Capítulo 5

Diseño
con probabilidades desiguales

1. Principio
2. Extracción sin reposición: cálculo de las probabilidades de inclusión
3. Algoritmos de extracción sin reposición
4. Aproximación y estimación de la varianza para los diseños con probabilidades desiguales

1. Principio

En determinados estudios, las unidades de la población contribuyen de manera muy desigual al total de la variable de interés.

Ejemplo: se desea estimar el total recursos gastados en servicios sociales en el conjunto de los ayuntamientos de los municipios de Catalunya.

Hay ayuntamientos de muy gran tamaño y ayuntamientos muy pequeños. Parece legítimo seleccionar de oficio los ayuntamientos muy grandes (por ejemplo, los de las ciudades de al menos 50000 ó 100000 habitantes) y seleccionar una muestra de los ayuntamientos restantes.

Seleccionar de oficio quiere decir dar a las unidades-ayuntamientos una probabilidad igual a 1 de pertenecer a la muestra.

Siguiendo esta lógica, se puede dar a los municipios una probabilidad de pertenecer a la muestra **proporcional a su tamaño (Proporcional To Size/PPS)**.

2. Enfoque general: estimador de Horvitz-Thompson

2.1 Expresión del estimador y de su varianza

Se parte de las probabilidades Π_α fijadas a priori

Π_α : Probabilidad de que la unidad α pertenezca a la muestra

El estimador de Horvitz-Thompson del total es igual a:

$$\hat{T} = \sum_{i \in s} \frac{y_i}{\Pi_i}$$

Estimador de las sumas dilatadas o *Π*-estimador

Es un estimador:

- lineal.
- sin sesgo si todos los Π_a son no nulos

Notas:

- los pesos no dependen de la muestra.
- el *Π*-estimador generaliza el muestreo aleatorio simple (*Π=n/N*) y el muestreo estratificado ($\Pi=n_h/N_h$ para las unidades del estrato *h*)

Varianza del estimador del total:

Se demuestra que la varianza del estimador de Horvitz-Thompson viene dada, cuando el tamaño de la muestra es fijo, por:

$$V\left(\hat{T}\right) = \frac{1}{2} \sum_{\substack{\alpha \neq \beta \\ \alpha = 1,\ldots,N \\ \beta = 1,\ldots,N}} \left(\Pi_\alpha \Pi_\beta - \Pi_{\alpha\beta}\right)\left(\frac{y_\alpha}{\Pi_\alpha} - \frac{y_\beta}{\Pi_\beta}\right)^2$$

Idea de partida del diseño con probilidades desiguales: se puede obtener un estimador muy preciso si:

$$\left(\frac{y_\alpha}{\Pi_\alpha} \cong \frac{y_\beta}{\Pi_\beta}\right)$$

2.2 Extracción sin reposición: Cálculo de las probabilidades de inclusión

Si existe un variable *X* muy correlacionada con *Y*, conocida sobre todas las unidades de la población, entonces se determinan los Π tales que

$$\left(\frac{X_\alpha}{\Pi_\alpha} = \frac{X_\beta}{\Pi_\beta}\right) \qquad \text{con} \qquad \sum_\alpha \Pi_\alpha = n$$

$$\Pi_\alpha = \frac{X_\alpha}{\sum_{\beta \in U} X_\beta} \cdot n$$

Nota: Π_α puede ser mayor que 1. entonces, las unidades con Π_α>1 son incluidas en la muestra con probabilidad 1; se vuelve a empezar el cálculo con las unidades que quedan (evidentemente, disminuye n y el total de X).

Quedan dos grupos, las unidades con probabilidad de inclusión igual a 1, y las unidades con probabilidad de inclusión inferior a 1 y estrictamente proporcional a X_α.

Así, el algoritmo debe seleccionar *n* individuos, con probabilidades de inclusión de orden 1 fijadas y tales que:

$$0 < \pi_\alpha \leq 1 \qquad \text{para todo } \alpha \in U \text{, tal que } \sum_{\alpha \in U} \pi_\alpha = n$$

Problema 1 Extracción sistemática

Se considera una población U compuesta de 6 hogares de respectivas dimensiones, 2, 4, 3, 9, 1 y 2. El tamaño del hogar, variable auxiliar *X* es el número de personas que le integran. Se selecciona 3 hogares sin reposición con probabilidad proporcional al tamaño.

Calculen las probabilidades de inclusión de primer orden

Nota: aquí no se indica la variable de interés.

Expresión del estimador, de su varianza y de la estimador de la varianza del estimador

$$\hat{T} = \sum_{i \in s} \frac{y_i}{\Pi_i}$$

La varianza del estimador de Horvitz-Thompson viene dada, cuando el tamaño de la muestra es fijo, por:

$$V\left(\hat{T}\right) = \frac{1}{2} \sum_{\substack{\alpha \neq \beta \\ \alpha=1,\ldots,N \\ \beta=1,\ldots,N}} \left(\Pi_\alpha \Pi_\beta - \Pi_{\alpha\beta}\right)\left(\frac{Y_\alpha}{\Pi_\alpha} - \frac{Y_\beta}{\Pi_\beta}\right)^2$$

Los $\pi_{a,b}$ dependen del método de extracción (del algoritmo empleado).

Estimación de la varianza del estimador del total:

$$\hat{V}\left(\hat{T}\right) = \frac{1}{2} \sum_{\substack{i \neq j \\ i \in s \\ j \in s}} \frac{\left(\Pi_i \Pi_j - \Pi_{ij}\right)}{\Pi_{ij}}\left(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j}\right)^2$$

Si todos los $\pi_{a,b}$ son no nulos, $\hat{V}\left(\hat{T}\right)$ estima $V\left(\hat{T}\right)$ sin sesgo

Estimador de Horvitz-Thompson de la media :

$$\hat{\bar{Y}} = \frac{\hat{T}}{N}$$

$$V\left(\hat{\bar{Y}}\right) = \frac{V\left(\hat{T}\right)}{N^2}$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{\hat{V}\left(\hat{T}\right)}{N^2}$$

Problema 1 Extracción sistemática

Se considera una población U compuesta de 6 hogares de respectivas dimensiones, 2, 4, 3, 9, 1 y 2. El tamaño del hogar, variable auxiliar *X* es el número de personas que le integran. Se selecciona 3 hogares sin reposición con probabilidad proporcional al tamaño.

1. Den las probabilidades de inclusión de primer orden
2. Supongan que el algoritmo escogido ha llevado a seleccionar los hogares 2 y 3, además del hogar que siempre pertenecerá a la muestra por tener una probabilidad de inclusión igual a 1; den una estimación del tamaño medio de los hogares de la población ¿Os sorprende el resultado?

El problema de los elefantes de Basu (1971)

The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weight ? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd.

Therefore, the owner plans to weigh Sambo and take 50*y* (where *y* is the present weight of Sambo) as an estimate of the total weight T = Y1 + Y2 + . . . + Y50 of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive samplings plan. "*How can you get an unbiased estimate of Y this way?"* protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate T?", asks the statistician. "Why ? The estimate ought to be 50y of course," says the owner.

Oh! No! That cannot possibly be right," says the statistician, "*I recently read an article in the Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "*What is the Horvitz-Thompson estimate in this case?"* asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of T is 100y/99 and not 50y." "And, how would you have estimated T," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?"

"According what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of T would then have been 4900y, where y is Jumbo's weight."

That is how the statistician lost his circus job (…and perhaps became teacher of statistics!).

3. Algoritmos de extracción

3.1 Generalidades

Se parte de los $\Pi_\alpha, \alpha \in \mathbf{U}$

El cálculo de los Π_α es una etapa que precede a la aplicación de todos los algoritmos.

Queda efectuar la selección mediante un diseño muestral. El problema se puede formular así: se busca un diseño muestral *p*(.) tal que:

$$\sum_{\alpha \in s} p(s) = \pi_\alpha$$

Existe una infinidad de soluciones

3.2 Propiedades de los algoritmos

Un buen algoritmo debe respetar los siguientes 4 criterios:

- exactitud: las probabilidades de inclusión de orden 1 son exactamente las indicadas
- tamaño fijo: el tamaño de la muestra, un vez aplicado el algoritmo, es de tamaño *n* fijado
- generalidad: el algoritmo se puede aplicar a todo juego de probabilidades de inclusión
- sin reposición: una unidad sólo se puede seleccionar una vez

Dada la infinidad de soluciones al problema planteado, se pueden escoger algoritmos que proporcionan probabilidades de orden 2 interesantes como las siguientes:

• las probabilidades de inclusión de orden 2 deben ser estrictamente positivas

• se deben verificar las condiciones de Sen-Yates-Grundy ($\pi_\alpha . \pi_\beta - \pi_{\alpha\beta} \geq 0$)

•El algoritmo debería proporcionar probabilidades de inclusión de orden 2 tales que la varianza del estimador del total sea siempre inferior al diseño con reposición

Además, el algoritmo debe ser fácil de aplicar:

• el algoritmo debe ser rápido y proceder a la extracción sin calcular los $p(s)$ para todos las muestras de tamaño n.

• el algoritmo debe ser secuencial y proceder a la extracción en una sola pasada.

Finalmente:

• sería interesante que los $p(s)$ así como las probabilidades de inclusión de orden 2 no dependan del orden de las unidades en el fichero de datos

Ningún método posee todas las propiedades deseables…

3.3 Extracción sistemática

Se pueden seleccionar las unidades a partir de algoritmos sistemáticos.
Método con tamaño fino, Madow (1949)

Algoritmo presentado mediante un ejemplo *n=4*

Ident. α	Π_α	v_α acumulados	Marca de selección
A1	0.1	0.1	
A2	0.4	0.5	▪
A3	0.2	0.7	
A4	0.3	1.0	
A5	0.6	1.6	▪
A6	0.1	1.7	
A7	0.9	2.6	▪
A8	0.8	3.4	▪
A9	0.4	3.8	
A10	0.2	4.0	

1. Al azar, se extrae un número en [0,1].
   Por ejemplo, se obtiene 0.222:
   Primera unidad selecionada: a_1 tal que

$$v_{\alpha_1 - 1} \leq u < v_{\alpha_1}$$

2. saltar de 1 en 1:
1.222
2.222
3.222

$$v_{\alpha_j - 1} \leq u + j - 1 < v_{\alpha_j}$$

Inconveniente: muchas probabilidades de orden 2 nulas

Solución parcial: sortear AL AZAR el fichero antes de extraer la muestra, pero quedan probabiliades de orden 2 nulas…y es muy difícil calcularlas

Marco muestral **ordenado según los P_a**
En el ejemplo anterior:

α	Π_α	v_α acumulados	Marca de selección
A1	0.1	0.1	
A6	0.1	0.2	
A3	0.2	0.4	▪
A10	0.2	0.6	
A4	0.3	0.9	
A2	0.4	1.3	▪
A9	0.4	1.7	
A5	0.6	2.3	▪
A8	0.8	3.1	
A7	0.9	4.0	▪

1. *al azar, se extrae un número entre 0 y 1 por ejemplo, se obtiene 0.222*

2. saltar de 1 en 1:
1.222
2.222
3.222

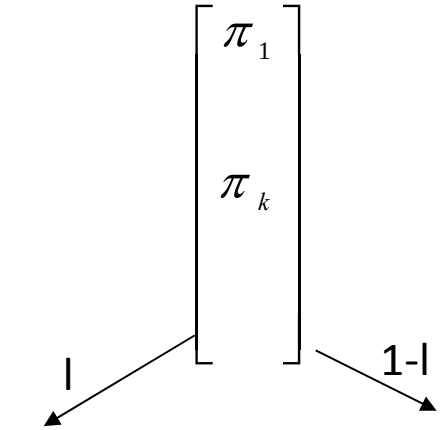
Inconveniente: todavía más probabilidades de orden 2 nulas

Pero: varianza del estimador (mucho) más pequeña, aunque muy difícil de calcular

3.4 Algoritmos de escisión

Principio

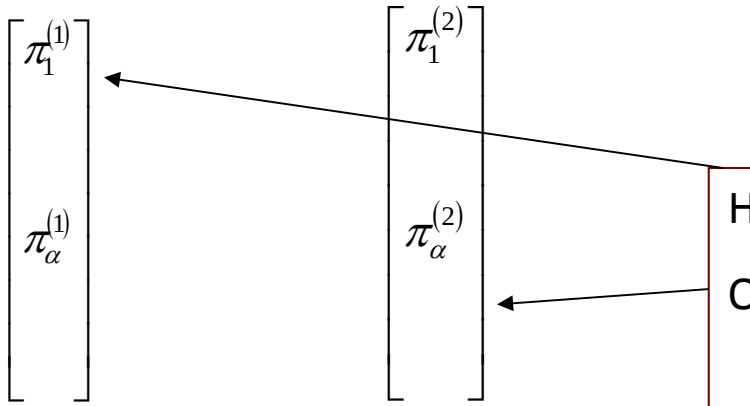
- Se ha escogido un juego de probabilidades π_α
- *Se escoge λ, 0<λ<1,* y dos juegos de probabilidades $\pi_\alpha^{(1)}$ y $\pi_\alpha^{(2)}$ de forma astuta": dividir el problema complejo en problemas menos complejos



$$\pi_\alpha = \lambda\pi_\alpha^{(1)} + (1-\lambda)\pi_\alpha^{(2)}$$

$$0 \le \pi_\alpha^{(1)} \le 1 \qquad 0 \le \pi_\alpha^{(2)} \le 1$$

$$\sum_{k\in U} \pi_\alpha^{(1)} = \sum_{k\in U} \pi_\alpha^{(2)} = n$$



Hay varios "0" o "1"

O bien π_α uniformes

Diseño con soporte mínimo

$\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(\alpha)} \ldots, \pi_{(N)}$ ordenados de forma creciente

$$\lambda = \min\left\{1 - \pi_{(N-n)}, \pi_{(N-n+1)}\right\}$$

$$\pi_{(\alpha)}^{(1)} = \begin{cases} 0 \; si \; \alpha \leq (N-n) \\ 1 \; si \; \alpha > (N-n) \end{cases}$$

Como mucho *N* etapas

$$\pi_{(\alpha)}^{(2)} = \begin{cases} \dfrac{\pi_{(\alpha)}}{1-\lambda} \; si \; \alpha \leq (N-n) \\ \dfrac{\pi_{(\alpha)} - \lambda}{1-\lambda} \; si \; \alpha > (N-n) \end{cases}$$

Escisión en diseños simples

$$\lambda = \min\left\{\pi_{(1)}\frac{N}{n}, \left(1-\pi_{(N)}\right)\frac{N}{N-n}\right\}$$

$$\pi_{(\alpha)}^{(1)} = \frac{n}{N}$$

$$\pi_{(\alpha)}^{(2)} = \frac{\pi_\alpha - \lambda\frac{n}{N}}{1-\lambda}$$

$$\pi_k = \lambda\frac{n}{N} + \left(1-\lambda\right)\frac{\pi_k - \lambda\frac{n}{N}}{\left(1-\lambda\right)}$$

En cada etapas, se elige entre un diseño simple y un diseño de probabilidades desiguales

Si se elige el diseño con prob. desiguales, entonces se tiene que escoger (*n*-1) unidades entre (*N*-1)

Como mucho (*N*-1) etapas

Condiciones de Yates-Grundy no satisfechas

$$si\ \lambda = \pi_{(1)}\frac{N}{n}\ entonces\ \pi_{(1)}^{(2)} = 0$$

$$si\ \lambda = \left(1-\pi_{(N)}\right)\frac{N}{N-n}\ entonces\ \pi_{(N)}^{(2)} = 1$$

4. Aproximación de la varianza

Aproximación de la varianza

$$V\left(\hat{T}\right)=\sum_{\alpha\in U}\frac{b_\alpha}{\pi_\alpha^2}\left(y_\alpha-y_\alpha^*\right)^2$$

$$y_\alpha^*=\pi_\alpha\frac{\sum_{\beta\in U}b_\beta y_\beta/\pi_\beta}{\sum_{\beta\in U}b_\beta}$$

$$b_\alpha=\frac{N\pi_\alpha\left(1-\pi_\alpha\right)}{N-1}$$

Estimación de la aproximación de la varianza

$$\hat{V}\left(\hat{T}\right)=\sum_{i\in s}\frac{c_i}{\pi_i^2}\left(y_i-\hat{y}_i^*\right)^2$$

$$\hat{y}_i^*=\pi_i\frac{\sum_{l\in s}c_l y_l/\pi_l}{\sum_{l\in s}b_l}$$

$$c_i=\frac{n\pi_i\left(1-\pi_i\right)}{n-1}$$

Problema 2 Cooperativas de consumo.

Se desea conocer el número medio de cooperativas de consumo en activo en una determinada comarca de 10 municipios. Se piensa que dicho número tiene relación estrecha con el número de habitantes, conocido gracias a un censo reciente.

Municipio	Nº habitantes
A	12000
B	15000
C	3000
D	40000
E	60000
F	10000
G	10000
H	25000
I	30000
J	120000

Dado dicho conocimiento, escoger un método de selección de la muestra pertinente. Sabiendo que desea una muestra de tamaño igual a 3, y que el número escogido entre 0 y 1, que se debe utilizar para arrancar la selección vale 0.7, decir cuál es la muestra seleccionada.

De cara a la práctica:

$$V\left(\hat{T}\right) = \frac{1}{2} \sum_{\substack{\alpha \neq \beta \\ \alpha=1,\ldots,N \\ \beta=1,\ldots,N}} \left(\Pi_\alpha \Pi_\beta - \Pi_{\alpha\beta}\right)\left(\frac{Y_\alpha}{\Pi_\alpha} - \frac{Y_\beta}{\Pi_\beta}\right)^2 = \sum_{\substack{\alpha=1,\ldots,N \\ \beta=1,\ldots,N}} \left(\frac{Y_\alpha}{\Pi_\alpha}\frac{Y_\beta}{\Pi_\beta}\right)\left(\Pi_{\alpha\beta} - \Pi_\alpha \cdot \Pi_\beta\right)$$

$$= \sum_{\substack{\alpha=1,\ldots,N \\ \beta=1,\ldots,N}} \left(\frac{Y_\alpha}{\Pi_\alpha}\frac{Y_\beta}{\Pi_\beta}\right)\Delta_{\alpha\beta}$$

$$\text{avec} \qquad \Delta_{\alpha\beta} = \left(\Pi_{\alpha\beta} - \Pi_\alpha \cdot \Pi_\beta\right)$$

Cálculo de la varianza exacta del estimador

$$V\left(\hat{T}\right)=V\left(\sum_{\alpha\in s}\frac{Y_\alpha}{\pi_\alpha}\right)$$

$$u \quad \text{vecteur} \quad \text{colonne} \quad \begin{pmatrix} \frac{Y_\alpha}{\pi_\alpha} \\ \\ \end{pmatrix} \qquad \text{vec teur} \quad \text{colonne} \quad 1=\begin{pmatrix} I_\alpha \\ \\ \end{pmatrix}$$

$$V\left(\hat{T}\right)=V\left(\sum_{\alpha\in s}\frac{Y_\alpha}{\pi_\alpha}\right)=V(u'1) \qquad V(u'1)=u'V(1)u=u'\Delta u$$

Estimación de la varianza del estimador

$$\hat{V}\left(\hat{T}(Y)\right)=\frac{1}{2}\sum_{\substack{i\neq j\\ i\in s\\ j\in s}}\frac{\left(\Pi_i\Pi_j-\Pi_{ij}\right)}{\Pi_{ij}}\left(\frac{y_i}{\Pi_i}-\frac{y_j}{\Pi_j}\right)$$