

# **Muestreo bietápico finalmente autoponderado con unidades seleccionadas necesariamente: modificación de las probabilidades de selección en la segunda etapa, para mantener la fracción de muestreo.**

Vicente Manzano Arrondo (1993)

## **Resumen**

En el presente trabajo, tras un proceso de introducción al contexto del muestreo bietápico con conglomerados de diferente tamaño, se aborda la situación no infrecuente de contar con unidades seleccionadas sin aleatorización en la primera etapa. Los objetivos de investigación concretos, pueden aconsejar contar necesariamente con determinadas unidades muestrales (por ejemplo, colectivos humanos de interés especial). Sin embargo, este proceder atenta contra el principio de la equiprobabilidad de las unidades elementales o de interés último, en la población. Al respecto, se deducen las expresiones de probabilidad que deben ser utilizadas durante la segunda etapa, para mantener finalmente los dos objetivos: equiprobabilidad de unidades elementales y selección necesaria de conglomerados. Así mismo, se acompaña y comenta el programa BIETAPN, que realiza los cálculos.

## **Preliminares**

En una muestra aleatoria simple de elementos, se cuenta con un marco completo o listado exhaustivo de todas las unidades elementales de la población. Mediante un procedimiento manual de azar, utilizando una tabla de dígitos aleatorios, o recurriendo a la generación de entidades pseudoaleatorias con un programa de ordenador, se procede a seleccionar  $m$  elementos del total  $N$  que compone la población de interés.

En muchas ocasiones, no es posible contar con un marco completo. Ocurre también con frecuencia que resulta demasiado costoso el proceso de selección de los elementos y, especialmente, la fase de recogida de datos durante el trabajo de campo. Cabe esperar una dispersión más o menos homogénea en todo el territorio geográfico que ocupa la población de interés; aspecto que multiplica los costes económicos y temporales.

Una estrategia de muestreo, que consigue salvar los dos inconvenientes anteriores, es el muestreo por etapas. El que nos interesa en el presente trabajo, es el llamado *muestreo bietápico*, por contar con dos momentos en la recogida de información y en los procesos de selección.

En una primera etapa, los elementos que sirven de base para la selección (unidades de muestreo) son conjuntos, agrupaciones o, más correctamente, *conglomerados* de unidades elementales. Un conglomerado puede ser un municipio, una sección censal, una manzana de edificios, etc... No obstante, los elementos que interesan finalmente, no son los conglomerados sino las unidades elementales, hacia las que se dirigen las inferencias que serán realizadas a partir de los análisis efectuados con la muestra. En ciencias sociales, la unidad elemental más frecuente es la persona, si bien, pueden interesar también la unidad familiar, o colectivos aún mayores.

En la segunda etapa, se realiza una investigación para obtener más información de las unidades elementales que forman los conglomerados ya seleccionados. Y, seguidamente, se realiza un nuevo muestreo, esta vez de unidades elementales, dentro de los conglomerados.

Mediante este procedimiento, es más viable solventar los problemas de marco, ya que no es necesario revisar o completar éste en los conglomerados no seleccionados. Por otro lado, el trabajo de campo es menos costoso, puesto que los encuestadores invierten menos tiempo y esfuerzo en desplazamientos; y resulta también más fácil la supervisión de la recogida de datos<sup>1</sup>.

---

<sup>1</sup> Al respecto, ver por ejemplo Azorín y Sánchez-Crespo (1986, cap. VII) y Cochran(1976, cap. IX).

## Equiprobabilidad y conglomerados de igual tamaño

La inmensa mayoría de los desarrollos estadísticos, especialmente al nivel de usuario final, consideran que todas las unidades elementales de la población poseen la misma probabilidad de ser seleccionadas en la muestra final. Si este supuesto no se cumple, es necesario recurrir a otros procedimientos de inferencia en las conclusiones. Incluso, muchas utilidades informáticas no permiten realizar estimaciones fuera de este supuesto.

El amplio abanico de modelos de muestreo, hacen prohibitiva la construcción de utilidades informáticas con un objetivo estadístico general, que contemplen todas las posibilidades de estimación. Así, por ejemplo, los comandos descriptivos del SPSS/PC+, como DESCRIPTIVES o FREQUENCIES<sup>2</sup>, utilizan la cuasivarianza como estimador de la varianza. En sentido estricto, esta sustitución es válida en el muestreo aleatorio de elementos con iguales probabilidades, bien sea en poblaciones finitas sin reposición, o en poblaciones prácticamente infinitas. El problema no es grave si consideramos que trabajar con muestras autoponderadas permite reducir considerablemente los errores provocados por alejamiento del modelo teórico de muestreo que sirve de base para la estimación estadística. Por otro lado, si bien las variaciones reales con respecto a este modelo de base producen error, éste puede llegar a ser despreciable si se le compara con otros errores, generados especialmente durante el trabajo de campo (Kish, 1965).

En un muestreo bietápico, donde los conglomerados poseen un mismo tamaño  $M$ , se consigue respetar la equiprobabilidad de las unidades elementales, sin necesidad de recurrir a ninguna estrategia especial. Sencillamente, se utiliza a un procedimiento de selección totalmente aleatorio<sup>3</sup>, en ambas etapas.

La probabilidad de selección  $p(c_i)$  de cualquier conglomerado  $c_i$ , suponiendo un total de  $k$  conglomerados, es  $1/k$ . En efecto, como:

---

<sup>2</sup> Ver, por ejemplo, los manuales de referencia del paquete estadístico, o Manzano (1992).

<sup>3</sup> En el presente trabajo no establecemos distinciones entre un muestreo simple aleatorio y un muestreo estratificado aleatorio con afijación proporcional. Ambos difieren sustancialmente en las expresiones de los errores típicos para los estimadores (menores en la estratificación). No obstante, nuestro interés no se centra en el proceso de estimación, sino en la equiprobabilidad de las unidades elementales, buscando muestras autoponderadas.

$$\sum_{i=1}^k p(c_i) = 1 \quad \text{y} \quad p(c_i) = p(c_j) = p$$

entonces:

$$1 = \sum_{i=1}^k p(c_i) = \sum_{i=1}^k p = kp \quad \text{y} \quad p = \frac{1}{k}$$

La probabilidad de que un elemento cualquiera  $u_j$ , sea seleccionado en la segunda etapa, es decir, una vez que lo ha sido su conglomerado  $c_i$ , es:

$$p(u_j) = p(c_i) p(u_j/c_i) = \frac{1}{k} \frac{1}{M} = \frac{1}{N}$$

donde  $N = kM$  es el tamaño de la población, medido en unidades elementales.

Vemos, pues, que todas las  $N$  unidades elementales de la población tienen una misma probabilidad de ser seleccionadas, la misma que si se realizara un muestreo aleatorio de elementos con iguales probabilidades.

## Conglomerados de diferente tamaño

La situación más usual, especialmente en el campo de las ciencias sociales, es contar con conglomerados de tamaños diferentes. Esta situación parece atender contra la equiprobabilidad. No obstante, ésta se mantiene siempre y cuando la probabilidad de selección de los conglomerados, durante la primera etapa, sea proporcional al tamaño de éstos.

La probabilidad de que un conglomerado  $c_i$  sea seleccionado en la primera etapa del muestreo, si se quiere considerar su tamaño  $n_i$  y asignarle una probabilidad proporcional, es:

$$p(c_i) = \frac{n_i}{N}$$

Como en todo espacio muestral, la suma de las probabilidades de los k conglomerados debe ser 1. Como el monto de todos los conglomerados forman el total poblacional ( $\sum_{i=1}^k n_i = N$ ), entonces:

$$\sum_{i=1}^k P(c_i) = \sum_{i=1}^k \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^k n_i = \frac{1}{N} N = 1$$

¿Cuál es la probabilidad de que un elemento  $u_j$  cualquiera, una vez que ha sido seleccionado su conglomerado  $c_i$ , sea obtenido en una extracción?. Si todos los elementos del mismo conglomerado tienen la misma probabilidad de ser seleccionados en una extracción, entonces:

$$P(u_j/c_i) = \frac{1}{n_i}, \quad \forall u_j \in c_i$$

Una vez seleccionado el conglomerado  $c_i$ , la suma de las probabilidades de selección para todos sus elementos debe ser 1:

$$\sum_{j=1}^{n_i} P(u_j/c_i) = \sum_{j=1}^{n_i} \frac{1}{n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} 1 = \frac{1}{n_i} n_i = 1$$

Con esta información, podemos responder a la pregunta ¿cuál es la probabilidad de que cualquier elemento, de entre todos los conglomerados, sea seleccionado en una extracción?:

$$P(u_j) = P(c_i) P(u_j/c_i) = \frac{n_i}{N} \frac{1}{n_i} = \frac{1}{N}$$

Si la muestra final tiene el tamaño  $m < N$ . ¿Cuál es la probabilidad de que un elemento cualquiera, de entre todos los conglomerados, se encuentre finalmente en la muestra?. Como posee m posibilidades con probabilidad  $1/N$  en cada una de ellas, entonces:

$$P(u_j) = \frac{m}{N}$$

La expresión anterior se denomina *fracción de muestreo*. Así pues, si todos los elementos de una población poseen la misma probabilidad de pertenecer finalmente a

la muestra, ésta probabilidad es la fracción de muestreo. En estos casos, en los que todas las unidades elementales *pesan* lo mismo, se consigue lo que llamamos **muestras autoponderadas**.

## **Probabilidades no proporcionales en la selección de conglomerados, durante la primera etapa**

En todo el proceso, la intención final ha consistido en realizar un muestreo de conglomerados, respetando la equiprobabilidad de las unidades elementales de la población. Para ello, es necesario que la probabilidad de selección de los conglomerados sea proporcional a su tamaño. No obstante, supongamos que se establece algún tipo de restricción que modifica las probabilidades de los conglomerados. Por ejemplo, unas zonas determinadas de la ciudad son seleccionadas necesariamente, dadas sus características peculiares, que interesan en función de los objetivos del estudio. Denominaremos en los sucesivos *conglomerados seleccionados necesariamente* a aquéllos cuya probabilidad de selección durante la primera etapa sea 1; es decir, a aquéllos que son seleccionados directamente, sin que medie ningún procedimiento aleatorio<sup>4</sup>.

Supongamos que el conglomerado  $c_h$  es seleccionado necesariamente; es decir,  $p(c_h)=1$ . En este caso, la probabilidad de selección, en una extracción, para cualquiera de sus  $n_h$  unidades elementales es:

$$p(u_j) = p(c_h) p(u_j/c_h) = 1 \frac{1}{n_h} = \frac{1}{n_h}$$

como  $n_h < N$ , la probabilidad de selección de cualquiera de los elementos del conglomerado  $h$  es superior al resto.

$$p(c_h) = \frac{1}{n_h} > \frac{1}{N} = p(c_{no\&h})$$

---

<sup>4</sup> Mirás (1985) trata este mismo concepto, en el contexto del muestreo estratificado de unidades elementales, con la denominación unidades que entran con certeza en la muestra (pág. 169).

Sea  $m_h$  el número de elementos del conglomerado  $c_h$  que forman parte de la muestra final. Cualquier elemento  $u_h$  de este conglomerado, posee una probabilidad de selección final de

$$P(u_h, j) = \frac{m_h}{n_h}$$

El resto de las unidades elementales de la población, tienen  $m - m_h$  posibilidades de ser seleccionadas y constituyen el  $N - n_h$  restante poblacional. Así pues, la probabilidad de que un elemento de un conglomerado no-h sea seleccionado finalmente, es:

$$P(u_{no\&h}, j) = \frac{m \& m_h}{N \& n_h}$$

Esta violación del principio de la equiprobabilidad tiene serias consecuencias en los procesos de inferencia, en el sentido de que complica innecesariamente los cálculos y exige recurrir a expresiones inusuales. Por otro lado, la comodidad durante el proceso de inferencia justifica débilmente el sacrificio de objetivos relevantes, como es nuestro caso: interesan unos determinados conglomerados.

## **Corrección de probabilidades para la selección de unidades elementales durante la segunda etapa**

Existe un recurso que permite cubrir ambos intereses: mantener los objetivos y salvar la equiprobabilidad. Si bien la selección de conglomerados no ha sido equitativa, dado que ésta es la primera etapa de una serie, existe el recurso de modificar las probabilidades de selección de los elementos dentro de los conglomerados.

La estrategia consiste, como puede resultar obvio, en igualar las expresiones para las probabilidades finales de selección de las unidades elementales, según pertenezcan o no al conglomerado seleccionado necesariamente. La intención se centra en despejar la expresión que define  $m_h$ , con el objetivo de asignarle el resultado de dicha expresión como la participación concreta del conglomerado  $c_h$ , en la muestra final  $m$ .

$$p(u_{h,j}) = \frac{m_h}{n_h} \cdot \frac{m \cdot m_h}{N \cdot n_h} \cdot p(u_{no\&h,j})$$

$$m_h (N \cdot n_h) = n_h (m \cdot m_h)$$

$$m_h (N \cdot n_h) \cdot n_h \cdot m_h = n_h \cdot m$$

$$m_h (N \cdot n_h \cdot n_h) = n_h \cdot m$$

$$m_h = m \frac{n_h}{N}$$

De esta forma, son necesarios  $m(n_h/N)$  elementos del conglomerado  $h$ , seleccionado necesariamente y, por tanto,  $m(1-n_h/N)$  del resto de los conglomerados seleccionados, en este último caso, aleatoriamente. Se observa que el conglomerado  $h$  mantiene la fracción de representación original ( $n_h/N$ ), mientras que el resto de los conglomerados, ya seleccionados, tienen una fracción de representación mayor, ya que el  $N'$  de la segunda etapa es mucho menor que el original, al restar el peso de los conglomerados no seleccionados.

La fracción de participación del conglomerado  $c_h$ , es decir, la razón entre sus elementos seleccionados finalmente y su tamaño, es:

$$m \left( \frac{n_h}{N} \right) \frac{1}{n_h} = \frac{m}{N}$$

Esta expresión, razón entre el tamaño de la muestra y el tamaño de la población, es lo que hemos denominado *fracción de muestreo*. Así pues, la fracción de muestreo se mantiene tras la segunda etapa, lo que permite aceptar el supuesto de la equiprobabilidad.

## Un ejemplo

Según el listado oficial del INE, actualizado al 1 de Enero de 1992, la provincia de Granada, cuenta con un total de 800.045 habitantes de derecho, que es utilizado como marco para el muestreo. Se desea obtener una muestra de 4500 habitantes y se decide obtenerla mediante muestreo bietápico a través de los municipios, como unidades muestrales de selección en la primera etapa.

Considerando los objetivos concretos de la investigación, se considera conveniente contar necesariamente con la capital en la etapa de selección aleatoria de conglomerados. La capital posee un tamaño de 259.702 habitantes.

En una selección aleatoria de conglomerados, con probabilidades proporcionales al tamaño, la probabilidad de selección de la capital granadina sería:

$$p(\text{Granada}) = \frac{259702}{800045} = 0,3246$$

No obstante, al contar necesariamente con este conglomerado, su probabilidad ha pasado a ser 1.

Son seleccionados aleatoriamente otros cinco municipios:

<u>Nombre</u>	<u>Tamaño</u>	<u>Porcentaje</u>
Alquife	1128	3.33
Cortes de Baza	2984	8.81
Almuñécar	20372	60.15
Juviles	217	0.64
La Zubia	9169	27.07
Total	33870	100.00

La participación que corresponde a la capital granadina es:

$$m_h = m \frac{n_h}{N} = 4500 \frac{259702}{800045} = 1460.7 \approx 1461$$

Así pues, un habitante de la capital, tiene una probabilidad de ser seleccionado de  $1461/259702=0'0056$ .

De los  $4500-1461=3039$  selecciones a realizar en los otros cinco municipios, en función de los porcentajes, se obtiene la siguiente tabla:

Nombre	Porcentaje	Participación
Alquife	3.33	101
Cortes de Baza	8.81	268
Almuñécar	60.15	1828
Juñeres	0.64	19
La Zubia	27.07	823
Total	100.00	3039

Un habitante de un pueblo de la provincia (que representa a  $800.045-259.702=540973$ ) tiene una probabilidad final de ser seleccionado de  $3039/540973=0'0056$ .

En otros términos, un habitante de la provincia de Granada, que tuviera noticias acerca de que se iba a realizar el estudio y acerca de este procedimiento, no modificaría en absoluto la probabilidad de ser seleccionado, mudándose de la capital hacia la provincia o viceversa.

## **Generalización al caso "selección necesaria de uno o más conglomerados"**

Sea  $r$  el número de conglomerados seleccionados necesariamente en la primera etapa del muestreo. Para cada uno de ellos, es aplicable la expresión

$$m_i = m \frac{n_i}{N}$$

donde

$n_i$  : tamaño poblacional del conglomerado  $c_i$

$m_i$  : participación muestral del conglomerado  $c_i$

$m$  : tamaño final de la muestra

N : tamaño de la población

El conjunto r de conglomerados seleccionados necesariamente, abarcará una participación de  $m_t$  unidades en la muestra final, con lo que restará  $m-m_t$  unidades elementales para el resto de los conglomerados, seleccionados aleatoriamente. El cálculo de  $m_t$  es:

$$m_t = \sum_{i=1}^r m \frac{n_i}{N} = \frac{m}{N} \sum_{i=1}^r n_i = m \frac{n_t}{N}$$

Supongamos el caso: lo que se ha considerado capital granadina en el ejemplo, es un macroconglomerado constituido, realmente, por los municipios A (78.945 h), B (113.053 h) y C (67.074 h).

La participación de cada conglomerado  $c_i$  seleccionado necesariamente (A, B, C)=, obedecerá a la expresión

$$m_i = m \frac{n_i}{N}$$

O bien:

$$m_i = m \frac{n_i}{N} = m \frac{n_i}{N} \frac{n_t}{n_t} = m \frac{n_t}{N} \frac{n_i}{n_t} = m_t \frac{n_i}{n_t}$$

Es decir, es factible partir de los resultados ya obtenidos para el macroconglomerado (A+B+C), de tamaño  $n_t$  y participación  $m_t$ .

Al conjunto formado por los tres conglomerados, seleccionados necesariamente en la primera etapa del muestreo, les corresponde la misma proporción de la muestra final, concretamente, 1461 unidades elementales. La siguiente acción consiste en repartir esta participación de forma proporcional al tamaño de cada conglomerado:

Municipio	Tamaño	Proporción	Participación
A	78945	30.47	445
B	113053	43.64	638

C	67074	25.89	378
Total	259072	100.00	1461

## Algoritmos de cálculo y selección

Los algoritmos expuestos hasta el momento, hacen referencia al cálculo de la frecuencia de participación  $m_i$  para cada conglomerado  $c_i$ , de forma que se consigue una muestra autoponderada. No obstante, es necesario tratar algunos aspectos más: características relevantes de la distribución de tamaños en los conglomerados y selección aleatoria de los  $c_i$ , proporcional a los tamaños  $n_i$ .

Es generalizado, en la bibliografía específica sobre teoría del muestreo, el consejo "*evitar, en la medida de lo posible, una excesiva dispersión en los tamaños de los conglomerados*". Además de efectos fáciles de intuir sobre las homogeneidad-heterogeneidad entre y dentro de los conglomerados, existen también consecuencias en las estimaciones, al aumentar los errores típicos de los estimadores (ver, al respecto, Mirás, 1985).

Una de las consecuencias negativas del tamaño desigual entre conglomerados, no resulta tan obvia. Lo usual en la fase de diseño de muestras, es encontrarse con un límite superior en el número de unidades a encuestar, bajo un criterio fundamentalmente económico. Por otro lado, las exigencias en la estimación, aconsejan no reducir el tamaño total de la muestra más allá de un mínimo, por debajo del cual, la significación de las conclusiones se debilita excesivamente. Se consigue, así, un intervalo admisible para el tamaño final de la muestra.

Kish (1965, pág.258) establece un criterio para controlar esta dispersión, en la fase en que deben decidirse los tamaños y número de conglomerados, contándose ya con un tamaño para toda la muestra, definido a través de un intervalo:

$$m = x(1 \pm t\alpha C) \quad C = \frac{V}{\sqrt{k}}$$

donde:

$m$ = tamaño de la muestra, expresado en el intervalo  $x \pm t\alpha C$

- x= tamaño media de la muestra, base para el intervalo
- $\alpha$ = medida de probabilidad.
- C= coeficiente de variación del tamaño muestral
- V= coeficiente de variación de Pearson sobre los tamaños de conglomerados
- k= nº de conglomerados

BIETAPN suministra el *intervalo de Kish* para la muestra, que corresponde a un  $\alpha=2.5$ . No obstante, el criterio de Kish es muy exigente y la mayoría de las veces en que se trabaja con conglomerados de tamaños diferentes, suministrará valores elevados para C y, por tanto, intervalos amplios.

El segundo aspecto que nos interesa, en este apartado sobre algoritmos de cálculo, es el procedimiento a seguir para la selección aleatoria aporportional de los conglomerados.

Una primera cuestión es la obtención de dígitos aleatorios, según una distribución uniforme, que sirva de base para la selección de los conglomerados. BIETAPN utiliza la generación por defecto de GWBASIC, en la versión que se muestra en este trabajo; así como un algoritmo de congruencia lineal, en la versión C. No obstante, éste es un tema que excede los objetivos del presente trabajo<sup>5</sup>

Con la generación de dígitos pseudoaleatorios resuelta, nuestra preocupación se centra en seleccionar los conglomerados no asignados necesariamente en la primera etapa.

Hansen y Hurwitz (1943) sugirieron una técnica aleatoria de selección, considerando el tamaño de los conglomerados, que podríamos denominar «método de las frecuencias acumuladas». Expuesto algorítmicamente:

---

<sup>5</sup> Dos trabajos clásicos sobre este tema son el de Tausworthe (1965), que se dedica monográficamente a la generación por recurrencia lineal; y el de Teichroew (1965), donde se realiza una recensión sobre procedimientos, pruebas de aleatoriedad y aplicaciones. Más en nuestros días, George Marsaglia tiene el don de generar algoritmos para crear números pseudoaleatorios; en la bibliografía se indica una obra significativa, de 1990.

1. Se construye una tabla de doble entrada. Los conglomerados se disponen en filas. Las columnas son: identificación del conglomerado, frecuencia absoluta y frecuencia absoluta acumulada.
2. Se obtiene un número aleatorio  $n_a$ , comprendido entre 1 y N.
3. Se comprueba en qué nivel se encuentra  $n_a$  en la columna de frecuencias acumuladas. Si  $F_{i-1} < n_a \leq F_i$ , el conglomerado  $c_i$  es seleccionado.
4. El proceso se repite hasta completar la selección de los k conglomerados necesarios.

Como se observa, este procedimiento contempla la reposición: el elemento  $c_i$  es devuelto a la urna (marco base) antes de la siguiente selección. La probabilidad de que un mismo  $c_i$  se repita en la muestra es menor cuanto mayor sea el número de conglomerados en la población y menor el número de éstos que son seleccionados. Sin embargo, tender a esta relación es contraproducente con respecto a la discusión que se ha abordado en el intervalo de Kish: disminuir el número de conglomerados hace ampliar el intervalo, al aumentar el valor del coeficiente C de variación del tamaño muestral.

BIETAPN considera un método de las frecuencias acumuladas corregido, que no contempla la reposición de las unidades  $c_i$  ya seleccionadas; es decir, establece un muestreo aleatorio proporcional sin reposición.

Lahiri (1951)<sup>6</sup>, propuso un procedimiento alternativo que podemos denominar «método del par aleatorio». Lo expondremos, igualmente, según un algoritmo:

1. Se registra el tamaño del conglomerado mayor,  $n_m$ .
2. Se generan dos números aleatorios  $i, j$  tal que:

$$1 \leq i \leq k \quad (k = \text{n}^\circ \text{ de conglomerados})$$

$$1 \leq j \leq n_m$$

---

<sup>6</sup> A method of sample selection providing unbiased ratio estimates. Bulletin of the international statistical institute. Vól 33 (1951), págs. 133-140. Citado por Azorín y Sánchez-Crespo (1986).

3. Si  $j \neq n_i$ , el conglomerado  $c_i$  es seleccionado y se continúa el proceso por el paso 2, hasta finalizar la selección del número preestablecido de conglomerados para la segunda etapa. Si  $j > n_i$ , no es seleccionado ningún conglomerado y se continúa por el paso 2.

Como se observa, este algoritmo contempla igualmente la reposición. Para establecer una selección proporcional sin reposición, hay que realizar una de las siguientes modificaciones:

1. En el paso 3, si  $j \neq n_i$ , añadir la condición de que  $c_i$  no haya sido seleccionado. En el caso de que  $j \neq n_i$  pero  $c_i$  ya se encuentra en la muestra, el comportamiento debe ser idéntico al de  $j > n_i$ .
2. En cada caso  $j \neq n_i$ ,  $c_i$  es seleccionado y, a su vez, eliminado de la lista. El flujo del algoritmo debe continuar por el paso 1, en lugar del 2. Así mismo,  $k$  debe decrecer en una unidad.

Los procedimientos de las frecuencias acumuladas y del par aleatorio son equivalentes en la práctica, como se puede demostrar algebraicamente mediante el cálculo de esperanzas matemáticas. Si algún procedimiento de simulación advirtiera alguna discrepancia, ésta sería achacable a una generación errónea de dígitos pseudoaleatorios, pero no a diferencias en los métodos.

## Utilización del programa BIETAPN.BAS

El objetivo del programa es realizar la selección de conglomerados y obtener la participación de las unidades seleccionadas, en la muestra final. Para ello, necesita la siguiente información:

1. Referente a los conglomerados:

. NUMC: Número de conglomerados que se van a utilizar como unidades de muestreo en la primera etapa (unidades primarias).

. NSEL: Número de conglomerados que se desean seleccionar.

- . TAMi: Tamaño del conglomerado  $c_i$ ; número de unidades elementales (o secundarias) de que se compone.
- . CODi: Código de modo de selección. Si se especifica 1, el programa entiende que el conglomerado  $c_i$  es seleccionado necesariamente en la primera etapa. Si se especifica 0, el programa entiende que el conglomerado  $c_i$  será sometido a una selección proporcional aleatoria.
- . MUESTRA: Tamaño final de la muestra, tras la selección de unidades elementales en la segunda etapa.

2. Referente a los archivos implicados:

- . ENT: nombre del archivo de disco en el que se encuentran los datos.
- . SAL: nombre del archivo de disco en el que se escribirán los resultados del proceso.
- . CONF: (Opcional) nombre del archivo de configuración que contiene toda la información que necesita BIETAPN para su funcionamiento.
- . COLANT: número de columnas de datos irrelevantes, situadas antes de las columnas TAM y COD, en ENT.
- . COLDES: número de columnas de datos irrelevantes, situadas después de las columnas TAM y COD, en ENT.

El archivo de datos (ENT), debe poseer la siguiente estructura:

```
(COLANT datos anteriores) COD1 TAM1 (COLDES datos posteriores)
..... COD2 TAM2 .....
.....
..... CODi TAMi .....
.....
..... CODk TAMk .....
```

Para utilizar BIETAPN es necesario contar en el entorno con el intérprete GWBASIC (si bien, otras versiones BASIC pueden ejecutarlo también). Lo

recomendable es crear un directorio BIETAPN y copiar GWBASIC.EXE, BIETAPN.BAS y el archivo ASCII de ejecución por lotes BIETAPN.BAT, que gestiona la utilización del programa y cuyo listado se acompaña también al final del trabajo.

Una vez situado en el directorio BIETAPN, basta con introducir<sup>7</sup>:

bietapn [CONF]

Si se tecllea únicamente "bietapn", el programa preguntará, al usuario toda la información que necesita. Si se incluye en la llamada, el nombre de un archivo CONF, la ejecución será automática, sin intervención del usuario.

Una vez en el programa BIETAPN, en formato interactivo, éste realizará siete peticiones de información, en el siguiente orden:

ENT

SAL

NUMC

NSEL

MUESTRA

COLANT

COLDES

Para evitar este proceso interactivo, se puede crear un archivo de disco, en código ASCII, cuya estructura interna está compuesta por siete líneas, cada una de las cuales posee las respectivas informaciones que necesita el programa para su ejecución. Este es el archivo de configuración que hemos codificado en el texto como CONF.

---

<sup>7</sup> No es necesario encontrarse en el directorio de BIETAPN, siempre que esta situación se contemple, modificando convenientemente BIETAPN.BAT. Para más información al respecto, consultar el manual de referencia del MS-DOS.

Por ejemplo:

1. Archivo de entrada (ENT), que llamamos "datos.uno":

```
01 0 11234
02 1 14180
03 0 17364
04 0 17384
05 0 13796
06 0 15263
07 0 15431
08 1 11912
09 0 17384
10 0 18374
11 0 16050
12 0 16909
```

2. Archivo de configuración (CONF), que llamamos "uno":

```
datos.uno
resulta.uno
12
6
3200
1
0
```

3. Llamada al programa:

```
bietapn uno
```

4. Archivo de salida (SAL), que llamamos "resulta.uno":

```
Bietápico : selección aleatoria de conglomerados
con unidades muestrales asignadas en primera etapa
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))
```

1. Información general desde «datos.uno»:

```
Número de conglomerados:          12
```

Unidades primarias: 6  
 Unidades secundarias: 3200  
 Columnas anteriores: 1  
 Columnas posteriores: 0

2. Sobre la distribución de tamaños:

Tamaño medio: 15440.08 unidades elementales  
 Varianza entre conglomerados: 4738130  
 Variación de Pearson: 306.872  
 Intervalo de Kish ( $t\alpha=2.5$ ): 0 <> 711890.5613718853  
 Tamaño total de la población: 185281

3. Etapa 1: Conglomerados seleccionados:

Existen 2 con selección necesaria

Orden	Número	Tamaño	Clave
))))	))))))	))))))	))))
1	2	14180	1
2	8	11912	1
3	1	11234	0
4	10	18374	0
5	4	17384	0
6	5	13796	0

4. Etapa 2: Selección de unidades elementales

Número	Tamaño	%Pobl.	%Muestra	Participación
))))))	))))))	))))))	))))))	))))))
2	14180	7.65	7.65	245
8	11912	6.43	6.43	206
1	11234	6.06	15.88	508
10	18374	9.92	25.97	831
4	17384	9.38	24.57	786
5	13796	7.45	19.50	624

## Efectos en el control de los códigos de selección

Si bien BIETAPN está creado para facilitar la selección de conglomerados en función de un tamaño y de su modo de selección, así como para calcular las participaciones respectivas que no afecten a la fracción de muestreo, puede ser utilizado en tres contextos diferentes:

1. Muestreo de conglomerados con tamaños desiguales y probabilidades de selección proporcionales al tamaño, con selección necesaria de algunas unidades primarias. Submuestreo (segunda etapa) de los conglomerados seleccionados, con la obtención final de una muestra autoponderada. Éste es el modelo comentado y al que obedece la existencia del presente trabajo.
2. Muestreo bietápico de conglomerados con probabilidades proporcionales a su tamaño. Este modelo surge cuando todos los códigos de modo de selección CODi son 0. Si, además, los tamaños de los conglomerados coinciden, la utilidad de BIETAPN se reduce a un muestreo simple aleatorio de conglomerados con iguales probabilidades. La etapa 2 se torna, entonces, irrelevante, ya que se entra en la repetición de un mismo registro informativo.
3. En el caso en que todos los códigos de modo de selección sean 1, estaremos en el contexto de un muestreo aleatorio estratificado con afijación proporcional y una sola etapa, ya que, en la primera, todos los conglomerados (en este caso, estratos) serían seleccionados. La utilidad se centra, pues, en la segunda etapa BIETAPN (única en este modelo), mediante la expresión:

$$m_i = m \frac{n_i}{N} \quad \text{Y} \quad \frac{m_i}{m} = \frac{n_i}{N}$$

La proporción del estrato en la población ( $n_i/N$ ), se mantiene en la muestra ( $m_i/m$ ), consiguiéndose la afijación proporcional.

Como puede deducirse de la expresión anterior, el modelo que se presenta en este trabajo, para abordar la selección necesaria de conglomerados en una primera etapa, puede ser tratado bajo otra perspectiva: formación de estratos de conglomerados y selección con afijación proporcional posterior, con la particularidad de que algunos estratos poseen una única unidad primaria.

## Listado del programa BIETAPN

(bietapn.bat)

```
@ECHO OFF
IF EXIST %1 GOTO HAY
GWBASIC BIETAPN
GOTO FIN
:HAY
GWBASIC BIETAPN < %1
:FIN
@ECHO ON
```

(bietapn.bas)

```
10 REM *****
20 REM * BIETAPN: bietápico con selecciones *
30 REM * asignadas en primera etapa *
40 REM *****
50 REM
60 REM          Lectura de las variables del entorno
70 REM
80 KEY OFF
90 RANDOMIZE TIME
100 INPUT "Archivo de entrada para los datos: ",ENT$
110 INPUT "Archivo de salida para los resultados: ",SAL$
120 INPUT "Número de conglomerados de la población: ",NUMC
130 INPUT "Número de conglomerados a seleccionar: ",NSEL
140 INPUT "Tamaño final de la muestra: ",MUESTRA#
150 INPUT "Número de columnas irrelevantes antes del código: ",COLANT
160 INPUT "Número de columnas irrelevantes después del código: ",COLDES
170 OPEN ENT$ FOR INPUT AS #1
180 OPEN SAL$ FOR OUTPUT AS #2
190 REM
200 REM          Presentación en archivo y primeras informaciones
210 REM
220 PRINT #2, "          Bietápico : selección aleatoria de
conglomerados"
230 PRINT #2, "          con unidades muestrales asignadas en primera
etapa"
240 PRINT #2, "
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))"
250 PRINT #2, : PRINT #2,
260 PRINT #2, " 1. Información general desde «";ENT$;"»:" : PRINT #2,
270 PRINT #2, "          Número de conglomerados: ",NUMC
280 PRINT #2, "          Unidades primarias:          ",NSEL
290 PRINT #2, "          Unidades secundarias:          ",MUESTRA#
300 PRINT #2, "          Columnas anteriores:          ",COLANT
310 PRINT #2, "          Columnas posteriores:          ",COLDES
320 REM
330 REM          Se completan los arrays, se leen los datos
340 REM          y se realizan los primeros cálculos
350 REM
360 DIM CONG#(NUMC), ORDEN(NUMC), CODIGO(NUMC)
370 DIM VALE(NSEL), TIENE#(NSEL), COD(NSEL)
380 FOR A=1 TO NUMC : ORDEN(A)=A : NEXT A
390 TOTAL#=0 : TCUA#=0 : HAY=NUMC
400 FOR A=1 TO NUMC
410     FOR B=1 TO COLANT : INPUT #1, NADA : NEXT B
```

```

420     INPUT #1, CODIGO(A), CONG#(A)
430     FOR B=1 TO COLDES : INPUT #1, NADA : NEXT B
440     TOTAL#=TOTAL#+CONG#(A)
450     TCUA#=TCUA#+CONG#(A)^2
460     NEXT A
470     CLOSE #1
480     REM
490     REM           Cálculos de estadísticos sobre
500     REM           los tamaños de los conglomerados
510     REM
520     MED=TOTAL#/NUMC
530     VAR=(TCUA#/NUMC)-MED^2
540     PEAR=VAR/MED
550     KISH#=MUESTRA#*PEAR*2.5/SQR(NUMC)
560     KISH1#=MUESTRA#-KISH# : IF KISH1#<0 THEN KISH1#=0
570     KISH2#=MUESTRA#+KISH#
580     PRINT #2, : PRINT #2,
590     PRINT #2, "     2. Sobre la distribución de tamaños:" : PRINT #2,
600     PRINT #2, "           Tamaño medio:";MED;"unidades elementales"
610     PRINT #2, "           Varianza entre conglomerados:",VAR
620     PRINT #2, "           Variación de Pearson:           ",PEAR
630     PRINT #2, "           Intervalo de Kish ( $\alpha=2.5$ ): ",KISH1#;"<>";KISH2#
640     PRINT #2, "           Tamaño total de la población:",TOTAL#
650     REM
660     REM           Se rellenan los arrays para la presentación de
selecciones
670     REM           Procedimiento de selección especial para los asignados
680     REM           y de Hansen y Hurwitz corregido para los aleatorios
690     REM
700     REM * ASIGNADOS *
710     ES=0 : POB#=0
720     FOR A=1 TO NUMC
730     IF CODIGO(A)=0 THEN POB#=POB#+CONG#(A) : GOTO 760
740     ES=ES+1
750     VALE(ES)=A : COD(ES)=1 : TIENE#(ES)=CONG#(A)
760     NEXT A
770     REM * ALEATORIOS *
780     NUMC=NUMC-ES
790     FOR A=ES+1 TO NSEL
800     CLAVE#=INT(RND(1)*POB#)+1
810     SUMA#=0
820     FOR B=1 TO NUMC
830     IF CODIGO(B)=0 THEN 850
840     IF B=NUMC THEN 920 ELSE 870
850     SUMA#=SUMA#+CONG#(B)
860     IF CLAVE#<SUMA# THEN 880
870     NEXT B
880     VALE(A)=ORDEN(B) : COD(A)=0 : TIENE#(A)=CONG#(B)
890     POB#=POB#-CONG#(B)
900     CONG#(B)=CONG#(NUMC) : ORDEN(B)=NUMC
910     NUMC=NUMC-1
920     NEXT A
930     REM
940     REM           Presentación de los conglomerados seleccionados
950     REM
960     PRINT #2, : PRINT #2,
970     PRINT #2, "     3. Etapa 1: Conglomerados seleccionados:"
980     PRINT #2, "           Existen";ES;"con selección necesaria" : PRINT #2,
990     PRINT #2, "           Orden     Número     Tamaño     Clave"
1000    PRINT #2, "           )))))  )))))))  )))))))  )))))))"
1010    FOR A=1 TO NSEL
1020    PRINT #2, USING "           #### ";A;

```

```

1030 PRINT #2, USING "      ### " ;VALE (A) ;
1040 PRINT #2, USING "      #####";TIENE# (A) ;
1050 PRINT #2, "      " ;COD(A)
1060 NEXT A
1070 REM
1080 REM                      Cálculos para las participaciones de los
seleccionados
1090 REM                      Simultáneamente, se presentan los resultados
1100 REM
1110 PRINT #2, : PRINT #2,
1120 PRINT #2, "      4. Etapa 2: Selección de unidades elementales"
1130 PRINT #2,
1140 PRINT #2, "          Número          Tamaño          %Pobl.          %Muestra
Participación"
1150 PRINT #2, "          )))))))          )))))))          )))))))          )))))))
)))))
1160 VAN#=0
1170 FOR A=1 TO NSEL
1180   IF COD(A)=0 THEN 1250
1190   PAR#=MUESTRA#*TIENE# (A) /TOTAL#
1200   VAN#=VAN#+PAR#
1210   PORM=PAR#*100/MUESTRA#
1220   PORP=TIENE# (A) *100/TOTAL#
1230   GUIA=A : GOSUB 1380
1240 NEXT A
1250 POB#=0 : QUEDA#=MUESTRA#-VAN#
1260 FOR B=A TO NSEL : POB#=POB#+TIENE# (B) : NEXT B
1270 FOR B=A TO NSEL
1280   PAR#=QUEDA#*TIENE# (B) /POB#
1290   PORP=TIENE# (B) *100/TOTAL#
1300   PORM=PAR#*100/MUESTRA#
1310   GUIA=B : GOSUB 1380
1320 NEXT B
1330 PRINT #2, : CLOSE #2
1340 PRINT : PRINT : SYSTEM
1350 REM
1360 REM          SUBROUTINAS
1370 REM
1380 PRINT #2, USING "      ### " ;VALE (GUIA) ;
1390 PRINT #2, USING "      #####";TIENE# (GUIA) ;
1400 PRINT #2, USING "      ##.##";PORP;
1410 PRINT #2, USING "      ##.## " ;PORM;
1420 PRINT #2, USING "      #####";PAR#
1430 RETURN
1440 REM
1450 REM Chento · 1993

```

## Nota sobre otras versiones

Los programas interpretados en BASIC, tiene la ventaja de generar poco código. El proyecto que se ha abordado en este trabajo no necesita de un listado excesivamente amplio de instrucciones. Prácticamente, la mitad del código se emplea en informar al usuario. Cuando el proyecto se hace más ambicioso, de forma que el tamaño del código fuente debe crecer sensiblemente, empieza a ser aconsejable abandonar GWBASIC y utilizar un código fuente más estructurado que, además, permita ser compilado.

Por esta razón, el programa cuyo listado se incluye en este trabajo tiene unos objetivos claros, pero limitados. El autor ha elaborado otra versión, BIETAPN.COM, compilada en C, cuyo tamaño de código hace prohibitiva su exposición escrita. Algunas diferencias con BIETAPN.BAS son:

1. Deja libre memoria RAM, utilizando el disco, lo que hace prácticamente ilimitada su capacidad para procesar cualquier conjunto de conglomerados, con cualquier tamaño.
2. Compilado y en C, es de ejecución mucho más rápida que la versión BIETAPN.BAS, a pesar de explotar preferentemente disco, en lugar de RAM.
3. Con fines de comprobación y simulación, permite escoger entre los procedimientos corregidos de Hansen y Hurwitz *versus* Lahiri, para la selección proporcional de conglomerados.
4. Opcionalmente, pueden variarse los parámetros del generador de dígitos pseudoaleatorios, puesto que es el mismo programa el que los genera (no una función del compilador).
5. Realiza controles de error e informa de éstos.

## Bibliografía

Azorín, Francisco; Sánchez-Crespo, José Luis. **Métodos y aplicaciones de muestreo**. Colección Alianza Universidad Textos. Madrid: Alianza Editorial, 1986.

Cochran, William G. **Técnicas de muestreo**. Edición original inglesa (no consta fecha). México: Compañía Editorial Continental, 1976.

Hansen, Morris H.; Hurwitz, William N. **On the theory of sampling from finite populations**. *The Annals of mathematical Statistics*. Vol. 14 (1943), págs 333-362.

Kish, Leslie. **Muestreo de encuestas**. Edición original inglesa, 1965. México: Trillas, 1982.

Manzano, Vicente. **Análisis estadísticos con el SPSS/PC+. Fundamentos de análisis, preliminares, estudios descriptivos y utilidades**. Madrid: RA-MA, 1992.

Mirás, Julio. **Elementos de muestreo para poblaciones finitas**. Madrid: Instituto Nacional de Estadística, 1985.

Tausworthe, Robert C. **Random numbers generated by linear recurrence modulo two**. *Mathematics of Computation*. Vol. 19 (1965), págs. 201-209.

Teichroew, Daniel. **A history of distribution sampling prior to the era of the computer and its relevance to simulation**. *Journal of the American Statistical Association*, Marzo 1965, págs. 27-49.