UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2010, Vol. 3, No. 1

# Afijación óptima de tamaños de muestra en muestreo aleatorio estratificado vía programación matemática

## Optimum allocation of sample sizes in stratified random sampling via mathematical programming

Alfonso Sánchez[a]
asanchez@ut.edu.co

Leonardo Solanilla[b]
lsolanilla@ut.edu.co

Jairo Clavijo[c]
jaclavijo@ut.edu.co

Alex Zambrano[d]
azambrano@ut.edu.co

**Resumen**

Algunos problemas de afijación en el muestreo son considerados resueltos por técnicas de programación matemática. La afijación óptima de tamaños de muestra en Muestreo Aleatorio Estratificado (M.A.E.) son considerados problemas de programación dinámica. En el caso multivariado el problema de programación convexa es fundamental para la identificación y los métodos para solucionarlos son indicados. En este articulo, se presentan los problemas de estimación de tamaños de muestra en muestreo aleatorio estratificado univariado y multivariado, siguiendo las ideas expuestas por Arthanari & Dodge (1981). Finalmente, se ilustra mediante un ejemplo el método de morral.

***Palabras clave*:** muestreo aleatorio estratificado, Knapsack, optimización, programación matemática.

**Abstract**

Some sampling allocation problems which can be solved by mathematical programming techniques are considered. Optimal allocation of sample sizes in Stratified Random Sampling (SI.), for example, can be regarded as a dynamic programming problem. In the multivariate case, the underlying convex–programming problem is stated and some solution methods are indicated. We have followed the illuminating

[a]Profesor de planta. Universidad del Tolima.
[b]Profesor de planta. Universidad del Tolima.
[c]Profesor de planta. Universidad del Tolima.
[d]Profesor cátedra. Universidad del Tolima.

ideas exposed in Arthanari & Dodge (1981). Finally, an example to illustrate the so-called Knapsack method is presented.

***Key words*:** mathematical programming, Knapsack, optimization, stratified random sampling.

# 1. Introducción

Tal como lo señala Arthanari & Dodge (1981, pp.216) en macroeconomía y macroeconomía se necesita establecer información a través de encuestas, enumeración de individuos, etc.

La teoría del muestreo ayuda a la selección de muestras de una población según ciertos mecanismos probabilísticos. Lo más simple es asignar igual probabilidad a cada unidad de la población en la muestra (Muestreo Aleatorio Simple, M.A.S.). Se asocia con o sin reemplazo. Otra forma simple es asignar probabilidades distintas con base en algún otro criterio en una medida de tamaño (Probabilidad Proporcional al Tamaño, P.P.T.).

Dentro del diseño de muestreo se consideran diversas maneras de seleccionar unidades y de calcular los tamaños de las muestras.

El problema de derivar información estadística sobre una característica de una población a partir de una muestra de datos puede formularse como un problema de optimización.

**Problema 1.**

$$\begin{aligned} \text{mín} \quad & \textit{Costo de Encuesta.} \\ \textbf{\textit{s. a.}} \quad & \textit{La pérdida de precisión quede dentro de ciertos límites.} \end{aligned}$$

Esta función objetivo depende del tamaño de la muestra, tamaño de las unidades de muestreo, el plan de muestreo y lo que busca la encuesta. Alternativamente se puede tomar el problema 1 de la siguiente manera:

$$\begin{aligned} \text{mín} \quad & \textit{La perdida de precisión.} \\ \textbf{\textit{s. a.}} \quad & \textit{El costo se acomode a un presupuesto.} \end{aligned}$$

Para nuestro trabajo estamos interesados en el problema concreto de afijación óptima de tamaños de muestra en M.A.E, ilustraremos el uso de la programación matemática para escoger el tamaño de muestra en diferentes situaciones.

# 2. Afijación óptima para el M.A.E. univariado

Se darán algunos conceptos y definiciones básicas del M.A.E., se recomienda la lectura de un libro clásico como Cochran (1977).

Sea $U = \{U_1, U_2, \ldots, U_N\}$ una población particionada en subpoblaciones o estratos. Las características de la población se infieren a partir de las muestras de cada uno de los estratos, explotando la ganancia de precisión de las estimaciones.

Sea $N_i$ el número de unidades del estrato $i$-ésimo y $\sum_{i=1}^{L} N_i = N$, donde $L$ es el número de estratos, $N$ es el número de unidades en la población, mientras que $n_i$ el tamaño de la muestra del $i$-ésimo estrato. Se asume que las muestras se toman independientemente en cada estrato.

El problema de afijación óptima es determinar los $n_i$, el objetivo en este problema es minimizar la varianza de la estimación de la característica de la población bajo estudio (ganancia de precisión), la restricción es el número de muestras tomadas (presupuesto disponible), esto con el fin de minimizar el costo total del muestreo para la precisión deseada.

Considérese la estimación insesgada de la media poblacional $\overline{Y}$, donde $Y$ es la característica bajo estudio. Sea $\overline{y}_i$ la estimación insesgada de la media $\overline{Y}_i$ (en el estrato $i$-ésimo), esto es,

$$\overline{y}_i = \frac{1}{n_i} \sum_{h=1}^{L} y_{ih}.$$

Sea $\overline{y}_{st}$ dada por

$$\overline{y}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{y}_i$$

el cual es una estimación insesgada de la media poblacional $\overline{Y}$.

La precisión de esta estimación está medida por la varianza de la estimación muestral. Así consideremos $V(\overline{y}_{st})$ definida como

$$V(\overline{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 V(\overline{y}_i)$$

donde

$$V(\overline{y}_i) = \frac{1}{N_i - 1} \sum_{h=1}^{N_i} \left(y_{ih} - \overline{Y}_i\right)^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right).$$

Ya definidos los conceptos anteriores, determinemos el problema de escoger $n_i$ $(i = 1, \ldots, L)$, tal que $\sum_{i=1}^{L} n_i = n$, y la $V(\overline{y}_{st})$ es mínima. Esto es

**Problema 2.**

$$\text{mín} \quad V(\overline{y}_{st}) = \sum_{i=1}^{L} W_i^2 S_i^2 x_i \quad \textit{(Función Objetivo)}$$

$$\textbf{\textit{s. a.}} \quad \sum_{i=1}^{L} n_i = n \quad \textit{(Función Lineal)} \tag{R1}$$

$$1 \leq n_i \leq N_i,\ n_i \in \mathbb{Z}^+ \quad (i = 1 \ldots, L) \tag{R2}$$

*donde* $W_i = N_i/N$, $S_i^2 = \sum_{h=1}^{N_i} \left(y_{ih} - \overline{Y}_i\right)^2 / (N_i - 1)$ *y* $x_i = 1/n_i - 1/N_i$.

Sea $a_i = W_i^2 S_i^2$ $(i = 1, \ldots, L)$, entonces la función objetivo se puede escribir de la siguiente manera

$$\sum_{i=1}^{L} W_i^2 S_i^2 x_i = \sum_{i=1}^{L} a_i x_i = \sum_{i=1}^{L} a_i \left( \frac{1}{n_i} - \frac{1}{N_i} \right) = \sum_{i=1}^{L} \frac{a_i}{n_i} - \sum_{i=1}^{L} \frac{a_i}{N_i}$$

Puesto que $\sum_{i=1}^{L} a_i/N_i$ es constante, entonces es suficiente considerar minimizar $\sum_{i=1}^{L} a_i/n_i$. Reescribiendo el problema 2 tenemos:

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{L} \frac{a_i}{n_i} = Z \quad (\textit{Función Objetivo}) \\ \textbf{\textit{s.a.}} \quad & \sum_{i=1}^{L} n_i = n \qquad \text{(R1)} \\ & 1 \leq n_i \leq N_i,\ n_i \in \mathbb{Z}^+ \quad (i = 1 \ldots, L). \qquad \text{(R2)} \end{aligned}$$

Notamos que la función objetivo es no lineal, además si olvidamos (R2) podemos usar *Multiplicadores de Lagrange* para encontrar $n_i$ de cada muestra. Recordemos que minimizar $V(\overline{y}_{est})$ es equivalente a minimizar:

$$f(n_1, \ldots, n_L) = \sum_{i=1}^{L} \frac{a_i}{n_i} \quad \textbf{\textit{s.a.}} \quad g(n_1, \ldots, n_L) = \sum_{i=1}^{L} n_i - n = 0.$$

Hallando $\nabla f$ y $\nabla g$ tenemos:

$$\nabla f = \left( -\frac{a_1}{n_1^2}, \ldots, -\frac{a_L}{n_L^2} \right), \quad \nabla g = (1, \ldots, 1) = \mathbb{1}.$$

Nótese que $\nabla g \neq 0$. Por la condición de Lagrange $\nabla f - \lambda \nabla g = 0$, bajo la restricción $g(n_1, \ldots, n_L) = 0$, es decir:

$$-\frac{a_i}{n_i^2} - \lambda = 0 \quad (i = 1, \ldots, L) \quad \textbf{\textit{s.a.}} \quad \sum_{i=1}^{L} n_i = n.$$

Considerando $\lambda$ negativo y solucionando las ecuaciones tenemos:

$$n_i = \sqrt{\frac{a_i}{\lambda}} \quad (i = 1, \ldots, L) \quad \textbf{\textit{s.a.}} \quad n = \frac{\sum_{i=1}^{L} \sqrt{a_i}}{\sqrt{\lambda}}.$$

Despejando de la restricción se tiene:

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{L} \sqrt{a_i}}{n}.$$

Por tanto

$$n_i = \frac{\sqrt{a_i}}{\sqrt{\lambda}} = \frac{n\sqrt{a_i}}{\sum_{k=1}^{L} \sqrt{a_k}}.$$

*Nota.* Se pueden presentar los siguientes problemas:

- $n_i > N_i$, entonces, aquí se tendría que redistribuir, ya que se presenta un sobremuestreo.
- $n \notin \mathbb{Z}^+$, entonces, se redondea.
- $n_i < 1$, entonces, tendríamos que tomar como mínimo una unidad en cada estrato.

## 2.1. Interpretación gráfica

Se ha notado que la función objetivo $Z$ es convexa si $a_i > 0$, para todo $i$. Nuestro interés es minimizar una función estrictamente convexa sobre una región convexa, creada por una ecuación lineal y $2L$ restricciones acotadas arriba y abajo. Cuando $L = 2$ la región factible y la región objetivo pueden ser observadas en la figura 1.

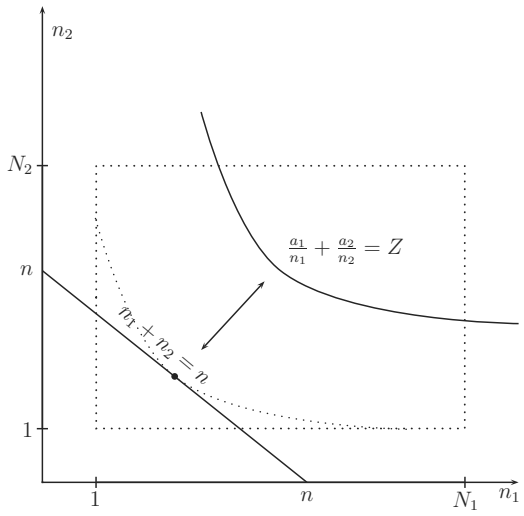


Figura 1: *Región factible y función objetivo cuando $N_1$ y $N_2$ son ambos más grandes que $n$.*

Fuente: (Arthanari & Dodge 1981)

La situación que se observa en la figura 1 es ideal, por un lado tanto $N_1$ y $N_2$ son más grandes que $n$ y además el óptimo es el punto de tangencia de la recta con la familia de hipérbolas y se encuentra dentro de la región factible.

La situación que se observa en la figura 2 no es ideal, puesto que el óptimo se encuentra fuera de la región factible, por lo cual el tamaño de muestra $n_1^* > N_1$. Un problema de sobremuestreo.

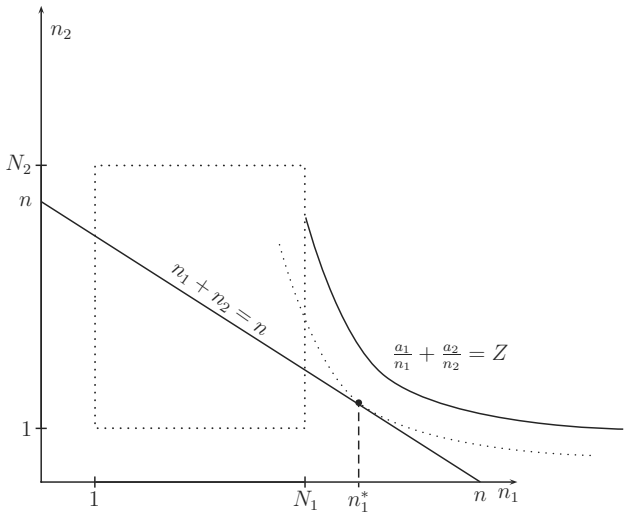


Figura 2: *Región factible y función objetivo cuando únicamente $N_2$ es más grande que $n$.*

Fuente: (Arthanari & Dodge 1981)

Tenemos una función objetivo no lineal $Z$ y una restricción de igualdad lineal

$$n_1 + \cdots + n_L = n$$

y restricciones acotadas arriba y abajo sobre $n_i$,

$$1 \leq n_i \leq N_i \quad (i = 1, \ldots, L) \quad n_i \in \mathbb{Z}^+$$

Un método de programación lineal que resuelve este problema es el método Knapsack que se describirá en la siguiente sección.

## 2.2. Método de iteración Knapsack

El método knapsack (que podríamos traducir como método del morral) es un método iterativo de programación matemática entera, el cual puede ser usado para resolver una función objetivo no lineal con restricciones lineales como se presenta en el problema 2.

Se define $f(k, r)$, donde $k$ maneja el estrato, $1 \leq k \leq L$, $r$ maneja el tamaño de la muestra en el estrato, $0 \leq r \leq N_i$, $r \in \mathbb{Z}^+ \cup \{0\}$, tal que

$$f(k,r) = \begin{cases} \text{mín} & \sum_{i=1}^{k} \frac{a_i}{n_i} \\ \boldsymbol{s.a} & \sum_{i=1}^{k} n_i = r,\ n_i \in \mathbb{Z}^+,\ 1 \leq n_i \leq N_i \quad (i = 1, \ldots, k) \end{cases}$$

Se describe a continuación este método encontrando sucesivamente $f(k, r)$ para cada estrato

**Paso 1:**

$$f(1,r) = \begin{cases} \text{mín} & \left\{\frac{a_1}{n_1}\right\} \\ \boldsymbol{s.a} & n_1 = r,\, 1 \leq n_1 \leq N_1 \end{cases}$$

solución:

$$f(1,r) = \begin{cases} \frac{a_1}{r} & 1 \leq r \leq N_1 \\ \infty & r < 1 \,\text{ó}\, r > N_1 \end{cases}$$

**Paso 2:**

$$f(2,r) = \begin{cases} \text{mín} & \left\{\frac{a_1}{n_1} + \frac{a_2}{n_2}\right\} \\ \boldsymbol{s.a} & n_1 + n_2 = r,\, 1 \leq n_i \leq N_i,\, n_i \in \mathbb{Z}^+,\, i = 1,2 \end{cases}$$

si fijamos $n_2$, la función anterior se reduce a:

$$\begin{aligned} f(2,r) &= \frac{a_2}{n_2} + \begin{cases} \text{mín} & \left\{\frac{a_1}{n_1}\right\} \\ \boldsymbol{s.a} & n_1 = r - n_2,\, 1 \leq n_1 \leq N_1,\, n_1 \in \mathbb{Z}^+ \end{cases} \\ &= \frac{a_2}{n_2} + f(1, r - n_2) \end{aligned}$$

pero como $n_2$ no es fijo

$$f(2,r) = \min_{n_2=1,\ldots,n} \left\{ \frac{a_2}{n_2} + f(1, r - n_2) \right\}$$

**Paso 3:**

$$f(3,r) = \begin{cases} \text{mín} & \left\{\frac{a_1}{n_1} + \frac{a_2}{n_2} + \frac{a_3}{n_3}\right\} \\ \boldsymbol{s.a} & n_1 + n_2 + n_3 = r,\, 1 \leq n_i \leq N_i,\, n_i \in \mathbb{Z}^+,\, i = 1,2,3 \end{cases}$$

si fijamos $n_3$, la función anterior se reduce a:

$$\begin{aligned} f(3,r) &= \frac{a_3}{n_3} + \begin{cases} \text{mín} & \left\{\frac{a_1}{n_1} + \frac{a_2}{n_2}\right\} \\ \boldsymbol{s.a} & n_1 + n_2 = r - n_3,\, 1 \leq n_1 \leq N_1,\, 1 \leq n_2 \leq N_2,\, n_1, n_2 \in \mathbb{Z}^+ \end{cases} \\ &= \frac{a_3}{n_3} + f(2, r - n_3) \end{aligned}$$

pero como $n_3$ no es fijo

$$f(3,r) = \min_{n_3=1,\ldots,n} \left\{ \frac{a_3}{n_3} + f(1, r - n_3) \right\}$$

**Etc.** Si conocemos $f(k, r)$ entonces:

$$f(k+1, r) = \min_{n_{k+1}=1,\ldots,n} \left\{ \frac{a_{k+1}}{n_{k+1}} + f(k, r - n_{k+1}) \right\}$$

**Paso final (L-ésima):** Ubicando el valor mínimo de $f(L, r)$, esa posición nos da $n_L$. Del anterior ($L$-1)-ésimo el valor $f(L-1, n-n_L)$ nos da el $n_{L-1}$. El ($L$-2)-ésimo $f(L-2, n-n_l-n_{L-1})$ nos da el $n_{L-2}$. Así hasta que finalmente encontramos el óptimo para $n_1$.

## 2.3. Ejemplo

La siguiente tabla tomada de Cochran (1977), muestra el número de habitantes (miles) de 64 grandes ciudades en los E.U. en el año 1930. Las ciudades están agrupadas en tres estratos. Supongamos que se desea tomar un total de 24 muestras entre los tres estratos.

Tabla 1: *Números de habitantes (en miles), para el año 1930, en 64 grandes ciudades de los E.U (por estrato)*

| 1 | | 2 | | | 3 | | | |
|---|---|---|---|---|---|---|---|---|
| 90.0 | 57.8 | 36.4 | 30.8 | 25.3 | 19.5 | 13.9 | 14.0 | 16.3 |
| 82.2 | 48.7 | 31.7 | 27.2 | 23.2 | 18.3 | 17.0 | 11.9 | 11.6 |
| 78.1 | 44.2 | 32.8 | 28.4 | 26.0 | 16.3 | 15.0 | 13.0 | 12.2 |
| 80.5 | 45.1 | 30.2 | 25.5 | 29.2 | 20.1 | 14.3 | 12.7 | 13.4 |
| 67.0 | 45.9 | 28.8 | 27.0 | | 14.7 | 11.3 | 10.0 | |
| 123.8 | 46.4 | 29.1 | 21.4 | | 16.4 | 11.5 | 10.7 | |
| 57.3 | 40.0 | 25.3 | 26.0 | | 14.3 | 12.3 | 11.4 | |
| 63.4 | 36.6 | 29.1 | 20.9 | | 16.9 | 15.4 | 11.1 | |

### 2.3.1. Método de multiplicadores de Lagrange

Calculamos los $n_i$ mediante el método de Lagrange (tabla 2).

Tabla 2: *Cálculos de los $n_i$ mediante el método de Lagrange*

| Estrato $i$ | $N_i$ | $\sum Y_i$ | $\overline{Y}_i$ | $S_i^2$ | $W_i$ | $a_i$ | $n_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 16 | 1007.0 | 62.94 | 538.43 | 0.25 | 33.65 | 17.05 |
| 2 | 20 | 554.3 | 27.71 | 14.24 | 0.31 | 1.39 | 3.47 |
| 3 | 28 | 395.5 | 14.13 | 7.33 | 0.44 | 1.40 | 3.48 |
| Total | 64 | | | | | | |

Las soluciones redondeadas son 17, 3, 4 (tabla 2). Hay un problema de sobre muestreo.

### 2.3.2. Aplicación del método Knapsack

Primero se calculan los $f(1,r)$, para $r = n_1$ factibles. Luego calculamos $f(2,r)$ y $f(3,r)$ (Tabla 3).

Tabla 3: *$f(k,r)$ para $k = 1,2,3$ y $r = 1,\ldots,24$*

| $r$ | $f(1,r)$ | $n_1$ | $f(2,r)$ | $n_2$ | $f(3,r)$ | $n_3$ |
|---|---|---|---|---|---|---|
| 1 | 33.65 | 1 | | | | |
| 2 | 16.83 | 2 | 35.04 | 1 | | |
| 3 | 11.22 | 3 | 18.22 | 1 | 36.44 | 1 |
| 4 | 8.41 | 4 | 12.61 | 1 | 19.62 | 1 |
| 5 | 6.73 | 5 | 9.80 | 1 | 14.01 | 1 |
| 6 | 5.61 | 6 | 8.12 | 1 | 11.21 | 1 |
| 7 | 4.81 | 7 | 7.00 | 1 | 9.52 | 1 |
| 8 | 4.21 | 8 | 6.20 | 1 | 8.40 | 1 |
| 9 | 3.74 | 9 | 5.50 | 2 | 7.60 | 1 |
| 10 | 3.37 | 10 | 4.90 | 2 | 6.90 | 2 |
| 11 | 3.06 | 11 | 4.43 | 2 | 6.20 | 2 |
| 12 | 2.80 | 12 | 4.06 | 2 | 5.60 | 2 |
| 13 | 2.59 | 13 | 3.75 | 2 | 5.14 | 2 |
| 14 | 2.40 | 14 | 3.50 | 2 | 4.76 | 2 |
| 15 | 2.24 | 15 | 3.27 | 3 | 4.46 | 2 |
| 16 | 2.10 | **16** | 3.05 | 3 | 4.20 | 2 |
| 17 | | | 2.87 | 3 | 3.97 | 3 |
| 18 | | | 2.71 | 3 | 3.74 | 3 |
| 19 | | | 2.57 | 3 | 3.52 | 3 |
| 20 | | | 2.45 | **4** | 3.33 | 3 |
| 21 | | | 2.38 | 5 | 3.17 | 3 |
| 22 | | | 2.33 | 6 | 3.03 | 3 |
| 23 | | | 2.30 | 7 | 2.92 | 3 |
| 24 | | | 2.28 | 8 | 2.80 | **4** |

Encontramos que $f(3,24) = 2.814977$ para $n_3 = 4$, con $r = 24-4 = 20$ y $k = 2$ tenemos un $f(2,20) = 2.467863$ para un $n_2 = 4$, finalmente con $r = 20-4 = 16$ y $k = 1$ tenemos $f(1,16) = 2.109619$ para un $n_1 = 16$. Por lo tanto, la asignación óptima de los tamaños de muestra es de 16, 4, 4.

Comparando la solución obtenida estimamos la varianza de algunos métodos (Tabla 4).

Tabla 4: *Comparación de varianzas*

| Asignación | $(n_1,n_2,n_3)$ | $V(\overline{y}_{est})$ |
|---|---|---|
| Método Knapsack | $(16,4,4)$ | 0.5841 |
| Asignación mediante Lagrange | $(17,3,4)$ | 0.5627 |
| M.A.S. | $(8,8,8)$ | 2.3410 |
| Proporcional | $(6,8,10)$ | 3.7127 |

La afijación que mejor minimiza la varianza es el método de los multiplicadores de Lagrange, pero incurrimos en sobremuestreo, mientras que el método de Knapsack no incurre en este problema y además minimiza muy bien la varianza (Tabla 4).

## 2.4. Asignación óptima con restricción de presupuesto

Supongamos que el costo por muestra difiere por los diferentes estratos. Sea $c_i$ el costo de una muestra en el $i$-ésimo estrato. Sea $C$ el presupuesto total disponible.

Podemos definir el siguiente problema de asignación óptima para el presupuesto como:

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{L} \frac{a_i}{n_i} = Z \\ \textbf{s. a.} \quad & \sum_{i=1}^{L} c_i n_i = C \\ & n_i \in \mathbb{Z}^+,\ 1 \leq n_i \leq N_i \quad (i = 1 \ldots, L). \end{aligned}$$

Una solución similar por el método Knapsack a este problema puede ser dada por

$$f(k,c) = \begin{cases} \text{mín} & \sum_{i=1}^{k} \frac{a_i}{n_i} \\ \boldsymbol{s.a} & \sum c_i n_i \leq c,\ 1 \leq n_i \leq N_i,\ n_i \in \mathbb{Z}^+ \quad (i = 1, \ldots, k) \end{cases}$$

la cual tiene la siguiente fórmula recursiva

$$f(k,c) = \min_{n_k = 1,\ldots,n} \left\{ \frac{a_k}{n_k} + f(k-1, c - c_k n_k) \right\}.$$

Para todo $c$ factible implica que $\sum_{i=1}^{k} c_i \leq c \leq C$. El método encontrará este óptimo $n_i$ exactamente como el ejemplo anterior.

**Problema 3.** *Similarmente se puede plantear minimizando el costo manteniendo la pérdida de precisión dentro de "ciertos" límites, es decir:*

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{L} c_i n_i = C \\ \boldsymbol{s.\ a.} \quad & \sum_{i=1}^{L} \frac{a_i}{n_i} \leq v \\ & 1 \leq n_i \leq N_i,\ n_i \in \mathbb{Z}^+ \quad (i = 1 \ldots, L). \end{aligned}$$

Hasta ahora se ha considerado una característica en estudio. Pero si el muestreo es multivariado, se deseará estudiar varias características, para el problema de la afijación óptima no es tan simple dar una aproximación. En la siguiente sección consideraremos minimizar el costo total para lograr prescribir la precisión de la estimación de varias características de la población.

## 3. Asignación óptima para el M.A.E. multivariado

Kohan & Khan (1967) han estudiando el problema de asignación óptima para el M.A.E. multivariado, de ello podemos describir lo siguiente. Suponemos $p$ características en estudio. Sea $Y_j$ la $j$-ésima característica estudiada. Sea $L$ números de

estratos y $N_i$ unidades en el $i$-ésimo estrato, donde $\sum_{i=1}^{L} N_i = N$. Sea $n_i$ número de muestras elegidas en el $i$-ésimo estrato $(i = 1, \ldots, L)$, las muestras aleatorias en los $L$ estratos es independiente. Sea $\overline{y}_{ij}$ la estimación insesgada de $\overline{Y}_{ij}$, dada por

$$\overline{y}_{ij} = \frac{1}{n_i} \sum_{h=1}^{n_i} y_{ijh}$$

donde $y_{ijh}$ es el valor observado de $Y_j$ en el $i$-ésimo estrato para la $h$-ésima unidad muestral. Un estimador insesgado de la media poblacional $\overline{Y}_j$ viene dado por

$$\overline{y}_{jst} = \frac{1}{N} \sum_{i=1}^{L} N_i \overline{y}_{ij}, \quad \text{para } j = 1, \ldots, p.$$

Teniendo esto, consideremos el problema de la afijación óptima. La precisión de la estimación es medida por la varianza estimada de la característica poblacional, para cada característica. Esto es

$$V_j = V(\overline{y}_{jst}) = \sum_{i=1}^{L} a_{ij} x_i,$$

donde $a_{ij} = W_i^2 S_{ij}^2$, $W_i = N_i/N$, $S_{ij}^2 = \sum_{h=1}^{N_i} (y_{ijh} - \overline{Y}_{ij})^2/(N_i - 1)$ y $x_i = 1/n_i - 1/N_i$.

Sea $C_i$ el costo del muestreo de las $p$ características en el $i$-ésimo estrato, donde el costo total sería:

$$K = \sum_{i=1}^{L} C_i n_i.$$

Asumamos $a_{ij}$, $C_i > 0\,(i = 1, \ldots, L,\ j = 1, \ldots, p)$. El problema de afijación puede plantearse como:

**Problema 4.**

$$\begin{aligned}
\text{mín} \quad & \sum_{i=1}^{L} C_i n_i = K \quad \textit{(Función Objetivo Lineal)} \\
\textbf{\textit{s. a.}} \quad & \sum_{i=1}^{L} a_{ij} x_i \leq v_j \quad (j = 1, \ldots, p) \qquad \textit{(Restricción No Lineal)} \\
& 0 \leq x_i \leq 1 - \frac{1}{N_i} \quad (i = 1, \ldots, L) \\
& 1 \leq n_i \leq N_i,\ n_i \in \mathbb{Z}
\end{aligned}$$

*donde $v_j$ es el error permitido en la $j$-ésima característica, es decir el límite superior del intervalo de confianza para $V_j$.*

El problema anterior es un problema de programación lineal entera pero con restricción no lineal. Introduciendo una nueva variable $X_i = 1/n_i$, $i = 1, \ldots, L$ para reescribir el problema 4 se puede reducir a la solución del siguiente problema.

**Problema 5.**

$$\begin{aligned} \text{mín} \quad & \sum_{i=1}^{L} \frac{C_i}{X_i} = K(\boldsymbol{X}) \quad \textit{(Función Objetivo No Lineal)} \\ \textbf{\textit{s. a.}} \quad & \sum_{i=1}^{L} a_{ij} X_i \leq b_j, \quad (j = 1, \ldots, p) \quad \textit{(Restricción Lineal)} \\ & \frac{1}{N_i} \leq X_i \leq 1, \quad (i = 1 \ldots, L) \end{aligned}$$

*donde* $\boldsymbol{X} = (X_1, \ldots, X_L)$, $b_j = v_j + \sum_{i=1}^{L} a_{ij}/N_i \, (j = 1, \ldots, p)$.

La función objetivo es estrictamente convexa puesto que $C_i/X_i$ es estrictamente convexo para $C_i > 0$. Las restricciones del problema 5 proporcionan una región factible convexa acotada, formada por inecuaciones lineales. La región no es vacía si $(1/N_1, \ldots, 1/N_L)$ es factible. Así, un óptimo $\boldsymbol{X} = (X_1^*, \ldots, X_L^*)$ existe. La convexidad estricta implica también la unicidad de la solución óptima. La solución al problema de optimización ocurre en la frontera de este convexo.

El problema 5 es un problema de programación convexa como el descrito en la sección 2. Para desarrollar las condiciones necesarias y suficientes es necesario un $\boldsymbol{X}$ optimal. Entre algunos métodos para solucionar el problema 5 tenemos: el método simplex, método de la dirección factible, método de la proyección del gradiente, método del corte del plano, método de linealización, entre otros (Bazaraa & Shatty 1979, Collatz & Wetterling 1975, Rustagi 1994).

Las soluciones mencionadas anteriormente al problema 5 son soluciones factibles que deben ser redondeadas. Un método que encuentra soluciones sin redondear se conoce como «branch bound», el cual se considera como una extensión al método Knapsack en forma univariada.

## 3.1. Interpretación gráfica

Consideremos a $L = 2$ (número de estratos). La función objetivo viene dada por

$$K = \frac{C_1}{X_1} + \frac{C_2}{X_2} = \frac{C_1 X_2 + C_2 X_1}{X_1 X_2}$$

Esta se puede reescribir de la siguiente manera

$$\left(X_1 - \frac{C_1}{K}\right)\left(X_2 - \frac{C_2}{K}\right) = \frac{C_1 C_2}{K^2},$$

lo que nos muestra una hipérbola rectangular centrada en $(C_1/K, C_2/K)$. Cuando $K$ varía, el centro siempre estará sobre la recta

$$\frac{X_2}{X_1} = \frac{C_2}{C_1},$$

el vértice de la rama positiva de la hipérbola rectangular $\big((C_1+\sqrt{C_1C_2})/K, (C_2+\sqrt{C_1C_2})/K\big)$, estará sobre la recta

$$\frac{X_2}{X_1} = \frac{C_2+\sqrt{C_1C_2}}{C_1+\sqrt{C_1C_2}}.$$

Las restricciones del problema 5 con los signos de igualdad representan las $p$ líneas rectas sobre el plano de dos dimensiones. Las cantidades $a_{ij}$ y $v_j$ son todas cantidades positivas, las pendientes de estas rectas son negativas y todas ellas están dadas en el primer cuadrante del plano $X_1X_2$ (ver Figura 3).

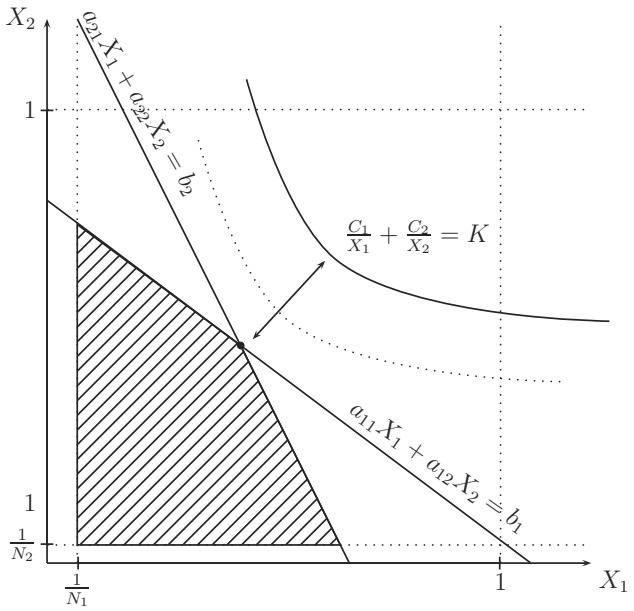


Figura 3: *Función objetivo y región factible.*
Fuente: (Arthanari & Dodge 1981)

Para obtener la asignación óptima debemos encontrar la hipérbola regular para algún valor $K$, tal que toque el limite de la región factible (ver Figura 3). En general, cuando tenemos $L$ estratos tenemos el siguiente teorema.

**Teorema 3.1.** *El punto de contacto del hiperplano*

$$\sum_{i=1}^{L} a_i X_i = b \quad (a_i, b > 0),$$

*con la función objetivo* $K = \sum_{i=1}^{L} C_i/X_i$ *es dada por* $\boldsymbol{X} = (X_1, \ldots, X_L)$, *donde*

$$X_i = \frac{b\sqrt{C_i a_i}}{a_i \sum_{j=1}^{L} \sqrt{C_j a_j}} \quad (i = 1, \ldots, L)$$

*Demostración.* Consideremos $L$ estratos y sea $X_i = 1/n_i\,(i = 1, \ldots, L)$. La función objetivo dada por

$$K = \sum_{i=1}^{L} \frac{C_i}{X_i}.$$

Las ecuaciones de los hiperplanos correspondientes a los contrastes lineales, describen la función de restricción dada por

$$g(X_1, \ldots, X_L) = \sum_{i=1}^{L} a_i X_i = b$$

Mediante el método de Lagrange tenemos la condición $\nabla K - \lambda \nabla g = 0$ para $g = 0$. Es decir,

$$\left(-\frac{C_1}{X_1^2}, \ldots, -\frac{C_L}{X_L^2}\right) - \lambda(a_1, \ldots, a_L) = 0,$$

tomando la $i$-ésima componente tenemos,

$$-\frac{C_i}{X_i^2} + \lambda a_i = 0 \quad (i = 1, \ldots, L)$$

despejando $X_i$ tenemos

$$X_i = \sqrt{\frac{C_i}{\lambda a_i}},$$

reemplazando $X_i$ en la función de restricción y tomando $\lambda$ negativo tenemos,

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{L} \sqrt{a_i C_i}}{b},$$

así,

$$X_i = \frac{b\sqrt{C_i a_i}}{a_i \sum_{j=1}^{L} \sqrt{C_j a_j}}.$$ □

Introduciendo el subíndice $j$ para las diferentes características, tenemos el correspondiente resultado para el $j$-ésimo hiperplano.

*Nota.* Lo primero que se intenta:

- Encontrar los puntos de contacto de $K$ con los $p$-hiperplanos.
- Escoger el punto de contacto que arroja el valor mínimo $K$.
- Verificar que $\boldsymbol{X}$ que minimiza $K$ es factible.
- $n_i = \frac{1}{X_i}$, redondear se es necesario.

Este método no es factible cuando hay muchas características y muchos estratos (lo cual en la realidad muy pocas veces ocurre).

## 3.2. Algoritmo Mejorado

Describiremos a continuación un algoritmo eficiente para determinar la afijación óptima en un M.A.E. multivariado.

**Paso 0:** descarte del conjunto de restricciones del problema 5 los contrastes que no son obligatorios, es decir, elimine los hiperplanos que introducen restricciones redundantes (Ver Figura 4). Esto implica encontrar los $j$ para cada vector $(b_j/a_{1j}, \ldots, b_j/a_{Lj})$ de los interceptos del $j$-ésimo hiperplano sobre los ejes que estrictamente dominan el vector correspondiente de cualquier otro $j$. Asuma que $I_1 = \{j \mid j = 1, \ldots, q\}$ es el conjunto de los $q$ subíndices de las restricciones no redundantes.

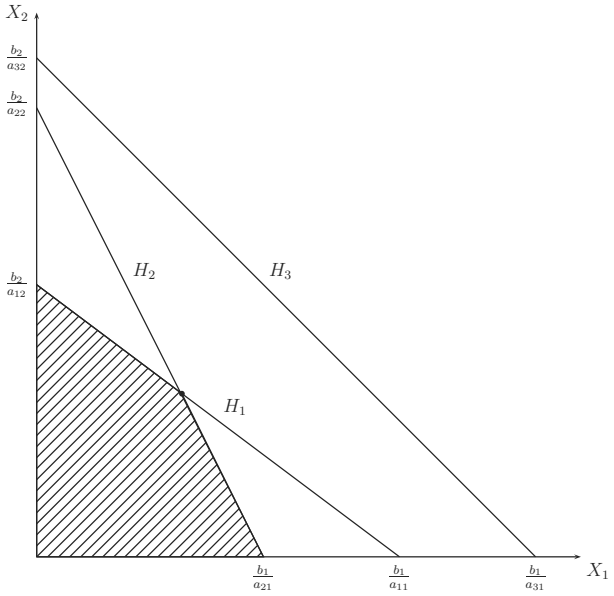


Figura 4: *Región factible con $H_3$ como restricción redundante.*

**Paso 1:** compute $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{Lj})$ para cada característica $j \in I_1$, usando el teorema 3.1, esto es,

$$X_{ij} = \frac{b_j \sqrt{C_i a_{ij}}}{a_{ij} \sum_{i=1}^{L} \sqrt{C_j a_{ij}}}.$$

**Paso 2:** encuentre $j^*$ tal que $\sum_{i=1}^{L} 1/X_{ij^*}$ sea el máximo de los $j \in I_1$.

Una forma alternativa para trabajar los pasos 2 y 3 se puede realizar construyendo la tabla 5.

Tabla 5: Pasos 2 y 3 del algoritmo para determinar la afijación óptima en un M.A.E. multivariado

| $j$ | $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{Lj})$ | Total $m_j = \sum_{i=1}^{L} \frac{1}{X_{ji}}$ |
|---|---|---|
| 1 | $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{L1})$ | Total $m_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $q$ | $\boldsymbol{X}_q = (X_{1q}, \ldots, X_{Lq})$ | Total $m_q$ |

Si $\boldsymbol{X}_{j^*}$ es factible, entonces redondear la solución. Si $\boldsymbol{X}_{j^*}$ tiene una componente que no es factible, entonces fijar

$$X_{j^*i} = \left\{ \frac{1}{N_i} \quad \text{ó } 1 \quad i\text{: tenga el problema} \right.$$

se elimina el estrato y se repiten los pasos con los restantes.

# 4. Conclusiones

- Los métodos de Programación Matemática, en particular la Programación Convexa, la cual visualiza problemas de afijación óptima de muestra, en muestreo aleatorio estratificado, permitiendo obtener resultados satisfactorios.
- El método de los multiplicadores de LaGrange, en un ejemplo particular presenta problemas de sobremuestreo, el cual puede ser evitado al replantear el problema, con un algoritmo de Programación Entera, conocido en la literatura como el problema de la mochila o el morral.
- El muestreo estratificado con múltiples caracteristicas o multivariados generaliza el método particular de la obtención de muestras en el caso univariado, por lo cual la Programación Convexa es util en las dos situaciones.

# 5. Agradecimientos

Los autores expresan los mas sentidos agradecimientos a los estudiantes que han venido participaron en los seminarios realizados por el grupo GELIMO de la Universidad del Tolima y también a la Oficina de Investigaciones y Desarrollo Científico de la misma Universidad, por su constante y valioso apoyo en el desarrollo de la Ciencia y la Tecnología.

Recibido: 10 de febrero de 2010
Aceptado: 12 de abril de 2010

# Referencias

Arthanari, T. S. & Dodge, Y. (1981), *Mathematical Programming in Statistics*, John Wiley & Sons, Inc., Canada.

Bazaraa, M. S. & Shatty, C. M. (1979), *Nonlinear Programming and Algorithms*, Wyley, New York.

Cochran, W. G. (1977), *Sampling Techniques*, 3 edn, Wiley, New York.

Collatz, L. & Wetterling, W. (1975), *Optimization Problems*, Springer - Verlag, New York.

Hartly, H. O. & Rao, J. N. K. (1968), 'A new estimation theory for sample surveys', *Biometrika* **55**(547).

Kohan, A. R. & Khan, S. (1967), 'Allocation in multivariate surveys: An analytical solution', *Journal of the Royal Statistical Society. Serie B (Methodological)* **29**(1), 115–125.

Rustagi, J. S. (1994), *Optimization Techniques in Statistics*, Academic Press, Inc., New York.

Wolfe, P. (1959), 'The simple method for quadratic programming', *Econometrica* .