QÜESTIIÓ, Vol 20, 1 pp. 45-70, 1996

DISEÑOS MUESTRALES π–EQUIVALENTES Y EQUIVALENTES DE PRIMER ORDEN

FERNÁNDEZ GARCÍA, FRANCISCO R.

MAYOR GALLEGO, JOSÉ A.

Universidad de Sevilla*

El comportamiento de un diseño muestral en relación al estimador de Horvitz-Thompson depende exclusivamente de sus probabilidades de inclusión de primer y segundo orden. Al ser, usualmente, mucho mayor el número de muestras de un diseño que el número de dichas probabilidades de inclusión, fijadas éstas, existen una gran cantidad de diseños que las satisfacen y que por ello proporcionan similares resultados en relación al mencionado estimador, siendo posible escoger entre los mismos aquellos que mejoren ciertos criterios adicionales. En este trabajo, relacionamos los diseños con las mismas probabilidades de inclusión con un poliedro convexo, e indicamos la forma de obtener diseños óptimos. Además, definimos los diseños muestrales equivalentes de primer orden como aquellos con las mismas probabilidades de inclusión de primer orden, lo que permite obtener diseños muestrales óptimos en una clase más amplia, eludiendo el problema de la determinación de las probabilidades de inclusión de segundo orden.

π–Equivalent and First Order Equivalent Sampling Designs.

AMS clasification: 62D05

Key words: muestreo, poblaciones finitas, estimador de Horvitz-Thompson, programación lineal.

* Departamento de Estadística e Investigación Operativa. Universidad de Sevilla. Tarfia. 41012 Sevilla.

- Article rebut el setembre de 1994.

- Acceptat l'octubre de 1995.

1. INTRODUCCIÓN

Los dos problemas centrales que estudia la teoría del muestreo en poblaciones finitas son,

- La elección del diseño muestral usado en la selección de la muestra.
- La elección del estimador más adecuado para estimar un parámetro poblacional.

Estos problemas están muy relacionados, de forma que carece de sentido estudiarlos independientemente. En efecto, la bondad de un estimador depende fuertemente del diseño empleado en la obtención de la muestra. Pensemos, por ejemplo, en la estimación de una razón de medias poblacionales mediante la razón de medias muestrales. Dicha estimación no es insesgada para un diseño muestral aleatorio simple, siéndolo en cambio para el diseño de Midzuno (1952). Véase Sukhatme *et al.* (1984).

Para establecer la notación, suponemos que en la población U, con N elementos que representamos por,

$$U = \{1, 2, \ldots, N\}$$

estamos interesados en estudiar un parámetro poblacional lineal del tipo,

$$\theta(Y) = \sum_{i \in U} a_i Y_i$$

siendo $Y_1, Y_2, \ldots, Y_N$ una variable de estudio cuantitativa definida sobre los elementos de la población.

Dado un diseño muestral, $d = (M, p(\cdot))$, es decir, un conjunto de muestras y una distribución de probabilidad sobre las mismas, y una muestra, m, perteneciente a dicho diseño, empleamos el estimador de Horvitz-Thompson (1952),

$$\widehat{\theta}(m) = \sum_{i \in m} a_i \frac{Y_i}{\pi_i}$$

Dicha estimación es insesgada y su varianza viene dada por la siguiente expresión,

$$V[\widehat{\theta}(m)] = \sum_{i,j \in U} \sum \Delta_{ij} a_i a_j \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} \qquad \text{donde} \quad \Delta_{ij} = \pi_{ij} - \pi_i \pi_j$$

y esta varianza puede estimarse mediante el siguiente estimador insesgado,

$$\widehat{V}[\widehat{\theta}(m)] = \sum_{i,j\in m}\sum \frac{\Delta_{ij}}{\pi_{ij}} a_i a_j \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}$$

supuesto que $\pi_{ij} > 0,\ \forall i \neq j$.

Como puede observarse, el estimador y el error dependen de los valores fijos para la población, $Y_1, \ldots, Y_N$, y también de las probabilidades de inclusión de primer y segundo orden,

$$\pi_i \quad = P[\ i \in m\ ] = \sum_{\substack{m\in M \\ m\ni i}} p(m) \qquad i \in U$$

$$\pi_{ij} \quad = P[i, j \in m] = \sum_{\substack{m\in M \\ m\ni i,j}} p(m) \qquad i \neq j \in U$$

que a su vez están determinadas por el diseño muestral utilizado en la selección de la muestra. Véase Fernández y Mayor (1994).

Podemos afirmar pues que, en relación al estimador de Horvitz-Thompson, dos diseños muestrales diferentes producen el mismo error de muestreo si sus probabilidades de inclusión de primer y segundo coinciden, lo que nos permite buscar diseños muestrales con similar comportamiento frente a dicho estimador y mejorando otros criterios adicionales.

Uno de los primeros trabajos sobre este tema se debe a Goodman y Kish (1950). Estos autores consideran la existencia de muestras **no preferidas**, a las que el diseño muestral debería asignarle probabilidades pequeñas. Avadhani y Sukhatme (1973) definen el concepto de diseño controlado como aquel en el cual las probabilidades de las muestras no preferidas no exceden de cierta cantidad predeterminada, indicando la forma de obtener tales diseños.

En otra línea, Wynn (1977), Foody y Hedayat (1976) y Hedayat (1979) estudian la estructura de convexidad del conjunto de los diseños muestrales con las mismas probabilidades de inclusión de primer y segundo orden, centrándose en la búsqueda de diseños de tamaño de soporte mínimo, es decir, con el menor número de muestras posibles.

Bellhouse (1984), realiza el estudio de diseños óptimos bajo algunos modelos de superpoblación y Rao y Nigam (1990, 1992) estudian procedimientos de obtención de diseños controlados con las mismas probabilidades de inclusión que el diseño muestral aleatorio simple y otros diseños de tipo ΠPS y PPS.

Como veremos, la búsqueda de diseños que mejoren ciertas propiedades suele dar lugar a la resolución de problemas de programación matemática por lo que es aplicable en la selección de unidades primarias en un muestreo multietápico donde puede ser de interés la selección de conglomerados con determinados criterios pero manteniendo el diseño muestral.

En la siguiente sección, estudiaremos los diseños muestrales $\pi-$equivalentes centrándonos en el aspecto práctico de la búsqueda de diseños de este tipo que mejoren ciertos criterios.

En la sección tercera estudiamos algunas familias especiales de diseños $\pi-$equivalentes, tanto en tamaño de muestra fijo como variable.

En la cuarta sección definimos los diseños equivalentes de primer orden e indicamos el procedimiento para buscar entre ellos los mejores diseños. En este sentido, introducimos en el conjunto de las muestras de un diseño un orden total especial que nos permite afirmar cuando un muestra es más informativa que otra. A partir de este orden, damos un criterio de mejora entre los diseños equivalentes de primer orden a uno dado.

2. DISEÑOS MUESTRALES $\pi-$EQUIVALENTES

Denotaremos por Π a la matriz del diseño, es decir, la matriz cuadrada de orden $N \times N$ cuyas componentes son la probabilidades de inclusión de segundo orden (con el convenio $\pi_{ii} = \pi_i$). Siguiendo la línea de los autores mencionados anteriormente, damos la siguiente definición.

Definición 1

Dados dos diseños muestrales, $d_1 = (M_1, p_1(\cdot))$ y $d_2 = (M_2, p_2(\cdot))$, con matrices de diseño respectivas $\Pi^{(1)}$ y $\Pi^{(2)}$, diremos que son diseños $\pi-$equivalentes si se cumple $\Pi^{(1)} = \Pi^{(2)}$.

Es conocido, Hedayat y Sinha (1991), que dos diseños $\pi-$equivalentes coinciden en la esperanza y la varianza del tamaño muestral, y en caso de ser ambos de tamaños fijos, dichos tamaños son iguales.

Dado un diseño muestral, $d = (M, p(\cdot))$ sobre U, podemos considerarlo como un punto del espacio producto I^q, siendo $q = 2^N$ é $I = [0,1]$, al enumerar todas las posibles muestras del diseño, desde 1 a 2^N, y asignarle a cada coordenada del punto la probabilidad correspondiente.

Suponemos que el conjunto de todas las muestras, en algún determinado orden, es,

$$M = \{m_1, m_2, \ldots, \ldots, m_q\}$$

y denotemos por $x = (x_1, x_2, \ldots, x_q)^t$ al punto mencionado, es decir, la distribución de probabilidad correspondiente, siendo $x_k = p(m_k)$, $k = 1, \ldots, q$. Si consideramos la función indicadora,

$$I_{ij}(k) = \begin{cases} 1 & i, j \in m_k \\ 0 & i, j \notin m_k \end{cases} \qquad \forall\, i, j \in U,\ k = 1, \ldots, q$$

se deberá verificar,

$$\begin{aligned} \sum_{k=1}^{q} x_k I_{ij}(k) &= \pi_{ij} \qquad \forall i \leq j \\ \sum_{k=1}^{q} x_k &= 1 \\ x_k &\geq 0 \qquad k = 1, \ldots, q \end{aligned}$$

Si denotamos,

$$A = \begin{pmatrix} I_{11}(1) & I_{11}(2) & \ldots & I_{11}(q) \\ I_{22}(1) & I_{22}(2) & \ldots & I_{22}(q) \\ \ldots & \ldots & \ldots & \ldots \\ I_{NN}(1) & I_{NN}(2) & \ldots & I_{NN}(q) \\ I_{12}(1) & I_{12}(2) & \ldots & I_{12}(q) \\ I_{13}(1) & I_{13}(2) & \ldots & I_{13}(q) \\ \ldots & \ldots & \ldots & \ldots \\ I_{N-1N}(1) & I_{N-1N}(2) & \ldots & I_{N-1N}(q) \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \ldots \\ \pi_N \\ \pi_{12} \\ \pi_{13} \\ \ldots \\ \pi_{N-1N} \\ 1 \end{pmatrix}$$

entonces las anteriores restricciones se pueden expresar como,

$$Ax = b \qquad x \geq 0$$

lo que representa un poliedro convexo, determinando todo punto del mismo un diseño con las mismas probabilidades de inclusión, π_{ij}. El número de restricciones que determinan este poliedro es,

$$r = N + \binom{N}{2} + 1 = \binom{N+1}{2} + 1$$

pero estas restricciones no son independientes si se desea controlar el tamaño muestral medio y su variabilidad ya que se verifican las relaciones,

$$\begin{aligned}
\sum_{i\in U} \pi_i &= \sum_{m\in M} n(m)p(m) = E[n(m)] \\
\sum_{\substack{i\in U \\ i\neq j}} \pi_{ij} &= \sum_{\substack{m\in M \\ m\ni j}} (n(m)-1)p(m) \qquad \forall j \in U \\
\sum\sum_{\substack{i,j\in U \\ i\neq j}} \pi_{ij} &= \sum_{m\in M} n(m)(n(m)-1)p(m)
\end{aligned}$$

En el caso de que consideremos diseños con muestras de tamaño fijo $n(m) = n$, las N restricciones correspondientes a las probabilidades de inclusión de primer orden se deducen a partir de las correspondientes a las probabilidades de inclusión de segundo orden pues en este caso,

$$\sum_{\substack{i\in U \\ i\neq j}} \pi_{ij} = \sum_{\substack{m\in M \\ m\ni j}} (n-1)p(m) = (n-1)\pi_j \qquad \forall j \in U$$

y por ello, al ser redundantes, pueden ser eliminadas quedando el poliedro determinado por $r = \binom{N}{2}$ restricciones siendo $q = \binom{N}{n}$ el número de muestras posibles.

Es interesante observar que si consideramos el diseño muestral aleatorio MAS(N,n) formado por todas las muestras posibles de tamaño fijo n con probabilidad uniforme existen numerosos diseños π–equivalentes con un número de muestras mucho menor, por ejemplo, cualquier vértice del poliedro.

Así pues, el poliedro de los diseños π–equivalentes a uno dado viene definido por una matriz, A, de dimensión $r \times q$, siendo sus elementos ceros ó unos según se valore la función indicadora en cada muestra. Observemos que definidas las N primeras filas correspondientes a los valores $\pi_{ii} = \pi_i$, $i \in U$, las restantes filas se obtienen automáticamente pues los elementos de la fila asociada a π_{ij} se calculan multiplicando, elemento a elemento, los de las filas correspondientes a π_i y π_j, por lo cual, la matriz A tiene entre sus filas una dependencia de "tipo producto".

Definido así el conjunto de los diseños π–equivalentes a uno dado, es lógico que la búsqueda de uno de ellos mejorando algún criterio adicional nos conduzca a un problema de programación matemática, en el cual las restricciones son lineales, por ello al diseño así obtenido le llamaremos **óptimo** en relación al criterio utilizado.

Ejemplo 1

Si deseamos encontrar un diseño $\pi-$equivalente a un MAS(8,3) , como indicábamos anteriormente, debemos utilizar las 56 muestras posibles del diseño original. El conjunto convexo que define todos los diseños $\pi-$equivalentes será,

$$\sum_{k=1}^{56} x_k I_{ij}(k) = \pi_{ij},\ i \leq j \ \text{ con } \pi_i = 3/8 \ \text{ y } \pi_{ij} = 3/28 \quad x_k > 0,\ k = 1,\ldots,56$$

Supongamos que nuestro problema se considera en un contexto real, siendo la población U las ocho provincias de Andalucía, de las que se quiere una muestra de tres provincias, y deseamos dar preponderancia a las muestras que tienen dos provincias marítimas y una del interior. Para ello definimos la función objetivo,

$$C = \sum_{k=1}^{56} c_k x_k$$

donde $c_k = 1$ si la muestra $k-$ésima no verifica la condición deseada, siendo cero en caso contrario.

De esta forma, la búsqueda del diseño óptimo se reduce a un problema de programación lineal. En caso de obtenerse un valor mínimo nulo habríamos obtenido un diseño $\pi-$equivalente al MAS(8,3), verificando las condiciones deseadas. Si el valor mínimo no fuera nulo, el diseño resultante sería el que más se aproxima a los deseos manifestados, aunque habría muestras no deseadas.

Para nuestro problema concreto, si usamos la codificación que aparece en la siguiente tabla,

ALMERÍA	GRANADA	MÁLAGA	CÁDIZ	HUELVA	SEVILLA	CÓRDOBA	JAÉN
1	2	3	4	5	6	7	8

y denotamos por X_{ijk} la probabilidad con la que la muestra $\{i,j,k\}$ aparece en el diseño, tendremos la siguiente función objetivo,

```
C = X123 + X124 + X125 + X134 + X135 + X145 + X167 + X168 + X178 + X234 + X235 + X245 + X267
  + X268 + X278 + X345 + X367 + X368 + X378 + X467 + X468 + X478 + X567 + X568 + X578 + X678
```

con cuya minimización pretendemos que las muestras no deseables tengan probabilidad nula, desapareciendo del diseño.

Las restricciones serán,

```
X123 + X124 + X125 + X126 + X127 + X128 + X134 + X135 + X136 + X137 + X138 +   X145 + X146 + X147 + X148
     + X156 + X157 + X158 + X167 + X168 + X178 = 3/8

X123 + X124 + X125 + X126 + X127 + X128 + X234 + X235 + X236 + X237 + X238 +   X245 + X246 + X247 + X248
     + X256 + X257 + X258 + X267 + X268 + X278 = 3/8

X123 + X134 + X135 + X136 + X137 + X138 + X234 + X235 + X236 + X237 + X238 + X345 + X346 + X347 + X348
     + X356 + X357 + X358 + X367 + X368 + X378 = 3/8

X124 + X134 + X145 + X146 + X147 + X148 + X234 + X245 + X246 + X247 + X248 + X345 + X346 + X347 + X348
     + X456 + X457 + X458 + X467 + X468 + X478 = 3/8

X125 + X135 + X145 + X156 + X157 + X158 + X235 + X245 + X256 + X257 + X258 + X345 + X356 + X357 + X358
     + X456 + X457 + X458 + X567 + X568 + X578 = 3/8

X126 + X136 + X146 + X156 + X167 + X168 + X236 + X246 + X256 + X267 + X268 + X346 + X356 + X367 + X368
     + X456 + X467 + X468 + X567 + X568 + X678 = 3/8

X127 + X137 + X147 + X157 + X167 + X178 + X237 + X247 + X257 + X267 + X278 + X347 + X357 + X367 + X378
     + X457 + X467 + X478 + X567 + X578 + X678 = 3/8

X128 + X138 + X148 + X158 + X168 + X178 + X238 + X248 + X258 + X268 + X278 + X348 + X358 + X368 + X378
     + X458 + X468 + X478 + X568 + X578 + X678 = 3/8

X123 + X124 + X125 + X126 + X127 + X128 =  3/28        X123 + X134 + X135 + X136 + X137 + X138 =  3/28
X124 + X134 + X145 + X146 + X147 + X148 =  3/28        X125 + X135 + X145 + X156 + X157 + X158 =  3/28
X126 + X136 + X146 + X156 + X167 + X168 =  3/28        X127 + X137 + X147 + X157 + X167 + X178 =  3/28
X128 + X138 + X148 + X158 + X168 + X178 =  3/28        X123 + X234 + X235 + X236 + X237 + X238 =  3/28
X124 + X234 + X245 + X246 + X247 + X248 =  3/28        X125 + X235 + X245 + X256 + X257 + X258 =  3/28
X126 + X236 + X246 + X256 + X267 + X268 =  3/28        X127 + X237 + X247 + X257 + X267 + X278 =  3/28
X128 + X238 + X248 + X258 + X268 + X278 =  3/28        X134 + X234 + X345 + X346 + X347 + X348 =  3/28
X135 + X235 + X345 + X356 + X357 + X358 =  3/28        X136 + X236 + X346 + X356 + X367 + X368 =  3/28
X137 + X237 + X347 + X357 + X367 + X378 =  3/28        X138 + X238 + X348 + X358 + X368 + X378 =  3/28
X145 + X245 + X345 + X456 + X457 + X458 =  3/28        X146 + X246 + X346 + X456 + X467 + X468 =  3/28
X147 + X247 + X347 + X457 + X467 + X478 =  3/28        X148 + X248 + X348 + X458 + X468 + X478 =  3/28
X156 + X256 + X356 + X456 + X567 + X568 =  3/28        X157 + X257 + X357 + X457 + X567 + X578 =  3/28
X158 + X258 + X358 + X458 + X568 + X578 =  3/28        X167 + X267 + X367 + X467 + X567 + X678 =  3/28
X168 + X268 + X368 + X468 + X568 + X678 =  3/28        X178 + X278 + X378 + X478 + X578 + X678 =  3/28
```

Observemos que las restricciones correspondientes a las probabilidades de inclusión de primer orden se pueden obtener a partir de las correspondientes a las de segundo orden. Por ejemplo, sumando las siete primeras restricciones para las probabilidades de inclusión de segundo orden se obtiene la primera para las de primer orden.

Así pues, para facilitar la resolución del problema podemos eliminar las ocho primeras restricciones. Una solución óptima del problema viene dada por el siguiente diseño,

m	$p(m)$	m	$p(m)$
$X125$	0.017857	$X238$	0.017857
$X128$	0.089286	$X247$	0.071429
$X135$	0.017857	$X256$	0.053571
$X137$	0.089286	$X346$	0.017857
$X145$	0.017857	$X348$	0.053571
$X146$	0.089286	$X356$	0.035714
$X156$	0.017857	$X357$	0.017857
$X157$	0.017857	$X358$	0.035714
$X158$	0.017857	$X457$	0.035714
$X234$	0.035714	$X458$	0.053571
$X236$	0.053571	$X678$	0.107143

con valor de la función objetivo,

$$C = 0.19642850$$

siendo para el diseño muestral aleatorio MAS(8,3),

$$C_{\text{MAS}} = \frac{26}{56}$$

habiéndose obtenido pues, una mejora en relación al muestreo aleatorio.

Para extraer una muestra, basta aplicar cualquier método de selección de un elemento con probabilidades variables directamente sobre el diseño obtenido, por ejemplo el método de las probabilidades acumuladas o el método de Lahiri (1951). Véase Sukhatme *et al.* (1984). Una vez obtenida la muestra, para realizar la estimación de un parámetro lineal, se aplicarán los estimadores usuales en el muestreo aleatorio. Así, si el parámetro es la media poblacional, la estimación se realizará mediante la media muestral y el error se calculará por las expresiones usuales basadas en la cuasivarianza.

Es importante notar que si queremos obtener diseños de tipo controlado, $\pi-$equivalentes a uno dado, basta añadir, a las restricciones ya consideradas, otras que acoten, en la medida deseada, las probabilidad de las muestras no preferidas. Es decir, si denominamos $W \subseteq M$ al subconjunto del espacio muestral formado por las muestras no preferidas, y queremos que la probabilidad de dichas muestras no exceda el valor $\alpha \in [0,1]$, basta añadir las restricciones,

$$x_k \leq \alpha \qquad \forall k \in W$$

Observemos también que para el caso de tamaño muestral fijo, todo diseño óptimo en relación a una función objetivo lineal tiene, a lo sumo, un tamaño de soporte $\binom{N}{2}$ ya que, considerado como punto del poliedro de los diseños $\pi-$equivalentes, es un punto extremo. Ello nos indica que aquellos diseños, en los que todas las muestras tienen probabilidades estrictamente positivas no son diseños óptimos. Similares consideraciones pueden hacerse para diseños de tamaño muestral variable.

No obstante, en caso de ser la función objetivo convexa, los mínimos están en el interior. Así, si consideramos la función cóncava,

$$F = -\sum_{k=1}^{q} x_k \ln x_k$$

que podemos denominar **entropía del diseño**, la minimización de su opuesta,

$$C = \sum_{k=1}^{q} x_k \ln x_k$$

dará lugar a diseños uniformes, si existen, o bien próximos a ellos.

3. DISEÑOS MUESTRALES ESPECIALES

Hay diseños muestrales que desempeñan un papel importante en la teoría del muestreo por representar situaciones especiales de los mismos, en relación a los valores de π_i y π_{ij}, en el conjunto de los diseños $\pi-$equivalentes. Como veremos los diseños que describimos representan la máxima independencia entre las unidades muestrales o bien la máxima equivalencia entre las mismas.

Definición 2

Decimos que un diseño muestral es $\pi-$**independiente** si,

$$\pi_{ij} = \pi_i \pi_j \ , \quad \forall i < j$$

Notemos que en este caso $\Delta_{ij} = 0 \ , i \neq j$ por lo que resulta lógico suponer que darán buenos estimadores del error de muestreo.

Si denotamos por I_i e I_{ij} a variables aleatorias que indican, respectivamente, si la unidad i y las unidades i y j están en la muestra, se tiene,

$$\pi_{ij} = E[I_{ij}] = E[I_i I_j] = E[I_i]E[I_j] = \pi_i \pi_j \qquad i < j$$

de donde se deduce que la elección de la unidad i es independiente de la elección de cualquier otra unidad, pues las variables aleatorias I_i son independientes, siendo su distribución de probabilidad,

$$P[I_i = 1] = \pi_i \qquad P[I_i = 0] = 1 - \pi_i \qquad \forall i \in U$$

por lo cual, el diseño muestral $\pi-$independiente más general es el llamado **diseño de Poisson**, es decir, $M = 2^U$, siendo la probabilidad asociada a una muestra,

$$p(m) = \prod_{i \in m} \pi_i \prod_{i \in U-m} (1 - \pi_i) \qquad (1)$$

La realización práctica de este diseño se puede llevar a cabo mediante el denominado **muestreo de Poisson** consistente en explorar secuencialmente la población, seleccionando cada elemento i con probabilidad π_i e independientemente de los demás.

El diseño de Poisson presenta la propiedad de maximizar la entropía del diseño,

$$-\sum_{m\in M} p(m)\ln p(m) = -\sum_{k=1}^{q} x_k \ln x_k$$

en la clase de diseños muestrales con probabilidades de inclusión dadas, y esta propiedad se mantiene, Hájek (1981), en los llamados **diseños de Poisson condicionados** definidos a partir de un subconjunto $M \subseteq 2^U$, por las probabilidades,

$$p_M(m) = \frac{p(m)}{p(M)} \qquad m \in M$$

siendo $p(M) = \sum_M p(m)$, $p_M(m) = 0$ si $m \notin M$, y $p(m)$ dada por (1).

Estos diseños de Poisson condicionados son importantes porque contienen, como casos particulares, a diseños muestrales clásicos. Así, para M formado por todas las muestras de tamaño fijo n, y $\pi_i = p_i = p$, $\forall i$, se tiene,

$$p_M(m) = \frac{p^n(1-p)^{N-n}}{\sum_{m\in M} p^n(1-p)^{N-n}} = \frac{p^n(1-p)^{N-n}}{\binom{N}{n} p^n(1-p)^{N-n}} = \frac{1}{\binom{N}{n}}$$

es decir, el diseño muestral aleatorio simple.

La realización práctica de estos métodos se puede llevar a cabo mediante procedimientos de aceptación-rechazo pero ello puede llegar a ser muy lento. Así, para el diseño de Poisson condicionado a tamaño fijo n, se tiene la siguiente cota dada por Hájek (1981),

$$P[n(m) = n] < \frac{1}{\sqrt{2\pi\lambda n}}$$

siendo λ una cantidad que verifica $\lambda \leq 1 - \pi_i$, $\forall i \in U$.

Por supuesto, existen métodos más directos para obtener una muestra aleatoria simple de tamaño n, incluso sin conocer el valor de N. En este sentido, proponemos el siguiente procedimiento, que denominamos de inserción, y que tiene en común con el muestreo de Poisson la forma de obtención de la muestra mediante una exploración secuencial de la población.

Algoritmo 1 (MÉTODO DE INSERCIÓN)

En este algoritmo, Θ representa una lista ordenada de elementos de la población usando como criterio de ordenación el indicado en el paso 2.

1. Hacer $i := 1$ y $\Theta := ()$.

2. Generar un número aleatorio $\alpha_i \sim \mathcal{U}[0,1)$, y según sea i en relación a n realizar las siguientes transformaciones sobre la lista Θ,

   - $i \leq n$. Introducir la unidad i en Θ, de forma que los elementos de dicha lista aparezcan ordenados por la magnitud del número aleatorio correspondiente.
   - $i > n$. Si $\alpha_i \leq \alpha_l = \max_{j \in \Theta} \alpha_j$ eliminar la unidad l de Θ e insertar en dicha lista la unidad i de forma que la lista permanezca ordenada por el criterio anteriormente indicado. En caso contrario, la lista Θ no se modifica.

3. Hacer $i := i + 1$. Si $i \in U$, ir al paso 2. En caso contrario, finalizar el proceso, formando los elementos en Θ la muestra final.

Este procedimiento asegura la obtención de una muestra perteneciente a un diseño aleatorio simple, a partir de una exploración secuencial de la población, sin requerir el conocimiento previo del tamaño de la población. Para probarlo, basta observar que si ordenáramos la población completa usando como criterio la magnitud de los números aleatorios, cualquiera de las $N!$ ordenaciones posibles tiene probabilidad $1/N!$, por lo cual, la probabilidad de obtener la muestra m será,

$$p(m) = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$$

Otro procedimiento para la obtención de muestras es el denominado **método de permutaciones**. Véase Hájek (1981).

Si imponemos que el tamaño muestral sea constante, los diseños muestrales $\pi-$independientes degeneran en diseños con una sola muestra.

Teorema 1

Dado un diseño muestral de tamaño fijo,

$$\text{si} \quad \exists j \quad \text{tal que} \quad \pi_{ij} = \pi_i \pi_j, \ \forall i \neq j$$

entonces j ha de pertenecer a todas las muestras del diseño, o no pertenecer a ninguna.

Demostración

Veamos que, en las condiciones del teorema, π_j es cero ó uno.

Si $\pi_j \neq 0$, como,

$$\sum_{\substack{i \in U \\ i \neq j}} \pi_{ij} = (n-1)\pi_j = \sum_{\substack{i \in U \\ i \neq j}} \pi_i \pi_j = \pi_j(n - \pi_j)$$

luego $\pi_j = 1$.

■

Del teorema anterior, imponiendo la condición a todos los pares de elementos de la población, se deduce el siguiente corolario.

Corolario 2 *Un diseño muestral $\pi-$independiente de tamaño fijo está formado por una sola muestra.*

Vemos pues que la propiedad de $\pi-$independencia, en diseños de tamaño fijo, da lugar a casos triviales. A continuación estudiamos otra propiedad interesante de los diseños muestrales y que para el caso de tamaño fijo, es verificada por diseños importantes.

Definición 3

Decimos que un diseño muestral es **simétrico** si verifica,

$$\pi_i = a,\ i = 1,\ldots,N \quad \text{y} \quad \pi_{ij} = b,\ i,j = 1,\ldots,N \quad i \neq j$$

Es decir, todas las unidades tienen las mismas probabilidades de inclusión de primer orden, y todas las parejas de unidades las mismas probabilidades de inclusión de segundo orden.

En este caso los niveles dados a los valores de las variables, $\{1,2,\ldots,N\}$, son independientes de las unidades, frente a la estimación, no proporcionando ninguna información, como nos dice el teorema siguiente,

Teorema 3

Dado un diseño muestral simétrico, cualquier otro diseño muestral obtenido del anterior al renumerar los niveles de las variables mediante una permutación, σ, en U y en todas las muestras del espacio, siendo $p(m_\sigma) = p(m)$, será un diseño simétrico $\pi-$equivalente al anterior.

Nótese que el teorema es cierto debido a ser un diseño simétrico pues de este modo es $\pi_i = \pi_{i_\sigma}$. Si además imponemos al diseño que sea de tamaño muestral fijo, tenemos,

Teorema 4

Todo diseño muestral simétrico de tamaño fijo es $\pi-$equivalente al MAS(N,n).

Demostración

Puesto que $\sum_{i\in U}\pi_i = n$, si π_i es constante debe ser $\pi_i = n/N$ y dado que $\sum_{j\neq i}\pi_{ij} = (n-1)\pi_i$, al ser π_{ij} constante se tendrá $(N-1)\pi_{ij} = (n-1)n/N$, de donde $\pi_{ij} = n(n-1)/N(N-1)$, por lo que el diseño es $\Pi-$equivalente al MAS(N,n).

■

Aunque esta familia de diseños $\pi-$equivalentes al MAS(N,n) puede caracterizarse con menor número de condiciones,

Teorema 5

Todo diseño muestral de tamaño fijo tal que las probabilidades de inclusión de segundo orden son constantes es $\pi-$equivalente al MAS(N,n).

Demostración

Como $\pi_{ij} = \lambda = cte$, será $(N-1)\lambda = (n-1)\pi_j$, por lo que también son constantes las probabilidades de inclusión de primer orden, y el teorema se deduce del anterior.

■

Ello nos dice que las $\binom{N}{2}$ restricciones,

$$\sum_{k=1}^{q} x_k I_{ij}(k) = \pi_{ij} \qquad i<j$$

asociadas a las π_{ij}, $i\neq j$ determinan dicho conjunto para los diseños $\pi-$equivalentes al MAS(N,n).

Los diseños $\pi-$equivalentes a un MAS(N,n), óptimos frente a una función lineal tienen a lo sumo un tamaño de soporte $\binom{N}{2}$, por la definición del poliedro, pero puede ocurrir que por la degeneración del mismo, el soporte llegue a ser incluso de tamaño $(N-1)N/(n-1)n$, como indica Hedayat (1979). Para estos soportes muestrales con menos elementos que el que proporciona el diseño MAS(N,n), las probabilidades asociadas a cada muestra no forman, en general, una distribución uniforme sobre el conjunto de todas ellas, como ocurre en el muestreo aleatorio simple, aunque a veces puede que exista un diseño $\pi-$equivalente uniforme sobre su

soporte, como se comprueba en el siguiente ejemplo, en el cual se indica un diseño uniforme π−equivalente a un MAS(7,3) con tamaño de soporte igual a la cota inferior $(N-1)N/(n-1)n = 7$,

m	124	235	346	457	561	672	713
$p(m)$	1/7	1/7	1/7	1/7	1/7	1/7	1/7

En caso de ser el diseño de tamaño variable y π−independiente, la simetría implica que $\pi_i = p$, $\forall i$ y $\pi_{ij} = p^2$, $\forall\ i < j$, es decir, π−equivalente a un diseño de Poisson con $p_i = p$, $\forall i$, es decir, al diseño de Bernoulli, siendo fácil probar que esta propiedad se mantiene imponiendo solamente la igualdad de las probabilidades de inclusión de segundo orden.

4. DISEÑOS MUESTRALES EQUIVALENTES DE PRIMER ORDEN

Ya hemos visto como en la clase de diseños muestrales π−equivalentes a uno dado es posible encontrar diseños óptimos en relación a determinado criterio.

De forma similar, dado un diseño muestral podemos buscar diseños óptimos en la clase de diseños con las mismas probabilidades de inclusión de primer orden que aquel. A estos diseños les denominamos **equivalentes de primer orden**.

Con este planteamiento, se buscan diseños óptimos en una clase más amplia que la de los π−equivalentes, por lo que en general se obtendrán diseños mejores en relación al criterio de optimalidad aplicado. No obstante, al no coincidir necesariamente la probabilidades de inclusión de segundo orden, los estimadores no tendrán la misma varianza.

Haciendo un planteamiento similar al caso de los diseños π−equivalentes, las restricciones relativas a las probabilidades de inclusión serían ahora,

$$\sum_{k=1}^{q} x_k I_{ii}(k) = \pi_i \qquad \forall i$$

con la imposición adicional, supuesto tamaño muestral fijo, de que se verifique,

$$\pi_{ij} \leq \pi_i \pi_j \qquad \forall i \neq j$$

para que el estimador de la varianza dado por la fórmula de Yates-Grundy-Sen, Yates y Grundy (1953) y Sen (1953),

$$\widehat{V}[\widehat{\theta}(m)] = -\frac{1}{2}\sum_{\substack{i,j\in m\\ i\neq j}}\sum \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(a_i\frac{Y_i}{\pi_i}-a_j\frac{Y_j}{\pi_j}\right)^2$$

sea no negativo. Véase Sukhatme *et al.* (1984). Ello nos obliga a introducir restricciones adicionales del tipo,

$$\sum_{k=1}^{q} x_k I_{ij}(k) - \pi_{ij} = 0 \qquad \forall i \neq j$$

originando así que las probabilidades de inclusión de segundo orden sean también variables del problema.

Este modo de trabajar con los diseños tiene como principal limitación el elevado tamaño de los espacios muestrales, lo que puede hacer inabordable el problema de programación matemática que se plantee, dependiendo por supuesto de los recursos computacionales disponibles. Sin embargo, presenta la ventaja de permitir la obtención de las probabilidades de inclusión de segundo orden, necesarias para la estimación del error, evitando así las dificultades de otros procedimientos clásicos para obtener dichas probabilidades, usualmente mediante expresiones asintóticas complicadas.

Adicionalmente se obtienen las probabilidades de elección de cada muestra con lo que se facilita su obtención.

Por todo ello, puede ser de utilidad en la selección de las unidades primarias en muestreos multietápicos. Seguidamente exponemos un ejemplo numérico que indica la realización práctica de lo anteriormente expuesto.

Ejemplo 2

Consideremos nuevamente la situación expuesta en el ejemplo 1 de este trabajo, pero ahora vamos a buscar el diseño óptimo entre los equivalentes de primer orden al MAS(8,3). Las restricciones relativas a las probabilidades de inclusión de primer orden serán,

$$\sum_{k=1}^{56} x_k I_{ii}(k) = 3/8 \qquad \forall i$$

y además tendremos,

$$\pi_{ij} \leq \pi_i\pi_j = 9/64 \qquad \forall i \neq j$$

para que el estimador de la varianza sea siempre no negativo.

De esta forma, usando la notación Xijk para la probabilidad de la muestra $\{i,j,k\}$, e Yij para π_{ij}, tendremos que añadir las restricciones,

$$\sum_{k=1}^{56} x_k I_{ij}(x_k) - \pi_{ij} = 0 \qquad \forall i \neq j$$

Observemos que si el diseño es de tamaño fijo,

$$\sum_{\substack{i \in U \\ i \neq j}} \pi_{ij} = (n-1)\pi_j$$

y en nuestro caso, al ser MAS(8,3), y fijar la probabilidades de inclusión de primer orden, obtenemos las siguientes restricciones,

$$\sum_{\substack{i \in U \\ i \neq j}} \pi_{ij} = \frac{6}{8}$$

Así pues, eliminando restricciones redundantes, el problema a resolver es minimizar,

```
C = X123 + X124 + X125 + X134 + X135 + X145 + X167 + X168 + X178 + X234 + X235 + X245 + X267
  + X268 + X278 + X345 + X367 + X368 + X378 + X467 + X468 + X478 + X567 + X568 + X578 + X678
```

Sujeto a las restricciones,

```
X123 + X124 + X125 + X126 + X127 + X128 - Y12 = 0
X124 + X134 + X145 + X146 + X147 + X148 - Y14 = 0
X126 + X136 + X146 + X156 + X167 + X168 - Y16 = 0
X128 + X138 + X148 + X158 + X168 + X178 - Y18 = 0
X124 + X234 + X245 + X246 + X247 + X248 - Y24 = 0
X126 + X236 + X246 + X256 + X267 + X268 - Y26 = 0
X128 + X238 + X248 + X258 + X268 + X278 - Y28 = 0
X135 + X235 + X345 + X356 + X357 + X358 - Y35 = 0
X137 + X237 + X347 + X357 + X367 + X378 - Y37 = 0
X145 + X245 + X345 + X456 + X457 + X458 - Y45 = 0
X147 + X247 + X347 + X457 + X467 + X478 - Y47 = 0
X156 + X256 + X356 + X456 + X567 + X568 - Y56 = 0
X158 + X258 + X358 + X458 + X568 + X578 - Y58 = 0
X168 + X268 + X368 + X468 + X568 + X678 - Y68 = 0

X123 + X134 + X135 + X136 + X137 + X138 - Y13 = 0
X125 + X135 + X145 + X156 + X157 + X158 - Y15 = 0
X127 + X137 + X147 + X157 + X167 + X178 - Y17 = 0
X123 + X234 + X235 + X236 + X237 + X238 - Y23 = 0
X125 + X235 + X245 + X256 + X257 + X258 - Y25 = 0
X127 + X237 + X247 + X257 + X267 + X278 - Y27 = 0
X134 + X234 + X345 + X346 + X347 + X348 - Y34 = 0
X136 + X236 + X346 + X356 + X367 + X368 - Y36 = 0
X138 + X238 + X348 + X358 + X368 + X378 - Y38 = 0
X146 + X246 + X346 + X456 + X467 + X468 - Y46 = 0
X148 + X248 + X348 + X458 + X468 + X478 - Y48 = 0
X157 + X257 + X357 + X457 + X567 + X578 - Y57 = 0
X167 + X267 + X367 + X467 + X567 + X678 - Y67 = 0
X178 + X278 + X378 + X478 + X578 + X678 - Y78 = 0

Y12+Y13+Y14+Y15+Y16+Y17+Y18 = 6/8
Y13+Y23+Y34+Y35+Y36+Y37+Y38 = 6/8
Y15+Y25+Y35+Y45+Y56+Y57+Y58 = 6/8
Y17+Y27+Y37+Y47+Y57+Y67+Y78 = 6/8

Y12+Y23+Y24+Y25+Y26+Y27+Y28 = 6/8
Y14+Y24+Y34+Y45+Y46+Y47+Y48 = 6/8
Y16+Y26+Y36+Y46+Y56+Y67+Y68 = 6/8
Y18+Y28+Y38+Y48+Y58+Y68+Y78 = 6/8

0 < Yij <= 9/64 para todo i,j
```

Una solución óptima del anterior problema viene dada por el siguiente diseño muestral,

m	$p(m)$	m	$p(m)$
$X126$	0.093750	$X247$	0.031250
$X127$	0.031250	$X248$	0.015625
$X136$	0.015625	$X258$	0.062500
$X148$	0.093750	$X346$	0.031250
$X156$	0.031250	$X347$	0.109375
$X157$	0.109375	$X358$	0.078125
$X236$	0.046875	$X456$	0.093750
$X237$	0.031250	$X678$	0.062500
$X238$	0.062500		

siendo las probabilidades de inclusión de segundo orden,

$$\Pi = \begin{pmatrix} \bullet & 0.125000 & 0.015625 & 0.093750 & 0.140625 & 0.140625 & 0.140625 & 0.093750 \\ & \bullet & 0.140625 & 0.046875 & 0.062500 & 0.140625 & 0.093750 & 0.140625 \\ & & \bullet & 0.140625 & 0.078125 & 0.093750 & 0.140625 & 0.140625 \\ & & & \bullet & 0.093750 & 0.125000 & 0.140625 & 0.109375 \\ & & & & \bullet & 0.125000 & 0.109375 & 0.140625 \\ & & & & & \bullet & 0.062500 & 0.062500 \\ & & & & & & \bullet & 0.062500 \\ & & & & & & & \bullet \end{pmatrix}$$

donde los puntos representan los elementos diagonales de la matriz del diseño.

El valor mínimo de la función objetivo resulta ser,

$$C = 0.062500$$

habiéndose obtenido pues una mejora con respecto a la búsqueda en la clase de diseños $\pi-$equivalentes.

Observemos que, en general, la solución del problema anterior no es única siendo posible escoger entre todas las existentes aquella que mejore otros criterior adicionales.

No es posible, en general, hacer afirmaciones sobre la precisión de la estimación con estas nuevas probabilidades de inclusión de segundo orden, no obstante, si ponderamos las muestras de forma adecuada, podemos conseguir una ganancia en la misma, como se demuestra en el siguiente apartado.

4.1. Muestras más informativas

Usualmente, entre las unidades de la población U existen ciertas relaciones lo que origina que unas sean más afines y otras menos de cara a la información que ofrecen, sin que este grado de afinidad puede llevarse a una estratificación aunque sí sea posible cuantificar el grado de proximidad.

Por ejemplo, si queremos realizar un estudio en Andalucía, en relación con la producción agrícola, ciertas comarcas tienen entre sí más afinidad que otras, aunque no estén próximas geográficamente. Así, las provincias o regiones que tienen producción oleícola no se parecen en su problemática a las que tienen producción temprana de productos que son suceptibles de exportación. Por ello, si queremos hacer una encuesta para estudiar el estado socioeconómico de la región, si en una muestra intervienen dos comarcas con la misma problemática será menos informativa que si interviniesen dos comarcas con distinta problemática.

Basándonos en la idea anterior, suponemos la existencia de una matriz de afinidad, $\mathcal{A}$, de dimensión $N \times N$, simétrica y cuyos elementos supondremos no negativos, de manera que el elemento a_{ij} representa la afinidad entre las unidades i y j. Esta matriz nos permite cuantificar el concepto de muestra más informativa. Para ello, dado un diseño con espacio muestral M, definimos la función,

$$A : M \longmapsto R^{+} \cup \{0\}$$

en la forma,

$$A(m) = \sum_{\substack{i,j \in m \\ i \neq j}} \sum a_{ij} \qquad \forall m \in M$$

Definición 4

Dadas dos muestras, $m, m' \in M$, diremos que m es más informativa que m', lo que denotamos $m \succeq m'$, si verifican, $A(m) \leq A(m')$.

Observemos que esta relación es reflexiva y transitiva pero no antisimétrica, es pues un preorden. Además, dadas dos muestras, m y m', se verifica $m \succeq m'$ ó $m' \succeq m$, luego esta relación es un preorden total.

Como el espacio muestral es finito, podemos hablar de las muestras de máxima información, es decir, del conjunto de muestras,

$$M_I = \{m \in M | A(m) = \min_{m' \in M} A(m')\}$$

sin embargo, la elección de una de estas muestras para realizar la estimación no resulta apropiada por incumplir la estrategia del muestreo probabilístico. No obstante, para evitar esta dificultad, podemos definir el concepto de **afinidad media** del diseño $d = (M, p(\cdot))$ como,

$$E[A(d)] = \sum_{m \in M} A(m)p(m)$$

y buscar entre los diseños equivalentes de primer orden a uno dado, los que minimizan dicha afinidad media. Este enfoque tiene una interpretación interesante, en efecto, observemos,

$$\begin{aligned} E[A(d)] &= \sum_{m \in M} A(m)p(m) = \sum_{m \in M} \sum_{\substack{i,j \in m \\ i \neq j}} \sum a_{ij} p(m) \\ &= \sum_{\substack{i,j \in U \\ i \neq j}} \sum a_{ij} \sum_{\substack{m \in M \\ m \ni i,j}} p(m) = \sum_{\substack{i,j \in U \\ i \neq j}} \sum a_{ij} \pi_{ij} \end{aligned}$$

Ello nos permite afirmar que al minimizar la afinidad media en el conjunto de los diseños equivalentes de primer orden a uno dado, tienden a estar menos representados los pares de elementos con mayor afinidad.

Dada una clase C, de diseños muestrales, a los diseños en C que hagan mínima la afinidad media, los denominaremos **diseños más informativos** en la clase C. Estos diseños, para tipos especiales de afinidad, son también óptimos en otro sentido, como indica el siguiente teorema.

Teorema 6

Dada una variable cuantitativa, $Y_1, Y_2, \ldots, Y_N$, definida sobre la población U, sea,

$$M = \max_{i \neq j} (Y_i - Y_j)^2 = (Y_{(N)} - Y_{(1)})^2$$

y consideremos la matriz de afinidad $\mathcal{A}$, dada por,

$$a_{ij} = K - (Y_i - Y_j)^2 \qquad \text{siendo} \quad K \geq M$$

Se verifica que el diseño más informativo en la clase de diseños de tamaño fijo equivalentes de primer orden al MAS(N,n), minimiza la varianza del estimador de Horvitz-Thompson para el total poblacional,

$$T(Y) = \sum_{i \in U} Y_i$$

Demostración

Sabemos que el estimador de Horvitz-Thompson para el total poblacional es,

$$\widehat{T}(m) = \sum_{i \in m} \frac{Y_i}{\pi_i}$$

y por ser el diseño de tamaño fijo, su varianza se puede expresar mediante la fórmula de Yates-Grundy-Sen,

$$[\widehat{T}(m)] = -\frac{1}{2} \sum_{\substack{i,j \in U \\ i \neq j}} \sum (\pi_{ij} - \pi_i \pi_j) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

Desarrollando la afinidad media, obtenemos,

$$\begin{aligned} E[A(d)] &= \sum_{\substack{i,j \in U \\ i \neq j}} \sum a_{ij} \pi_{ij} = \sum_{\substack{i,j \in U \\ i \neq j}} \sum (K - (Y_i - Y_j)^2) \pi_{ij} \\ &= K \sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} - \sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} (Y_i - Y_j)^2 = Kn(n-1) - \sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} (Y_i - Y_j)^2 \end{aligned}$$

lo que nos dice que minimizar la afinidad media es equivalente a maximizar la cantidad,

$$\sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} (Y_i - Y_j)^2$$

Por otra parte, si el diseño es equivalente de primer orden al MAS(N,n), se tendrá para la varianza,

$$\begin{aligned} V[\widehat{T}(m)] &= -\frac{1}{2} \left[\sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} \left(\frac{Y_i}{n/N} - \frac{Y_j}{n/N} \right)^2 - \sum_{\substack{i,j \in U \\ i \neq j}} \sum (Y_i - Y_j)^2 \right] \\ &= \frac{1}{2} \sum_{\substack{i,j \in U \\ i \neq j}} \sum (Y_i - Y_j)^2 - \frac{N^2}{2n^2} \sum_{\substack{i,j \in U \\ i \neq j}} \sum \pi_{ij} (Y_i - Y_j)^2 \end{aligned}$$

Luego el diseño más informativo es de mínima varianza.

■

Observemos que en la práctica, los valores de la variable de estudio, $Y_1, \ldots, Y_N$, no son conocidos, por lo que puede emplearse una variable auxiliar con valores $X_1, \ldots, X_N$, relacionada con la variable de estudio, como es habitual en el muestreo.

REFERENCIAS

[1] **Avadhani, M.S.** y **Sukhatme, B.V.** (1973). "Controlled sampling with equal probabilities and without replacement". *Internat. Statist. Rev.*, **41**, 175–182.

[2] **Bellhouse, D.R.** (1984). "A review of optimal designs in survey sampling". *The Canadian Journal of Statistics*, **12**, 53–65.

[3] **Foody, W.** y **Hedayat, A.** (1976). "On theory and application of BIB designs with repeated blocks". *Ann. Statist. Assoc.*, **5**, 932–945.

[4] **Fernández, F.R.** y **Mayor, J.A.** (1994). *Muestreo en poblaciones finitas: curso básico.* P.P.U. Barcelona.

[5] **Goodman, R.** y **Kish, L.** (1950). "Controlled selection — a technique in probability sampling". *J. Amer. Statist. Assoc.*, **45**, 350–372.

[6] **Hájek, J.** (1981). *Sampling from a Finite Population.* Marcel Dekker, Inc. New York.

[7] **Hedayat, A.** (1979). "Sampling Designs with Reduced Support Sizes". *Optimizing Methods in Statistics.* Rustagi, J. (Ed.). Academics Press, New York.

[8] **Hedayat, A.** y **Sinha, B.** (1991). *Desing and Inference in Finite Population Sampling.* John Wiley & Sons, Inc. New York.

[9] **Horvitz, D.G.** y **Thompson, D.J.** (1952). "A generalization of sampling without replacement from a finite universe". *J. Amer. Statist. Assoc.*, **47**, 663–685.

[10] **Lahiri, D.B.** (1951). "A method of sample selection providing unbiased ratio estimates". *Bulletin of the International Statistical Institute*, **33**, 133–140.

[11] **Midzuno, H.** (1952). "On the sampling system with probability proportionate to sum of sizes". *Annals of the Institute of Statistical Mathematics*, **3**, 99–107.

[12] **Rao, J.N.K.** y **Nigam, A.K.** (1990). "Optimal controlled sampling designs". *Biometrika*, **77**, 807–814.

[13] **Rao, J.N.K.** y **Nigam, A.K.** (1992). "Optimal controlled sampling: a unified approach". *Internat. Statist. Rev.*, **60**, 89–98.

[14] **Sen, A.R.** (1953). "On the estimate of the variance in sampling with varying probabilities". *J. Indian Soc. Agric. Statist.*, **5**, 119–127.

[15] **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S.** y **Asok, C.** (1984). *Sampling Theory of Surveys Applications.* Tercera edición. Iowa State University Press. Ames. Iowa.

[16] **Yates, F.** y **Grundy, P.M.** (1953). "Selection without replacement from within strata with probability proportional to size". *J. Roy. Statist. Soc.*, **B15**, 253–261.

[17] **Wynn, H.P.** (1977). "Convex sets of finite population plans". *The Annals of Statistics*, **5**, 414–418.

ENGLISH SUMMARY:

π–EQUIVALENT AND FIRST ORDER EQUIVALENT SAMPLING DESIGNS

Fernández García, Francisco R. and Mayor Gallego, José A.

In order to estimate the parameter,

$$\theta(Y) = \sum_{i \in U} a_i Y_i$$

over the population $U = \{1, 2, \ldots, N\}$, we can use the Horvitz-Thompson estimator,

$$\widehat{\theta}(m) = \sum_{i \in m} a_i \frac{Y_i}{\pi_i}$$

where m is a sample from a sampling design $d = (M, p(\cdot))$. This estimator and its variance depend only on the first and second order inclusion probabilities, so we can look for sampling designs improving additional criteria in the class of designs with the same inclusion probabilities.

Usually, searching for these designs implies the resolution of mathematical programming problems. So, we study the sampling designs with equal inclusion probabilities and the practical method of finding a design with given inclusion probabilities, improving some utility criteria.

Let be Π the design matrix, that is,

$$\Pi = \{\pi_{ij}\}_{1 \leq i,j \leq N}$$

with $\pi_{ii} = \pi_i$. Two sampling designs, $d_1 = (M_1, p_1(\cdot))$ and $d_2 = (M_2, p_2(\cdot))$, with design matrices $\Pi^{(1)}$ and $\Pi^{(2)}$, respectively, are said π-equivalent if $\Pi^{(1)} = \Pi^{(2)}$.

We can consider a sampling design as a point belonging to the product space I^q with $q = 2^N$ and $I = [0, 1]$. The sampling space is,

$$M = \{m_1, m_2, \ldots, \ldots, m_q\}$$

and $x = (x_1, x_2, \ldots, x_q)^t$ is the above mentioned point, that is to say, the probability distribution, with $x_k = p(m_k)$, $k = 1, \ldots, q$. We denote by,

$$I_{ij}(k) = \begin{cases} 1 & i, j \in m_k \\ 0 & i, j \notin m_k \end{cases} \qquad \forall\, i, j \in U,\ k = 1, \ldots, q$$

we have,

$$\begin{aligned} \sum_{k=1}^{q} x_k I_{ij}(k) &= \pi_{ij} \qquad \forall i \leq j \\ \sum_{k=1}^{q} x_k &= 1 \\ x_k &\geq 0 \qquad k = 1, \ldots, q \end{aligned}$$

that is to say, a convex polyhedron whose elements are π-equivalent sampling designs.

Notice that in order to obtain π-equivalent controlled sampling designs, it is sufficient to add some constraints bounding the probabilities of the non-preferred samples. So, let be W the subset of the sampling space of the non-preferred samples, if we want the probabilities of these samples not to be greater than $\alpha \in [0, 1]$, it is sufficient to add the constraints,

$$x_k \leq \alpha \qquad \forall k \in W$$

Also, for a given sampling design, we can consider the class of sampling designs with the same first order inclusion probabilities. These designs are named **first order equivalent designs**. Thus, we can search for optimal designs in the class of first order equivalent designs, obtaining, in general, better designs in relation to the optimality criteria. Nevertheless, the second order inclusion probabilities are not the same and therefore the estimators do not have equal variances.

The posed problem has the following constraints for the inclusion probabilities,

$$\sum_{k=1}^{q} x_k I_{ii}(k) = \pi_i \qquad \forall i$$

with the additional constraints, if we have a fixed sample size,

$$\pi_{ij} \leq \pi_i \pi_j \qquad \forall i \neq j$$

in order that the Yates-Grundy-Sen's variance estimator,

$$\widehat{V}[\widehat{\theta}(m)] = -\frac{1}{2} \sum_{\substack{i,j \in m \\ i \neq j}} \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(a_i \frac{Y_i}{\pi_i} - a_j \frac{Y_j}{\pi_j} \right)^2$$

may be non negative. This forces the additional constraints, to be introduced,

$$\sum_{k=1}^{q} x_k I_{ij}(k) - \pi_{ij} = 0 \qquad \forall i \neq j$$

that is to say, the second order inclusion probabilities are also problem variables.

Usually, between the units in the population there are some relations of affinity, with respect to the information that they provide. Using this idea, let be $\mathcal{A}$ an affinity matrix, with dimensions $N \times N$, symmetric and non negative, where the element a_{ij} is the affinity between the units i and j. Given a sampling design with sampling space M, we define the function,

$$A : M \longmapsto R^+ \cup \{0\}$$

as,

$$A(m) = \sum_{\substack{i,j \in m \\ i \neq j}} \sum a_{ij} \qquad \forall m \in M$$

So, for two samples, $m, m' \in M$, we say that m is more informative than m', and we denote it by $m \succeq m'$, if they verify $A(m) \leq A(m')$.

This relation is reflexive and transitive but not antisymmetric, so, it is a preorder. Furthermore, for two samples m and m', they verify $m \succeq m'$ or $m' \succeq m$, thus, the relation is a total preorder.

Since the sampling space is finite, we can consider the samples providing maximal information, that is to say, the set,

$$M_I = \{m \in M | A(m) = \min_{m' \in M} A(m')\}$$

In order to estimate the parameter, the choice of one of these samples is not appropriate because it does not fulfil the probabilistic sampling strategy. Nevertheless, to avoid this difficulty, we define the **expected affinity** of the design $d = (M, p(\cdot))$ as,

$$E[A(d)] = \sum_{m \in M} A(m)p(m)$$

and minimizing the expected affinity by searching into the first order equivalent sampling designs. This approach has an interesting interpretation. We note that,

$$\begin{aligned} E[A(d)] &= \sum_{m \in M} A(m)p(m) = \sum_{m \in M} \sum_{\substack{i,j \in m \\ i \neq j}} \sum a_{ij} p(m) \\ &= \sum_{\substack{i,j \in U \\ i \neq j}} \sum a_{ij} \sum_{\substack{m \in M \\ m \ni i,j}} p(m) = \sum_{\substack{i,j \in U \\ i \neq j}} \sum a_{ij} \pi_{ij} \end{aligned}$$

Thus, minimizing the expected affinity in the class of the first order equivalent sampling designs, the pairs of units with high affinity have less second order inclusion probability.

For a class of sampling designs, C, the designs with minimal expected affinity will be named **most informative designs**. These designs are also optimum with other criteria and with special types of affinity as the following theorem states,

Theorem

For a quantitative variable, $Y_1, Y_2, \ldots, Y_N$, defined over the population U, let be,

$$M = \max_{i \neq j} (Y_i - Y_j)^2 = (Y_{(N)} - Y_{(1)})^2$$

and let us consider the affinity matrix $\mathcal{A}$, with elements,

$$a_{ij} = K - (Y_i - Y_j)^2 \qquad K \geq M$$

Then, the most informative design in the class of fixed sample size designs first order equivalent to SRS(N,n) minimizes the variance of the Horvitz-Thompson estimator for the population total,

$$T(Y) = \sum_{i \in U} Y_i$$