CONSTRUCCIÓN DE POST-ESTRATOS ÓPTIMOS MULTIVARIADOS PARA ESTUDIOS PSICOLÓGICOS

Sira Allende*, Carlos Bouza*[1], Dante Covarrubias**, Nicanor Guerra*** & Lakshma Singh****
*Departamento de Matemática Aplicada, Universidad de La Habana, Cuba.
**Escuela de Matemáticas, Universidad Autónoma de Guerrero, México
****Universidad de Las Palmas de Gran Canaria, España
****Bhat & Sarkar Informatic Consultors, India

RESUMEN

Se utilizan métodos de optimización estocástica para determinar post estratos óptimos en el caso multivariado. Los métodos son analizados y se aplican en 4 muestras obtenidas en estudios psicológicos. Los resultados validan la conveniencia de su uso en términos de precisión relativa respecto al uso del muestreo simple aleatorio.
PALABRAS CLAVE: Post estratificación, programación estocástica, ganancia en precisión. Precisión relativa

ABSTRACT

Stochastic optimization methods are used for determining optimal post strata for the multivariate case. The methods are analyzed and are applied to the study of 4 samples obtained in psychological studies. The results validate the convenience of using then in terms of the relative precision of them with respect to simple random sampling.
KEY WORDS: Post stratification, stochastic programming, gain in accuracy, relative precision

1. INTRODUCCIÓN

Siguiendo la teoría clásica de muestreo, el estadístico selecciona muestras de individuos, negocios, y otras unidades de muestreo para obtener información deseada acerca del comportamiento de la población. La selección de muestras se hace siguiendo un esquema de probabilístico de selección de un marco de lista, también llamado marco de muestreo, de la población objetivo. Esto es de particular importancia al desear hacer estudio usando estratos. Estos no están delimitados de antemano en muchas de las aplicaciones y es necesario nacer una inversión en determinarles o tomar una muestra y construirlos a posteriori. Al proceso de muestreo bajo las anteriores condiciones se le llama post estratificación.

El problema multivariado ha recibido poca atención no siendo así con el problema de la fijación del tamaño de muestra en el caso de trabajar con varias variables de interés. Allende y. Bouza (1987), Allende et al (2008), Allende ·et al (2012) han tratado el problema de la estratificación y post estratificación optima en el caso multivariado. Los interesados en los problemas de afijación pueden ver Khan y. Ahsan (2003), Kokan y Khan(1967) y Lavallèe, (1987).

En este trabajo abordamos este problema para dar solución a varios casos de interés en el área de la psicología. En estos los usos de los métodos convencionales de clustering no arrojan resultados satisfactorios dados los intereses de los psicólogos centrados en obtener grupos homogéneos en un sentido no necesariamente en un sentido metric. Los métodos desarrollados permiten determinar grupos que garantizan minimizar la varianza de las estimaciones al utilizar muestras. Estos plantean problemas de programación estocástica. Dos métodos son considerados aplicados en 4 problemas y los resultados analizados.

Los grupos obtenidos son post estratos y estos pueden ser usados para identificar las características de nuevos individuos, al clasificarles en uno de ellos, o para desarrollar otros estudios estadísticos basados en muestreos.

2. EL MUESTREO ESTRATIFICADO

2.1. El caso univariado

[1] bouza@matcom.uh.cu

Con el objeto de mejorar las precisiones podemos utilizar un agrupamiento (clustering) de los elementos más parecidos entre si. La población *U* está dividida en subpoblaciones que no se solapan. Esto es está particionada en subpoblaciones homogéneas tales que

$$U = \cup_{1 \le k \le K} U_k, \qquad U_k \cap U_h = \emptyset, \forall k \neq h$$

Dentro de cada U_k se hace una selección aleatoria utilizando un diseño muestral garantizándose la independencia de la selección inter-grupo. Este método es muy utilizado pues permite realizar estimaciones con una precisión especificada para cada estrato U_k. Entonces se divide la población de *N* individuos, en *K* subpoblaciones, estratos, atendiendo a criterios que puedan ser importantes en el estudio, de tamaños respectivos $N_1, ..., N_k$, y

$$N = \sum_{k=1}^{K} N_k$$

En cada una de estas subpoblaciones se selecciona un muestra $s_k \subseteq U_k$ de tamaño n_k.

$$n = \sum_{k=1}^{K} n_k$$

Este es muy usado dado que la estratificación provee de una base teórica para reducir el costo por observación al estratificar la población en grupos convenientes.

La información de las muestras aleatorias simples de cada estrato constituye una muestra global. La estratificación permite obtener estimaciones de parámetros poblacionales para subgrupos de la población pues obtenemos información sobre:

- Toda la población
- Cada estrato
- La relación entre estratos.

Si las mediciones dentro de cada estrato son homogéneas, la estratificación garantiza un pequeño error de estimación.

Al encarar la aplicación de este modelo se abren una serie de problemas como son:

- Como distribuir el tamaño de la muestra *n* entre los *K* estratos.
- Si no están definidos los estratos: establecer las características a utilizar y fijar cuántos estratos debe haber.

Evidentemente, cada individuo pertenece a un estrato y solo a uno. La muestra estratificada es

$$s = \cup_{1 \le k \le K} s_k, \qquad s_k \subseteq U_k, \qquad s_k \cap s_h = \emptyset, \qquad \forall k \neq h$$

Al determinar un diseño $P(s_k)$ para cada $k=1, ...,$K el diseño muestreo estratificado está dado por

$$P(s) = \prod_{k=1}^{K} P(s_k),\ s_k \subseteq U_k$$

El parámetro (media poblacional o total) puede ser expresado mediante una combinación lineal de los de los estratos. Veamos el caso de la media.

$$\mu_Y = \frac{\sum_{i \in U} Y_i}{N} = \frac{\sum_{j=1}^{K} \sum_{t \in U_j} Y_i}{N} = \frac{\sum_{j=1}^{K} N_i \mu_{Y(j)}}{N} = \sum_{j=1}^{K} W_i \mu_{Y(j)}$$

$W_j = \frac{N_j}{N}$ es la probabilidad de que al escoger usando MAS un $i \in U$ este sea tal que $i \in U_j$. Para explicitar la pertenencia a los estratos utilizaremos la doble indización para denotar que $i \in U_j$ a Y_{ij}. Sabemos que

es insesgado si usamos MAS por lo que un estimador insesgado de μ_Y es:

$$\bar{y}_e = \sum_{j=1}^{K} W_j \bar{y}_j$$

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j}, \qquad n_j = |s_j|$$

Si seleccionamos en forma independiente de estrato a estrato

$$V(\bar{y}_e) = \begin{cases} \sum_{j=1}^{K} W_j^2 \frac{\sigma_j^2}{n_j} & si\ se\ usa\ MASCR\ en\ cada\ U_j \\ \sum_{j=1}^{K} W_j^2 \frac{(1-f_j)S_j^2}{n_j}, & si\ se\ usa\ MASSR\ en\ cada\ U_j \end{cases}$$

Es claro que

$$\sigma_j^2 = \frac{\sum_{i=1}^{N_j} (Y_{ij} - \mu_{Y(j)})^2}{N_j}, \qquad S_j^2 = \frac{\sum_{i=1}^{N_j} (Y_{ij} - \mu_{Y(j)})^2}{N_j - 1}, \qquad f_j = \frac{n_j}{N_j}$$

La estimacion insesgada del error es, si se usa el mismo tipo de MAS,

$$\hat{V}(\bar{y}_e) = \begin{cases} \sum_{j=1}^{K} W_j^2 \frac{s_j^2}{n_j} & si\ se\ usa\ MASCR\ en\ cada\ U_j \\ \sum_{j=1}^{K} W_j^2 \frac{(1-f_j)s_j^2}{n_j}, & si\ se\ usa\ MASSR\ en\ cada\ U_j \end{cases}$$

siendo $s_j^2 = \frac{\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n_j - 1}$

Otro problema es de fijar criterios par afijar los tamaño de la muestra (afijación).

Los procedimientos más utilizados para la estratificación de la muestra:

- Afijación simple o igualitaria: se reparte la muestra total en partes iguales para cada estrato, esta es $n_j = n/K$.
- Afijación proporcional al tamaño: la muestra se reparte proporcionalmente a la población de cada estrato, sea $n_j = nW_j$.
- Afijación óptima o no proporcional: se considera la mayor o menor heterogeneidad dentro de cada estrato y opcionalmente los costos encuestar.

El costo de una encuesta es medido por la función:

$$C = C_o + \sum_{j=1}^{K} c_j n_j$$

C_o representa los costos que no dependen de la evaluación de las unidades seleccionadas y c_j el de entrevistar una unidad de U_j.

2.2. Ganancia en precisión

El uso de este método tiene una justificación adicional en el hecho de que generalmente se obtiene un estimado más preciso de la media poblacional al utilizar MAE que el que se obtiene a través de un MAS. Cuánto se gana con la estratificación, dependerá de la constitución de los estratos. Comparando con la varianza de la media aritmética de una muestra seleccionada mediante MAS tenemos que

$$V(\bar{y}) = V_{MAS} = \frac{1}{n}\sum_{j=1}^{K} W_j\,{\sigma_j}^2 + \frac{1}{n}\sum_{j=1}^{K} W_j\left(\mu_{Y(j)} - \mu_Y\right)^2$$

Como el segundo término es positivo siempre preferimos el uso del MAE. El incremento en precisión es mayor si los estratos son muy diferentes entre si en términos de sus medias. Una vez seleccionada la muestra podemos estimar la ganancia en precisión. Al hacer la selección de la muestra podemos calcular

$$\hat{V}(\bar{y}_e) = \sum_{j=1}^{K} {W_j}^2\frac{{s_j}^2}{n_j} - \sum_{j=1}^{K} {W_j}^2\frac{{s_j}^2}{N}$$

Podemos hacer las comparaciones teóricas necesarias para establecer una estimación de la ganancia debida al estratificar.

2.3. Post-estratificación

En ocasiones tenemos un muestra seleccionada mediante MAS pero tenemos interés en el análisis de una variable en los estratos. Suponemos que estos están definidos y que seleccionamos $s \subseteq U$ y hacemos a posteriori la partición $s = \cup_{1\le j\le K} s_{Pj}, \#s_{Pj} = m_j, \sum_{j=1}^{K} m_j = n$. Un estimador de la media poblacional, si N_j es conocido, $j=1,\ldots,K$, está dado por

$$\bar{y}_{Pe} = \sum_{j=1}^{K} W_j\frac{\sum_{i=1}^{m_j} y_{ij}}{m_j}$$

Bajo el uso de MASSR para seleccionar s tenemos que

$$E(\bar{y}_{Pe}) = \sum_{j=1}^{K} W_j E\left(\frac{\sum_{i=1}^{m_j} y_{ij}}{m_j}\right) = \sum_{j=1}^{K} W_j\mu_{Y(j)} = \mu_Y$$

Pero

$$V(\bar{y}_{Pe}) = \sum_{j=1}^{K} {W_j}^2 V\left(\frac{\sum_{i=1}^{m_j} y_{ij}}{m_j}\right) = \sum_{j=1}^{K}\frac{{W_j}^2 S^2_{Y(j)}}{m_j} - \sum_{j=1}^{K}\frac{W_j\ \ S^2_{Y(j)}}{N}$$

Si aceptamos la aproximación en Series de Taylor, W_j conocida, $j=1,\ldots,n$.

$$E\left(\frac{1}{m_j}\right) \cong \frac{1}{nW_j} + \frac{1-W_j}{n^2 W_j^2}$$

entonces

$$E(V(\bar{y}_{Pe})) \cong \begin{cases} V_P + \sum_{j=1}^{K}\dfrac{(1-W_j)\ \ S^2_{Y(j)}}{n^2}\ si\ SR \\ V_P + \sum_{j=1}^{K}\dfrac{(1-W_j)\ \ \sigma^2_{Y(j)}}{n^2}\ Si\ CR \end{cases}$$

El segundo término en la expresión derivada es el incremento en el error provocado por haber post-estratificado.

$$E(V(\bar{y}_{Pe})) - V_{MAS} = \sum_{j=1}^{K}\frac{(1-W_j)\ \ \sigma^2_{Y(j)}}{n^2} - \frac{1}{n}\sum_{j=1}^{K} W_j\left(\mu_{Y(j)} - \mu_Y\right)^2$$

mide el incremento en el error de estimación debido al uso de la post estratificación.

3. CONSTRUCCION DE ESTRATOS

3.1 Antecedentes

Dalenius en la década del 50 del pasado siglo se planteó el problema de construir estratos que garantizasen minimizar la varianza de la estimación de la media poblacional. Detalles de los resultados

diversos derivados en esta línea pueden verse en Dalenius-Hodges (1959), Bethel (1989) y en libros como Cochran (1981).

Usando los criterios clásicos de la estadística paramétrica podemos decir que tenemos una fdp *f(t)* tal que

$$W_i = \Pr ob(U_i) = \int_{U_i} f(t)dt = \int_{y(t-1)}^{y(t)} f(t)dt$$

Consideramos que U_i esta definido por cotas en la variable de interés *Y*. $Y \in [y(0), y(K)]$. Dada la continuidad $P(Y=y(i))=0$. Entonces tomando

$$M(i) = \bar{Y}_i \,/ U_i \text{ es determinado por } [y(i-1), y(K)]$$

$$W_i M(i) = \int_{U_i} tf(t)dt = \int_{y(i-1)}^{y(i)} tf(t)dt \Rightarrow \bar{Y} = \sum_{i=1}^{K} W_i M(i)$$

$$W_i^2 \sigma^2{}_i = \int_{y(i-1)}^{y(i)} t^2 f(t)dt - W_i M^2(i) \Rightarrow V(\bar{y}_e) = \sum_{i=1}^{K} \frac{\int_{y(i-1)}^{y(i)} t^2 f(t)dt - W_i M^2(i)}{n_i}$$

Por tanto

Entonces al fijar los tamaños de muestra se puede plantear el problema de optimización es un problema de particionamiento en *K* subintervalos del intervalo [*y(0), y(K)]*. *El* problema de optimización es

$$Min\{V(\bar{y}_e)\} = Min_{(y(0),\ldots,y(K))}\{\sum_{i=1}^{K} [\int_{y(i-1)}^{y(i)} t^2 f(t)dt - W_i M^2(i)]\}$$

3. 2. Construcción de Estratos óptimos multivariados: Programa Estocástico con restricciones probabilísticas

Allende-Bouza (1987) estudiaron el problema de determinar estratos óptimos usando información multivariada. Se plantearon resolver el problema de optimización. Información auxiliar se consideraba conocida para cada unidad de U. En el caso que se estudia solo tenemos la información brindada por la muestral y los estratos son formados a posteriori.

El interés es formar estratos tales que las variables medidas tengan valores muy similares entre si dentro de cada estrato. Tomemos X_t como el valor del ítem *t* y $\mu_t(h)$ su media en el estrato U_h y su media en *U* como μ_t. La dispersión en la población está dada por

$$\sum\nolimits_{q \in U} \frac{(X_{qt} - \mu_t)^2}{N} = \sigma_t^2$$

Dentro de cada estrato U_h es

$$\sum\nolimits_{q \in U_h} \frac{\left(X_{qt} - \mu_t(h)\right)^2}{N_h} = \sigma_t^2(h)$$

Allende et al (2012) consideraron que el decisor (DM) fija a su conveniencia un conjunto de puntos $\Re^M$. Ellos representan puntos que caracterizan el comportamiento del problema de interés, en este caso la asertividad. Estos son considerados como puntos de atracción de las conductas prototípicas y se tratan como centroides. Cada uno es denotado como

$$\theta_r = [\theta_{r1}, \ldots, \theta_{rM},]^T \in \Theta = \{\theta_r,\ r=1,..,R,\ R \geq H\}.$$

En ese trabajo propusieron el uso de la función de costo

$$C_{qr} = \sum_{t=1}^{M} (X_{qt} - \theta_{rt})^2 = \sum_{t=1}^{M} V_{qtr}$$

para cada individuo $q \in U$. Esto permite determinar clústeres donde los individuos están cerca del correspondiente centroide y los centroides están la más alejados posibles. Esto se traduce en que los estratos así construidos poseerán una varianza pequeña.

El objetivo que enfrentamos es usar la información obtenida en la muestra para establecer una partición sobre *U*. En este caso trabajamos con información muestral. El modelo de optimización propuesto por Allende et al (2012) es uno de Programación Estocástica y utiliza resultados desarrollados por Albareda-Sambola et. al. (2000).

Utilizándoles tenemos el problema de optimización estocástico equivalente

$$PS: min\left\{\sum_{q=1}^{N}\sum_{t=1}^{M}\sum_{r=1}^{R} P_t V_{qtr} X_{qr} = \sum_{q=1}^{N}\sum_{r=1}^{R} \vartheta_{qr} X_{qr}\right\}$$

$$sujeto\ a: \begin{cases} \sum_{r=1}^{R} X_{qr} = 1, \forall q \\ Prob\left(\sum_{q=1}^{N} \gamma_{qr} X_{qr} \le b_r\right) \ge 1-\alpha.\ \forall r \\ \sum_{r=1}^{R} Z_r \le KM, \quad K \le R \\ X_{qr} \in \{0,1\}, \quad \forall q, \forall r \\ Z_r \in \{0,1\}, \quad \forall r \end{cases}$$

$$X_{qr} = \begin{cases} 1 & si \quad i \in s \cap U_r \\ 0 & si \quad i \notin s \cap U_r \end{cases}$$

$$Z_r = \begin{cases} 1 & si \quad \theta_r \quad es \quad un \quad centroide \quad selecciona do \\ 0 & en \quad otro \quad caso \end{cases}$$

$$\gamma_{qr} = \begin{cases} 1\ si\ q \in s \cap U_r \\ 0\ en\ otro\ caso \end{cases}$$

P_t es un peso que evalúa la importancia de la variable *t*. Note que γ_{qr} es una variable Bernoulli con parámetro $E(\gamma_{qr})=Prob\ (u_q \in s \cap U_r)$ y que para r fijo, $Prob\ (q \in s \cap U_r) = W_r = N_r / N$. Entonces la solución de este problema determina que estratos tienen altas probabilidades y calcular valores exactos de W_r posteriormente.

Ahora se puede determinar un problema de asignación generalizado asociado:

$$PAS: min\left\{\sum_{q=1}^{N}\sum_{r=1}^{R} \gamma_{qr} \vartheta_{qr} X_{qr}\right\}$$

$$sujeto\ a: \begin{cases} \sum_{r=1}^{R} X_{qr} = 1, \forall q \\ \sum_{q=1}^{N} \gamma_{qr} X_{qr} \leq b_r, \forall r \\ X_{qr} \in \{0,1\}, \quad \forall q, \forall r \end{cases}$$

$$\gamma_{qr} = \begin{cases} 1\ si\ q \in s \cap U_r \\ 0\ en\ otro\ caso \end{cases}$$

En este modelo b_r representa el número máximo de individuos en U_r y γ_{qr} es el "recurso" necesario para el individuo q para "hacer la tarea r". En el contexto de la estatificación óptima multivariada el significado es:

- b_r es una cota superior del número de individuos que se incluirían en U_r.
- La restricción primera debe ser una igualdad para garantizar que cada individuo sea clasificado solo en un estrato.
- Como la asignación a un estrato no depende de r $\gamma_{qr}=\gamma_{q\bullet}$, $\forall q=1,\ldots,N$ y

$$\sum_{q=1}^{N} \gamma_{q\bullet} \leq \sum_{r=1}^{R} b_{r\bullet}$$

Haneveld et. al. (1999) discutieron este problema en detalle y quedó claro que nuestro problema tiene al menos una solución factible. En este marco tenemos que un estrato particular con una colección de individuos con un comportamiento notable puede ser descrito usando la demanda

$$\sum_{q=1}^{N} \gamma_{q\bullet} X_{qr} = \gamma_{(r)}$$

Entonces la asignación efectiva $N(r)= \sum^{N}_{q=1} x_{qr}$, $\forall r=1,..,R$ es obtenida de la muestra s al determinar los post-estratos.

El problema es transformado, ver Allende et al (2012) para detalles técnicos en

$$PAS-T: min\left\{\sum_{q\in s} \sum_{r=1}^{R} \vartheta_{qr} X_{qr}\right\}$$

$$sujeto\ a: \begin{cases} \sum_{r=1}^{R} X_{qr} = 1, \forall q \in s \\ Prob\left(\gamma(r) = \sum_{q\in s} \gamma_{q\bullet} X_{qr} \leq b_r\right) \geq 1-\alpha, \forall r \\ X_{qr} \in \{0,1\}, \quad \forall q \in s, \forall r \end{cases}$$

Este un problema de la programación estocástica con restricciones probabilísticas y $\gamma(r)$, por ser la suma de variables independientes con distribución Bernoulli con esperanza W_r, es una Binomial siendo

$$E(\gamma(r)|s) = \left(\sum_{q\in s} X_{qr}\right) W_r = n(r) W_r$$

Por tanto:

$$\Pr ob\left\{\sum_{q\in s} \gamma_{q\bullet} X_{qr} \geq b_r\right\} = \sum_{h=br+1}^{n(r)} \binom{n(r)}{h} W_r^{\ h} \left[1-W_r\right]^{n(r)-h}$$

El DM debe interactuar para fijar W_r fiando la proporción de individuos que espera haya en U_r . En nuestro caso hay indicios de las proporcionalidades de los niveles de asertividad razonables bajo ciertas condiciones de la población

La obtención de los cuantiles para un orden determinado por α es sencilla pues estamos trabajando con una binomial. O sea que estos serán

$$\lambda\{b_r , W_r , \alpha\}=Max\{h \in Z / Prob\ (b_r) \leq \alpha\}$$

La restricción *r-ésima se satisface si* $n(r) \leq \lambda\{b_r , W_r , \alpha\}$

El programa determinístico equivalente es entonces :

PD1:

$$Min\left\{\sum_{r=1}^{R}\sum_{q\in s} \vartheta_{qr}x_{qr} \Big| \sum_{r=1}^{R} X_{qr} \geq 1, \forall q \in s, \qquad \sum_{q\in s} X_{qr} \leq \lambda(b_r, W_r, \alpha), \forall r = 1,.., R\right\}$$

Si se relaja el problema y no imponemos la condición de entero de $\lambda\{b_r , W_r , \alpha\}$ podemos usar un modelo de transporte.

3.3. Construcción de Estratos óptimos multivariados :Programa Estocástico multietápico

Podemos remodelar el problema usando el problema general multietápico *0-1* mixto de amplio uso es

$$PM: min\left\{\sum_{t\in T} \vartheta_t z_t \ sujeto\ a\ C_t^* z_{t-1} + C_t \ \ z_t = d_t, \forall t \in T, \ \ z_t \in Z \subseteq \Re^n\right\}$$

C_t y C^*_t son matrices de restricciones para las etapas t-1 y t respectivamente, d_t un vector. Note que como ϑ_t es una variable aleatoria el problema es estocástico dado que es generado por una evento aleatorio caracterizado por la selección de un muestra s∈S, S es el espacio muestral. Cada muestra fija un escenario y ella permite observar ϑ(s), C(s), C*(s) y d(s). Siguiendo el modelo de Alonso et al (2003) este puede ser representado por

$$PM(R): RP = min\left\{\sum_{s\in S} p(s) \sum_{t\in T} \vartheta_t(s) z_{t(s)} \ sujeto\ a\ C_t^*(s) z_{t-1}(s) + C_t \ \ (s) z_t(s) = d_t(s), \right.$$
$$\left. z_t(s) \in Z \subseteq \Re^n \ ; \ \forall s \in S \ , \forall t \in T; \right\}$$

En nuestro caso la verosimilitud p(s) es la probabilidad de observar la muestra s. Al usar MSA esta es constante.

En el marco utilizado $z_t(s)$ toma en cuenta los cambios generados por los elementos desconocidos y las decisiones se ajustan en cada etapa al obtener nueva información para *t>1*. Estas variables son particionadas en variables de estado (de decision), las que denotaremos y_t, y las de recursos, y^*_t. Entonces, siguiendo a Escudero et al. (2007), podemos denotar $z_t \equiv (y_t\ y^*_t)$, $c_t \equiv (a_t, b_t)$, $C_t \equiv (A_t, B_t)$, $C^*_t \equiv (A^*_t, B^*_t)$. Entonces las restricciones son representadas por

$$A_t(s)y_{t-1} + B_t(s)y^*_{t-1} + A^*_t(s)y_t + B^*_t(s)y_t = d_t$$

El valor óptimo de la función objetivo la podemos escribir como

$$Z_{EV} = min \sum_{t \in T} E(a_t) y_t + E(b_t) y_t^*$$

Sujeto a :

$E(A_t(s))y_{t-1} + E(B_t(s))y^*_{t-1} + E(A^*_t(s))y_t + E(B^*_t(s))y_t = E(d_t)$ para todo $t \in T$

$y_t \in Y \subseteq \Re^{m(1)}$ $y^*_t \in Y^* \subseteq \Re^{m(2)}$, para todo $t \in T$

El problema exige que tanto Y como Y* sean no vacíos y cerrados. Las esperanzas son halladas respecto a la medida de probabilidad utilizada para seleccionar la muestra. Denotamos la solución óptima por (y_{to}, y^*_{to})

El valor esperado en t , $t \in T$, dadas las variables de decisión y_e, $e=1,\ldots,t-1$, es

$$EEV(t) = \begin{cases} RP \\ sujeto\ a \\ y_1(s) = y_{1o} \quad ; \forall s \in S \\ \vdots \\ y_{t-1}(s) = y_{t-1o} \end{cases}$$

El valor de la solución estocástica en t es $VSS(t)=RP-EEV(t)$ y en todo programa estocástico de este tipo $VSS(t+1) \leq VSS(t) \leq 0$, ver Escudero et al (2007).

En nuestro caso podemos escribir, dado que usamos MAS, $p(s)=\lambda$ para todo $s \in S$ el problema como

$$PS(D): min \left\{ \sum_{t \in T} \sum_{s \in S} p(s) \sum_{q=1}^{N} \sum_{b=1}^{M} \sum_{r=1}^{R} P_b V_{qbtr}(s) X_{qtr}(s) = \lambda \sum_{t \in T} \sum_{s \in S} \sum_{q=1}^{N} \sum_{r=1}^{R} \vartheta_{qtr}(s) X_{qtr}(s) \right\}$$

$$sujeto\ a: \begin{cases} \sum_{r=1}^{R} X_{qtr} = 1, \forall q, \forall t \in T \\ \vartheta_{qtr}(s) X_{q(t-1)r}(s) + \vartheta_{qtr}(s) X_{qtr}(s) = d(t) \\ \sum_{r=1}^{R} Z_{tr} \leq KM, \qquad K \leq R \\ X_{qtr} \in \{0,1\}, \qquad \forall q, \forall r, \forall t \\ Z_r \in \{0,1\}, \qquad \forall r \end{cases}$$

$$X_{qtr} = \begin{cases} 1 & si \quad i \in s \cap U_r \\ 0 & si \quad i \notin s \cap U_r \end{cases}$$

$$Z_{tr} = \begin{cases} 1 & si \quad \theta_r \quad es \quad un \quad centroide \quad selecciona do \\ 0 & en \quad otro \quad caso \end{cases}$$

Un arreglo consecuente nos lleva a que se pueda aplicar la solución dinámica promedio que desarrollara Escudero et al (2007) para resolver PR(M). Este es descrito como sigue:

Paso 1. Resolver $Z_{EV}(t)$ y guardar (y_{to}, y^*_{to})

Paso 2. Hacer $t=t+1$ para una nueva muestra s^*

Paso 3. Si $t<T$ regresar al paso 1

Paso 4 La estratificación óptima es la obtenida.

4. DETERMINACIÓN DE LOS ESTRATOS

Existente diversos problemas en los estudios poblacionales en los que los psicólogos desean conformar grupos homogéneos a partir de determinadas conceptiones expresadas en la forma de centroides.

El destacado estadístico Francis Galton fue quien fijara por vez primera ideas claves sobre las diferencias en personalidad. Sugirió que las actitudes relevantes de los individuos son codificables mediante adjetivos. Posteriormente Gordon Allport y H. S. Odbert así como, Raymond Cattell hicieron estudios para determinar un número de adjetivos que describían rasgos observables y relativamente permanentes. Rasgos de personalidad distintivos (esfera de personalidad) fueron identificados y estos dieron pie a construir tests de personalidad para estos rasgos. La estadística fue aplicada para validar sus hipótesis. La técnica aplicada fue la entonces novedosa del análisis factorial.

Trabajos posteriores sentaron las bases teóricas para desarrollos ulteriores fijándose actualmente clasificándose los individuos como:

1. **Extrovertidos:** Son sociables y disfrutan de la compañía de otros, evitan la soledad tienden a experimentar emociones positivas. Son asertivos.
2. **Introvertidos**: Son reservados e independientes Entre sus rasgos positivos está el de ser asertivos, cordiales y amables así como su búsqueda de emociones, emociones positivas. Son negativos su tendencia a la ansiedad, depresión, impulsividad y vulnerabilidad.
3. **Abiertos:** Son imaginativos, estéticamente sensibles y con una intensa vida interna. Entre sus rasgos positivos están su sociabilidad, asertividad y la búsqueda de emociones positivas. Son negativos su tendencia a la convencionalidad y conservadurismo político.
4. **Concienzudos:** Son caracterizados por su auto-control por lo que son planificados, organizados, confiables y escrupulosos. Por otra parte son voluntariosos y amantes del éxito.
5. **Amables:** Son altruistas, considerados, confiados y solidarios pero también egocéntricos, escépticos y competitivos
6. **Neuróticos:** Son inestables emocionalmente, tienden a tener una percepción sesgada hacia las situaciones negativas que hacen que sientan excesivamente emociones negativas. Su conducta no es estable teniendo poca tolerancia ante situaciones de estrés.

Como se ve el término asertivo califica a algunas de las personalidades. La asertividad es aquella habilidad personal que permite expresar de forma adecuada las emociones frente a otra persona sin hostilidad ni agresividad. Una persona asertiva va a expresar directa y adecuadamente sus opiniones y sentimientos (tanto positivos como negativos) en cualquier situación social. La asertividad es el resultado de una serie de conductas, aprendidas o adquiridas, que permite a que un individuo oriente sus decisiones sin lastimar su ego y personalidad haciéndolo sin dañar lo derechos a los demás. Es por ello importante evaluar la asertividad de colectivos para conocer los rasgos de sus componentes en su individualidad. Por ello existen diversas pruebas de ella.

Se realizó un amplio estudio en un colectivo de 81 estudiantes de la Facultad de Psicología de la Universidad de La Habana sobre asertividad. El test constaba de 26 ítems. Se hicieron clústeres utilizando métodos convencionales, ver Díaz et al (2012) para establecer comportamientos en los grupos para hacer posteriormente estudios sobre la pertenencia a ellos y las políticas de formación de asertividad diferenciada. Estos no fueron satisfactorios pues aparecían grupos que no satisfacían los intereses de la investigación al no estar suficientemente diferenciados.

Cairo Valcárcel et al. (2000) analizaron 589 sujetos, 286 hombres y 303 mujeres mediante la prueba de las matrices progresivas de Raven para establecer regularidades en el aprovechamiento de sus múltiples alternativas. Se deseaba establecer el índice de dificultad y discriminación de los 60 ítems de la

prueba. Había 4 grupos etarios. La clasificación debía determinar los aspirantes que serian contratados de acuerdo a los requerimientos de los cargos. Una alternativa a la solución obtenida en este trabajo se encuentra en la determinación de grupos homogéneos para hacer estudios posteriores sobre asimilación de las habilidades del puesto al caracterizar la personalidad de estos su pertenencia a un cierto post estrato..

Cairo Valcárcel et al. (2001, 2002) estudiaron las características diagnósticas del test de tachado de la prueba de atención de la Batería de Diagnóstico Neuropsicológico de la Universidad de La Habana (DNUH versión 2000R) en la exploración de un trastorno que sólo es detectado cuando su manifestación rebasa determinados límites en los cuales se pone en evidencia, en el comportamiento del individuo, una falla atencional con relación a algún suceso que debió percibir, responder u orientar y no lo hizo. Sin embargo, en la mayoría de las ocasiones este hecho pasa desapercibido y con mucha frecuencia se le atribuye a otras causas. Este estudio se desarrolló con niños de nivel secundario (7mo.,8vo, y 9no. grado) dirigido a la "negligencia visual", mediante una tarea de tachado o cancelación de letras dispuestas. La muestra constó de 792 sujetos de los cuales el 51,0% fueron varones (404) y el 49,0% hembras (388) de ellos 36.7% de séptimo grado (291:124 mujeres y 167 hombres) mientras que el 28,9%(229: 120 mujeres y 109 hombres) y el 34,4% (272: 144 mujeres y 128 hombres) fueron de octavo y noveno grado respectivamente. El número de ítems fue de 391. Se llevaron cabo clasificaciones partir de percentiles de las distribuciones marginales. Se hace de interés el formar grupos homogéneos para elaborar inferencias sobre estos que sean lo más precisas posible.

Los trastornos del dormir son un problema que afecta la personalidad de los individuos. Similarmente el embarazo lo hace en las mujeres. Una muestra de 1120 fue analizada a través de 22 ítems además de considerar su edad (entre 19 y 40 años), ocupación y escolaridad. El 50% de la mujeres estaban embarazadas (dividas en primerizas y multíparas), estando en distintos periodos de gravidez (de 16 semanas a mas de 28), y el resto no lo estaban. La frecuencia de los trastornos del dormir fue también medida. El interés es conformar estratos para establecer estudios de las variables involucradas dentro de ellos y caracterizar la problemática en forma particularizada.

Para evaluar la ganancia en precisión debida al uso de los criterios de postestratificación evaluamos, en el caso del problema con restricciones probabilísticas

$$V_{PECR(s)} = min\left\{\sum_{q\in s} \sum_{r=1}^{R} \vartheta_{qr} X_{qr}\right\}$$

Para cada muestra y computamos la media y varianza de la solución en 100 muestras generadas aleatoriamente

$$V(1) = \frac{1}{100}\sum_{s=1}^{100} V_{PECR(s)} \quad , \hat{\sigma}^2\big(V(1)\big) = \frac{1}{100}\sum_{s=1}^{100} (V_{PECR(s)} - V(1))^2$$

Usando la densidad empírica obtenida calculamos el primer y tercer cuartil. $V_{0,25}(1)$ y $V_{0,75}(1)$. La comparación con el MAS fue desarrollado calculando

$$PR(1) = \frac{V(1)}{\sum_{j=1}^{J} P_j V_{MAS(j)}}$$

J es el número de ítems analizados y P_j la importancia asignada a este. Los resultados obtenidos en la simulación aparecen en la tabla 1.

	PR(1)	$\hat{\sigma}\ (V(1))$	V(1)	$V_{0,25}(1)$	$V_{0,75}(1)$
Asertividad	0,64	0,19	7,45	4,67	9,36
Matrices de Raven	0,41	0,22	187,92	74,78	190,74
Negligencia visual	0,66	0,37	17,89	7,80	22,37
Trastornos del dormir	0,92	0,17	15,29	12,95	35,70

Tabla 1. Comportamiento de la estratificación óptima usando PAS-T

Como se ve hay grandes ganancias en precisión pues PR(1) obtiene valores bastante lejanos de 1 excepto para los trastornos en el dormir. Por otra parte su desviación estándar es bastante similar. Las distribuciones parecen ser muy asimétricas respecto al valor promedio de V(1) en todos los casos.

Para el problema multietápico trabajamos con

$$V_{ME}(T) = min\left\{\sum_{t\in T}\ \sum_{s\in S} p(s)\sum_{q=1}^{N}\ \sum_{b=1}^{M}\ \sum_{r=1}^{R} P_b V_{qbtr}(s) X_{qtr}(s)\right\}$$

Entonces para cada *T* la precisión relativa es

$$PR(ME{:}T) = \frac{V_{ME}(T)}{\sum_{j=1}^{J} P_j V_{MAS(j)}}$$

Consideramos *T=2,3,5*.

Los resultados obtenidos en los experimentos realizados se brindan en la tabla 2.

	PR(ME:2)	PR(ME:3)	PR(ME:5)
Asertividad	0,721	0,796	0,849
Matrices de Raven	0,975	0,968	0,977
Negligencia visual	0,915	0,922	0,943
Trastornos del dormir	0,879	0,893	0,949

Tabla 2. Comportamiento de la estratificación óptima usando PSD

Como se ve el modelo multietápico genera ganancias menores que el de restricciones probabilísticas. Además el incremento de las etapas no incrementa la precisión relativa.

REFERENCIAS

[1] ALBAREDA-SAMBOLA, E. & E. FERNÁNDEZ (2000): The stochastic generalized assignment problem with bernoulli demands. **TOP**, *8, 165-190.*

[2] ALLENDE, S.M. & C. BOUZA (1987): Optimization criteria for multivariate strata construction, En "Approximation en Optimization" (**Lecture Notes in Mathematics 1353**), 227-233. Springer, & N. York.

[3] ALLENDE , S., C. N. BOUZA, L. PEDREIRA & L. SINGH (2008): The solution of optimal post-stratification with multiple auxiliary information: an stochastic optimization approach. **Revista Investigación Operacional**, 29, 140-149.

[4] ALLENDE, S., C. BOUZA y· D. COVARRUBIAS (2012): Optimal post-stratification for the study of the sustainability: an application to the monitoring of diversity in sierra de guerrero. **Ann Oper Res**. DOI 10.1007/s10479-012-1154-x

[6] BETHEL, J. (1989): Sample allocation in multivariate surveys. **Survey Methodology**, 15, 47-57.

[7] BOUZA HERRERA, CARLOS N., SIRA M. ALLENDE ALONSO, EDUARDO CAIRO VALCÁRCEL & LUIS PEDREIRA ANDRADE (2009): **La afijación multicriterio de la muestra en el modelo estratificado multivariado**. (eds. J. C. Leyva, E. Aviles y J.J. Zepeda) Plaza Y Valdés, Spain, 2009,383-404.

[8] BÜHLER, A, y DEUTLER, M. (1975): Optimal stratification grouping by dynamic programming. **Metrika,** 22, 161-175.

[9] CAIRO VALCÁRCEL E., E. CAIRO MARTINEZ, C. BOUZA y T. PONCE SOLOZÁBAL (200): Algunas características y posibilidades del test de matrices progresivas de Raven. **R. Cubana de Psicología**, 17, 95-105.

[10] CAIRO VALCÁRCEL, E., E. IJALBA, R. GÓMEZ LOZANO, R. MARÍN LLANES, L. C. DEVIA COLLAZOS, y C. BOUZA HERRERA (2001): Inatención visual unilateral. **Revista Cubana De Psicología** I8, 101-119.

[11] CAIRO VALCÁRCEL, E., E. IJALBA, L.K.. OBLER, GÓMEZ LOZANO, R. MARÍN LLANES, L. C. DEVIA COLLAZOS, y C. BOUZA HERRERA (2002): Bias in visual attention behavior in anormal school aged poplation. **Brain Cogn**, 48, 291-296.

[12] COCHRAN, W.G. (1981): **Técnicas del muestreo**, CECSA, México

[13] DALENIUS, T. y J. L. HODGES (1959): Minimum variance stratification. **J. Amer. Statist. Assoc.** 54, 88-101.

[14] DIAZ, A. C. BOUZA y CAIRO E. (2012): The use of clustering in the tests "are you assertive". **10th International Conference on Operations Research**. La Habana.

[15] HANEVELD, W., K. KLEIN & M.H. van der KLERK (1999): Stochastic integer programming: general models and algorithms. **Ann of Oper. Res**. 85, 39-57

KHAN M.G.M. & M. J. AHSAN (2003): A note on optimum allocation in multivariate stratified sampling. **S. Pac. J. Nat. Sci**., 21, 91- 95.

[16] KOKAN, A.R., & KHAN, S.U., (1967): oOptimum allocation in multivariate surveys: an analytical solution. **J.Roy. Stat. Soc.,** Ser. B, 29, 115-125

[17] LAVALLÈE, P (1987): **Some contributions to optimum stratification**. tesis de MSC., Carleton University, Ottawa.

[18] SALAZAR GONZALEZ, JUAN J- (2001): **Programación Matemática**. Ed. Diaz Santos. Madrid.

[19] SCHENEEBERGER, H- (1985): maxima, minima und sattelepunkte bei optimaler schichtung und optimaler aufteilung. **Allgemeines Statis. ARCHIV**., 69, 286-297.