
La variabilidad debida a los conglomerados seleccionados
La variabilidad debida a las unidades finales seleccionadas, en la muestra de conglomerados

Muestreo polietápico

Consideraciones generales

- Si los conglomerados tienen tamaños desiguales lo habitual no es seleccionarlos mediante MAS sino con probabilidad proporcional a alguna medida de tamaño
- Si el parámetro a estimar se refiere a una media o total de los unidades secundarias
- Estratificar por *tamaño* las UPE
- Cantidad mínima de UPE

Muestreo polietápico

Consideraciones generales

- Si los conglomerados tienen tamaños desiguales lo habitual no es seleccionarlos mediante MAS sino con probabilidad proporcional a alguna medida de tamaño
- Si el parámetro a estimar se refiere a una media o total de los unidades secundarias
- Estratificar por *tamaño* las UPE
- Cantidad mínima de UPE

Muestreo polietápico

Consideraciones generales

- Si los conglomerados tienen tamaños desiguales lo habitual no es seleccionarlos mediante MAS sino con probabilidad proporcional a alguna medida de tamaño
- Si el parámetro a estimar se refiere a una media o total de los unidades secundarias
- Estratificar por *tamaño* las UPE
- Cantidad mínima de UPE

Muestreo polietápico

Consideraciones generales

- Si los conglomerados tienen tamaños desiguales lo habitual no es seleccionarlos mediante MAS sino con probabilidad proporcional a alguna medida de tamaño
- Si el parámetro a estimar se refiere a una media o total de los unidades secundarias
- Estratificar por *tamaño* las UPE
- Cantidad mínima de UPE

Muestreo polietápico

Consideraciones generales

- Al estimar el total de una variable Y la selección pps será efectiva si hay alta correlación (positiva) entre Y y el tamaño de los conglomerados, en cada estrato. Ejemplos donde no pasa eso.
- Si M_i (tamaño estimado al hacer la selección) y M_i^*, tamaño al momento del relevamiento difieren.

Muestreo polietápico

Consideraciones generales

- Al estimar el total de una variable Y la selección pps será efectiva si hay alta correlación (positiva) entre Y y el tamaño de los conglomerados, en cada estrato. Ejemplos donde no pasa eso.
- Si M_i (tamaño estimado al hacer la selección) y M_i^*, tamaño al momento del relevamiento difieren.

Muestreo bi etápico MAS(estratificado) / MAS(estratificado

Implementación en *survey*

En el caso de un MAS (estratificado) en la primera etapa y en las etapas subsiguientes (estratificado) podemos darle a **survey** la información exacta del diseño para incorporarla en la estimación de los errores de muestreo:

Por ejemplo en un muestreo estratificado en la primera etapa, MAS en cada estrato y MAS en la segunda etapa en cada conglomerado seleccionado:

diseno $\leftarrow$ svydesign(id= $\sim$ Id1+Id2, strata=$\sim$ Estrato1, weights=$\sim$ pondera, fpc=$\sim$ fpc1+fpc2, data=df)

donde Id1,Id2, fpc1 y fpc2 pueden ser variables de *df*: identificadores de las unidades de primera y segunda etapa, y fpc1 y fpc2 la cantidad de unidades del estrato de primera etapa y de segunda etapa respectivamente

Selección con reposición en la primera etapa

Con probabilidades de extracción P_i

Si en la primera etapa n UPE se seleccionan con reposición con probabilidades p_i, $\sum_i p_i = 1$ entonces

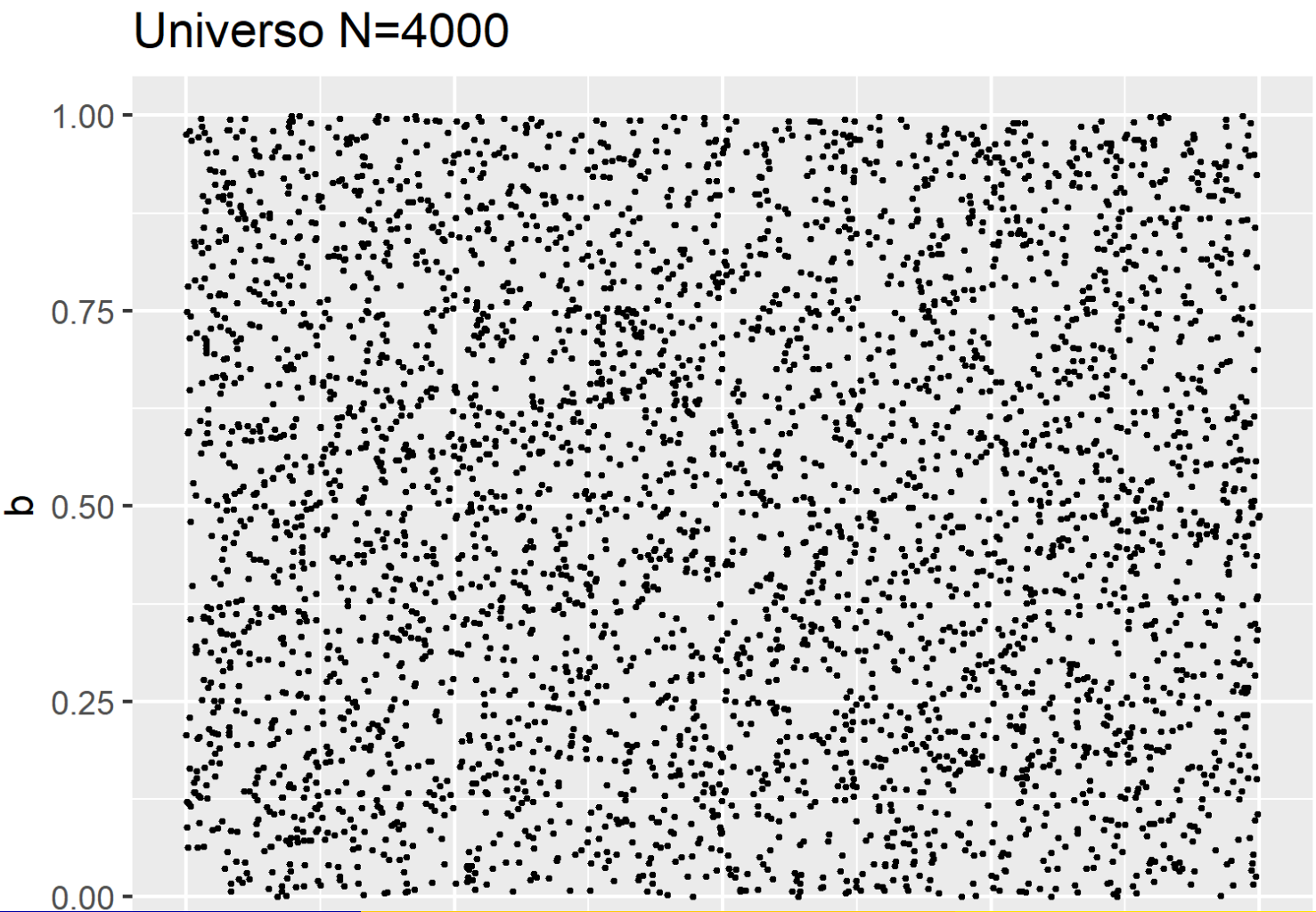
$$\hat{t}_{pwr} = \frac{1}{n} \cdot \sum_{i \in s} \frac{\hat{t}_i}{p_i}$$

es un estimador insesgado del total de Y y un estimador insesgado de $\hat{t}_{pwr}$ es

$$\hat{V}(\hat{t}_{pwr}) = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \in s} (\frac{\hat{t}_i}{p_i} - \hat{t}_{pwr})$$

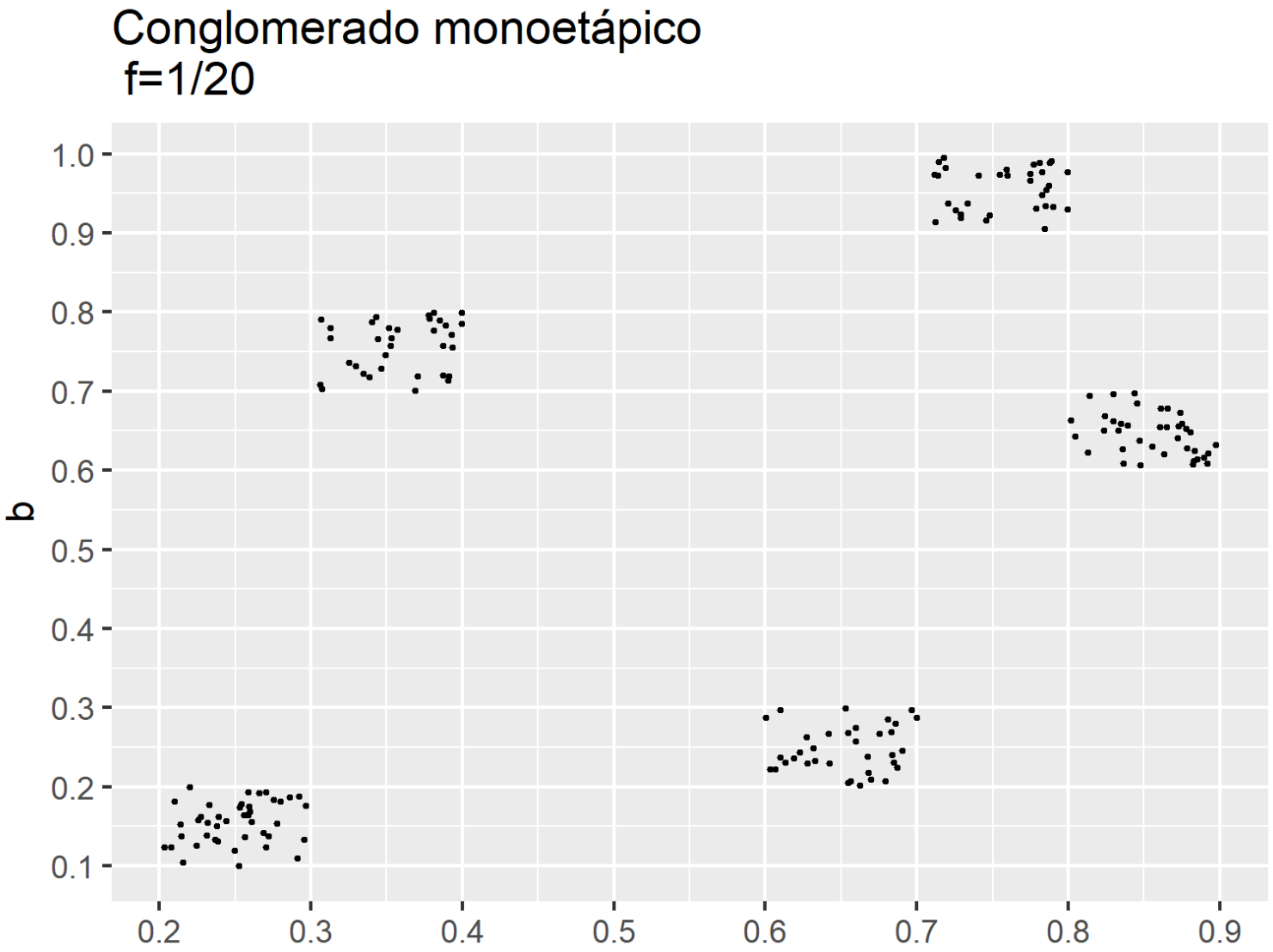
Visualización del muestreo monoetápico, polietápico, estratificado y MAS

Universo



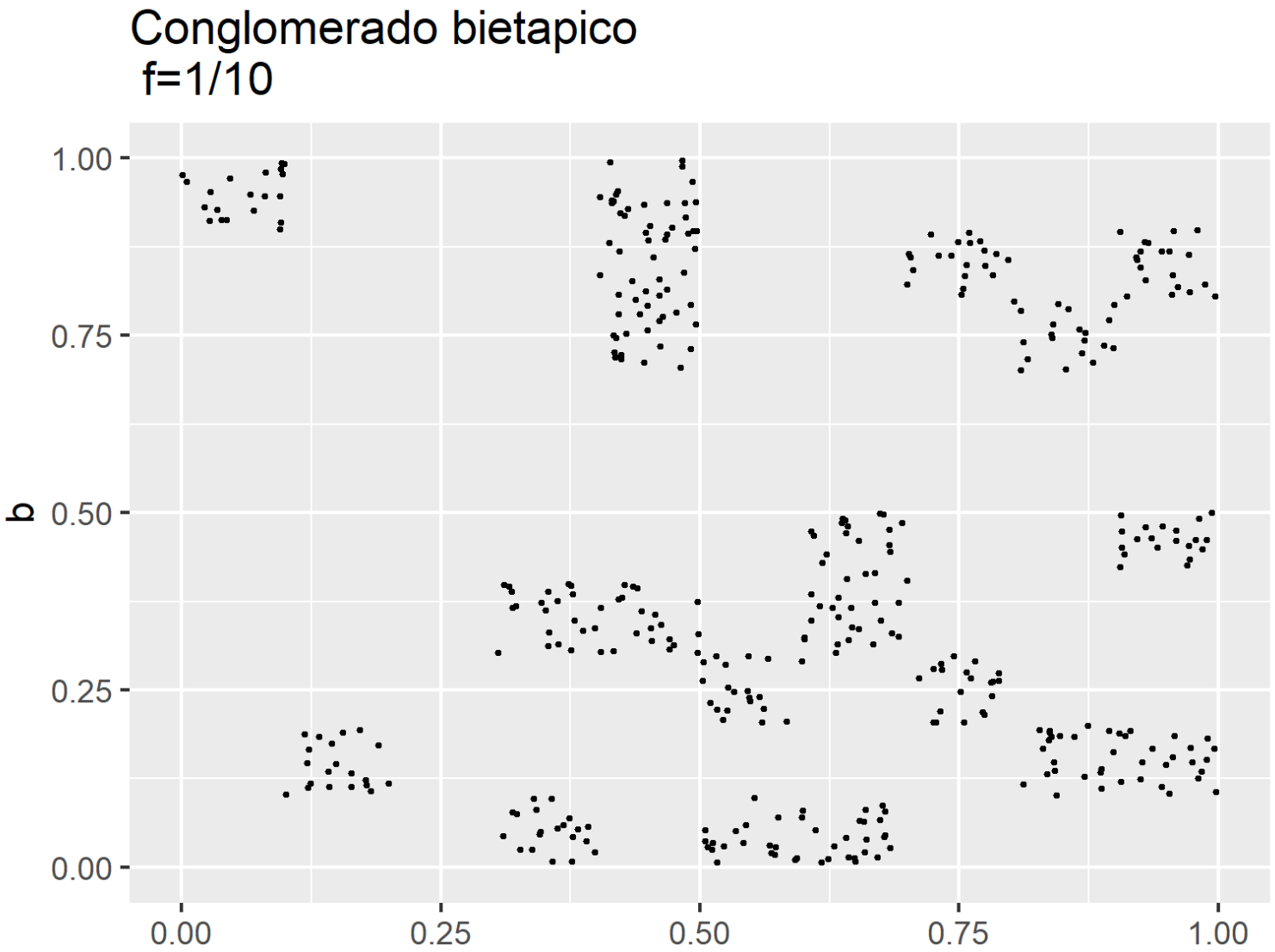
Muestreo monoetápico, polietápico, estratificado y MAS

Muestreo monoetápico - n=10



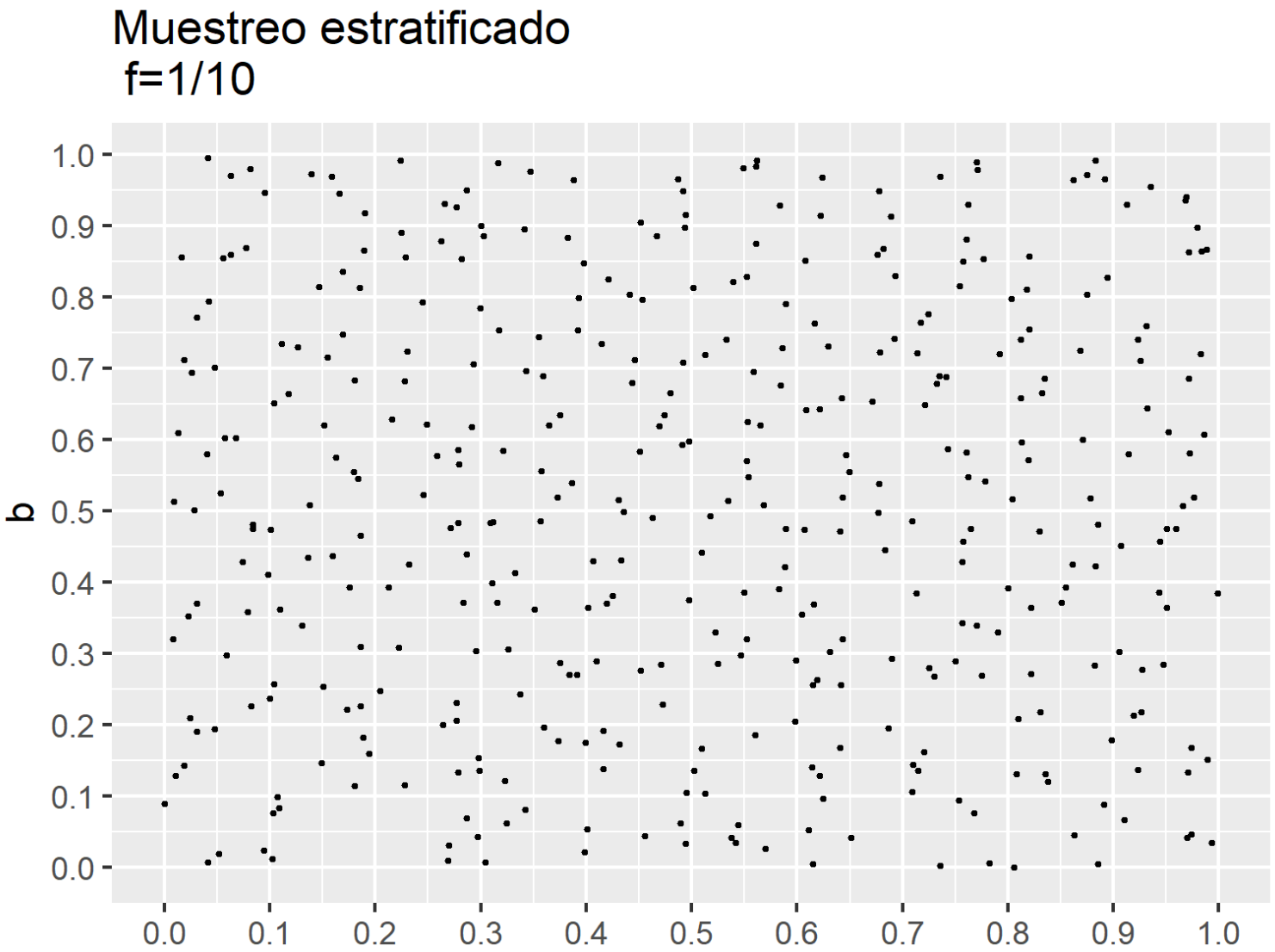
Muestreo monoetápico, polietápico, estratificado y MAS

Muestreo bietápico - n=20, m=20



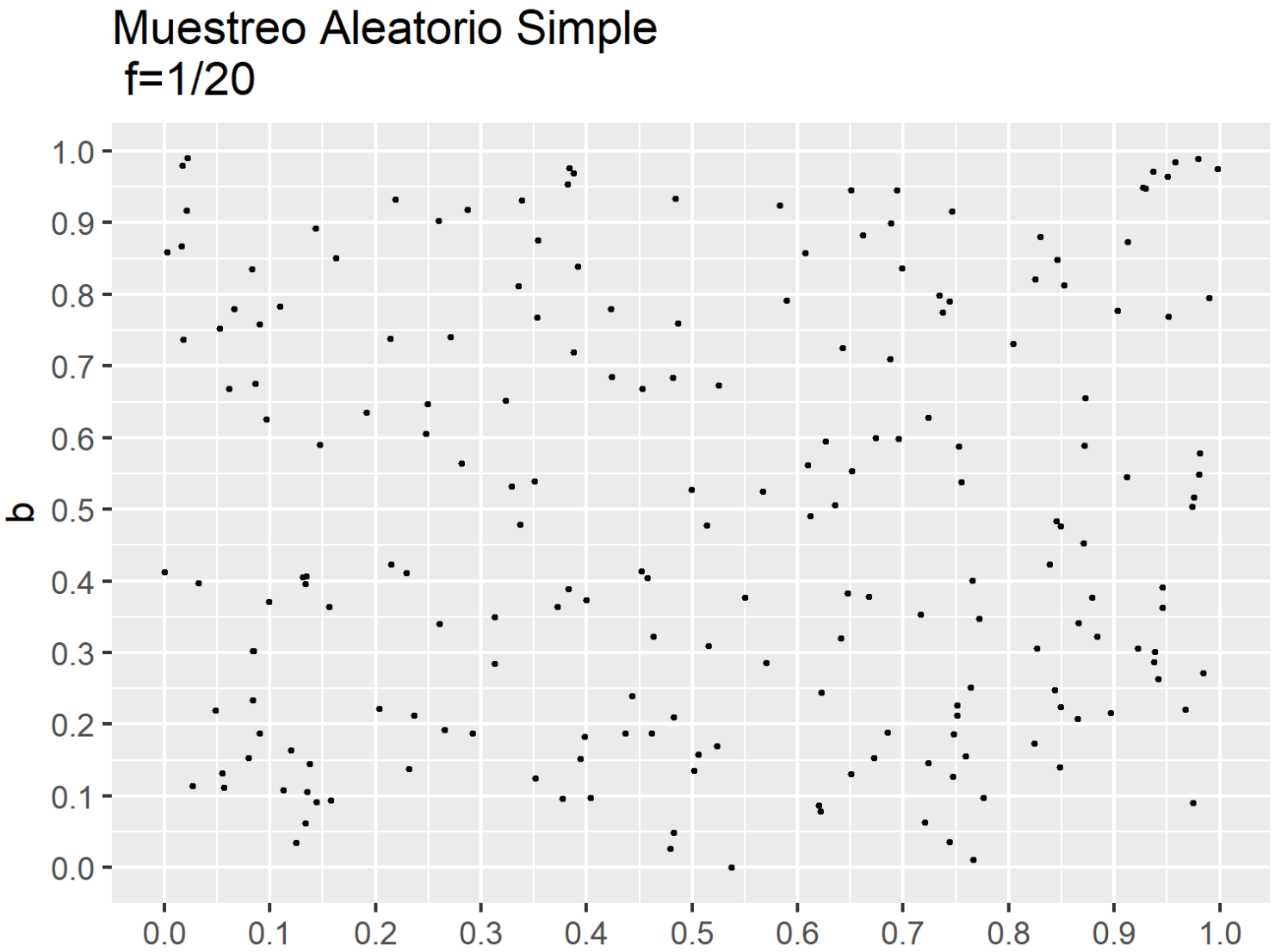
Muestreo monoetápico, polietápico, estratificado y MAS

Muestreo estratificado, H=100, nh=4



Muestreo monoetápico, polietápico, estratificado y MAS

Muestreo aleatorio simple - n=400



Muestreo polietápico

UPE autorepresentadas

- En la práctica si π_i cercano a 1 se la fuerza a autorepresentada
- Como declarar un *survey* un diseño con unidades de primera etapa autorepresentadas?
  - Las UPE autorepresentadas se declaran como estratos
  - Y las unidades de segunda etapa en esas UPE serán declaradas como unidades de primera etapa

Muestreo polietápico

UPE autorepresentadas

- En la práctica si π_i cercano a 1 se la fuerza a autorepresentada
- Como declarar un *survey* un diseño con unidades de primera etapa autorepresentadas?
  - Las UPE autorepresentadas se declaran como estratos
  - Y las unidades de segunda etapa en esas UPE serán declaradas como unidades de primera etapa

Muestreo polietápico

UPE autorepresentadas

- En la práctica si π_i cercano a 1 se la fuerza a autorepresentada
- Como declarar un *survey* un diseño con unidades de primera etapa autorepresentadas?
  - Las UPE autorepresentadas se declaran como estratos
  - Y las unidades de segunda etapa en esas UPE serán declaradas como unidades de primera etapa

Muestreo polietápico

UPE autorepresentadas

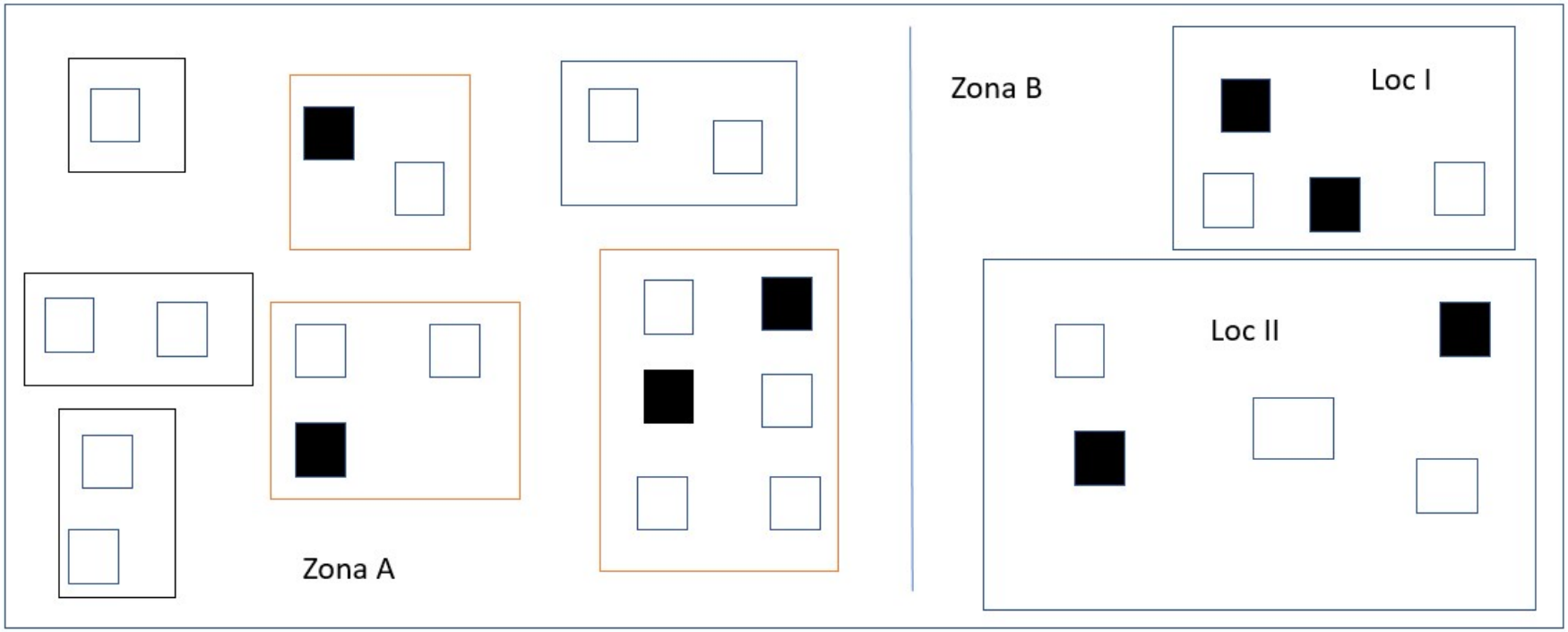
- En la práctica si π_i cercano a 1 se la fuerza a autorepresentada
- Como declarar un *survey* un diseño con unidades de primera etapa autorepresentadas?
  - Las UPE autorepresentadas se declaran como estratos
  - Y las unidades de segunda etapa en esas UPE serán declaradas como unidades de primera etapa

Muestreo polietápico

UPE autorepresentadas. Ejemplo

Universo: 9 localidades y 27 radios. En la muestra: dos localidades autorep. Del resto de localidades, tres selec. aleatoriamente. Uno o dos radios selec. por localidad. De hecho el diseño es este:

- Unidad de primera etapa: En zona A, localidades . En zona B, Radios .
- Estratos de primera etapa: Tres estratos: Zona A, Localidad I y Localidad II.



Muestreo polietápico

Estratos con solo una UPE seleccionada (o relevada)

Si en la muestra hay estratos con sólo una UPE, que no es autorepresentada, en general los soft no pueden estimar la varianza:

- Colapsar el estrato con otro similar
- Omitir el estrato para el cálculo de la varianza pero no para las estimaciones

Opciones de **survey**:
options(survey.lonely.psu="adjust")
options(survey.lonely.psu="remove")
options(survey.lonely.psu="certainty")
options(survey.lonely.psu="fail")

Muestreo polietápico

Estratos con solo una UPE seleccionada (o relevada)

Si en la muestra hay estratos con sólo una UPE, que no es autorepresentada, en general los soft no pueden estimar la varianza:

- Colapsar el estrato con otro similar
- Omitir el estrato para el cálculo de la varianza pero no para las estimaciones

Opciones de **survey**:
options(survey.lonely.psu="adjust")
options(survey.lonely.psu="remove")
options(survey.lonely.psu="certainty")
options(survey.lonely.psu="fail")

Muestreo polietápico

Ejercicio

En cierta localidad se seleccionó una MAS de seis alumnos mediante un diseño bi etápico: El colegio con mayor matrícula (colegio A) fue seleccionado con probabilidad 1. Del resto (colegios B) se seleccionaron dos colegios mediante muestreo aleatorio simple. En la segunda etapa, una muestra aleatoria simple de dos alumnos en cada colegio. La tabla siguiente presenta los resultados. Estimar con **survey** el total de Y y la varianza del estimador, dando la información de ambas etapas. Rta: (532 ; 19736)

Tipo	**Colegio**	**F1**	**Alumno**	**F2**	**Y**
A	1	1	1	3	20
A	1	1	2	3	24
B	2	4	3	5	6
B	2	4	4	5	8
B	3	4	5	6	2
B	3	4	6	6	3