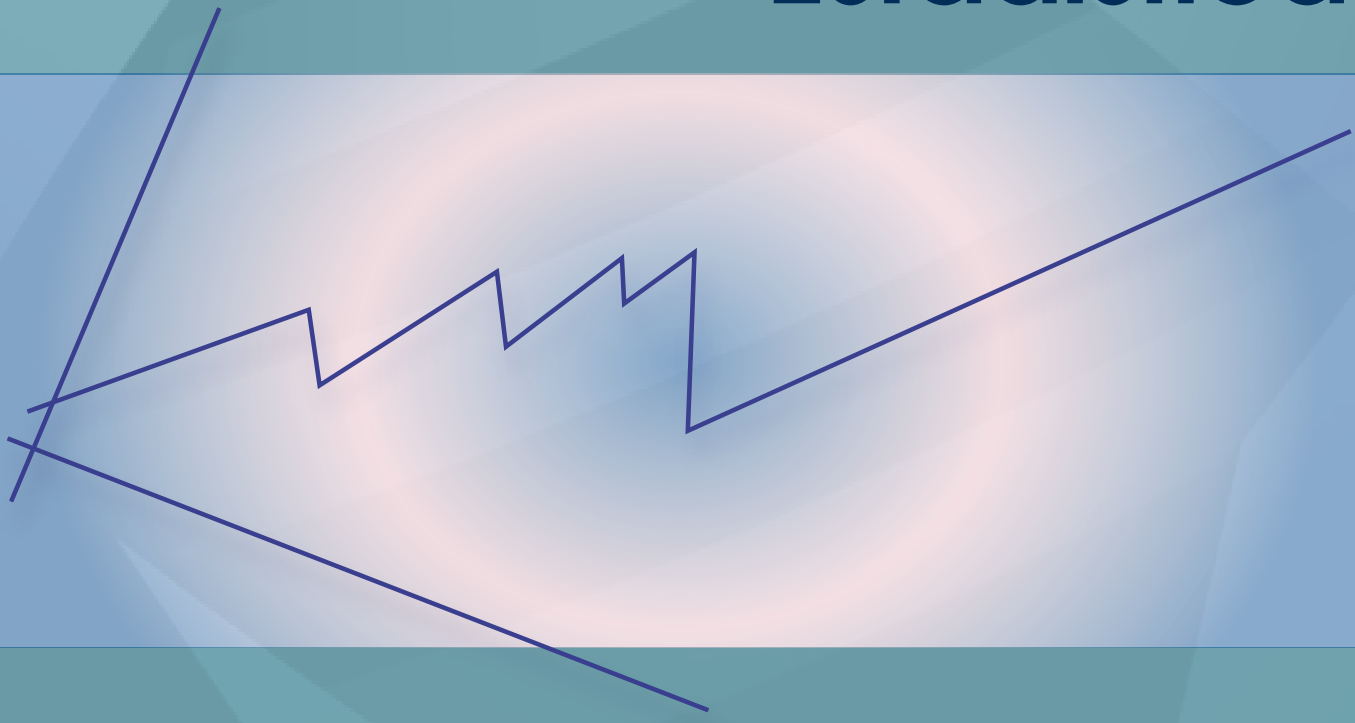
Instituto Nacional de Estadística y Geografía

Asociación Mexicana de Estadística

# Memoria del XXV Foro Nacional de Estadística



AME

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Instituto Nacional de Estadística y Geografía

25 años

# Memoria del XXV Foro Nacional de Estadística

AME

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

**Memoria del XXV Foro**
**Nacional de Estadística**

Impreso en México

# Presentación

Este libro está constituido por 20 artículos que forman la Memoria del XXV Foro Nacional de Estadística, evento que se llevó a cabo en las instalaciones del Instituto Nacional de Salud Pública, en la ciudad de Cuernavaca, Morelos, del 22 al 24 de septiembre de 2010.

La Sección 1 contiene tres artículos y es acerca de temas generales. El primero, de Patricia Romero Mares y Raúl Rueda, invita a preguntarse si, en el caso del muestreo de poblaciones finitas para hacer inferencias, se debe escoger el enfoque basado en el modelo o en el diseño. El segundo, de Alexander von Eye y Patrick Mair, analiza la pérdida de información debido a la dicotomización de variables y presentan un ejemplo empírico donde este proceso no produce tal pérdida. El tercer trabajo, de María Guadalupe Russell Noriega y Enrique Villa Diharce, se refiere al uso de cartas de control en el área de salud.

La Sección 2 es de estadística bayesiana y contiene dos artículos. En el primero, Gabriel Núñez Antonio y Eduardo Gutiérrez Peña presentan "Un Modelo Bayesiano para Datos Longitudinales Circulares". Por su parte, el trabajo de Olga Vladimirovna Panteleeva, Humberto Vaquera Huerta y Eduardo Gutiérrez González trata de los estimadores de máxima verosimilitud en "Modelos de Mezclas Finitas Univariadas".

La Sección 3 está dedicada a problemas de inferencia estadística y lo forman cuatro artículos. En el primero, Félix Almendra Arao nos habla de "Una prueba de no inferioridad Basada en Estimadores de Proporción Contraídos"para la comparación de grupos vía la estimación de la desviación estándar de la diferencia de los estimadores de proporción. María D. Kantúm Chim y José A. Villaseñor Alva, se refieren a "La prueba de Bondad de Ajuste R para la Distribución Exponencial", donde R es la razón de dos estimadores del parámetro de escala de la distribución exponencial. Por su parte, Agustín Jaime García Banda, Luis Cruz Kuri e Ismael Sosa Galindo, utilizan el programa Mathematica para mostrar con ejemplos sencillos "Una Aproximación Binomial con Tres parámetros". El último trabajo de esta sección es el de Luis Cruz-Kuri, Agustín Jaime García Banda e Ismael Sosa Galindo, "Utilización de Procesos de Ramificación para el Estudio del Desarrollo de una Epidemia".

La Sección 4 contiene cuatro artículos sobre el tema de muestreo. Martín H. Félix Medina discute la "Estimación de Totales y Medias en el Muestreo por Bola de Nieve en Presencia de Probabilidades de Nominación Heterogéneas". A continuación, Alberto Manuel Padilla Terán nos habla de "Cotas para la varianza, efecto del diseño y coeficiente de variación de proporciones en el muestreo por conglomerados en dos etapas con tamaños iguales". El tercer trabajo es de Javier Suárez Espinosa "Formulación natural del tamaño de muestra para el caso del Muestreo por Conglomerados en dos Etapas". Esta sección termina con el artículo de Fernando Velasco Luna y Mario Miguel Ojeda Ramírez, "Caracterización del BLUP de la media poblacional finita $Y_j$ en estimación en áreas pequeñas (Small Area Estimation)".

La Sección 5 se dedica a las aplicaciones y está compuesta por siete artículos. El primero es de aplicaciones a las ciencias de la salud: Fidel Ulín Montejo, Jorge A. Pérez Chávez y Rosa Ma. Salinas Hernández discuten un "Análisis de Confiabilidad para Tiempos de Eficacia Analgésica en Pacientes con Cólico Renoureteral". El siguiente trabajo es sobre cuestiones atmosféricas, en el que Nahun Israel Loya Monares, Hortensia J. Reyes Cervantes y Francisco J. Ariza Hernández hablan de la "Modelación de fenómenos atmosféricos usando Procesos de Poisson No Homogéneos". El artículo de Soraida Nieto Murillo, Blanca Rosa Pérez Salvador y José Fernando Soriano Flores, "Credit Scoring: Una Aplicación de la Estadística", presenta una aplicación en finanzas.

Posteriormente vienen dos trabajos de aplicaciones a cuestiones sociales: El primero, de Blanca Rosa Pérez Salvador, "Un modelo de series de tiempo para describir la demanda en grupos escolares con seriación estricta". El segundo, de Alfredo Cuevas Sandoval, Flaviano Godínez Jaimes y Sulpicio Sánchez Tizapa, "Estudio de factores que influyen en la resistencia de los morteros formulados para reparación de vivienda de interés social en la zona costera de Guerrero".

Finalmente, se discuten dos aplicaciones a la agronomía: Emilio Padrón Corral, Armando Muñoz Urbina, Haydée de la Garza Rodríguez e Ignacio Méndez Ramírez, en el trabajo "Relación entre Ecuaciones Estructurales y Correlación Canónica en un Experimento con Guayule", y Lorena Alonso, Dante Covarrubias y Carlos N. Bouza, con "Muestreo por conjuntos ordenados (Ranked Set Sampling) y su aplicación en población de maguey silvestre".

A nombre de la Asociación Mexicana de Estadística, agradecemos al Instituto Nacional de Salud Pública todo su entusiasmo y dedicación en la organización de este Foro. También estamos en deuda, una vez más, con el Instituto Nacional de Estadística y Geografía por su apoyo para la publicación del presente volumen. Finalmente, nuestro reconocimiento y gratitud a todos y cada uno de los colegas que amablemente accedieron a revisar los trabajos que aquí se presentan.

**El Comité Editorial**

Juan González Hernández

Juan Morales Velasco

Elida Estrada Barragán

# Índice general

## Sección I. Temas Generales



## Sección II. Estadística Bayesiana



## Sección III. Inferencia Estadística

## Sección IV. Muestreo



## Sección V. Aplicaciones

# Sección I

# Temas Generales

Memoria FONE XXV (2010), Sección I, 3–9

# Inferencias basadas, ¿en modelo o en diseño?

Patricia Romero Mares[a], Raúl Rueda Díaz del Campo[b]
*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas*
*Universidad Nacional Autónoma de México*

## 1. Introducción: un poco de historia

El muestreo es la única actividad estadística en donde los procesos de inferencia son completamente distintos a las otras áreas. Dado un mecanismo aleatorio que permite seleccionar muestras aleatorias de una población finita, las inferencias se basan en la distribución inducida por este mecanismo, independientemente de la estructura de la población de donde las muestras son obtenidas. Al final del Siglo XIX, la teoría de muestreo estaba en sus inicios, para algunos el censo era el único medio para estudiar poblaciones finitas.

La "revolución muestral" se inició con un artículo de J. Neyman publicado en 1934 y cuyo título podría traducirse como: "Sobre dos diferentes aspectos del método representativo: el método de muestreo estratificado y el método de selección dirigida" [selección subjetiva dependiente del investigador].

El objetivo para Neyman era: "Dada una población hipotética, encontrar la distribución de ciertas características en muestreo repetido". Y añade: "La solución consiste en determinar ciertos intervalos, a los que llamaré intervalos de confianza y que, suponemos, contienen a los valores poblacionales con una "probabilidad"de error que no sea mayor a $1-\epsilon$, donde $\epsilon$ es cualquier número tal que $0 < \epsilon < 1$. A este número $\epsilon$ le llamo el coeficiente de confianza".

Esta propuesta dice más de lo que parece: Si estamos interesados en una característica $X$ de una población $P$, los métodos de muestreo y de estimación propuestos nos permiten, para cada muestra $S$ que obtengamos, asignar un intervalo de confianza $(X_1(S), X_2(S))$ tal que la frecuencia de errores en la proposición

$$X_1(S) \leq X \leq X_2(S)$$

[a]patricia@sigma.iimas.unam.mx
[b]pinky@sigma.iimas.unam.mx

no exceda el límite $1 - \epsilon$ fijado de antemano, sin importar cuáles son las propiedades desconocidas de la población [una solución no paramétrica].

La proposición de confianza no es una proposición probabilista sino una proposición sobre el comportamiento frecuentista basado en todas las muestras aleatorias posibles que pueden ser obtenidas [¿violación al principio de verosimilitud?]. Como tal, las únicas probabilidades que aparecen, son las inducidas por la aleatoriedad en la selección de la muestra (Cornfield, 1944) y la proposición de confianza será válida independientemente de la estructura de la población.

Horvitz y Thompson (1952) es otro trabajo fundamental en el desarrollo del muestreo. En él se presenta un tratamiento elegante sobre muestreo probabilista que completa las bases para hacer inferencias usando la distribución en el muestreo [*p-distribution*] e introducen el uso de probabilidades diferentes de selección. Sospechan que a pesar de poder encontrar el mejor estimador lineal para cada diseño muestral, no existe el estimador óptimo para todos los diseños. Godambe (1955) confirma la conjetura de Horvitz-Thompson. Demuestra que no existe un estimador lineal insesgado que tenga varianza mínima uniformemente para todas las poblaciones. Como consecuencia de esto, ninguna comparación empírica puede ser concluyente.

Resumiendo, los fundamentos tradicionales de la teoría del muestreo están basados casi en su totalidad en el trabajo fundamental de Neyman. El diseño y las inferencias están basadas en intervalos de confianza calculados de acuerdo con la aleatorización realizada, dándole protección al estadístico en repetidas aplicaciones de todos los diseños muestrales.

Sin meternos en demasiados detalles, es bien sabido [en el mundo bayesiano, si se quiere] que las inferencias basadas en muestras no observadas, violan el principio de verosimilitud: *Todas las inferencias deben estar basadas sólo en la información contenida en la función de verosimilitud*. Basu (1969) demuestra, como consecuencia de los principios de suficiencia y verosimilitud, que condicional a la muestra observada, las inferencias sobre los parámetros poblacionales, son independientes del diseño muestral. Todo esto lleva a preguntarnos:

¿Porqué la forma de hacer inferencia cuando tenemos poblaciones finitas debe ser tan distinta a como hacemos inferencia en el resto de la estadística?

# 2. Inferencias en poblaciones finitas

## 2.1. Introducción y notación

Supongamos que tenemos una población finita $P = \{X_1, \ldots, X_M\}$, donde cada $X_j$ puede ser un vector en $\Re^k$. Queremos hacer inferencias sobre funciones de esta población, a partir de una muestra $z_m = \{x_1, \ldots, x_m\}$. Algunas funciones de interés son la media y la matriz de varianzas poblacionales.

Por simplicidad, diremos que un *diseño muestral* es un conjunto de muestras, $z_m$, de $P$ con una medida de probabilidad asignada, de manera que $\sum_{z_m} P(z_m) = 1$, que denotaremos como $P_D$. De acuerdo a Little y Rubin (1983), existen tres enfoques relevantes, para hacer inferencias en poblaciones finitas:

1. Inferencias basadas en diseño, en las que las proposiciones probabilísticas están basadas en la distribución de la selección muestral [Teoría clásica de muestreo].

2. Inferencias basadas en modelo, en donde se supone que los valores de la población fueron generados por un modelo con [hiper] parámetros fijos. [Modelado frecuentista de superpoblaciones].

3. Inferencia bayesiana. En general, se supone un modelo jerárquico.Como los principios de verosimilitud y suficiencia son naturales en la perspectiva bayesiana, no existen "inferencias bayesianas basadas en diseño".

   [Algunas referencias para el segundo punto son Royall (1970), Thompson (1988) y Valliant *et al.* (2000)].

## 2.2. Inferencias por diseño

Están basadas en la distribución del diseño muestral, pues las variables de muestreo $X$ son tratadas como cantidades fijas [¡ahora sí son fijas!]. La medida de probabilidad usada para hacer inferencias es $P_D$.

De acuerdo a Neyman (1934), para hacer inferencias sobre la cantidad de interés $Q = Q(X)$ se siguen los siguientes pasos:

1. **Elegir un estimador** $\hat{q} = \hat{q}(X)$, **que sea** (aproximadamente) $P_D$**-insesgado** para $Q$. [Nótese que $\hat{q}$ es una variable aleatoria debido a la *aleatoriedad* del proceso de selección].

2. **Elegir un estimador** $\hat{v} = \hat{v}(X)$, **que sea** (aproximadamente)$P_D$**-insesgado** para la varianza de $\hat{q}$. Es decir, $\hat{v}$ es un estimador de la varianza del estimador de $Q$.

3. Las inferencias se basan en suponer una **aproximación a normalidad para muestras grandes**; por ejemplo, un intervalo al 95 % de confianza para $Q$ está dado por $\hat{q} \pm 1.96\sqrt{\hat{v}}$

Ejemplo. Queremos estimar la media poblacional $\bar{X}_p$ en un diseño muestral estratificado formado con $J$ estratos de tamaño $M_j$ cada uno. La cantidad de interés es $Q = \bar{X} = \sum_{j=1}^{J} W_j \bar{X}_j$, donde $W_j = M_j/M$ es la proporción de la población que pertenece al $j$-ésimo estrato y $\bar{X}_j$ la media [poblacional] del $j$-ésimo estrato. Suponemos que una muestra aleatoria [simple] de tamaño $m_j$ es seleccionada en cada estrato $j$ y que todas las unidades seleccionadas son observadas.

Así, el estimador para $\bar{X}_p$ es la media muestral estratificada

$$\hat{q} = \bar{X}_{st} = \sum_{j=1}^{J} W_j \bar{x}_j,$$

que es una media ponderada de unidades muestrales [estimador lineal].

El estimador de la varianza de la media estratificada estimada es

$$\hat{v}_{st} = \sum_{j=1}^{J} W_j^2 \left(1 - \frac{m_j}{M_j}\right) \frac{s_j^2}{m_j}$$

donde $s_j^2$ es la varianza muestral en el estrato $j$. Estos dos estimadores son la base para calcular intervalos de confianza. Por ejemplo, un intervalo de 95 % de confianza para $\bar{X}_p$ está dado por

$$\bar{x}_{st} \pm 1.96\sqrt{\hat{v}_{st}}. \tag{1}$$

## 2.3. Inferencias por modelo

Desde la perspectiva Bayesiana, suponemos el siguiente modelo

$$X_{ij} \sim N\left(X_j|\theta_j, h_j\right) \quad \forall i, j \tag{2}$$

$$\text{y} \quad p(\theta_j, h_j) \propto h_j^{-1} \quad \forall j \tag{3}$$

no es difícil demostrar que la distribución final de $\bar{X}_p$ tiene como valor esperado a $\bar{X}_{st}$ y varianza

$$\sum_{j=1}^{J} W_j^2 \left(1 - \frac{m_j}{M_j}\right) \frac{s_j^2}{m_j - 3},$$

por lo que intervalos de probabilidad tendrán la misma cobertura que los frecuentistas.

La ventaja del modelado, es que podemos pensar en estructuras más complejas, como por ejemplo

$$\begin{aligned} p(X_{1j}, \ldots, X_{m_j j}) &= \prod_{j=1}^{J} \prod_{i=1}^{m_j} N(X_{ij}|\theta_j, h_j) \\ p(\theta_1, \ldots, \theta_J | h_1, \ldots, h_J) &= \prod_{j=1}^{J} N(\theta|\mu, h_j\tau) \\ \text{y} \quad p(h_1, \ldots, h_J) &\propto h_1^{-1} \cdots h_J^{-1}, \end{aligned}$$

que expresa de manera natural la idea de superpoblaciones: las subpoblaciones o estratos se consideran una muestra aleatoria de una población mayor, regida por los hiperparámetros $(\mu, \tau)$.

# 3. Ejemplos

En los siguientes ejemplos una población es simulada, de manera que conocemos el valor "real"de $\bar{X}_p$. La población es dividida en $J$ subpoblaciones.

Para cada ejemplo, se calcula el contenido frecuentista de los intervalos generados en cada propuesta.

En el primer ejemplo, se generan 15 subpoblaciones Normales formando verdaderos estratos, esto es, distintos entre si y con poca varianza al interior. El segundo ejemplo es una situación extrema: se generan uniformes reales en el intervalo [0,300000], este rango se divide en 15 intervalos de distintas longitudes con los que se generan las 10 poblaciones; por ejemplo, una población es el intervalo extremo derecho, otra el extremo izquierdo, otra una mezcla del tercero y cuarto intervalos, de tal manera que las subpoblaciones no se traslapen y tengan mucha varianza al interior. Finalmente, para el ejemplo 3, se generan 12 subpoblaciones Poisson con diferentes parámetros que varían desde 0.48 hasta 11.21.

En los tres casos, se uso (1) como solución frecuentista. La solución bayesiana fue (2) para el ejercicio 1 y para el 2; para el ejercicio 3 se supuso

$$\begin{aligned} X_{ij} &\sim Po\left(X_j|\lambda_j\right) \quad \forall i \\ \Pi\left(\lambda_j\right) &\propto \lambda_j^{-1/2} \end{aligned}$$

Los intervalos son del 90 % de confianza.

| Cobertura | Diseño | Modelo |
|---|---|---|
| Ejemplo 1 | 89.9 | 90.8 |
| Ejemplo 2 | 79.2 | 81.4 |
| Ejemplo 3 | 88.7 | 88.3 |

# 4. Conclusiones

¿MODELO O DISEÑO?

# Bibliografía

Basu, D. (1969), "Role of sufficiency and likelihood", *Sankhya* **A31**, 441–454.

Cornfield, J. (1994), "On samples from finite populations.", *J. Amer. Statist. Assoc.* **39**, 236–239.

Godambe, V. (1955), "A unified theory of sampling from finite populations", *J. Royal Statist. Soc. B* **17**, 267–278.

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

Little, R.J. y Rubin, D. (1983), "Discusión de six approaches to enumerate survey sampling", *por K.R.W. Brewer y C.E. Särndal, Proceedings of the Symposium on Incomplete Data, Washington, D.C.: National Academy of Sciences* .

Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection", *J. Roy. Statist. Soc.* **97**, 558–625.

Royall, R. (1970), "On finite population sampling under certain linear regression models", *Biometrika* **57**, 377–387.

Thompson, M. (1988), "Superpopulation models", *Encyclopedia of Statistical Sciences* **9**, 93–99.

Valliant, R., Dorfman, A. y Royall, R. (2000), "Finite population sampling and inference: A prediction approach", *New York: Wiley* .

Memoria FONE XXV (2010), Sección I, 11–19

# On the effects of dichotomizing information

Alexander von Eye[a]
*Michigan State University and University of Vienna*

Patrick Mair
*WU Vienna University of Economics and Business*

## 1. Introduction

Categorizing data is defined as "changing a continuous variable to a categorical form" Durkalski y Berger (2004, p. 239). A special case of categorization is dichotomization which involves changing a continuous or categorical variable to a categorical form that possesses two categories. The attitudes toward dichotomizing data are extreme. On one side, there are researchers in marketing or in clinical psychology who regularly and routinely dichotomize data. Reasons given for this habit include statements that decisions have to be made that are dichotomous. Examples of such decisions include whether a patient is sent to the therapist or not, whether a candidate is hired or not, and whether a product is purchased or not. On the other side, there are recommendations by statisticians who argue strongly against dichotomizing. For example, Royston et al. (2006) call the practice of dichotomizing continuous predictors in multiple regression a bad idea. Similarly, Streiner (2002) speaks about the "heartbreak of dichotomizing continuous data" (p. 262). Irwin y McClelland (2003) conclude that "dichotomization has only negative consequences and should be avoided' (p. 366), and van Belle (2008) advises "do not dichotomize unless absolutely necessary" (p. 138). The author further notes that the asymptotic relative efficiency will always be reduced when data are dichotomized.

Turning the page, Westfall (2011) shows that instances exist in which dichotomizing actually improves the power of statistical testing. In a different context, Petrie y Sabin

[a]voneye@msu.edu

(2009) note that "dichotomizing the dependent variable ... may simplify the fitting and interpretation of the statistical model" (p. 96).

In one word, it is easy to find strong arguments in support of dichotomizing data, and it is equally easy to find strong arguments against dichotomizing MacCallum et al. (2002). The number of studies in which researchers dichotomize data is legion. This is counter the routine arguments against dichotomizing which include the statements that information is lost, statistical power is lost, and that the relative efficiency of statistical testing suffers. In this contribution, we do not intend to take sides. Instead, we pursue two goals. First, we describe the effects of dichotomizing on the information that variables carry Schuster (2009). For this description, we use the tools of information theory. Second, we present an example of empirical data analysis that did not suffer when variables were dichotomized.

## 2. Measures of Entropy

In this section, we briefly review the measures of entropy Weaver y Shannon (1949). *Entropy* is defined as the amount of information that is carried by a variable. This definition can be used for continuous as well as categorical variables. For categorical variables, one specifies

$$H(F) = -\sum_i p_i \log p_i,$$

where $p_i$ is the probability of the $i$-th category. For the entropy of a continuous univariate distribution, one specifies

$$H(F) = -\int f(x) \log f(x) d(x).$$

This measure is also known as *differential entropy.*

The remainder of this article is structured as follows. In the next section, we present illustrations of the effects of categorizing and dichotomizing data. The first two examples consider equiprobable and non-equiprobable scores. The third example considers asymmetric distributions. In the fourth example, the structure of empirical data is modeled once before and once after dichotomizing the variables.

# 3. The Effects of Categorizing/Dichotomizing Data

## 3.1. Example 1: Reducing the Number of Categories

In this example, it is shown how the information carried by a variable with equiprobable scores is reduced by categorization. It is shown that the amount of information lost depends on both the number of originally possible scores and the number of categories created by categorization. To show this, a string of 50 scores was created, and the entropy of this string was calculated as the entropy of a uniform distribution, that is, a distribution with the maximum entropy for each score $i$. Two ratios were calculated.

The first is $H(F)_i/H(F)_{50}$ for $i < 50$. This ratio indicates the portion of information left after categorizing the 50 scores into i categories (Ratio 1). For $i = 2$, one obtains the amount of information left when the distribution of 50 scores is dichotomized.

The second ratio is $H(F)_{i-1}/H(F)_i$. This ratio indicates the portion of entropy left after the distribution of $i$ scores was reduced to $i - 1$ categories (Ratio 2). For $i - 1 = 2$, one obtains the amount of information left when 3 categories are reduced to 2. Figure 1 displays the results of this simulation.

Figure 1 illustrates that, by definition, the amount of entropy (variation) increases with the number of categories (curve of circles). Note, again, that, in this example, all values (categories) are equiprobable. Ratio1 shows that the reduction of 50 values to 49 results in a loss of information that is minimal. Specifically, the amount of entropy left after this transformation is 0.995 of that of the original information. In contrast, dichotomizing 50 equiprobable values leaves one with 17.72 % of the original entropy. This reflects the maximum loss that can result if a uniform distribution with 50 scores is dichotomized. Naturally, if the original uniform distribution has fewer different scores, information loss from dichotomization is smaller. For example, dichotomizing three categories leaves one with 63.09 % of the original entropy

Ratio 2 shows that the amount of remaining entropy that results from reducing the number of values by 1 decreases in a non-linear fashion as the number of values decreases. As for Ratio 1, the reduction of 50 values to 49 leaves one with 99.5 % of the original information, and the reduction of 3 scores to 2 leaves one with 63.09 % of the information contained in three equiprobable scores.

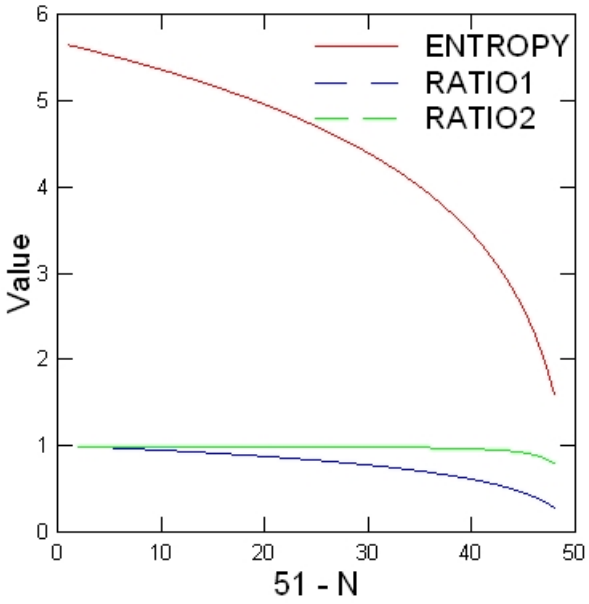



**Figura 1:** Information loss from categorizing variables.

## 3.2. Example 2: Reducing the Number of Categories in Uniform Versus Non-uniform Distributions

In this example, we compare the amount of information lost in a uniform distribution with that from a normal distribution. Figure 2 shows first, that the distribution of the entropy components reflect the shape of the underlying distribution. As the information carried by a non-uniform distribution is less than the information carried by a uniform distribution, the effects of categorization, especially dichotomization, can be expected to be less dramatic. In the present example, 17 scores from a normal distribution were used. The remaining information is, after dichotomizing the 17 intervals of the normal distribution, 32.66 % of the information contained in the 16-interval original distribution. For a uniform distribution with 17 scores, the amount of entropy after dichotomization is only 23.81 %. The portion of remaining information is larger when normal distributions are dichotomized that have fewer than 17 different scores. Examples of this include the popular 5-score Likert scales.

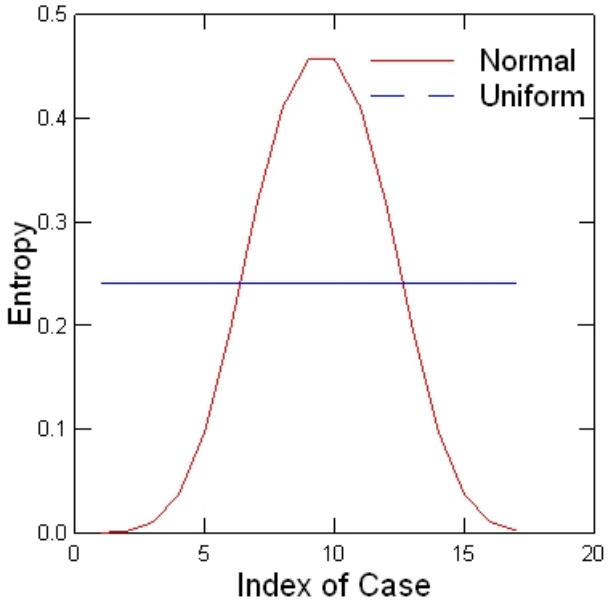



**Figura 2:** Information in a uniform distribution and a normal distribution.

## 3.3. Example 3: Reducing the Number of Categories in Asymmetric Distributions

For the following illustration, scores of a standard $\chi^2$ distribution with $df = 15$ were split into 8 equal intervals. The smallest score was 5; the largest was 40; step size was 5. The resulting intervals were equidistant on the $\chi^2$ scale and came with the probabilities (from 5 to 40) 0.9921, 0.8197, 0.4514, 0.1719, 0.0499, 0.0119, 0.0025, and 0.0005. For each of the 8 intervals, the corresponding entropy was calculated. The sum of the entropy components of this $\chi^2(15)$ distribution was 2.1183. The entropy of 8 equiprobable scores is 3, a difference of 29.39 %.

Figure 3 illustrates, again, that entropy components reflect the shape of the underlying distribution. As the information carried by an asymmetric distribution is less than the information carried by a symmetric distribution, the effects of categorization can be expected to be even less dramatic than the effects of categorizing a uniform distribution. In the present example with 8 different scores from a $\chi^2(15)$ distribution, the remaining information is, after dichotomizing the 8 intervals of the $\chi^2$ distribution, 47.21 % of the information contained in the 8-interval original distribution. Had a uniform distribution with eight different scores

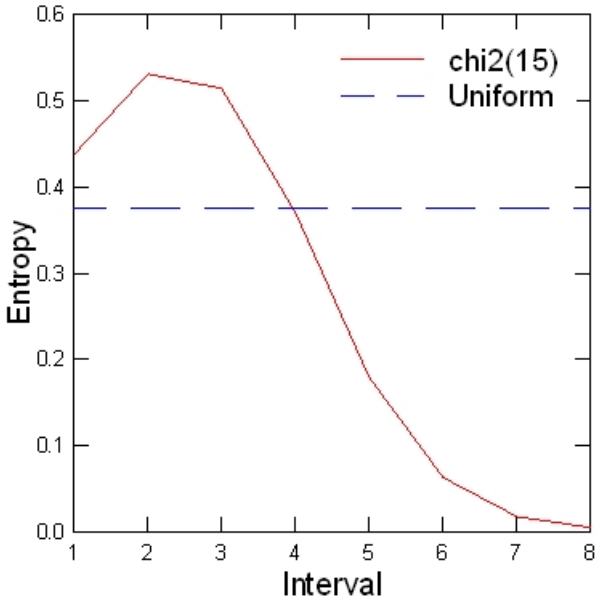



**Figura 3:** Information in a uniform distribution and a $\chi^2(15)$ distribution.

been dichotomized, the remaining entropy would have been 33.33 %, a portion of 70.60 %. The portion of remaining information will be larger when asymmetric distributions are dichotomized that have fewer than 17 different scores. Examples of this include, again, 5-score Likert scales.

## 3.4. Example 4: On the Effects of Dichotomization on the Representation of the Structure of Data

For the following illustration, we use empirical data. We ask whether dichotomizing will always have the effect that data characteristics of interest become invisible. The data we use stem from a study on positive youth development Lerner et al. (2005). We use information about delinquent behavior and positive youth development (PYD) in a sample of 675 male and female adolescents, and ask whether delinquent behavior can be predicted from PYD. We estimate the same recursive latent variable model twice. For the first model, we use the original raw data. For the second model, we dichotomized the raw data at the median. Figure 4 shows the results for the model that uses continuous data.

Clearly, the model describes the data well. Now, if dichotomizing reduces the information

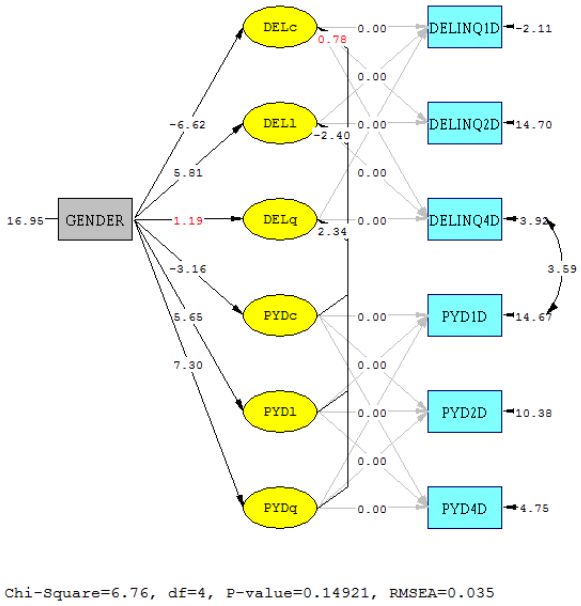



**Figura 4:** Structural model of the relationship between PYD and delinquent behavior.

in a data set considerably, it would be hard to fit the same model to the dichotomized data. In the present example, this was not the case. Specifically, the exact same model fit very well again ($\chi^2(4) = 9.64$; $p = 0.047$; $RMSEA = 0.049$). The only difference between the two models was that the path from Gender to the curvature of delinquent behavior now is significant (it was nonsignificant in the model with the continuous data).

# 4. Discussion

There is a price to be paid for categorizing data. This price is a loss of information (variation). If this variation is reliable and interpretable, dichotomizing eliminates it, to a degree. In other instances, the information that researchers are interested in, e.g., the magnitude of the $Y$ scores as predictable from $X$ may come through even after or only after dichotomization. This result is reflected in studies that show that power is neither systematically nor always lost by way of dichotomization Teng et al. (2009).

We conclude that data can carry information that reflects multiple characteristics. If researchers focus on characteristics that are clouded by variation that may reflect a dif-

ferent characteristic, elimination of this variability can result in a situation in which the characteristic of interest becomes accessible. Categorization can be a method of eliminating information that a researcher is not interested in. Almost always, categorization results in loss of information. However, not all the information that is lost leads to a loss of relevant information.

# Bibliografía

Durkalski, V. y Berger, V. W. (2004), Categorizing data, *in* B. S. Everitt y D. C. Howell, eds, "Encylopedia of Statistics in Behavioral Science, Vol. 1", Wiley, Chichester, UK, pp. 239–242.

Irwin, J. R. y McClelland, G. H. (2003), "Negative consequences of dichotomizing continuous predictor variables", *Journal of Marketing Research* **40**, 366–371.

Lerner, R. M., Lerner, J. V., Almerigi, J., Theokas, C., Phelps, E., Gestsdóttir, S., Naudeau, S., Jelicic, H., Alberts, A. E., Ma, L., Smith, L. M., Bobek, D. L., Richman-Raphael, D., Simpson, I., Christiansen, E. D. y von Eye, A. (2005), "Positive youth development, participation in community youth development programs, and community contributions of fifth-grade adolescents: Findings from the first wave of the 4-h study of positive youth development", *Journal of Early Adolescence* **25**, 17–71.

MacCallum, R. C., Zhang, S., Preacher, K. J. y Rucker, D. D. (2002), "On the practice of dichotomizing quantitative variables", *Psychological Methods* **7**, 19–40.

Petrie, A. y Sabin, C. (2009), *Medical Statistics at a Glance*, 3rd edn, Wiley, New York.

Royston, P., Altman, D. G. y Sauerbrei, W. (2006), "Dichotomizing continuous predictors in multiple regression: A bad idea", *Statistics in Medicine* **25**, 127–141.

Schuster, D. (2009), Bemerkungen zur Dichotomisierung metrischer Variablen. Paper presented at the 9. Tagung der Fachgruppe Methoden und Evaluation, Bielefeld, Germany.

Streiner, D. L. (2002), "Breaking up is hard to do: the heartbreak of dichotomizing continuous data", *Canadian Journal of Psychiatry* **48**, 429–430.

Teng, C.-H., Fedorov, V., Mannino, F. y Zhang, R. (2009), "Statistical power consideration of dichotomizing continuous outcomes". http://biometrics.com/wp-content/uploads/2009/12/Dichotomization2009.ppt.

van Belle, G. (2008), *Statistical Rules of Thumb*, Wiley, New York.

Weaver, W. y Shannon, C. E. (1949), *The Mathematical Theory of Communication*, University of Illnois, Urbana, IL.

Westfall, P. H. (2011), "Improving power by dichotomizing (even under normality)", *Statistics in Biopharmaceutical Research* . Forthcoming.

Memoria FONE XXV (2010), Sección I, 21–28

# Uso de cartas de control en el área de la salud*

María Guadalupe Russell Noriega[a]
*Universidad Autónoma de Sinaloa, Culiacán, Sin.*

Enrique Villa Diharce[b]
*Centro de Investigación en Matemáticas, Guanajuato, Gto.*

## 1. Introducción

El Control Estadístico de Procesos (CEP) es una filosofía, una estrategia y un conjunto de métodos para el mejoramiento de sistemas y procesos. El enfoque del CEP está basado en el aprendizaje a partir de los datos y se fundamenta en la teoría de la variabilidad (causas especiales y causas aleatorias). Todos los procesos, incluyendo todos los aspectos del cuidado médico, se suponen sujetos a variación aleatoria intrínseca (causa común). El propósito de las cartas de control es distinguir entre variación aleatoria y variación por causas especiales, la cual surge de factores externos al proceso. Las herramientas del CEP, como las cartas de control propuestas en 1924 por Shewhart (1926), han sido utilizadas ampliamente en la industria de manufactura

La construcción de una carta de control se basa en la distribución de la característica de interés, visualizando el conjunto de observaciones de las característica como una serie cronológica que nos ayuda a entender, controlar y mejorar los procesos.

La aplicación de las cartas de control en el contexto clínico es reciente y posee retos especiales, Winkel and Zhanh (2007). Por ejemplo, al considerar un proceso determinado de la práctica clínica, aún un pequeño cambio en la media puede ser de gran interés ya que este puede indicar, un incremento en una tasa de mortalidad y además de que los pacientes

*Este trabajo fue realizado con el auspicio del Proyecto PROFAPI2010/123
[a]mgrussell@uas.uasnet.mx
[b]villadi@cimat.mx

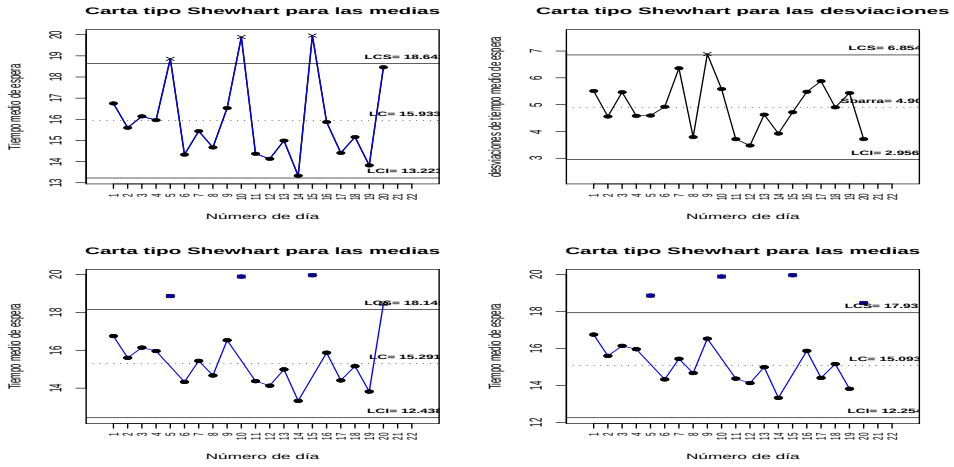



**Figura 1:** Cartas de control de medias y desviaciones del Ejemplo de tiempos de espera.

varían considerablemente, no es posible controlar las entradas del proceso, es decir el número y tipo de paciente. Debido a lo anterior, se dice que existe una gran heterogeneidad entre los pacientes y al monitorear, por ejemplo, la mortalidad en el caso de cirugías, debemos tomar en cuenta algunos factores de riesgo como: enfermedades previas, edad, género, nivel socioeconómico, tipo de cirugía, día de la semana en que se realiza, si es programada o urgente y otros.

En el área de la salud, las cartas de control han venido a ayudar y mejorar algunos de los procesos clínicos y administrativos, como son: infecciones y complicaciones quirúrgicas, lesiones de pacientes, monitoreo de presión arterial, control de glicemia, proporción de pacientes con neumonía debido al uso de ventiladores, enfermedades cardiacas, desempeños de asistencia médica o sanitaria.

**Ejemplo de tiempos de espera** En una clínica, durante veinte días se toman muestras al azar de 30 pacientes (que deambulaban en la sala de espera) y se mide su tiempo de espera, que inicia cuando el paciente llega a la clínica y termina cuando el médico lo recibe. Con el fin de evaluar la calidad del servicio, se estudian los tiempos de espera de los pacientes, utilizando cartas de control de medias y desviaciones. En la Figura 1, tenemos en la parte superior izquierda, la carta de las medias de los veinte dias y en la parte superior derecha tenemos la carta de desviaciones. En esta podemos ver que la variabilidad dentro de las muestras esta controlada. En cambio en la primer gráfica observamos que los puntos 5,10 y

15 se separan significativamente del resto, lo cual nos dice que hay una causa especial que esta actuando esos días. Al eliminar estos tres puntos recalculamos la carta de medias y obtenemos la gráfica inferior izquierda en donde notamos que hay el punto, correspondiente al último día, se encuentra por encima del límite de control superior, esto es, se encuentra fuera de control. Al eliminar este último día del análisis, recalculamos nuevamente la carta de control y obtenemos la carta de medias que se encuentra en la parte inferior derecha de la Figura 1. En esta carta notamos que los quince puntos incluidos en el estudio, se encuentran en control. Estos puntos corresponden a un proceso de espera estable. Las muestras 5, 10, 15 y 20 corresponden a los días viernes, en los cuales se tiene un número mayor de pacientes que requieren el servicio de cardiología lo cual requiere de un mayor tiempo de atención y el cardiólogo solo presta atención los viernes.

# 2. Cartas de control en cirugías

## 2.1. Cartas CUSUM

Las cartas CUSUM fueron propuesta por Page (1954) y se desarrollaron en la industria para aumentar la capacidad de detección de cambios pequeños persistentes en los procesos. En los últimos años se han aplicado las cartas CUSUM en medicina, en el monitoreo de procesos como: Tiempos de espera de atención y también, resultados (Éxito o Fracaso) de cirugías Steiner et. al.(2000). La CUSUM se basa en una prueba secuencial tipo Wald utilizada para elegir entre dos posibles hipótesis bajo un conjunto de resultados observados de la característica de interés $W_t$ y un límite de control o intervalo de decisión $H = h\sigma$ definido como un múltiplo de la desviación estándar del proceso.

En la industria los procesos tienen mayor estabilidad que en el caso del área de la salud, por ejemplo, los pacientes que reciben un tratamiento presentan diferentes riesgos y de aquí la importancia de considerar los riesgos propios del paciente en el monitoreo del proceso.

## 2.2. Monitoreo de un proceso de cirugías

Se ilustra el monitoreo de los resultados de cirugías mediante un proceso CUSUM, para los siguientes casos:

*Sin incluir el riesgo preoperatorio del paciente (CUSUM).*

*Incluyendo el riesgo pre-operatorio del paciente (CUSUM-RA).*

Expresamos los dos estados del proceso de cirugías mediante las siguientes hipótesis:

$H_0 \;:\; p = p_0$ Tasa de fallas de las cirugías, cuando el proceso está en control

$H_A \;:\; p = p_A$ Tasa de fallas de la cirugías, cuando el proceso está fuera de control, con $p_0 < p_A$.

La carta CUSUM puede utilizarse para detectar un incremento o decremento de la mortalidad o ambos. Para detectar un incremento en la mortalidad, tomamos,

$$S_j = \text{máx}\{0, S_{j-1} + W_j\}.$$

Mientras que para detectar un decremento en la mortalidad tomamos,

$$S_j = \text{mín}\{0, S_{j-1} - W_j\}.$$

Considerando $S_0 = 0$, y una métrica de peso $W_j$, que mide la evidencia a favor de $H_A$, respecto a $H_0$.

Sea $y_t$, el resultado de la cirugía en el paciente $t$, donde los resultados posibles son 1 (muere) y 0 (sobrevive). Sea $f(y_t|p)$ la función de verosimilitud de $p$ dado $y_t$,

$$f(y_t|p) = p^{y_t}(1-p)^{1-y_t}$$

El diseño de la carta CUSUM esta determinado por la elección del peso $W_t$ y el límite de control $h$.

La elección óptima de los pesos se basa en el logaritmo de la razón de verosimilitudes (Moustakides (1986)).

$$W_t = \log\left[\frac{f(y_t|p_A)}{f(y_t|p_0)}\right] = \log\left[\frac{p_A^{y_t}(1-p_A)^{1-y_t}}{p_0^{y_t}(1-p_0)^{1-y_t}}\right]$$

Considerando que la CUSUM prueba secuencialmente $H_0 : p = p_0$ contra $H_0 : p = p_A$, tenemos que la función de peso para este ejemplo es:

$$W_t = \begin{cases} \log(p_A/p_0) & \text{si } y_t = 1 \\ \log[(1-p_A)/(1-p_0))] & \text{si } y_t = 0. \end{cases}$$

Cuando diseñamos una carta para detectar incrementos en las tasas de fallas de la cirugía, los pesos asociados con las fallas serán positivos. Si consideramos el riesgo pre-operatorio, el

cual varía de un paciente a otro, es apropiado hacer un ajuste debido a tal riesgo. Para cada paciente se estima previamente el índice de Parsonnet, et. al. (1989) y posteriormente se obtiene la mortalidad dentro de los 30 días posteriores a la cirugía, con el siguiente modelo,

$$\text{logit(p)} = \beta_0 + V,$$

con $V$ la puntuación del índice de Parsonnet, donde $V = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. Las variables $X_1, X_2, ..., X_n$, son los factores de riesgo del paciente, como: edad, género, peso, enfermedades previas, además de otros. Los valores de los parámetros $\beta_1, \beta_2, ..., \beta_n$ se estiman mediante un análisis de regresión, considerando un gran número de resultados de operaciones realizadas anteriormente. Parsonnet et. al. (1989), utilizaron 14 factores de riesgo y realizaron el análisis de regresión considerando 3500 operaciones realizadas con anterioridad.

Así, se obtiene la probabilidad de muerte estimada, considerando que,

$$\text{logit(p)} = \log[p/(1-p)].$$

La inclusión del riesgo se logra considerando las razones de momios:

$$OR_0 = \frac{p_{t0}/(1-p_{t0})}{p_t/(1-p_t)}, \qquad OR_A = \frac{p_{tA}/(1-p_{tA})}{p_t/(1-p_t)}.$$

Mientras que los momios bajo $H_0$ son $p_{t0}/(1-p_{t0})$, los momios bajo $H_A$ son $p_{tA}/(1-p_{tA})$. Consideramos ahora, que el procedimiento CUSUM prueba repetidamente las hipótesis,

$$H_0 : OR = OR_0, \text{ contra } H_A : OR = OR_A.$$

Se obtiene que la función de pesos para el paciente $t$ es:

$$W_t = \begin{cases} \log\left(\frac{(1-p_t+OR_0 p_t)OR_A}{(1-p_t+OR_A p_t)OR_0}\right) & \text{si } y_t = 1 \\ \log\left(\frac{(1-p_t+OR_0 p_t)}{(1-p_t+OR_A p_t)}\right) & \text{si } y_t = 0. \end{cases}$$

la correspondiente carta CUSUM se identifica ahora como CUSUM-RA dado que esta considera el riesgo.

La carta CUSUM diseñada para detectar una duplicación de los momios de muerte, considera $OR_0 = 1$ y $OR_A = 2$ de manera que monitoreamos los incrementos en las probabilidades de fallas En este caso, la función de pesos para el paciente $t$, es

$$W_t = \begin{cases} \log\left(\frac{2}{(1+p_t)}\right) & \text{si } y_t = 1 \\ \log\left(\frac{1}{(1+p_t)}\right) & \text{si } y_t = 0. \end{cases}$$

La carta CUSUM para detectar una reducción a la mitad de los momios de muerte, o mejor dicho un decremento en las probabilidades de fallas de la cirugía, considera $OR_0 = 1$ y $OR_A = .5$.

## 2.3. Ejemplo del proceso de cirugías

La Figura 2 contiene las cartas CUSUM y CUSUM-RA, para el ejemplo del proceso de cirugías, bajo un escenario simulado, donde la probabilidad de fallas de las cirugías es .10 en los primeras 100 observaciones y posteriormente se incrementa a .40 en las siguientes 100 observaciones. se considera además una mortalidad adicional (correspondiente al riesgo pre-operatorio) con un patrón constante dado por una distribución Uniforme(.05, .55) para la contribución aditiva a la mortalidad post-operatoria. La carta se diseña para detectar un aumento en la razón de momios de 1.00 a 2.25, de fallecimiento después de la cirugía. En las gráficas se observa que las sumas acumuladas sin considerar el riesgo pre-operatorio son mayores, lo cual se debe a que la variabilidad de los pacientes se atribuye a la variabilidad del proceso, produciendo así un incremento en las falsas alarmas identificadas. La gráfica CUSUM con ajuste de riesgo muestra una señal de cambio muy cerca de la observación 100, como se espera que ocurra y en cambio, la gráfica sin ajuste de riesgo, muestra la señal de cambio muy temprano, alrededor de la observación número 20.

Con el fin de apreciar el impacto del riesgo pre-operatorio en las cartas CUSUM, simulamos otro proceso de cirugías con los mismos parámetros que el anterior, excepto para el riesgo pre-operatorio, que ahora se simula considerando un patrón dado por una distribución Uniforme(.05, .25), esto es, ahora estamos considerando contribuciones menores de riesgo pre-operatorio que en el primer caso. El efecto de esta diferencia en el riesgo pre-operatorio lo apreciamos al comparar las cartas CUSUM sin ajuste del riesgo pre-operatorio. En la figura 3, la gráfica sin riesgo pre-operatorio tarda más en mostrar la señal de cambio que en la carta de la Figura 2, ahora la señal de cambio se presenta después de la observación número 40.

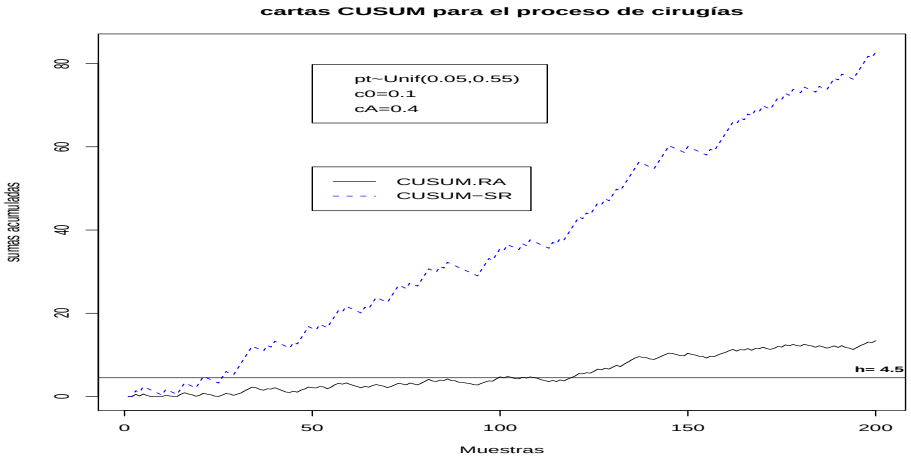



**Figura 2:** Cartas de control CUSUM y CUSUM-RA, para el Ejemplo de cirugías, bajo un escenario simulado, con una contribución promedio de riesgo pre-operatorio de .30.

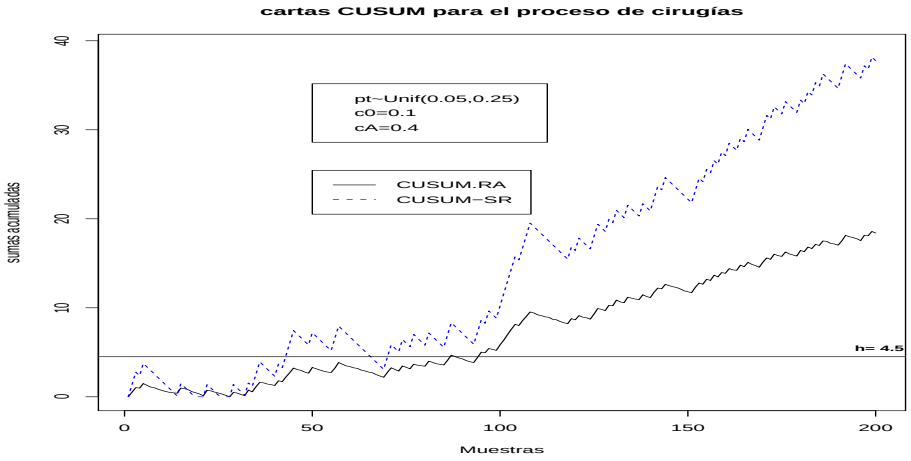



**Figura 3:** Cartas CUSUM y CUSUM-RA, bajo un escenario simulado, con una contribución promedio de riesgo pre-operatorio de .15.

# 3. Conclusiones

Se ha discutido una carta CUSUM que permite monitorear el desempeño de cirugías, considerando el riesgo pre-operatorio de cada paciente, de manera que las falsas alarmas no se presenten con mucha frecuencia, y con la capacidad de detectar cambios sustanciales en la tasa de fallas a partir de un momento determinado. Más especificamente la carta CUSUM sin considerar el riesgo de los pacientes, da señales incorrectas de un crecimiento en la tasa de mortalidad (falsas alarmas) identificando posiblemente que el proceso de cirugías está fuera de control cuando en realidad el crecimiento de la tasa de mortalidad se debe a probabilidades de riesgo pre-operatorio grandes.

Como trabajo futuro se hace necesario cuantificar la capacidad de la propuesta para detectar cambios especificos, por medio de la comparación de las longitudes promedio de corrida (ARL), así mismo hace falta considerar otros factores o covariables de riesgo que influyen en el resultado de las cirugías, los cuales dependeran del tipo de cirugías bajo estudio.

# Bibliografía

Moustakides, G. (1986), "Optimal stopping times for detenting chances in distributions", *Annals of Statistics* **14**, 1379–1387.

Page, E. (1954), "Continuous inspection schemes", *Biometrika* **41**, 100–114.

Parsonnet, V., Dean, D. y Bernstein, A. (1989), "A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease", *Circulation* **779(suppl 1)**, 1–12.

Shewhart, W. (1926), "Quality control charts", *ATT Tech Journal* **5**, 593–606.

Steiner, S., Cook, R., Farewell, V. y Treasure, T. (2000), "Monitoring surgical performance using risk adjusted cumulative sum chart", **1**, 441–452.

Winkel, P. y Zang, N. (2007), *Statistical Development of Quality in Medicine*, J. Wiley.

# Sección II

# Estadística Bayesiana

Memoria FONE XXV (2010), Sección II, 31–36

# Un modelo bayesiano para datos longitudinales circulares

Gabriel Núñez Antonio[a]
*Departamento de Estadística*
*Universidad Carlos III de Madrid, España*

Eduardo Gutiérrez Peña[b]
*Departamento de Probabilidad y Estadística*
*IIMAS,UNAM, México*

## 1. Introducción

La metodología para analizar datos longitudinales aún se encuentra en continuo desarrollo, particularmente cuando se trabaja con modelos que permiten distribuciones de probabilidad más generales que la normal, para las medidas repetidas. En la literatura se han propuesto varios enfoques para el análisis de datos longitudinales en el caso de una respuesta escalar. En contraste, las propuestas para modelar relaciones de dependencia cuando las medidas repetidas son datos circulares son limitadas. La metodología propuesta en este trabajo de investigación para el análisis longitudinal de datos circulares está basada en un modelo Normal bivariado bajo proyecciones. En esta propuesta cada componente de la distribución Normal bajo proyección se especifica a través de un modelo lineal de efectos mixtos.

## 2. El modelo LCP

Para facilitar la exposición, suponemos que $n_i = n \ \forall \ i = 1, \ldots, N$. Así, se tienen $N$ sujetos o individuos en estudio con $n$ observaciones angulares, $\theta_{ij}$, $j = 1, \ldots, n$, para cada sujeto o individuo $i$. Si se tuvieran datos completos, de la forma $R_{ij}(\cos\Theta_{ij}, \text{sen}\Theta_{ij})$, es decir, si se pudieran observar realizaciones de las variables $R_{ij}$ se estaría en condiciones de hacer

[a]gab.nunezantonio@gmail.com
[b]eduardo@sigma.iimas.unam.mx

inferencias para los parámetros involucrados en el modelo. El problema es que sólo las direcciones $\{\theta_{ij}\}$ son observables. Este problema se puede atacar introduciendo adecuadamente variables latentes $R_{ij}$ a través de la transformación definida por $\boldsymbol{Y}_{ij} = R_{ij}(\cos\Theta_{ij}, \text{sen}\Theta_{ij})'$, de la siguiente manera.

$$\begin{pmatrix} \text{individuo} & & \text{observaciones} & \\ 1 & \theta_{11} & \dots & \theta_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ i & \theta_{i1} & \dots & \theta_{in} \\ \vdots & \vdots & \vdots & \vdots \\ N & \theta_{N1} & \dots & \theta_{Nn} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \boldsymbol{Y}_{11} & \dots & \boldsymbol{Y}_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ i & \boldsymbol{Y}_{i1} & \dots & \boldsymbol{Y}_{in} \\ \vdots & \vdots & \vdots & \vdots \\ N & \boldsymbol{Y}_{N1} & \dots & \boldsymbol{Y}_{Nn} \end{pmatrix}$$

con

$$\boldsymbol{Y}_{ij} = \begin{pmatrix} Y_{ij}^{I} \\ Y_{ij}^{II} \end{pmatrix} = R_{ij} \times \begin{pmatrix} \cos\theta_{ij} \\ \text{sen}\theta_{ij} \end{pmatrix} \sim N_2(\boldsymbol{\mu}_{ij}\ , \boldsymbol{I}), \qquad \begin{matrix} i = 1, \cdots, N, \\ j = 1, \cdots, n, \end{matrix}$$

donde $R_{ij} = ||\boldsymbol{Y}_{ij}||$ y $\boldsymbol{\mu}_{ij} = \begin{pmatrix} \mu_{ij}^{I} \\ \mu_{ij}^{II} \end{pmatrix}$, con $\mu_{ij}^{k} = (\boldsymbol{x}_{ij}^{k})'\boldsymbol{\beta}^{k} + (\boldsymbol{z}_{ij}^{k})'\boldsymbol{b}_{i}^{k}$, $\forall\ \ k \in \{I, II\},\ i = 1, \ldots, N$ y $j = 1, \ldots, n$. Aquí, $\boldsymbol{x}_{ij}^{k}$ y $\boldsymbol{z}_{ij}^{k}$ son los vectores de covariables asociados a los efectos fijos y aleatorios, respectivamente, de la componente $k$-ésima de $\boldsymbol{\mu}$, para el individuo $i$ en la ocasión $j$.

Así, para los datos aumentados y considerando cada componente $k \in \{I, II\}$, se propone el siguiente modelo denominado LCP, denominación debido a sus componentes de tipo *longitudinal*, para datos *circulares*, basado en una distribución obtenida bajo una *proyección* radial.

- Nivel I. Para cada individuo $i$,

$$\boldsymbol{Y}_{i}^{k} |\ \boldsymbol{\beta}^{k}, \{\boldsymbol{b}_{i}\}^{k} \sim N_n(\boldsymbol{X}_{i}^{k}\boldsymbol{\beta}^{k} + \boldsymbol{Z}_{i}^{k}\boldsymbol{b}_{i}^{k}, \boldsymbol{I}), \qquad i = 1, \cdots, N,$$

  es decir, dado $\boldsymbol{\beta}^{k}$ y $\{\boldsymbol{b}_{i}\}^{k}$, $\boldsymbol{Y}_{i}^{k} = \boldsymbol{X}_{i}^{k}\boldsymbol{\beta}^{k} + \boldsymbol{Z}_{i}^{k}\boldsymbol{b}_{i}^{k} + \boldsymbol{\varepsilon}_{i}^{k}$, $\forall\ \ i = 1, \cdots, N$, donde $\boldsymbol{\varepsilon}_{i}^{k} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

- Nivel II. $\boldsymbol{\beta}^{k}$ y $\boldsymbol{b}_{i}^{k}$ se consideran vectores independientes, $\forall\ \ i = 1, \cdots, N$, con $\boldsymbol{b}_{i}^{k} | \boldsymbol{\Omega}^{k} \sim N_q(\boldsymbol{0}, \boldsymbol{\Omega}^{k})\quad \forall\ \ i = 1, \cdots, N$ y $\boldsymbol{\beta}^{k} \sim N_p(\boldsymbol{0}, A^{k})$.

- Nivel III. $\boldsymbol{\Omega}^k \sim Wi(v^k, B^k), \quad v^k \geq q^k$, donde $q^k$ es la dimensión del vector $\boldsymbol{b}_i^k$. En esta parametrización $E(\boldsymbol{\Omega}^k) = v^k (B^k)^{-1}$

Se debe observar que el nivel I del modelo propuesto para cada observación $ij$-ésima se puede ver como $f(\boldsymbol{Y}'_{ij} = r_{ij}(\cos\theta, \operatorname{sen}\theta) | \, \boldsymbol{\beta}^I, \boldsymbol{\beta}^{II}, \{\boldsymbol{b}_i\}^I, \{\boldsymbol{b}_i\}^{II}, \boldsymbol{x}_{ij}^I, \boldsymbol{x}_{ij}^{II}, \boldsymbol{z}_{ij}^I, \boldsymbol{z}_{ij}^{II}) = N_2(\cdot | \boldsymbol{\mu}_{ij}, \boldsymbol{I})$.

# 3. Inferencias vía métodos MCCM

Una vez aumentado los datos, para estar en condiciones de llevar a cabo inferencias para los parámetros del modelo usando técnicas MCCM, como el muestreo de Gibbs, se deben especificar todas las correspondientes densidades condicionales completas. Así, si $\boldsymbol{D_n} = \{(r_{11}, \theta_{11}), \cdots, (r_{Nn}, \theta_{Nn})\}$ es un conjunto de observaciones del modelo $N_2(\boldsymbol{\mu}_{ij}, \boldsymbol{I})$, se pueden obtener de manera relativamente fácil las densidades condicionales completas de todos los parámetros involucrados. Sin embargo, la implementación directa de métodos MCCM usando estas condicionales presenta problemas de convergencia. Por lo anterior, para tener un mejorar desempeño en la implementación de los métodos MCCM en este trabajo se consideró una propuesta basada en la simulación por bloques. Un algoritmo que implementa un esquema de Gibbs-Metropolis con esta metodología por bloqueo se puede especificar de la siguiente manera.

- Para cada componente $k$, $k \in \{I, II\}$:

  Muestrear $\boldsymbol{\beta}^k$ y $\{\boldsymbol{b}_i\}$ de $f(\boldsymbol{\beta}^k, \{\boldsymbol{b}_i\}^k | \boldsymbol{\Omega}^I, \boldsymbol{\Omega}^{II}, \boldsymbol{D_n}) = f(\boldsymbol{\beta}^k, \{\boldsymbol{b}_i\}^k | \boldsymbol{\Omega}^k, \boldsymbol{D_n})$, simulando $\boldsymbol{\beta}^k$ de $f(\boldsymbol{\beta}^k | \boldsymbol{\Omega}^I, \boldsymbol{\Omega}^{II}, \boldsymbol{D_n}) = f(\boldsymbol{\beta}^k | \boldsymbol{\Omega}^k, \boldsymbol{D_n})$, y $\boldsymbol{b}_i^k$ de $f(\boldsymbol{b}_i^k | \boldsymbol{\beta}^I, \boldsymbol{\beta}^{II}, \Omega^I, \Omega^{II}, \boldsymbol{D_n}) = f(\boldsymbol{b}_i^k | \boldsymbol{\beta}^k, \Omega^k, \boldsymbol{D_n})$, para toda $i = 1, \ldots, N$.

  Muestrear $\boldsymbol{\Omega}^k$ de $f(\boldsymbol{\Omega}^k | \boldsymbol{\beta}^I, \boldsymbol{\beta}^{II}, \{\boldsymbol{b}_i\}^I, \{\boldsymbol{b}_i\}^{II}, \boldsymbol{D_n}) = f(\boldsymbol{\Omega}^k | \{\boldsymbol{b}_i\}^k, \boldsymbol{D_n})$. Muestrear $R_{ij}$ de $f(r_{ij} | \boldsymbol{\beta}^I, \boldsymbol{\beta}^{II}, \{\boldsymbol{b}_i\}^I, \{\boldsymbol{b}_i\}^{II}, \ \boldsymbol{\Omega}^I, \ \boldsymbol{\Omega}^{II}, \{\theta_{ij}\})$ $\forall$ $i = 1, \ldots, N$ y $\forall$ $j = 1, \ldots, n$. Este último paso se lleva a cabo a través de un algoritmo de Metropolis.

**Ejemplo**. Se consideró un conjunto de datos tomados de Song (2007) que consisten en las direcciones de 65 *talitrus saltator* (pulgas de mar) después de ser liberados secuencialmente en 5 ocasiones. Algunas covariables registradas en el estudio incluyen la velocidad del viento, la dirección azimuth del sol (Sun) y medidas oculares con las cuales se elabora un índice de simetría ocular (Eye). La dirección del viento fue transformada en cuatro categorías

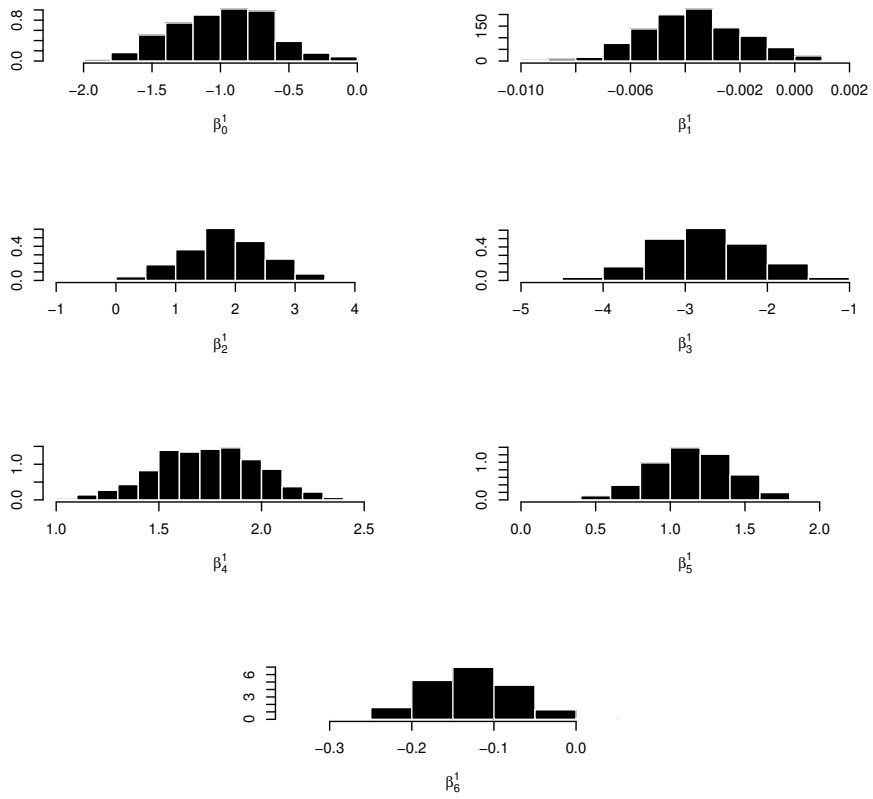



**Figura 1:** *Distribuciones finales de los parámetros en la componente I del modelo LCP para los datos de* talitrus saltators.

dependiendo del lugar donde provenía el viento (OS para offshore, LSE para longshore-east, LSW para longshore-west y onshore). Para analizar estos datos, en este trabajo se considera un modelo LCP con

$$\begin{aligned} \boldsymbol{\mu}_{ij}^{I} &= \beta_0^I + \beta_1^I Sun + \beta_2^I Eye + \beta_3^I OS + \beta_4^I LSW + \beta_5^I LSE + \beta_6^I Tiempo \qquad (1) \\ \boldsymbol{\mu}_{ij}^{II} &= \beta_0^{II} + \beta_1^{II} Sun + \beta_2^{II} Eye + \beta_3^{II} OS + \beta_4^{II} LSW + \beta_5^{II} LSE + \beta_6^{II} Tiempo + b_{0i} \\ & \quad i = 1, \ldots, 65. \end{aligned}$$

Las Figuras 1 y 2 muestran las respectivas distribuciones finales para todos los parámetros de las componentes I y II del modelo LCP definido en (1). En la Tabla 1 se muestran los intervalos de credibilidad al 95 % para los parámetros de las dos componentes del modelo LCP. Los resultados sugieren que el efecto de $Sun$ y los efectos de $Eye$, $OS$ y $LSW$ no son

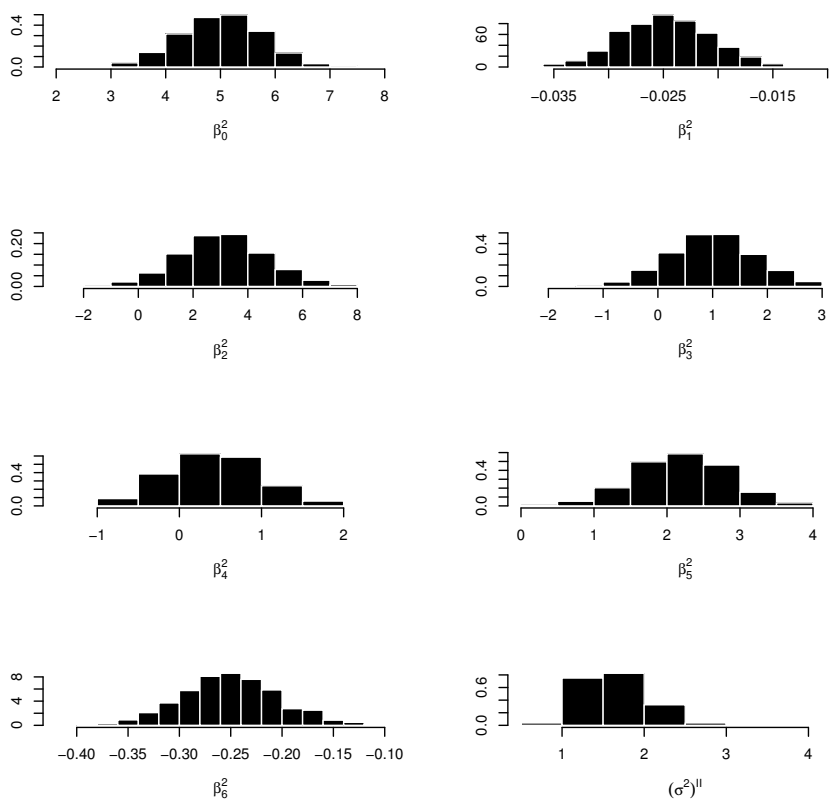



**Figura 2:** *Distribuciones finales de los parámetros en la componente II del modelo LCP para los datos de* talitrus saltators.

relevantes para $\boldsymbol{\mu}^I$ y $\boldsymbol{\mu}^{II}$, respectivamente. Como el parámetro $\sigma^2$ es diferente de cero, la inclusión del efecto aleatorio asociado resulta necesario. Lo anterior indica una presencia de heterogeneidad entre los escapes de los *talitrus saltators*.

# 4. Conclusiones

Los modelos existentes en la literatura para el análisis de datos longitudinales de tipo direccional asumen una distribución de probabilidad von Mises para la variable respuesta y atacan el problema desde un enfoque semiparamétrico como los métodos GEE o de modelos lineales generalizados. Sin embargo, estos enfoques presentan diversos problemas de maximización numérica. Por su parte, el modelo LCP es relativamente fácil de analizar bajo la metodología

| | Componente I | Componente II |
|---|---|---|
| $\beta_0$ | (-1.7041 , -0.2797) | (3.5300 , 6.4445) |
| $\beta_1(Sun)$ | (-0.0069 , 0.0002) | (-0.0326 , -0.0167) |
| $\beta_2(Eye)$ | (0.5108 , 3.1228 ) | (-0.2894 , 6.1439) |
| $\beta_3(OS)$ | (-4.0097 , -1.6492) | (-0.5842 , 2.4310) |
| $\beta_4(LSW)$ | (1.2645 , 2.2534) | (-0.6257 , 1.5565) |
| $\beta_5(LSE)$ | (0.6042 , 1.6781) | (0.9985 , 3.4888) |
| $\beta_6(Tiempo)$ | (-0.2260 , -0.0277) | (-0.3420 , -0.1565) |
| $\sigma^2$ | | ( 0.9825 , 2.4505 ) |

**Cuadro 1:** *Intervalos finales de credibilidad al* 95 % *para cada uno de los componentes del modelo LCP para los datos de* talitrus saltators.

propuesta y no pierde aplicabilidad práctica comparado con los modelos que asumen una distribución von Mises para la variable circular. Adicionalmente, esta metodología ofrece la flexibilidad de realizar inferencias de manera directa para otros parámetros asociadas al modelo, por ejemplo, aquéllos obtenidos bajo alguna reparametrización.

# Bibliografía

Borgioli, C., Martelli, M., Porri, F., D'Elia, A., Marcheti, G. M. y Scapini, F. (1999), "Orientation in *talitrus saltator* (montagu): trends in intrapopulations variability related to enviromental and intrinsic factors", *Journal of Experimental Marine Biology and Ecology* **238**, 29–47.

Song, X.-K. P. (2007), *Correlated Data Analysis: Modeling Analytics, and Applications*, Springer: New York.

Memoria FONE XXV (2010), Sección II, 37–48

# EMV en modelos de mezclas finitas univariadas

Olga Vladimirovna Panteleeva[a], Humberto Vaquera Huerta
*Colegio de Postgraduados campus Montecillo*

Eduardo Gutiérrez González
*Instituto Politécnico Nacional*

## 1. Introducción

Los modelos de mezclas finitas se encuentran en la investigación pesquera, economía, medicina, psicología, antropología, botánica, agricultura, zoología, la vida de pruebas y la confiabilidad. Por citar algunos ejemplos tenemos la aplicación de Chuangmin Liu *et al.*, Liu et al. (2002), en donde se propone un modelo de mezclas finitas de la distribución Weibull para describir las distribuciones del diámetro de especies mixtas en masas forestales; en la parte de confiabilidad tenemos la aplicación realizada por Laura Attardi *et al.*, Attardi et al. (2005), para modelar los tiempos de garantía de las cajas de velocidad de una marca de carros que estaban fallando; otra aplicación a la confiabilidad se muestra en el trabajo presentado por Rufo *et al.*, Rufo et al. (2009), en donde estudian el tiempo de vida de los líquidos aislantes bajo dos niveles de alto voltaje; en la parte de medicina tenemos la aplicación realizada por Farcomeni & Nardi, Alessio y Nardi (2010), en donde se aplica un modelo de mezclas para los tiempos de supervivencia de pacientes después del trasplante de un órgano; en los ejemplos médicos no podia faltar una aplicación al estudio del cancer como lo muestran Lambert *et al.*, Lambert et al. (2010); otra aplicación de los modelos de mezclas finitas la podemos apreciar en los tiempos de vida de componentes electrónicos estudiados por Razali & Salih, Razali y Salih (2009); etcétera.

Uno de los principales problemas en las aplicaciones de mezclas finitas se refiere a la estimación de sus parámetros, puesto que éstos suelen ser una cantidad considerable y por

[a]panteleevaolga@yahoo.com.mx

la estructura de las mezclas surge el problema de identificabilidad, como lo muestra Kadane, Kadane (1974), la complejidad de este problema también fue revisada por Crawford, Crawford (1994), por su parte Titterington *et al.*, Titterington et al. (1985), también escribieron obre el problema de la identificabilidad en los modelos de mezclas. El problema de la identificabilidad para las mezclas ha tenido resultados positivos como los obtenidos por Teicher, Teicher (1960), quien muestra que el modelo de mezclas finitas con distribución Poisson es identificable, pero en el caso de la distribución binomial no es identificable si la cantidad de ensayos es menor al doble menos uno de la cantidad de componentes en la combinación convexa de la mezcla. Similarmente Yakowitz & Spragins, Yakowitz y Spragins (1968), mostraron que los modelos de mezclas finitas con distribución binomial negativa son identificables.

## 1.1. Planteamiento del problema

Sean $X_1, \ldots, X_m$ $m$ variables aleatorias con el mismo soporte y vector de parámetros $\boldsymbol{\theta}_i$, para $i = 1, \ldots, m$, respectivamente. Se supone que los parámetros tienen el mismo espacio paramétrico, aunque en general las distribuciones de las variables pueden ser de diferentes familias (conservando el soporte), se suele mezclarlas cuando pertenecen a la misma familia, pero proponiendo diferentes parámetros.

Denotemos por $f_{X_i}(\cdot)$ a la función de densidad de la variable aleatoria $X_i$ para $i = 1, \ldots, m$, entonces la función de densidad del modelo de mezclas finitas se muestra en (1)

$$f(x; \boldsymbol{\theta}) = \sum_{i=1}^{m} \pi_i f_{X_i}(x; \boldsymbol{\theta}_i) \tag{1}$$

en donde $\pi_i$ son valores no negativos, llamados pesos de las variables aleatorias, tales que $\sum_{i=1}^{m} \pi_i = 1$ y $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ con $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$.

Uno de los principales problemas en los modelos de mezclas finitas se refiere a la estimación de los parámetros, este problema en general se complica con el aumento de la cantidad de parámetros, puesto que a diferencia de las funciones de verosimilitud de los modelos habituales las expresiones no se simplifican.

En el trabajo se presenta un bosquejo sobre la evolución de los métodos que se utilizan con mayor frecuencia en el problema de la estimación de parámetros para el problema de mezclas finitas. Además, aprovechando que los valores de los parámetros cuando se trata de funciones de densidad tienen sus mayores variaciones dentro de intervalos que pueden ser

acotados con respecto a los valores de su distribución de frecuencias, se propone un algoritmo de búsqueda de los estimadores de máxima verosimilitud.

## 2. Marco teórico

El problema de estimar los parámetros de una mezcla de dos densidades normales tiene una larga historia y empieza con el estudio de Pearson, Pearson (1894), con el que por primera vez los modelos de mezclas fueron introducidas. Pearson utilizó el método de momentos para estimar los cinco parámetros de dos componentes de la mezcla normal univariada, similarmente al inicio del siglo XX Charlier, Charlier (1906), también utilizó el método de momentos. Así en las primeras décadas del siglo 20 este método era el más utilizado, aunque no se probó su eficiencia. Charlier & Wicksell, Charlier y Wicksell (1924), extendieron los trabajos realizados en esta época al caso de componentes normales bivariadas, mientras que Doetsch Doetsch (1928) lo extendió al caso de la mezcla normal univariada con más de dos componentes. Por otro lado, Strömgren, Strömgren (1934), consideró el uso de cumulantes, mientras que Rao C.R., Rao (1948), también utilizo el método de momentos considerando el uso de $k$ estadísticos. En trabajos más recientes, Blischke, Blischke (1978) llevo a cabo la estimación de los parámetros del modelo de mezclas con dos distribuciones binomiales mientras que Cohen, Cohen (1967) mostró cómo podría resolverse la ecuación de Pearson, Pearson (1894), por medio de un proceso iterativo que implica solución de ecuaciones cúbicas con la raíz única negativa. Por otro lado, Tan & Chang, Tan y Chan (1972) mostraron que el método de momentos se puede obtener en forma cerrada para modelos de mezclas de dos normales con la misma varianza, superando en la estimación al método de máxima verosimilitud para este problema. Así, el interés en este método de estimación para mezclas de normales ha sido renovado con los trabajos de Lindsay & Basak, Lindsay y Basak (1993), entre otros. Algunos estadísticos de renombre también utilizaron los estimadores de momentos para estudiar los modelos de mezclas, entre otros se tiene a Rao Rao (1952), Hasselblad Hasselblad (1966), Cox Cox (1966), Day Day (1969) y Behboodian Behboodian (1970).

Por su parte Preston, Preston (1953), mostró un método gráfico para el caso simple de dos componentes de un modelo de mezclas de dos normales con varianzas iguales, éste se basa en el método de momentos y el uso de un diagrama de la curtosis-sesgo. Otro método gráfico se tiene con la llamada curva $\lambda$ propuesta por primera vez por el matemático Doetsch,

Doetsch (1928). Medgyessy Medgyessy (1961) propuso la metodología $\lambda$ a la comunidad de la estadística aplicada. Trabajos más recientes de Tarter & Lock Tarter y Lock (1993) describieron este método de estimación de la curva ($\lambda$), la cual principalmente resulta ser un método gráfico de descomposición de mezclas. Otro método gráfico para estimar los parámetros de un modelo de mezclas fue presentado por Tarter & Silvers, Tarter y Silvers (1975), quienes realizaron un procedimiento gráfico basado en las propiedades de la función de densidad de Gauss bivariada, mientras que Ghikara & Register desarrollaron una técnica de clasificación numérica basada en métodos de asistencia computacional para el despliegue de los datos.

En el caso de los estimadores de máxima verosimilitud de un modelo de mezclas como el mostrado en (1), bajo una realización $x_1, x_2, \ldots, x_n$ de tamaño $n$ el problema es considerablemente complejo porque la función de verosimilitud o log-verosimilitud están dadas

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{j=1}^{n} \sum_{i=1}^{m} \pi_i f_{X_i}(x_j; \boldsymbol{\theta}_i) \quad \text{o} \quad \ell(\boldsymbol{\theta}|\mathbf{x}) = \sum_{j=1}^{n} \log \left( \sum_{i=1}^{m} \pi_i f_{X_i}(x_j; \boldsymbol{\theta}_i) \right) \tag{2}$$

El problema de encontrar los valores de $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ con $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ y $\sum_{i=1}^{m} \pi_i = 1$ para los que (2) se hacen máximos es bastante complejo por la gran cantidad de parámetros que se tienen que estimar. Los métodos de momentos y MV fueron investigados por Rao Rao (1952), Hasselblad Hasselblad (1966), Behboodian Behboodian (1970), Cohen Cohen (1967), Cox Cox (1966). Por ejemplo, Cohen Cohen (1967), demostró que en la mayoría de los casos las estimaciones de MV son más precisas que las de momentos.

En la práctica uno de los problemas que tienen los EMV reside en la gran cantidad de cálculos computacionales que son requeridos, de tal forma que algunos autores como Hasselblad Hasselblad (1966), Day Day (1969), Wolfe Wolfe (1970), Hosmer Hosmer (1973) propusieron métodos iterativos para calcular los EMV. Además en sus obras estos autores demostraron que se necesita un tamaño de muestra grande o que los componentes estén bien separados a fin de obtener estimaciones confiables de máxima verosimilitud. Recientemente Razali & Salih Razali y Salih (2009) presentaron un trabajo en donde mezclan dos distribuciones Weibull una con 3 parámetros y la otra distribución con dos parámetros, para la estimación utilizaron MV.

Dempster *et al.* Dempster et al. (1977) desarrollaron el algoritmo EM para el cálculo de los EMV de las medias de una mezcla en proporciones conocidas de un número finito de

poblaciones normales univariadas con varianzas conocidas. Para el uso del algoritmo EM se pueden consultar por ejemplo Ganesalengam & McLachlan, O´Neill y Aitkin, entre muchos otros. Recientemente Sultan *et al.* Sultan et al. (2007) estimaron los parámetros en una mezcla de dos distribuciones Weibull inversas via el algoritmo EM, realizando demostraciones numéricas y cálculos a través de simulaciones Monte Carlo. El algoritmo EM es un proceso iterativo que se mejora con la introducción del algoritmo EM generalizado, GEM.

Los algoritmos vistos para calcular o aproximar los EMV tienen ciertos problemas teóricos en su aplicación algunos de ellos están descritos en el trabajo de Wu Wu (1983). Por ejemplo para la convergencia de las sucesiones de los parámetros que en ellos aparecen no garantizan la convergencia a un máximo, ni siquiera a un máximo local, ya que estos resultados nos conducen a los puntos estacionarios, que pueden ser máximos, mínimos o puntos silla, puede verse el trabajo de Murray en donde la aplicación del algoritmo EM a una distribución normal bivariada conduce a un punto silla. Además, tampoco aseguran la convergencia de la sucesión $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=0}^{\infty}$. Un estudio detallado sobre la convergencia de los algoritmos GEM puede ser revisado en Wu Wu (1983), quien propone varios resultados, que llevan a las condiciones de convergencia de los algoritmos GEM.

# 3. Método propuesto de recorridos para calcular EMV

En la mayoría de los casos la influencia en la variación de la función de densidad por los parámetros de escala y forma no varía considerablemente a partir de ciertos valores de los parámetros. En el caso del parámetro de localidad éste se puede acotar por medio del histograma de los datos y las condiciones que deba cumplir en la función objetivo. Por lo tanto, teniendo las variables acotadas, se puede proceder a realizar búsquedas exhaustivas del valor óptimo de la función objetivo en dichos intervalos, misma que se puede hacer por recorridos programados en cada variable pero llevaría mucho tiempo de cómputo, por ejemplo si se tienen 5 variables y en cada una se realiza un recorrido de 100 subdivisiones se requiere evaluar la función objetivo $100^5 = 10^{10}$ veces. Con esto se reduce la región de búsqueda y se realiza lo mismo en esta región. Para reducir el tiempo de cómputo se realiza la búsqueda como en simulación Monte Carlo, en donde se generan valores aleatorios uniformes para cada variable dentro de su acotación y posteriormente se evalúa la función objetivo en cada uno de ellos, para que finalmente se elija el valor que mejor aproxime al óptimo de la función

objetivo. Por ejemplo con $10^5$ evaluaciones se tendría una muy buena aproximación y como únicamente se está evaluando la función el tiempo de cálculo es muy pequeño. Las ventajas de este método es que únicamente se pide la continuidad de la función en la región de búsqueda condición que con los otros métodos no es suficiente. Véase el siguiente algoritmo.

1. El espacio paramétrico $\boldsymbol{\Theta}$ se acota en forma cerrada en cada parámetro, formando un conjunto compacto.

2. Se programa en el Proyecto R una búsqueda aleatoria de $n$ valores uniformes para cada parámetro en el conjunto compacto.

3. Se evalúa la log-verosimilitud en cada generación aleatoria y se ordenan sus valores en forma no decreciente, registrando los valores de los estimadores correspondientes.

4. Con los valores de los estimadores obtenidos, repetir los pasos anteriores, disminuyendo el segmento de búsqueda para cada parámetro. Se repite el proceso hasta que se obtenga una diferencia pequeña en los valores de la log-verosimilitud.

# 4. Aplicaciones

En esta sección se revisan dos aplicaciones del método de recorridos para aproximar los EMV en modelos de mezclas finitas, mostrando que con este método se pueden mejoran los valores de los EMV obtenidos con los métodos iterativos EM o GEM.

## 4.1. Aplicación 1. Weibull

En esta parte se revisa un ejemplo del modelo de mezclas Weibull con tres parámetros y función de densidad dada en (3)

$$f(t;\alpha,\beta,\gamma) = \frac{\alpha}{\beta}\Big(\frac{t-\gamma}{\beta}\Big)^{\alpha-1}\exp\left(-\Big(\frac{t-\gamma}{\beta}\Big)^{\alpha}\right) \text{ para } 0<\gamma\leq t<\infty. \tag{3}$$

en donde $\alpha,\beta,\gamma$ son parámetros de forma, escala y localidad, respectivamente. Por otro lado, la función log-verosimilitud del modelo de mezclas para una realización $t_1,\ldots,t_n$ se muestra en (4)

$$\ell(\boldsymbol{\theta}|\mathbf{t}) = \sum_{j=1}^{n}\log\big(pf_1(t_j;\boldsymbol{\theta}_1)+(1-p)f_2(t_j;\boldsymbol{\theta}_2)\big). \tag{4}$$

en donde $\boldsymbol{\theta} = (p, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ con $\boldsymbol{\theta}_i = (\alpha_i, \beta_i, \gamma_i)$ $i = 1, 2$.

Revisaremos el artículo de Murthy *et al.* (2004) "Combining Two Weibull Distributions Using a Mixing Parameter", sobre datos de tiempos de vida de 20 componentes electrónicos mostrados en la tabla 1.

| $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.03 | 2 | 0.12 | 3 | 0.22 | 4 | 0.35 |
| 5 | 0.73 | 6 | 1.79 | 7 | 1.25 | 8 | 1.41 |
| 9 | 1.52 | 10 | 1.79 | 11 | 1.8 | 12 | 1.94 |
| 13 | 2.38 | 14 | 2.4 | 15 | 2.87 | 16 | 2.99 |
| 17 | 3.14 | 18 | 3.17 | 19 | 4.72 | 20 | 5.09 |

**Cuadro 1:** Tiempos de vida de los 20 componentes.

En este artículo se fijaron valores del parámetro $p$ y se obtuvieron los correspondientes valores por el algoritmo EM para los restantes parámetros, obteniendo la mayor log-verosimilitud $-30.12164$ con $p = 0.15$ y $\gamma = 0$ fijos y $\alpha_1 = 38.82419046$, $\beta_1 = 3.11847693$, $\alpha_2 = 1.78908595$ y $\beta_2 = 1.08591208$.

Para el método de recorridos se utiliza la función log-verosimilitud de la mezcla dada en (4), primeramente deben acotarse los valores de los parámetros. En el caso del parámetro de localidad, vemos que éste no puede tomar valores mayores a los de la realización y no hay desplazamientos negativos, en este caso los valores que puede tomar el estimador de $\gamma$ están entre 0 y mín$\{t_i\}$. Por otro lado, los parámetros de escala y forma no son muy grandes, entonces se consideran entre 0.1 y 10. Con estas acotaciones se obtiene un valor de la función log-verosimilitud $-29.42596936$ mayor al obtenido con EM. Realizando 4 recorridos de tal forma que se reduzca el conjunto compacto de búsqueda se obtienen los resultados de los 4 recorridos en la tabla 2.

## 4.2. Aplicación 2. Exponencial

Para el caso de la distribución exponencial, cuya función de densidad está dada en (5)

$$f(t; \beta) = \frac{1}{\beta} \exp\Big( - \frac{t}{\beta} \Big) \text{ para } 0 \leq t < \infty. \tag{5}$$

| Recorrido | $\alpha_1$ | $\beta_1$ | $\gamma_1$ | $\alpha_2$ | $\beta_2$ | $\gamma_2$ | $p$ | log-veros. |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.3685018 | 0.6880545 | 0.0299527 | 1.8649952 | 2.6501640 | 0.0120913 | 0.4768 | −29.4260 |
| 2 | 0.3368656 | 0.2682135 | 0.0299976 | 2.0961695 | 2.4795924 | 0.0123601 | 0.2540 | −26.0738 |
| 3 | 0.5171664 | 0.4872062 | 0.0300000 | 2.2605242 | 2.8294988 | 0.0159228 | 0.2585 | −24.7369 |
| 4 | 0.3956516 | 0.4185487 | 0.0299998 | 2.1819940 | 2.7690040 | 0.0002076 | 0.2906 | −24.5692 |

**Cuadro 2:** Resultados de los estimadores durante los 4 recorridos.

donde $\beta > 0$ es el parámetro de escala. Luego, la función log-verosimilitud (6) para esta mezcla con $\boldsymbol{\theta} = (p, \beta_1, \beta_2)$ está dada para una realización $t_1, \ldots, t_n$

$$\ell(\boldsymbol{\theta}|\mathbf{t}) = \sum_{j=1}^{n} \log\left(pf_1(t_j;\beta_1) + (1-p)f_2(t_j;\beta_2)\right). \tag{6}$$

Para la aplicación de la mezcla de exponenciales se tomaron 34 datos que se presentaron en un estudio realizado por Sultan & Ebrahimi, Sultan et al. (2007), sobre los tiempos (en minutos) to break down of an insulating fluid under two level of voltage stress. Los datos se muestran en la tabla 3.

| $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ | $i$ | $t_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.19 | 6 | 0.38 | 11 | 0.85 | 16 | 0.77 | 21 | 0.24 | 26 | 0.10 | 31 | 0.10 |
| 2 | 0.59 | 7 | 0.99 | 12 | 0.66 | 17 | 1.39 | 22 | 0.37 | 27 | 0.52 | 32 | 1.58 |
| 3 | 0.18 | 8 | 0.52 | 13 | 0.26 | 18 | 2.80 | 23 | 0.03 | 28 | 0.12 | 33 | 8.42 |
| 4 | 0.35 | 9 | 0.18 | 14 | 4.33 | 19 | 36.18 | 24 | 0.70 | 29 | 0.19 | 34 | 11.73 |
| 5 | 1.47 | 10 | 1.65 | 15 | 19.15 | 20 | 0.35 | 25 | 0.28 | 30 | 0.77 | 35 | |

**Cuadro 3:** Times (minutes) to break down of an insulating fluid under two level of voltage stress.

Utilizando el algoritmo EM en el artículo obtuvieron que la mayor log-verosimilitud $-52.719$ se obtiene con $p = 0.475$, $\beta_1 = 0.47186$ y $\beta_2 = 7.42850$. Para el método de recorridos se utiliza la función log-verosimilitud de la mezcla dada en (6), acotando los valores de los parámetros. Ambos parámetros se acotan entre 0.01 a 20. Con estas acotaciones se obtiene un valor de la función log-verosimilitud $-47.7906$ mayor al obtenido con el algoritmo EM, mientras que los valores de los EMV aproximados son $\hat{\beta}_1 = 0.6169861$, $\hat{\beta}_2 = 13.6321185$ y $\hat{p} = 0.8183199$. Posteriormente se vuelven acotar los intervalos para los estimadores de los

parámetros de 0.01 a 1 para $\beta_1$ y de 10 a 15 para $\beta_2$, realizando el segundo recorrido se obtuvo un valor de la función log-verosimilitud de $-47.7820930$, mientras que los valores de los EMV aproximados son $\hat{\beta}_1 = 0.6050877$, $\hat{\beta}_2 = 12.8453908$ y $\hat{p} = 0.8149910$. No se realizan más recorridos porque no resultan aumentos sustanciales en la función log-verosimilitud.

# 5. Conclusiones

Por el método propuesto de recorridos se obtuvieron mejoras con respecto al algoritmos EM. En el caso de la mezcla de Weibul´s se obtuvo un valor de log-verosimilitud de $-24.5692$ que representa una mejora del 22.4 % con respecto del valor de la log-verosimilitud $-30.12164$ obtenida con el algoritmo EM. En el caso de la mezcla de exponenciales se obtuvo un valor de verosimilitud $-47.78209$, que representa una mejora del 9.4 % con respecto del valor de la log-verosimilitud $-52.719$ obtenida con el algoritmo EM. En ambos resultados se probó que el método propuesto proporciona mejores aproximaciones que el algoritmo EM.

# Bibliografía

AL-Hussaini, E. K. y Sultan, K. S. (2001), "Reliability and hazard based on finite mixture models", *Elsevier, Amsterdam* **20**.

Alessio, F. y Nardi, A. (2010), "A two-component weibull mixture to model early and late mortality in a bayesian framework".

Attardi, L., Guida, M. y Pulcini, G. (2005), "A mixed-weibull regression model for the analysis of automotive warranty data", pp. 265–273.

Behboodian, J. (1970), "On a mixture of normal distributions", pp. 215–7.

Blischke, W. R. (1978), *Mixtures of distributions*, Vol. 1, In International Encyclopedia of Statistics,New York:The Free Press, Kruskal , W.H.and Tanur, J.M. (Ed).

Charlier, C. V. L. (1906), "Researches into the theory of probability", *Bd. 1 Sec. 2* .

Charlier, C. V. L. y Wicksell, S. D. (1924), "On the dissection of frequency functions", *Bd. 18, No. 6* .

Cohen, A. C. (1967), "Estimation in mixtures of two normal distributions", pp. 15–28.

Cox, D. (1966), "Note on the analysis of mixed frequency distributions", pp. 39–47.

Crawford, S. (1994), "An aplication of the laplace method to finite mixture distributions", *Journal of the American Statistical Assosiation* **89**, 259–267.

Day, N. (1969), "Estimating the components of a mixture of normal distributions", *December* pp. 463–474.

Dempster, A. P., Laird, N. M. y Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the em algorithm", pp. 1–38.

Doetsch, G. D. (1928), "elimination des dopplerefferkts aus spektroskopischefeinstrukturen und exakte bestimmung der komponenten", *Zeitschriftf für Physik* **49**, 705–730.

Hasselblad, V. (1966), "Estimation of parameters for a mixture of normal distributions", pp. 431–444.

Hosmer, D. W. (1973), "A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample", pp. 761–770.

Kadane, J. (1974), *The role of identification in Bayesian theory*, In Studies in Bayesian Econometrics and Statistics, Editor S. Fienberg & A. Zellner, New York: American Elsevier.

Lambert, P. C., Weston, C. L. y Thompson., J. R. (2010), "Estimating the cure fraction in population-based cancer studies by using finite mixture models", pp. 35–55.

Lindsay, B. y Basak, P. (1993), "Multivariate normal mixtures:a fast, consistent method of moments", *Journal of the American Statistical Assosiation* **88**, 468–476.

Liu, C., Zhang, L., Craig, J., Dale, D., Solomon y Gove, J. H. (2002), "by the society of american foresters", *Forest Science* **48(4)**, 653–661.

Medgyessy, P. (1961), "Decomposition of superpositions of functions", *Budapest: Publishing House of the Royal Statistical Society B.* **51**, 127–138.

Pearson, K. (1894), "Contributions to the mathematical theory of evolution", pp. 71–110.

Preston, E. J. (1953), "A graphical method for the analysis of statistical distributions into two normal components", pp. 460–464.

Rao, C. R. (1948), "The utilization of multiple measurements in problems of biological classification", pp. 159–203.

Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*.

Razali, A. M. y Salih, A. A. (2009), "Combining two weibull distributions using a mixing parameter", pp. 296–305.

Rufo, M. J., Martín, J. y Pérez, C. J. (2009), "Local parametric sensitivity for mixture models of lifetime distributions", pp. 1238–1244.

Strömgren, B. (1934), "Tables and diagrams for dissecting a frequency curve into components by the half-invariant method", *Skandinavian Aktuarietidskr* **17**, 7–54.

Sultan, K. S., Ismail, M. A. y Al-Moisheer, A. S. (2007), "Mixture of two inverse weibull distributions: Properties and estimation", pp. 5377 – 5387.

Tan, W. Y. y Chan, W. C. (1972), "Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters a mixture of two normal densities", pp. 702–708.

Tarter, M. E. y Lock, M. D. (1993), *Model-Free Curve Estimation*, London: Chapman & Hall.

Tarter, M. E. y Silvers, A. (1975), "Implementation and application of bivariate gaussian mixture decomposition", *Journal of the American statistical Association* **70**, 47–55.

Teicher, H. (1960), "On the mixture of distributions", *Annals of Mathematical Statistics* **34**, 1265–1269.

Titterington, D. M., Smith, A. F. M. y Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New-York: Willey.

Wolfe, J. H. (1970), "Pattern clustering by multivariate mixture analysis", pp. 329–50.

Wu, C. F. J. (1983), "On the convergence properties of the em algorithm", pp. 95–103.

Yakowitz, S. J. y Spragins, J. D. (1968), "On the identifiability of finite mixtures", *Annals of Mathematical Statistics* **39**, 209–214.

# Sección III

# Inferencia Estadística

Memoria FONE XXV (2010), Sección III, 51–58

# Una prueba de no inferioridad basada en estimadores de proporción contraídos*

Félix Almendra Arao[a],

## 1. Introducción

La comparación de grupos aparece en forma natural en diversos contextos, una forma de comparación de grupos que ha tenido auge a últimas fechas, se basa en el uso de pruebas de no inferioridad las cuales, también son llamadas pruebas de equivalencia unilateral; éstas son procedimientos estadísticos que permiten verificar si existe evidencia muestral de que un tratamiento nuevo no es sustancialmente inferior, en términos de eficacia, que un tratamiento cuya eficacia es bien conocida y el cual es considerado como un tratamiento estándar. Cuando la medida de discrepancia es la diferencia de proporciones se conocen varias pruebas. Almendra-Arao y Sotres-Ramos (2009) realizaron una comparación de siete pruebas asintóticas de no inferioridad con diversas correcciones por continuidad. Para las configuraciones de tamaños de muestra, niveles de significancia nominal, márgenes de no inferioridad y correcciones por continuidad que se analizaron en dicho trabajo, la mejor prueba, con base en el comportamiento de los tamaños de prueba y potencias resultó ser la prueba definida por Mietinen y Nurminen (1985) y Farrington y Manning (1990), con la corrección por continuidad de Hauck-Anderson. Una dificultad que presenta el uso de esta prueba se debe a que el cálculo del correspondiente estadístico de prueba es complicado en la práctica. El presente trabajo introduce una nueva prueba de no inferioridad cuyo estadístico de prueba es más fácil de calcular que este último; esta nueva prueba usa estimadores de proporción contraídos, ver por ejemplo Böhning y Viwatwongkasen (2005), en la estimación de la desviación estándar de la diferencia de los estimadores de proporción . La nueva prueba presentada en

*Este trabajo fue realizado con el apoyo parcial de SNI-CONACYT , COFAA-IPN, EDI-IPN y por el proyecto SIP-IPN 20100130.

[a]falmendra@ipn.mx

este trabajo se comparó con aquélla dada en Mietinen y Nurminen (1985) y Farrington y Manning (1990), misma que de aquí en adelante se denotará mediante $T_1$, la comparación fue realizada con base en los tamaños de prueba y en las potencias. Para realizar la comparación se consideraron los tamaños de muestra $5 \leq n_1 = n_2 = n \leq 100$ y los niveles de significancia nominal $\alpha = 0.025, 0.05$ para los márgenes de no-inferioridad $d_0 = 0.05, 0.1, 0.15$.

## 2. El modelo

Sean $X_i$ variables aleatorias binomiales independientes, con parámetros $(n_i, p_i)$, $i = 1, 2$; donde $p_1$ y $p_2$ representan las probabilidades de éxito bajo el tratamiento estándar y el nuevo, respectivamente.

Consideremos el problema de prueba de hipótesis $H_0 : p_1 - p_2 \geq d_0$ vs $H_a : p_1 - p_2 < d_0$. El espacio muestral y el espacio de parámetros son $\chi = \{0, ..., n_1\} \times \{0, ..., n_2\}$ y $\Theta = [0, 1]^2$, respectivamente.

## 3. Una prueba conocida con buen comportamiento

La estadística $T_1$ se define mediante $T_1(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\check{\sigma}_1}$ donde $\hat{p}_i$ es el estimador de máxima verosimilitud de $p_i$ y $\check{\sigma}_1 = \sqrt{\frac{\check{p}_1(1-\check{p}_1)}{n_1} + \frac{\check{p}_2(1-\check{p}_2)}{n_2}}$ donde $\check{p}_i$ es el estimador de máxima verosimilitud de $p_i$, restringido bajo la hipótesis nula, ver Mietinen y Nurminen (1985) y Farrington y Manning (1990).

Hay evidencia de que esta prueba tiene un buen comportamiento en términos del tamaño de prueba, ver por ejemplo, Miettinen y Nurminen (1985), Farrington y Manning (1990), Chan (1998), Chan(2003) y Sotres-Ramos et al. (2010).

## 4. El nuevo procedimiento de prueba

Agresti y Caffo (2000), consideraron el problema de estimación del parámetro binomial de proporción $p$, dichos autores notaron que el estimador de máxima verosimilitud $\hat{p} = X/n$ no es una buena elección como un estimador de $p$ cuando el tamaño de muestra es pequeño. Ellos propusieron como una alternativa para estimar la proporción $p$ al estimador $\hat{p} = (X + 1)/(n + 2)$ y verificaron que este estimador trabaja bien también para comparación de proporciones de dos muestras.

Böhning y Viwatwongkasen (2005) sugirieron usar la clase de estimadores de forma paramétrica $\tilde{p}_c = (X + c)/(n + 2c)$ con $c \geq 0$ . Sea $\phi(n) = \lfloor n/25 \rfloor + 1$, donde $\lfloor x \rfloor$ denota al mayor entero que es menor o igual que $x$ y sea además al estimador $\tilde{p}_{\phi(n)}$ el cual en lo sucesivo, por comodidad, será denotado simplemente por $\tilde{p}$, es decir, $\tilde{p} = (X + \varphi(n))/(n + 2\varphi(n))$, se usará este estimador de proporción contraído para la construcción del nuevo estadístico de prueba que se propone en este trabajo: $T_2(X_1, X_2) = \frac{\tilde{p}_1 - \tilde{p}_2 - d_0}{\tilde{\sigma}_2}$ donde $\hat{p}_i$ es el estimador de máxima verosimilitud de $p_i$ y $\tilde{\sigma}_2 = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}$.

# 5. Comparación de las pruebas

Se compararon las pruebas $T_1$ y $T_2$ mediante los tamaños de prueba y las potencias.

La comparación de los tamaños de prueba se realizó de la siguiente forma. Se usó la fórmula (3) dada en Almendra-Arao (2009), para obtener una aproximación del tamaño de prueba. La aproximación del tamaño de prueba, se efectuó, reemplazando el intervalo $[d_0, (1 + d_0)/2]$ por el conjunto de puntos $I_c = \{d_0 + \Delta i : i = 0, 1, 2, ..., (1 - d_0)/\Delta\}$ con $\Delta = 0.001$. Dada una prueba $T$, el tamaño de prueba aproximado correspondiente al nivel de significancia nominal $\alpha$ y a los tamaños de muestra $n_1, n_2$ será denotado por $\alpha_a(T, n_1, n_2)$; para el caso $n_1 = n_2 = n$, la notación será $\alpha_a(T, n)$; cuando no se requiera especificar el tamaño de muestra, será denotado mediante $\alpha_a(T)$ o $\alpha_a$. Se compararon los tamaños de prueba aproximados, $\alpha_a$, para todas las configuraciones de tamaños de muestra, márgenes de no inferioridad y niveles de significancia nominal que se consideraron en este trabajo. Nótese que la fórmula (3) de Almendra-Arao (2009) usada para calcular el tamaño de prueba, es una fórmula exacta, luego entonces la única aproximación que se hace para calcular el tamaño de prueba es al reemplazar el intervalo $[d_0, (1 + d_0)/2]$ por el conjunto discreto $I_c$.

Sea $\Theta_1 = \{(p_1, p_2) : p_1 - p_2) < d_0\}$, se denotará mediante $\beta_i(p_1, p_2)$ al valor de la función potencia de la prueba $T_i$ en el punto $(p_1, p_2) \in \Theta_1$.

Posteriormente se compararon las potencias de las pruebas de la manera siguiente: se realizaron las comparaciones de $\beta_1(p_1, p_2)$ con $\beta_2(p_1, p_2)$ en una malla del conjunto $\Theta_1$ con incremento $\Delta' = 0.01$, específicamente, la comparación de dichas potencias se realizó en la malla $\Theta_{1D} = \{(p_1, p_2) : p_1 = \Delta' i, p_2 = max(p_1, -d_0) + \Delta' i : i = 0, 1, ..., (1 - max(p_1 - d_0))/\Delta'\}$.

Debido a la naturaleza discreta de las variables aleatorias, las pruebas no necesariamente

tienen tamaño igual a $\alpha$, pero por tratarse de pruebas exactas, su tamaño es $\leq \alpha$. Por tal motivo la comparación de potencias se torna complicada. De acuerdo con Mehta and Hilton (1993) "las características de dos pruebas discretas a ser comparadas no son comparables directamente, puesto que cada prueba puede tener un tamaño de prueba distinto, para un nivel de significancia especificado", además, de acuerdo con Berger (1994) "para comparar de manera significativa la potencia de dos pruebas, necesitamos hacer aproximadamente iguales las probabilidades de error tipo I". Tomando en cuenta ambas observaciones, la comparación de las potencias de las pruebas fue efectuada de la siguiente forma: para cada valor de $\alpha$ (0.025, 0.05) y para cada valor de $d_0$ (0.05, 0.10, 0.15) se compararon las potencias de las pruebas $T_1$ y $T_2$ solamente para aquellos tamaños de muestra $5 \leq n \leq 100$, para los cuales $\mid \alpha_a(T_1, n) - \alpha_a(T_2, n) \mid \leq 0.0001$. La comparación de las potencias se realizó en los puntos con $\Delta' = 0.01$, para efectos prácticos solamente se consideraron aquellos puntos donde al menos una de las potencias a comparar fue mayor o igual que 0.70.

## 5.1. Comparación de los tamaños de prueba

Como se trata de pruebas exactas, por definición son conservadoras, por tal motivo es natural tratar de cuantificar su grado de conservadurismo, para ello se clasifican los tamaños de prueba aproximados de acuerdo a su pertenencia o no al intervalo $[0.8\alpha, \alpha]$.

Para los tamaños de muestra pequeños ($5 \leq n \leq 30$) las pruebas no presentan grandes diferencias, el caso donde hay una mayor discrepancia entre ellas es cuando $d_0 = 0.15, \alpha = 0.025$ donde para $T_1$ el porcentaje de tamaños de prueba que pertenecen a $[.8\alpha, \alpha]$ es de 61.54 % y para $T_2$ es 84.62 %, mientras que las correspondientes desviaciones medias con respecto a $\alpha$ son 0.00448 y 0.00245, respectivamente.

Para los tamaños de muestra moderados ($31 \leq n \leq 100$), las dos pruebas presentan resultados prácticamente iguales en cuanto al porcentaje de valores que pertenecen al intervalo $[.8\alpha, \alpha]$ y a los valores de la desviación media relativa respecto al nivel de significancia nominal. Como un ejemplo, se considera $n_1 = n_2 = n = 60$, $\alpha = 0.05$, $d_0 = 0.10$. Para $T_1$ se tiene que $\alpha_a(T_1) = 0.04946$ y dicho máximo se obtiene en los puntos $(p_1, p_2) = (0.409, 0.309)$ y $(p_1, p_2) = (0.691, 0.591)$. Para $T_2$ se tiene que $\alpha_a(T_2) = 0.04407$ y el máximo se alcanza en $(p_1, p_2) = (0.369, .269)$ y $(p_1, p_2) = (0.631, 0.531)$.

## 5.2. Comparación de las potencias

La comparación de las potencias se realizó solamente para los tamaños de muestra $n$ que cumplieron la condición $\mid \alpha_a(T_A, n) - \alpha_a(T_B, n) \mid \leq 0.0001$.

Consideremos el caso $d_0 = 0.05$, $\alpha = 0.025$. El porcentaje de puntos de $\Theta_{1D}$ en los cuales $Potencia(T_1) < Potencia(T_2)$ es elevado, pero no hay gran diferencia en las potencias, ya que la media de las diferencias $Potencia(T_1) - Potencia(T_2)$ es, en general menor o igual que 0.0095, el cual tratándose de potencias, es un valor bastante pequeño y recordemos además que se están comparando las potencias de aquéllos puntos de la malla $\Theta_{1D}$ en los cuales al menos una de las potencias fue mayor o igual que 0.70. En los puntos para los cuales $Potencia(T_1) > Potencia(T_2)$, los porcentajes son menores notándose que la media de las diferencias $Potencia(T_1) - Potencia(T_2)$ es, en todos los casos menor que 0.00001, excepto para $n = 23$, donde tal diferencia fue 0.00170. Así, para el caso $d_0 = 0.05$, $\alpha = 0.025$, la diferencia entre las potencias de las dos pruebas es insignificante y por tanto se pueden considerar en términos prácticos como pruebas de igual potencia.

De la misma forma se analizaron todos los casos: $d_0 = 0.05, 0.10$; $\alpha = 0.025, 0.05$. En todos ellos se obtuvo la misma conclusión: la diferencia entre las potencias de las dos pruebas es insignificante y por tanto se pueden considerar como pruebas de igual potencia.

# 6. Ejemplos

En esta sección se analizan los datos reportados en un par de artículos: Rodary et. al. (1989) y Bernard et al. (2002). Dicho análisis se efectuó con cuatro pruebas: la prueba asintótica de Blackwelder ($T^a_{Bw}$), $T_1$ y su versión asintótica ($T^a_1$) y con $T_2$. Una prueba se considerará adecuada para aplicarse si su tamaño está muy cercano el nivel de significancia nominal para el cual fue construida y preferentemente por debajo de dicho nivel.

El primer ejemplo que se analiza se encuentra en Rodary et. al. (1989); donde los autores presentaron un ensayo clínico aleatorizado de 164 niños para los cuales el análisis estadístico se basó en un margen de no inferioridad de $d_0 = 0.10$ y un nivel de significancia nominal $\alpha = 0.05$. Las proporciones de éxito correspondiente a los grupos de quimioterapia y radiación fueron 83/88 y 69/76, respectivamente. Con la prueba asintótica $T^a_1$, Chan (1998) analizó dichos resultados, el tamaño de prueba que obtuvo fue 0.0578. En la tabla 1 se

presentan los resultados de los cálculos para analizar estos datos.

**Tabla 1**: Constantes críticas, tamaños de prueba y valor del estadístico para los datos de Rodary et al. (1989).

| Prueba | Constante crítica | $T(69, 83)$ | ¿Rechaza $H_0$? | Tamaño de la prueba |
|---|---|---|---|---|
| $T^a_{Bw}$ | $-1.64485$ | $-3.27229$ | Sí | $0.11543 = 2.31\alpha$ |
| $T^a_1$ | $-1.64485$ | $-2.95715$ | Sí | $0.05777 = 1.16\alpha$ |
| $T_1$ | $-1.70597$ | $-2.95715$ | Sí | $0.04778 = 0.95\alpha$ |
| $T_2$ | $-1.65234$ | $-2.72605$ | Sí | $0.04985 = 0.99\alpha$ |

La prueba $T^a_{Bw}$ es demasiado liberal pues su tamaño es más del doble del nivel de significancia nominal. Los tamaños de las otras tres pruebas se mantienen razonablemente cercanos al nivel nominal por lo cual se pueden considerar como apropiadas en este caso. El tamaño de la prueba $T^a_1$ es mayor que el nivel nominal, además, la prueba cuyo tamaño está más cercano al nivel nominal es $T_2$, por lo cual se recomienda el uso de esta prueba en este ejemplo. Es importante resaltar que el tamaño de la prueba puede calcularse antes de realizar el experimento; mientras que el valor $T(69, 83)$ naturalmente, se calcula después del experimento.

El segundo ejemplo analizado fue presentado por Bernard et al. (2002), cuyos autores reportan el resultado de un ensayo de no-inferioridad para comparar la pristinamicina oral contra el régimen estándar de penicilina para tratar erisipela en adultos, el ensayo clínico fue efectuado con 204 pacientes, 102 en cada uno de los tratamientos; el análisis estadístico fue realizado para un margen de no inferioridad $d_0 = 0.10$ y un nivel de significancia nominal $\alpha = 0.05$. Las proporciones de éxito para la pristinamicina y penicilia fueron 83/102 and 68/102, respectivamente. Los resultados de los cálculos realizados para el análisis de los datos de este ejemplo se presentan en la tabla 2.

**Tabla 2**: Constantes críticas, tamaños de prueba y valor del estadístico para los datos de Bernard et al. (2002).

| Prueba | Constante crítica | $T(68, 83)$ | ¿Rechaza $H_0$? | Tamaño de la prueba |
|---|---|---|---|---|
| $T^a_{Bw}$ | $-1.64485$ | $-4.08114$ | Sí | $0.10549 = 2.11\alpha$ |
| $T^a_1$ | $-1.64485$ | $-3.98089$ | Sí | $0.05173 = 1.03\alpha$ |
| $T_1$ | $-1.71727$ | $-3.98089$ | Sí | $0.04883 = 0.98\alpha$ |
| $T_2$ | $-1.70889$ | $-3.96845$ | Sí | $0.04884 = 0.98\alpha$ |

La situación para este ejemplo es similar a la del ejemplo anterior, la prueba $T_{Bw}^{a}$ también es demasiado liberal. Los tamaños de las otras tres pruebas se mantienen razonablemente cercanos al nivel nominal por lo cual se pueden considerar como razonables para aplicarse en este caso. El tamaño de la prueba $T_1^a$ es mayor que el nivel nominal, en este caso las dos pruebas exactas tienen el mismo tamaño, debido a lo cual se recomienda el uso de cualquiera de ellas, además, tomando en cuenta la sencillez de aplicación se sugiere el uso de $T_2$.

# 7. Conclusiones

Para todas las configuraciones de tamaños de muestra, niveles de significancia nominal y márgenes de no inferioridad estudiados, se puede establecer que el comportamiento de los tamaños de prueba aproximados de la prueba $T_1$, en términos generales, es ligeramente mejor que el de $T_2$. Por otra parte, en lo relativo a la comparación de potencias, prácticamente son pruebas de igual potencia. Además, una ventaja de la prueba nueva $T_2$ sobre la prueba $T_1$ es su facilidad de cálculo. Lo anterior permite recomendar el uso de $T_2$. Finalmente, en forma específica, en los dos ejemplos analizados es también recomendable el uso de $T_2$.

# Bibliografía

Agresti, A. y Caffo, B. (2000), "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures", *The American Statistician* **54**, 280–288.

Almendra-Arao, F. (2009), "A study of the asymptotic classical non-inferiority test for two binomial proportions", *Drug Information Journal* **43**, 567–571.

Almendra-Arao, F. y Sotres-Ramos, D. (2009), "Comparison of some non-inferiority asymptotic tests for two independent proportions", *Agrociencia* **43**, 163–172.

Berger, R. (1994), "Power comparison of exact unconditional tests for comparing two binomial probabilities", *Institute of Statistics, Mimeo Series* **2266**.

Bernard, P., Chosidow, O. y Vaillant, L. (2002), "Oral pristinamycin versus standard penicillin regimen to treat erysipelas in adults: randomised, non-inferiority, open trial", *British Medical Journal* **325(7369):864**.

Böhning, D. y Viwatwongkasen, C. (2005), "Revisiting proportion estimators", *Stat. Methods Med. Res* **14**, 1–23.

Blackwelder, W. (1982), "Proving the null hypothesis in clinical trials", *Controlled Clinical Trials* **3**, 345–353.

Chan, I. S. F. (1998), "Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies", *Stat. Med.* **17**, 1403–1413.

Chan, I. S. F. (2003), "Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods", *Stat. Methods Med. Res.* **12**, 37–58.

Farrington, C. y Manning, G. (n.d.), "Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk", *Statistics in Medicine* **9**, 1447–1454.

Mehta, C. R. y Hilton, J. F. (1993), "Exact power of conditional and unconditional tests: Going beyond the 2x2 contingency table", *The American Statistician* **47**, 91–98.

Miettinen, O. y Nurminen, M. (1985), "Comparative analysis of two rates", *Stat. Med* **4**, 213–226.

Röhmel, J. (2005), "Problems with existing procedures to calculate exact unconditional p-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them", *Biom. J.* **47**, 37–47.

Röhmel, J. y Mansmann, U. (1999), "Unconditional nonasymptotic one sided tests for independent binomial proportions when the interest lies in showing noninferiority and or superiority", *Biom. J.* **41**, 149–170.

Rodary, Com-Nougue, C. y Tournade, M. F. (1989), "How to establish equivalence between treatments: a one-sided clinical trial in paediatric oncology", *Stat. Med.* pp. 593–598.

Sotres-Ramos, D., Almendra-Arao, F. y Ramírez-Figueroa, C. (2010), "Exact critical values for farrington-manning noninferiority exact test", *Drug Information Journal* **44**, 159–164.

Memoria FONE XXV (2010), Sección III, 59–65

# La prueba de bondad de ajuste $R$ para la distribución exponencial

María D. Kantún Chim[a]
*Universidad Autónoma de Yucatán*

José A. Villaseñor Alva
*Colegio de Postgraduados*

## 1. Introducción

La distribución exponencial es de gran importancia por sus múltiples aplicaciones, entre las cuales se encuentra la teoría de confiabilidad y el análisis de supervivencia, por lo cual, desde 1960 se han desarrollado varias pruebas de bondad de ajuste con muestras de datos completos y censurados. Diversos autores han reportado estudios acerca de la comparación de pruebas existentes, entre los más interesantes se encuentran los de D'Agostino y Stephens (1984), Spurrier (1984) y Ascher (1990). En estos estudios han sobresalido como las pruebas de bondad de ajuste más potentes la prueba de Hollander y Proschan (1972) y la de Cox y Oakes (1984). Kantún *et al.* (2005) propusieron una prueba de bondad de ajuste basada en la razón de dos estimadores del parámetro de escala de la distribución exponencial, a este estadístico se le ha denominado $R$ y tiene la característica de ser simple de calcular. En el presente trabajo se estudia la distribución asintótica del estadístico $R$ y se proporciona un análisis comparativo en términos de la potencia de $R$ y las otras pruebas más potentes mencionadas anteriormente respecto a algunas alternativas.

[a]kchim@uady.mx

# 2. Marco teórico

La función de densidad exponencial con parámetro de escala $\beta$, se define como

$$f(x) = \frac{1}{\beta} \exp\{-x/\beta\} I_{(0,\infty)}(x), \quad \beta > 0. \tag{1}$$

Con base en una muestra aletoria $X_1, X_2, ..., X_n$ se desea probar la hipótesis nula $H_0$: $X_1, X_2, ..., X_n$ proviene de una distribución exponencial contra la alternativa $H_A$: $X_1, X_2, ..., X_n$ no proviene de una distribución exponencial.

La estadística $R$ es la razón del rango entre dos veces el semirango, es decir,

$$R = \frac{Y_n - Y_1}{Y_n + Y_1} \tag{2}$$

donde $Y_1 = min(X_1, X_2, ..., X_n)$ y $Y_n = max(X_1, X_2, ..., X_n)$ (Kantún *et al.*, 2005).

La prueba propuesta es: Rechazar $H_0$ si $R < r_1$ o $R > r_2$. Si se desea una prueba de tamaño $\alpha^*$ entonces los valores de $r_1$ y $r_2$ deben ser tales que:

$$\alpha^* = P(R < r_1 \ o \ R > r_2 | H_0).$$

Una manera de seleccionar a los valores de $r_1$ y $r_2$ es:

$$P(R < r_1 | H_0) = P(R > r_2 | H_0) = \frac{\alpha^*}{2}.$$

Usando el método de jacobianos se obtiene la función de densidad del estadístico $R$ bajo $H_0$ (ver detalles en Kantún *et al.*, 2005) dada por:

$$f_R(r) = 2n(n-1) \sum_{k=0}^{n-2} \binom{n-2}{k} \frac{(-1)^{n-2-k}}{[r(n-2-2k)+n]^2}, \ 0 < r < 1. \tag{3}$$

Es posible obtener las constantes críticas usando la expresión (3) mediante una rutina en S-Plus para tamaños de muestra $n \leq 45$; sin embargo, para $n > 45$ la distribución (3) resulta ser degenerada y no es posible encontrar las constantes críticas. Por lo tanto, se requiere estudiar la distribución asintótica del estadístico $R$ bajo $H_0$.

## 2.1. Distribución asintótica del estadístico $R$ bajo $H_0$

Para obtener la distribución asintótica de $R$, a continuación se presentan algunos resultados de la teoría de valores extremos. Para esto se requiere la siguiente notación.

Se define el ínfimo de la función de distribución $F(x)$ como $\alpha(F) = inf(x : F(x) > 0)$ y puede ser $-\infty$ o finito. Se define el supremo de $F(x)$ como $\omega(F) = \sup(x : F(x) < 1)$ y puede ser $+\infty$ o finito. Los siguientes teoremas aparecen en el libro de Galambos (1987); aquí los presentamos usando su misma notación.

**Teorema 1**. Para $F$ una función de distribución, supóngase que $\int_t^{\omega(F)} (1-F(y))dy < +\infty$ para $\alpha(F) < t < \omega(F)$ y sea $R(t) = (1 - F(t))^{-1} \int_t^{\omega(F)} (1 - F(y))dy$. Supóngase que para todo número real $x$ cuando $t \to \omega(F)$, $\lim \frac{1-F(t+xR(t))}{1-F(t)} = e^{-x}$. Entonces existen sucesiones $a_n$ y $b_n > 0$ tales que cuando $n \to +\infty$, $\lim P(Y_n < a_n + b_n x) = H_{3,0}(x)$, donde $H_{3,0}(x) = exp(-e^{-x})$, $-\infty < x < +\infty$, y las constantes de normalización pueden ser seleccionadas como $a_n = inf(x : 1 - F(x) \leq \frac{1}{n})$ y $b_n = R(a_n)$.

A la función de distribución $H_{3,0}(x)$ se le llama la distribución Gumbel de valores extremos.

**Teorema 2**. Sea $F$ una función de distribución tal que $\alpha(F)$ es finita. Supóngase que la función de distribución $F^*(x) = F(\alpha(F) - \frac{1}{x})$, $x < 0$ satisface $\lim_{t\to-\infty} \frac{F^*(tx)}{F^*(t)} = x^{-\gamma}$ entonces existen sucesiones $c_n$ y $d_n > 0$ tales que, cuando $n \to +\infty$, $\lim P(Y_1 < c_n + d_n x) = L_\gamma(x)$, donde $L_\gamma(x) = [1 - exp(-x^\gamma)]I_{(0,\infty)}(x)$ y las constantes de normalización $c_n$ y $d_n$ pueden ser seleccionadas como $c_n = \alpha(F)$ y $d_n = \sup(x : F(x) \leq \frac{1}{n}) - \alpha(F)$.

**Teorema 3**. Sean $X_1, X_2, ..., X_n$ variables aleatorias i.i.d con función de distribución $F(x)$. Supóngase que $F(x)$ es tal que existen sucesiones $a_n, c_n, b_n > 0$ y $d_n > 0$ para las cuales, cuando $n \to +\infty$, $\lim F^n(a_n + b_n x) = H(x)$ y $\lim[1 - F(c_n + d_n x)]^n = 1 - L(x)$ existen y son no degenerados. Entonces, cuando $n \to +\infty$, $\lim P(Y_1 < c_n + d_n x, Y_n < a_n + b_n y) = L(x)H(y)$.

Note que este teorema establece que las estadísticas extremas $Y_1$ y $Y_n$ son asintóticamente independientes.

Para obtener la distribución asintótica de $R$ bajo $H_0$, sean $X_1, X_2, ..., X_n$ una muestra aleatoria de la distribución exponencial con parámetro $\beta = 1$ y sean $Y_1 = \text{mín}\{X_1, X_2, ..., X_n\}$ y $Y_n = \text{máx}\{X_1, X_2, ..., X_n\}$. Por el Teorema 1, cuando $F$ es la distribución exponencial con parámetro $\beta = 1$ se deduce que las constantes de normalización de $Y_n$ son $b_n = \beta$ y $a_n = \beta \log n$ y que la distribución asintótica de $Y_n^* = \frac{Y_n - \beta log n}{\beta}$ es una distribución Gumbel.

Por el Teorema 2, cuando $F$ es la distribución exponencial con parámetro $\beta = 1$ resulta

$\gamma = 1$, de donde las constantes de normalización son $c_n = 0$ y $d_n = \beta log(\frac{n}{n-1})$ y por lo tanto la distribución asintótica de $Y_1^* = \frac{Y_1}{\beta log(\frac{n}{n-1})}$ es una distribución exponencial con $\gamma = 1$.

Por el Teorema 3, $Y_n^*$ y $Y_1^*$ son asintóticamente independientes. Por lo tanto, la estadística $R$ en función de las estadísticas $Y_1^*$ y $Y_n^*$ tiene la siguiente representación:

$$R = \frac{Y_n^* - log(\frac{n}{n-1})Y_1^* + logn}{Y_n^* + log(\frac{n}{n-1})Y_1^* + logn}. \tag{4}$$

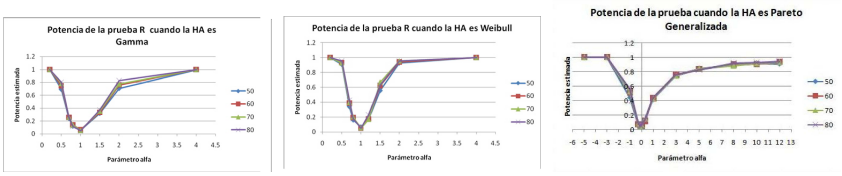
Ya que la distribución de $R$ bajo $H_0$ no depende del parámetro de escala de la distribución exponencial, es posible obtener una aproximación de la distribución asintótica de $R$ por simulación para cualquier valor dado de $\beta$ y $n \geq 40$. Para esto se generan 100,000 valores de $Y_n^*$ y $Y_1^*$ de las distribuciones Gumbel y exponencial con $\beta = 1$ respectivamente para cada $n$ tal que $40 \leq n \leq 100$. Debido a que $Y_n^*$ y $Y_1^*$ son aproximadamente independientes con $n \geq 40$, usando (4), se obtienen 100,000 valores simulados de $R$ para los cuales se considera su distribución empírica. De esta aproximación de la distribución de $R$ son obtenidos los cuantiles deseados, los cuales se presentan en la tabla siguiente.

**Cuadro 1:** Cuantiles de la distribución aproximada de $R$ bajo $H_0$.

| n/α* | 0.005 | 0.0125 | 0.025 | 0.05 | 0.95 | 0.975 | 0.9975 | 0.995 |
|---|---|---|---|---|---|---|---|---|
| 40 | 0.9249438 | 0.9407346 | 0.9508643 | 0.9606125 | 0.9993948 | 0.9997220 | 0.9999717 | 0.9999489 |
| 41 | 0.9280263 | 0.9422297 | 0.9521279 | 0.9623234 | 0.9994273 | 0.9997200 | 0.9999703 | 0.9999448 |
| 42 | 0.9300197 | 0.9428659 | 0.9528024 | 0.9634462 | 0.9994397 | 0.9997334 | 0.9999681 | 0.9999419 |
| 43 | 0.9352178 | 0.9459585 | 0.9557908 | 0.9645209 | 0.9994695 | 0.9997367 | 0.9999692 | 0.9999413 |
| 44 | 0.9365233 | 0.9465977 | 0.9549837 | 0.9649044 | 0.9994387 | 0.9997222 | 0.9999800 | 0.9999607 |
| 45 | 0.9369316 | 0.9476583 | 0.9567420 | 0.9658221 | 0.9994873 | 0.9997390 | 0.9999711 | 0.9999434 |
| 46 | 0.9393547 | 0.9491919 | 0.9579217 | 0.9670481 | 0.9994849 | 0.9997559 | 0.9999739 | 0.9999521 |
| 47 | 0.9378710 | 0.9510388 | 0.9589333 | 0.9678956 | 0.9994801 | 0.9997467 | 0.9999748 | 0.9999404 |
| 48 | 0.9419727 | 0.9535112 | 0.9613524 | 0.9696795 | 0.9995490 | 0.9997723 | 0.9999764 | 0.9999580 |
| 49 | 0.9434110 | 0.9529726 | 0.9618531 | 0.9694187 | 0.9995321 | 0.9997639 | 0.9999749 | 0.9999541 |
| 50 | 0.9416622 | 0.9542318 | 0.9629397 | 0.9700904 | 0.9995215 | 0.9997649 | 0.9999810 | 0.9999580 |
| 51 | 0.9458177 | 0.9566193 | 0.9643325 | 0.9711364 | 0.9995608 | 0.9997732 | 0.9999810 | 0.9999564 |
| 52 | 0.9475626 | 0.9580434 | 0.9650521 | 0.9719679 | 0.9995348 | 0.9997792 | 0.9999714 | 0.9999491 |
| 53 | 0.9505039 | 0.9594366 | 0.9655600 | 0.9725136 | 0.9995565 | 0.9997883 | 0.9999805 | 0.9999592 |
| 54 | 0.9496014 | 0.9586462 | 0.9662206 | 0.9730199 | 0.9995458 | 0.9997872 | 0.9999848 | 0.9999706 |
| 55 | 0.9528875 | 0.9598026 | 0.9665039 | 0.9735951 | 0.9996063 | 0.9998252 | 0.9999767 | 0.9999647 |
| 56 | 0.9507033 | 0.9602019 | 0.9679849 | 0.9743102 | 0.9995898 | 0.9997991 | 0.9999835 | 0.9999619 |
| 57 | 0.9531620 | 0.9622203 | 0.9686761 | 0.9755174 | 0.9996076 | 0.9997939 | 0.9999834 | 0.9999638 |
| 58 | 0.9543648 | 0.9643363 | 0.9700433 | 0.9756736 | 0.9995905 | 0.9998004 | 0.9999810 | 0.9999538 |
| 59 | 0.9571849 | 0.9640531 | 0.9704936 | 0.9762712 | 0.9996329 | 0.9998232 | 0.9999838 | 0.9999713 |
| 60 | 0.9545005 | 0.9652102 | 0.9712866 | 0.9767215 | 0.9996461 | 0.9998293 | 0.9999854 | 0.9999661 |
| 65 | 0.9611992 | 0.9687679 | 0.9738703 | 0.9788069 | 0.9996706 | 0.9998387 | 0.9999820 | 0.9999681 |
| 70 | 0.9633572 | 0.9699817 | 0.9753146 | 0.9802744 | 0.9996952 | 0.9998579 | 0.9999841 | 0.9999705 |
| 75 | 0.9661060 | 0.9725464 | 0.9778542 | 0.9823497 | 0.9997200 | 0.9998571 | 0.9999854 | 0.9999678 |
| 80 | 0.9694547 | 0.9749439 | 0.9790822 | 0.9832719 | 0.9997525 | 0.9998800 | 0.9999876 | 0.9999742 |
| 85 | 0.9724468 | 0.9769940 | 0.9810645 | 0.9846209 | 0.9997587 | 0.9998823 | 0.9999873 | 0.9999778 |
| 90 | 0.9729400 | 0.9790986 | 0.9822895 | 0.9859184 | 0.9997601 | 0.9998843 | 0.9999872 | 0.9999777 |
| 95 | 0.9760659 | 0.9797396 | 0.9832907 | 0.9866703 | 0.9997958 | 0.9998915 | 0.9999870 | 0.9999732 |
| 100 | 0.9763734 | 0.9809906 | 0.9844967 | 0.9873318 | 0.9998044 | 0.9999030 | 0.9999935 | 0.9999806 |

## 2.2. Estudio de Potencia

El método de Monte Carlo es empleado para analizar la potencia de la prueba. Las distribuciones alternativas empleadas para este análisis son: Gamma$(\alpha, \beta)$, Weibull$(\alpha, \beta)$ y Pareto Generalizada$(\alpha, \beta)$, donde $\alpha$ es el parámetro de forma y $\beta$ es el parámetro de escala. Se generaron m.a. de estas distribuciones para diferentes valores de $\alpha$ y $\beta = 1$, considerando tamaños de muestra de n=50, 60, 70, 80, y un tamaño de prueba $\alpha_* = 0.05$. Los resultados son presentados en las siguientes gráficas. Cuando $\alpha = 1$ en los casos gamma y Weibull y cuando el parámetro de forma $\alpha$ de la distribución Pareto generalizada es igual a cero se tiene la hipótesis nula y se puede observar que la potencia es cercana a 0.05, lo cual es congruente con el hecho de que las pruebas son de tamaño 0.05. Por otro lado, note que cuando $\alpha$ se aleja de la hipótesis nula, la función de potencia va creciendo a uno.




## 2.3. Estudio comparativo

Para realizar un estudio comparativo con respecto a la potencia de la prueba $R$, se consideran las pruebas de Cox & Oakes (1984), la prueba HP de Hollander & Proschan (1972) y la prueba Q de Wong & Wong (1979). Estas pruebas son descritas por Kantún *et al.* (2005). Se emplearon las alternativas Gamma$(\alpha, \beta)$, Weibull$(\alpha, \beta)$ y Pareto Generalizada$(\alpha, \beta)$, donde $\alpha$ es el parámetro de forma y $\beta$ es el parámetro de escala. El proceso es realizado con un tamaño de muestra n=50, $\beta = 1$ y diferentes valores del parámetro de forma $\alpha$. El tamaño de la prueba es $\alpha^* = 0.05$. Los resultados son presentados en los siguientes cuadros. Note que cuando $\alpha = 1$ para el caso de las alternativas gamma y Weibull y $\alpha = 0$ para el caso Pareto Generalizada, que es la hipótesis nula, las potencias de las pruebas Cox y HP son cercanas a 0.05 como era de esperarse; sin embargo, $R$ y Q exceden por poco ese valor. Cuando $\alpha$ se aleja de la hipótesis nula, se observa que la prueba de Cox es la más potente.

Potencias estimadas de las pruebas *R*, Q, Cox, y HP cuando la hipótesis alternativa es Gamma(α, β=1) con *n*=50 y α*=0.05 (1000 repeticiones).

| PRUEBAS/α | 0.2 | 0.5 | 0.7 | 0.8 | 1.0 | 1.5 | 2.0 | 4.0 | 8.0 |
|---|---|---|---|---|---|---|---|---|---|
| R | 1 | 0.693 | 0.241 | 0.119 | 0.058 | 0.313 | 0.691 | 0.997 | 1.000 |
| Q | 1 | 0.781 | 0.323 | 0.163 | 0.064 | 0.001 | 0.000 | 0.000 | 0.000 |
| COX | 1 | 0.971 | 0.537 | 0.232 | 0.057 | 0.555 | 0.955 | 1.000 | 1.000 |
| HP | 1 | 0.945 | 0.461 | 0.206 | 0.047 | 0.459 | 0.912 | 1.000 | 1.000 |

Potencias estimadas de las pruebas *R*, Q, Cox, y HP cuando la hipótesis alternativa es Weibull(α,β=1) con *n*=50 y α*=0.05 (1000 repeticiones).

| PRUEBAS/α | 0.2 | 0.5 | 0.7 | 0.8 | 1.0 | 1.2 | 1.5 | 2.0 | 4.0 | 8.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| R | 1.000 | 0.907 | 0.342 | 0.162 | 0.070 | 0.159 | 0.567 | 0.921 | 1.000 | 1.000 |
| Q | 1.000 | 0.952 | 0.441 | 0.235 | 0.065 | 0.007 | 0.001 | 0.000 | 0.000 | 0.000 |
| COX | 1.000 | 0.100 | 0.883 | 0.525 | 0.045 | 0.355 | 0.952 | 0.100 | 1.000 | 1.000 |
| HP | 1.000 | 0.999 | 0.800 | 0.464 | 0.045 | 0.289 | 0.908 | 0.100 | 1.000 | 1.000 |

Potencias estimadas de las pruebas *R*, Q, Cox, y HP cuando la hipótesis alternativa es Pareto Generalizada (α, β=1) con *n*=50 y α*=0.05 (1000 repeticiones).

| Pruebas/α | -5.0 | -3.0 | -1.0 | -0.3 | -0.1 | -0.01 | 0.01 | 0.3 | 1.0 | 3.0 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R | 1.000 | 0.997 | 0.442 | 0.067 | 0.051 | 0.055 | 0.052 | 0.128 | 0.408 | 0.729 | 0.822 |
| Q | 1.000 | 0.999 | 0.567 | 0.133 | 0.059 | 0.062 | 0.052 | 0.029 | 0.010 | 0.007 | 0.003 |
| COX | 1.000 | 1.000 | 0.989 | 0.404 | 0.101 | 0.052 | 0.054 | 0.298 | 0.946 | 1.000 | 1.000 |
| HP | 1.000 | 1.000 | 0.876 | 0.213 | 0.073 | 0.050 | 0.057 | 0.256 | 0.988 | 1.000 | 1.000 |

# 3. Conclusiones

La estadística $R$ es libre del parámetro de escala $\beta$ de la distribución exponencial por lo que no es necesario estimarlo para obtener la distribución de $R$ bajo $H_0$.

La potencia estimada de la prueba $R$ aumenta conforme el tamaño de muestra aumenta cuando los datos provienen de las distribuciones Gamma, Weibull y Pareto Generalizada, lo cual proporciona evidencia de que la prueba $R$ puede ser consistente.

Los resultados comparativos de potencia obtenidos para la prueba $R$ dan evidencia de su robustez respecto a cualquier alternativa al compararla con las pruebas más potentes para la distribución exponencial: Hollander & Proschan y la de Cox & Oakes, las cuales han sido desarrolladas para detectar alternativas con función de riesgo monótona, por lo que deberían tener ventaja sobre la prueba $R$ propuesta ya que las alternativas consideradas tienen esta característica (ver Hollander y Proschan, 1972).

De los resultados presentados se observan evidencias de que la prueba Q de Wong & Wong no es recomendable para probar exponencialidad.

# Bibliografía

Ascher, S. (1990), "A survey of tests for exponentiality", *Commun. Statist. Theory Meth* **19 (5)**, 1811–1825.

Cox, D. y Oakes, D. (1984), *Analysis of Survival Data*, first edn, Champan and Hall.

D'Agostino, R. y Stephens, M. (1984), *Goodness-Of-Fit Techniques*, Marcel Dekker.

Galambos, J. (1987), *The asymptotic Theory of Extreme Order Statistics*, second edn, Robert E. Krieger.

Hollander, M. y Proschan, F. (1972), "Testing whether new is better than used", *The Annals of Mathematical Statistics* **13 (4)**, 1136–1146.

Kantún, D., Villaseñor, J. y Vaquera, H. (2005), "Una prueba para exponencialidad basada en la razón de dos estimadores", *Agrociencia* **39 (1)**, 81–92.

Spurrier, J. (1984), "An overview of test for exponentiality", *Commun. Statistic. Theor. Meth.* **13 (13)**, 1635–1654.

Wong, P. y Wong, S. (1979), "An extremal quotient test for exponential distribution", *Metrika* **26**, 1–4.

Memoria FONE XXV (2010), Sección III, 67–73

# Una aproximación binomial con tres parámetros

Agustín Jaime García Banda[a], Luis Cruz-Kuri, Ismael Sosa Galindo
*Universidad Veracruzana*

## 1. Introducción

Un método común para mejorar la precisión de una aproximación es la construcción de una expansión asintótica. Sin embargo, en la práctica puede consumir más tiempo y es menos conveniente que calcular los valores de una distribución conocida. Un enfoque alternativo es modificar una distribución de aproximación común introduciendo algunos nuevos parámetros los cuales pueden ser usados para llevar a cabo un mejor ajuste. El uso de distribuciones comunes puede hacerlo fácil para evitar la necesidad de la programación especializada al usar los paquetes estadísticos estándar para modelar datos. Una de las modificaciones más simples es cambiar, y este enfoque trabaja bien en el caso de Poisson. Para un gran número de eventos raros independientes, la distribución del número de ellos que ocurren es frecuentemente aproximada a una distribución de Poisson. Si algunos de los eventos no son de hecho tan raros, la aproximación es probablemente baja: el número de eventos esperados que ocurren podrían no estar tan cerca de la varianza. Pero son iguales a la distribución de Poisson. Una manera fácil para tratar este problema es introducir un cambio añadiendo o substrayendo una constante de la variable aleatoria de Poisson. Esto genera esencialmente dos parámetros que pueden ser ajustados para igualar los primeros dos momentos. (Sujeto a la restricción de que el desplazamiento corresponda a un número entero). El desplazamiento también se le llama una traslación o un centrado. Uno de los propósitos de este tipo de trabajo consiste en investigar el efecto de desplazamiento aplicado a una distribución con dos parámetros. Es claro que el desplazamiento cambia la media de la distribución pero ni cambia ni su varianza ni sus momentos de orden superior centrados en la media. En parte del trabajo se utiliza

[a]jaimegarciabanda@yahoo.com

una aproximación binomial con desplazamiento para la suma de variables de Bernoulli. De interés principal para aplicaciones estadísticas se considera el caso en que las variables son independientes. En la literatura a la distribución de la suma de variables independientes de Bernoulli, con probabilidades de éxito no necesariamente iguales, se le denomina una distribución *Poisson-Binomial*.

Uno de los propósitos del presente trabajo es el de estudiar generalizaciones de la familia de la distribución Binomial, la cual como se sabe esta parametrizada por dos cantidades: n=Número de pruebas de Bernoulli y p=Probabilidad de éxito para cada prueba de Bernoulli. También se sabe que estas pruebas son independientes y que la variable X que obedece el modelo anterior es la correspondiente suma de indicadores independientes como variables aleatorias. En símbolos,

$$x = \sum_{j=1}^{n} I_j$$

Una generalización posible utiliza la formula anterior, pero ahora las variables indicadoras, tienen probabilidades de éxito no necesariamente iguales. Es decir, sean $I_1, ..., I_n$ variables aleatorias independientes con:

$$I_j = \begin{cases} 1, & p_j \\ 0, & 1 - p_j \end{cases}$$

La estrategia algebraica utiliza el desarrollo algebraico de un producto. Con instrucciones del programa *Mathematica*, tales como las que se señalan en la sección 2, se pueden realizar los cálculos requeridos. Por ejemplo, explícitamente, para n=10, digamos, la expresión se puede generar con la instrucción

Expand[($p_1$*t+$q_1$)*($p_2$*t+$q_2$)*...*($p_{10}$*t+$q_{10}$)].

# 2. Utilización del paquete *Mathematica*

## 2.1. Instrucciones básicas

Con el apoyo del paquete *Mathematica*, se plantean en esta sección algunas funciones matemáticas para generar probabilidades de acuerdo a algún modelo específico y para mostrar gráficamente las distribuciones, las cuales varían según los parámetros del estudio. Aunque el

coeficiente combinatorio es una función que ya se encuentra definida dentro del programa *Mathematica*, pueden usarse funciones más elementales para su cálculo; para ilustrar, con la instrucción que sigue, la cual usa la función factorial básica se puede definir un coeficiente combinatorio como sigue

$$combi[x\_, y\_] := Factorial[x]/(Factorial[y] * Factorial[x - y]) \tag{1}$$

Enseguida, se definen dos familias de funciones, a saber, *la familia de la distribución binomial* y *la familia de la distribución hipergeométrica*, las cuales usan la instrucción en (1) para obtener los valores correspondientes de probabilidad. La familia de la distribución binomial, la cual está parametrizada por las cantidades n y p, donde n es un entero positivo y p un número real en [0,1], queda especificada por

$$binom[n\_, p\_, x\_] := combi[n, x] * p^x * (1 - p)^{n-x} \tag{2}$$

De manera análoga, la familia de la distribución hipergeométrica, la cual esta parametrizada por las cantidades N, $N_1$ y n, y que sirve para modelar un esquema de muestreo sin remplazo de una población de tamaño N, donde hay $N_1$ elementos que tienen cierta característica que genéricamente la distinguimos por el color "rojo", digamos. Es conveniente utilizar la codificación N= npob, $N_1$ = nrojo. Se tiene,

$$hipergeom[n\_,npob\_,nrojo\_,x]:=$$

$$(combi[nrojo, x) * combi[npob - nrojo, n - x])/combi[npob, n] \tag{3}$$

Al seleccionar un miembro específico de una familia de distribuciones, digamos la Binomial, se procede a especificar los valores de los parámetros que intervienen. Como ilustración, tómese n=20 y p=0.7. La instrucción en *Mathematica* genera una tabla que muestra las probabilidades que se requieren.

$$Table[binom[20, .7, x], x, 0, 20]//MatrixForm \tag{4}$$

Los valores resultantes fluctúan desde $3.48678x10^{-11}$ hasta el 0.000797923, y por razones de espacio se omite la tabla. Para observar en forma gráfica la distribución así generada, se utiliza una instrucción como la que se presenta a continuación.

$$ListPlot[Out[8], PlotStyle- > PointSize[.03], RGBColor[1, 1, 1]] \tag{5}$$

Con lo cual se obtiene la distribución binomial con los parámetros n=20 y p=0.7. La instrucción que sigue, al ejecutarla proporciona las probabilidades hipergeométricas para un esquema de muestreo en que el tamaño de la población es N=100, teniendo $N_1$=70 elementos con una característica genérica denominada color "rojo"; el tamaño de la muestra es n=20.

$$ListPlot[Table[hipergeom[20,100,70,x],x,0,20],PlotStyle\rightarrow$$

$$PointSize[0.03], RGBColor[0,0,1], Background- > RGBColor[0.8,0.7,0]] \quad (6)$$

Con lo cual se obtiene la distribución hipergeometrica con parámetros n=20 y $N_1$=70 Para comparar las dos distribuciones se usa la instrucción en (7) que presenta las dos gráficas en una sola figura. Ver figura 1.

$$Show[Out[9], Out[10], Background- > GrayLevel[0]] \quad (7)$$

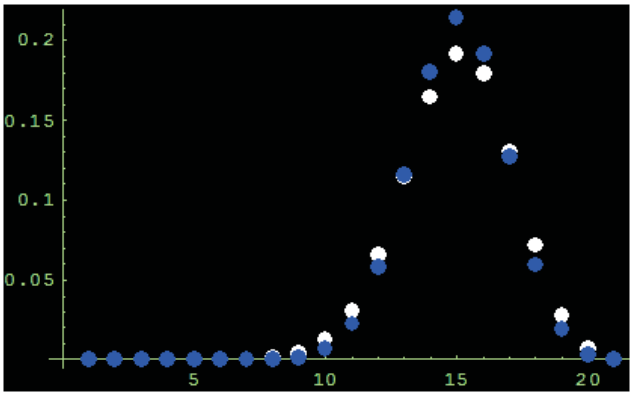



**Figura 1:** Gráfica de dos distribuciones: hipergeométrica y binomial

## 2.2. Instrucciones avanzadas

Para obtener la distribución de X se tienen a disposición con el programa *Mathematica* dos procedimientos: uno que es de tipo simbólico y que utiliza el desarrollo de un producto como potencias de una variable auxiliar t, y el otro que corresponde a una operación de *convolución*. (Nota: recuérdese que el producto de las funciones características de varias distribuciones corresponde a la función característica de la suma de variables aleatorias independientes). La instrucción en *Mathematica* es del tipo DiscreteConvolve[f, g, n, m]. Por consideraciones de espacio, se omiten los detalles.

# 3. Aspectos estadísticos

Para el estudio de distintas distribuciones de probabilidad, se puede uno apoyar en los momentos de las variables aleatorias que obedecen dichos modelos. Desde la perspectiva numérica, con la instrucción que sigue, *Mathematica* permite calcular momentos de cualquier orden k para cualquier distribución discreta f(x), x=0,1,2,3,... .

$$momento[f\_, k\_] := \sum_{x=0}^{1000} x^k \times f$$

Nótese que el límite superior de la sumatoria en 1000 fácilmente puede cambiarse a cualquier cantidad finita que se desee. Por ejemplo, para la distribución binomial con parámetros n = 20, p = 0.8, el momento de tercer orden (con la instrucción: momento[binom[20,0.8,x], 3]), produce un resultado igual a 4247.68. Con la instrucción anterior, se pueden generar momentos desde el primer orden hasta el tercero, digamos para una distribución dada, por ejemplo, la binomial con n = 20, p = 0.7.

$$Table[momento[binom[20, 0.7, x], k], k, 1, 3]//MatrixForm \tag{8}$$

Tabla 1. Momentos hasta el tercer orden para dos distribuciones

| Distribución Binomial con n=20, p=0.7 | Hipergeométrica, n=20, N=100, $N_1$=70 |
|---|---|
| 14.0 | 14.0 |
| 200.2 | 199.394 |
| 2918.72 | 2885.71 |

De manera análoga, con la instrucción en (10), se generan los momentos correspondientes para la distribución hipergeométrica con N=100, $N_1$= 70, n=20.

$$Table[momento[hipergeom[20, 100, 70., x], k], k, 1, 3]//MatrixForm \tag{9}$$

Las instrucciones que siguen calculan los momentos hasta el tercer orden para otras distribuciones y la finalidad es comparar sus respectivos momentos.

$$Table[momento[hipergeom[20, 500, 350., x], k], k, 1, 3]//MatrixForm \tag{10}$$

Tabla 2. Momentos hasta el tercer orden para tres distribuciones hipergeométricas.

| Hipergeométrica con N=500, $N_1$=350, n=20 | Hipergeométrica con N=1000, $N_1$=700, n=20 | Hipergeométrica con N=5000, $N_1$=3500, n=20 |
|---|---|---|
| 14.00000000000 | 14.00000000000 | 14.00000000000 |
| 200.0400801603 | 200.120120120 | 200.18403681 |
| 2912.79064635 | 2912.45974833 | 2918.0686557 |

$$Table[momento[hipergeom[20, 1000, 700., x], k], k, 1, 3]//MatrixForm \tag{11}$$

$$Table[momento[hipergeom[20, 5000, 3500., x], k], k, 1, 3]//MatrixForm \tag{12}$$

La inspección de las tablas 1 y 2 permite ver como los momentos de las distribuciones hipergeométricas se van aproximando a los correspondientes momentos de la distribución binomial, donde n se mantiene fija en 20 y N y $N_1$ van creciendo de tal forma que $N/N_1$ se aproxima a 0.7.

# 4. Comentarios adicionales

En la literatura, la distribución de la suma de variables aleatorias independientes de Bernoulli con probabilidades de éxito no necesariamente idénticas, se llama una distribución Binomial-Poisson. Esta distribución es extensamente aplicable y se ha estudiado ampliamente; asimismo, se han desarrollado varias aproximaciones. En el presente trabajo se ha intentado realizar estudios de esta familia de distribuciones (dentro de límites permisibles) mediante el apoyo del programa *Mathematica*. En particular se han considerado resultados para la aproximación Binomial donde el número de pruebas es igual al número de variables de Bernoulli y la probabilidad de éxito se ha escogido de tal manera que iguale en la medida de lo posible al primer momento. Otros autores han tratado a la aproximación Binomial como una aproximación de dos parámetros. Se tienen n resultados tanto a nivel teórico como a nivel numérico. También es posible, aunque no se hace en el presente trabajo, aplicar aproximaciones binomiales con dos parámetros a situaciones donde se tienen variables dependientes.

# Bibliografía

Olkin, I., Derman, C. y Gleser, L. (1973), *A Guide to Probability Theory and Application*, Holt, Rinehart and Winston.

Pekoz y et al (2009), "A three-parameter binomial aproximation", *J. Appl. Prob.* **46**, 1073–1085.

Memoria FONE XXV (2010), Sección III, 75–82

# Utilización de procesos de ramificación para el estudio del desarrollo de una epidemia

Luis Cruz-Kuri[a], Agustín Jaime García Banda, Ismael Sosa Galindo
*Universidad Veracruzana*

## 1. Introducción

Se considerarán algunos modelos estocásticos, tales como los procesos de *Galton-Watson*, tanto para un solo tipo de estado individual como para muchos tipos. Estos procesos, aunque permiten estudiar el desarrollo de una epidemia, tienen ciertas limitaciones, las cuales se pueden evitar mediante la utilización de modelos más generales, tales como los *procesos de ramificación*. Se dará una breve descripción de los procesos de referencia y se presentarán algunos comentarios sobre resultados de aproximación del desarrollo de una epidemia por medio de los procesos de ramificación.

## 2. Modelos epidemiológicos básicos de un solo tipo: SIR

Se supone que la población es cerrada y de tamaño N; así que no se consideran nacimientos, defunciones y migraciones en el modelo. Inicialmente se supone que hay m *individuos infecciosos* y N-m *susceptibles*. Se asume que la población es homogénea y que sus elementos se mezclan de manera aleatoria; es decir, la probabilidad de que exista contacto entre dos individuos no depende de quienes de estos se consideren y además se supone que todos los integrantes tienen las mismas características. En cada contacto entre un individuo infeccioso y uno susceptible se trasmite la enfermedad contagiosa. Cada integrante se pone en contacto con un individuo dado, en instantes de tiempo que obedecen un proceso de Poisson con

[a]kruz1111@yahoo.com.mx

parámetro $\beta$/N. A $\beta$ se le puede llamar la *tasa de infección* o *de contacto*. Se supone que la *tasa de recuperación* para los individuos infecciosos es una constante $\alpha$. Un elemento se dice que pertenece a la categoría R (de los casos removidos) si su periodo como transmisor del contagio ha terminado. Finalmente, todos los miembros de la población se distribuyen de manera independiente e idéntica en lo correspondiente a los periodos en que son transmisores de la enfermedad y la variable aleatoria asociada tiene un valor esperado $\iota$ con varianza finita $\sigma^2$ . El proceso así descrito se denota con $E_{N-m,m}(\beta, \iota)$. Si se hace referencia a la palabra "tasa" en un marco estocástico se entiende a la densidad de un proceso unidimensional de Poisson. Se asume que la tasa de recuperación para los individuos infecciosos es una constante $\alpha$. Por consideraciones de espacio se omiten detalles.

# 3. Procesos de Galton-Watson

Históricamente, esta familia de procesos estocásticos surge de consideraciones como las que siguen. Sean, $p_0, p_1, p_2, ...$, las respectivas probabilidades de que un hombre tenga 0, 1, 2, 3, ... hijos. Sea cada uno de ellos con probabilidades iguales de tener hijos propios, y así sucesivamente. ¿Cuál es la probabilidad de que la línea de los varones se extinga después de R generaciones?, y más generalmente ¿cuál es la probabilidad para cualquier número dado de descendientes en la línea de varones en cualquier generación dada? El proceso original de Galton-Watson y sus generalizaciones tienen conexiones con trabajos que se remontan a Niels Abel sobre ecuaciones funcionales y la iteracción de funciones, y de varias líneas de desarrollo de los procesos estocásticos. Por ejemplo, existe una interesante conexión entre el modelo de Galton-Watson y los así llamados *procesos de nacimiento y muerte*, introducidos en una forma especial por Yule (1954) en un estudio de la tasa de formación de nuevas especies. Estas estructuras, en lugar de los animales individuales, son los objetos que se multiplican. Hay también conexiones con la teoría de la radiación cósmica. Estos problemas biológicos y físicos han requerido tratamientos por modelos matemáticos más elaborados que los del proceso de Galton-Watson, lo cual es tema del presente trabajo. Con pocas excepciones se tratarán únicamente modelos estocásticos en los cuales se supone que los distintos objetos se reproducen de manera independiente unos de otros. Esta es una limitación severa para cualquier aplicación a problemas biológicos, aunque existen situaciones, las cuales se señalarán, donde la suposición de independencia parece razonable. Para muchos fenómenos

de interés en la física, este supuesto puede considerarse realista aunque, naturalmente los modelos son siempre imperfectos en otras formas.

## 3.1. Definición del proceso Galton-Watson

Considérense objetos que pueden generar elementos adicionales de la misma naturaleza; ellos pueden ser hombres o bacterias o virus que se reproducen por métodos biológicos familiares, o neutrones de una reacción en cadena. Un conjunto inicial de objetos, al cual se le denominará la *generación 0*, tienen hijos que son llamados de la *primera generación*; sus hijos constituyen la *segunda generación*, y así sucesivamente. El proceso es afectado por eventos aleatorios. En este trabajo se escoge la descripción matemática posible de tal situación, correspondiente al modelo de Galton-Watson. Ante todo se lleva un registro solamente de los tamaños de las generaciones sucesivas, y no de los tiempos en los que los objetos individuales nacen o de los cuáles son sus relaciones familiares entre los individuos. Se denotará con $Z_0, Z_1, Z_2$, ... a los números de individuos, de la generación 0, de la primera, de la segunda, etc., respectivamente (algunas veces podemos interpretar a $Z_0, Z_1$, ... como los tamaños de una población en una sucesión de puntos en el tiempo).

## 3.2. Funciones generadoras

Se hará repetido uso de la función generadora de probabilidades f(s) dada por

$$f(s) = \Sigma p_k s^k \text{ , donde s es una variable compleja.}$$

Las iteraciones de f(s) se definirán por medio de

$$f_0(s) = s, f_1(s) = f(s) \text{ y para n = 2, 3, ..., } f_n(s) = f[f_{n-1}(s)])$$

Se puede verificar que cada una de las iteraciones constituye una función generadora de probabilidades, y que las relaciones que siguen son consecuencia de las ecuaciones anteriores.

$$f_{m+n}(s) = f_m[f_n(s)](m, n = 0, 1, ...) \text{ y en particular, } f_{n+1}(s) = f_n[f(s)].$$

## 3.3. Momentos de Zn

Sean $m = EZ_1$ y $\sigma^2 = \text{varianza}(Z_1) = EZ_1^2 - m^2$. Una parte del resultado que sigue refleja el carácter multiplicativo del proceso. La demostración puede verse, e.g., en Harris (1989).

**Teorema 1**. El valor esperado $EZ_n$ es $m^n$, n = 0, 1, ... Si $\sigma^2$ representa la varianza de $Z_1 < \infty$, entonces la varianza de $Z_n$ está dada por las fórmulas que usan $\sigma^2$ y m como sigue.
$Var(Z_n) = n\sigma^2$ si m = 1 y $Var(Z_n) = m^{n-1}\frac{m^n-1}{m-1}\sigma^2$ si $m \neq 1$.

## 3.4. La función generadora y la probabilidad de extinción

Para una variable aleatoria discreta R se define su función generadora $f_R(s)$ por medio de

$$f_R(s) := \text{E}(s^R) = \Sigma\ \text{P(R= k)}s^k$$

donde la suma recorre los índices k desde 0 en adelante. Se puede asegurar que, si s≤1, entonces la serie es convergente. Con la función generadora se puede establecer una ecuación funcional que permite calcular la probabilidad de extinción del proceso (de la epidemia en el caso que aquí se trata). Por consideraciones de espacio, no se presentan los detalles, aunque es pertinente mencionar que el tratamiento es estándar. A manera de ejemplo, supóngase que la variable aleatoria R solo puede tomar los valores 0, 1, 2 y 3 con probabilidades 0.1, 0.4, 0.3 y 0.2, respectivamente; entonces la función generadora de R está dada por la relación f(s) = 0.1 + 0.4s + $0.3s^2$ + $0.2s^3$ . Para tal proceso de ramificación, puede mostrarse que la probabilidad de extinción a la larga, está dada por la solución de la ecuación f(s) = s . En el intervalo [0,1] la función f es convexa y hay una solución distinta de 1 siempre que E(R) > 1; en caso contrario, la única solución está dada por s=1. Para la presente ilustración, E(R)= 0×0.1+1×0.4+2×0.3+3×0.2 = 1.6. Así que, la probabilidad de extinción es < 1 y está dada por la raíz≠1 de la ecuación 0.1+ 0.4s + $0.3s^2$ + $0.2s^3$ = s. Por lo tanto, la probabilidad de extinción, al resolver la ecuación cuadrática correspondiente, es igual aproximadamente a 0.4 . (*Interpretación*: De cada 100 veces que se inicie el proceso con un ancestro, habrá 40 ocasiones en que su descendencia tarde o temprano quedará extinta y 60 ocasiones en que crecerá sin límite, i.e., "explotará".)

### 3.5. Procesos de Galton -Watson con varios tipos

En cascadas de rayos cósmicos, los electrones producen *fotones* y estos a su vez producen *electrones*. La reproducción de *bacterias* puede producir una *forma mutante* que se comporta de manera diferente. Los objetos están caracterizados por una variable continua, tal como posición o *edad*. Como una aproximación pueden formarse grupos [o tipos]. El proceso en tiempo discreto está dado por la sucesión $Z_0, Z_1, Z_2$, ... de vectores aleatorios donde $Z_n$ = vector del número de objetos de n-ésima generación de tipo i (i = 1, ..., k).

## 4. Procesos de ramificación

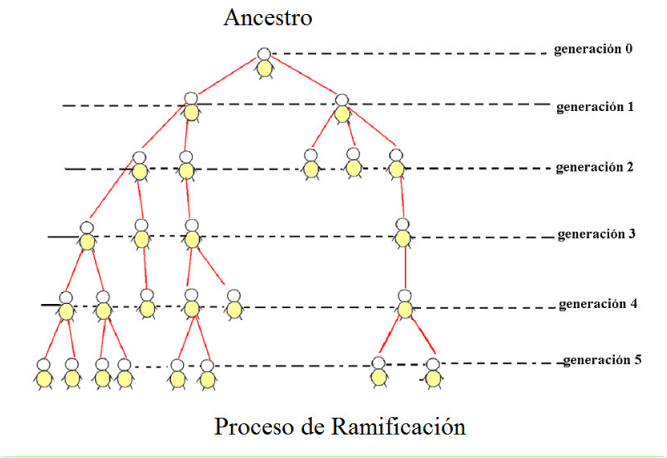



**Figura 1:** Generaciones del proceso.

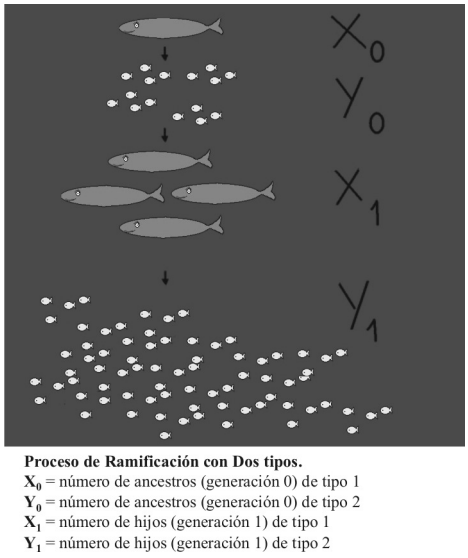



**Figura 2:** Proceso de ramificación con dos tipos.

### 4.1. Un proceso de ramificación simple en tiempo discreto

Una representación esquemática de un proceso de ramificación se presenta en la figura 3 que sigue. Los números dentro de los círculos indican la generación de la célula o el número de divisiones que ha sufrido. Se empieza con una población de células sin división en el tiempo 0. En cada unidad de tiempo, cada organismo se divide con probabilidad $\gamma$, sobrevive sin dividirse con probabilidad $\delta$ y muere con probabilidad $\alpha = 1 - \gamma - \delta$. En un tiempo posterior

$t$, reordenando a las células de acuerdo a su contenido CFSE (marcador fluorescente *Carboxy Fluorescein Succinimidyl Ester* para seguir la división celular) permite que el número de células en cada generación sea estimado. La metodología de las funciones generadoras permite calcular los momentos de la distribución de probabilidad de tales recuentos paso a paso en el tiempo dada la información sobre los números de células en cada generación, para etapas anteriores. La figura 2 arriba ilustra el caso de un proceso de ramificación con dos tipos de individuos (peces en este caso).

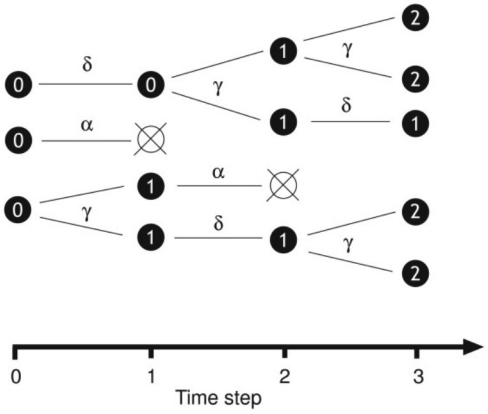



**Figura 3:** Un proceso de ramificación simple que indica las divisiones de una célula.

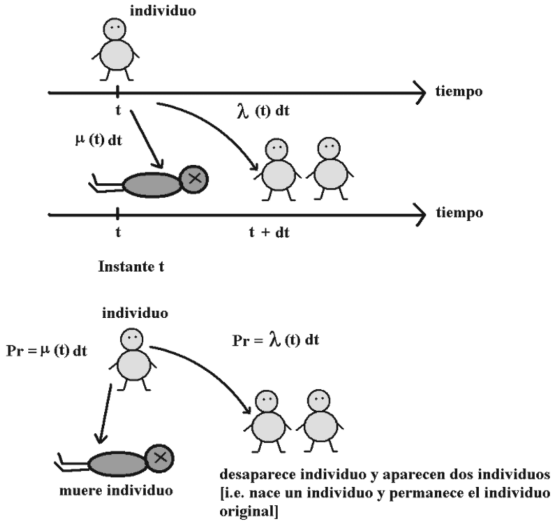



**Figura 4:** Proceso de nacimiento y muerte en tiempo continuo.

## 4.2. Procesos de nacimiento y muerte

El proceso markoviano de ramificación más familiar es el *proceso de nacimiento y muerte*, en el que se supone que cualquier objeto que existe en el instante t tiene una probabilidad $\mu$(t) dt de morir en el intervalo de tiempo (t, t + dt) y una probabilidad $\lambda$(t) dt de desaparecer y ser reemplazado por dos nuevos objetos durante el mismo periodo (este último evento puede pensarse como equivalente a generar el nacimiento de un objeto adicional). Se tienen las relaciones que siguen.

$p_i$(t) = 0 (para i = 1, o para i > 2)
$p_0$(t) = $\mu$(t) / [$\lambda$(t) + $\mu$(t)]
$p_2$(t) = $\lambda$(t) / [$\lambda$(t) + $\mu$(t)]

## 4.3. Modelación de una epidemia con un proceso de nacimiento y muerte

El proceso de nacimiento y muerte no parece ser un buen modelo para la propagación de una epidemia en una población finita, ya que cuando ha sido infectada una proporción grande de la población, no se puede suponer que la tasa de nuevas infecciones es independiente de la historia pasada. Sin embargo, para etapas tempranas de una epidemia se puede usar el proceso estocástico anterior y un modelo determinista para representar fases posteriores.

## 4.4. Proceso general de ramificación

En los procesos de ramificación de mayor interés en las ciencias físicas y biológicas, un objeto queda caracterizado por un parámetro x, el cual describe su posición, edad, energía, o una combinación de estos factores. Sea X el conjunto de tipos posibles de los objetos. Se tiene entonces una distribución puntual $\omega$ sobre X . Aquí, $\omega = (x_1, n_1 \mid x_2, n_2 \mid ... \mid x_k, n_k)$ donde $\omega$ indica que se tienen $n_1$ objetos de tipo $x_1, n_2$ de tipo $x_2$, ..., y $n_k$ de tipo $x_k$. La relación entre la epidemia para una población grande, cuyos individuos se mezclan de manera aleatoria, y los árboles genealógicos, puede explicarse en forma intuitiva como sigue. Los individuos que introducen una infección en una población se ven ahora como los "ancestros". Durante sus periodos como contagiosos, estos se ponen en contacto con un número aleatorio de individuos escogidos de manera uniforme en la población e infectan a esos miembros (si estos todavía son susceptibles). Los individuos recientemente contagiados pueden verse como los individuos de la primera generación. Durante su periodo como infecciosos, los individuos de la primera generación se ponen en contacto con otros integrantes de acuerdo a la misma ley que siguieron los ancestros. Si la población es muy grande, la probabilidad de que un individuo infeccioso se ponga en contacto con un participante que ya ha sido infectado durante la primera etapa de la epidemia (la cual puede tener una duración bastante grande) es muy pequeña, y, por consiguiente, el progreso de la propagación del contagio puede describirse por medio de un

proceso de ramificación.

# 5. Comentarios adicionales

La duración de una epidemia no puede ser estudiada de manera realista mediante un modelo determinista porque al principio y al final del proceso de contagio las fluctuaciones aleatorias juegan un papel muy importante. Para algunas enfermedades infecciosas no es suficiente utilizar tres estados de salud (susceptible, infeccioso, y removido) sino que también se requiere tomar un estado de latencia o más de un estado de contagio en consideración. La función generadora tiene propiedades útiles, dentro de las cuales se encuentran la obtención de los momentos de la variable aleatoria mediante el cálculo de las derivadas de $f_R(\mathrm{s})$, de los órdenes correspondientes a los momentos requeridos. Asimismo, con la función generadora se puede establecer una ecuación funcional que permite calcular la probabilidad de extinción del proceso (de la epidemia para el caso aquí discutido). Por consideraciones de espacio, no se presentan los detalles, aunque es pertinente señalar que el tratamiento es estándar.

# Bibliografía

Harris, T. E. (1989), *The Theory of Branching Processes*, Dover Books.

Kuri, L. C. y et al (2011), "Algunos modelos epidemiológicos estocásticos", *Memorias del Tercer Congreso Internacional de Biometría* pp. 98–102.

Trapman, J. P. (2006), *On stochastic models for the spread of infections*, PrintPartners Ipskamp, Enschede.

# Sección IV

# Muestreo

Memoria FONE XXV (2010), Sección IV, 85–90

# Estimación de totales y medias en el muestreo por bola de nieve en presencia de probabilidades de nominación heterogéneas*

Martín H. Félix Medina[a]
*Escuela de Ciencias Físico-Matemáticas de la Universidad Autónoma de Sinaloa*

## 1. Introducción

El Muestreo por Bola de Nieve es un método que se ha propuesto para muestrear poblaciones de difícil detección, tales como poblaciones de drogadictos, niños de la calle, trabajadoras sexuales, etc. En este método se selecciona una muestra inicial de miembros de la población de interés, y se les pide a las personas que fueron seleccionadas que nominen a otros miembros de la población objetivo. A las personas que fueron nominadas se les puede pedir que nominen a otras personas, y el proceso de nominación puede continuar de esta manera hasta que se satisfaga alguna regla de terminación del muestreo.

Félix-Medina y Thompson (2004) desarrollaron una variante del muestreo por bola de nieve y propusieron estimadores máximo verosímiles (EMVs) del tamaño poblacional derivados bajo el supuesto de que las probabilidades de nominación no dependen de los individuos nominados, es decir, que son homogéneas. Posteriormente, Félix-Medina *et al.* (2009) debilitaron el supuesto de homogeneidad y desarrollaron EMVs del tamaño poblacional bajo el supuesto de probabilidades de nominación heterogéneas. En este trabajo se considera el problema de estimar el total y la media de una variable respuesta, tal como gasto mensual en drogas, edad de inicio de consumo de drogas o ingreso mensual. Se proponen estimadores de momentos de estos parámetros, los cuales se obtienen bajo el supuesto de que la variable respuesta no está asociada con los valores de las probabilidades de nominación.

*Trabajo realizado con apoyo del proyecto PIFI-2008-25-08 asignado a la UAS por la SEP.
[a]mhfelix@uas.uasnet.mx

## 2. Diseño muestral, notación y modelos probabilísticos

Al igual que en Félix-Medina y Thompson (2004), supondremos que una parte $U_1$ de la población de interés $U$ está cubierta por un marco muestral de $N$ sitios $A_1, \ldots, A_N$, tales como parques, hospitales o cruceros de calles. De este marco se selecciona una muestra aleatoria simple sin reemplazo de $n$ sitios, que denotaremos por conveniencia como $S_A = \{A_1, \ldots, A_n\}$, y que no significa que la muestra esté compuesta por los primeros $n$ sitios del marco muestral. A las personas de la población de interés que pertenecen, de acuerdo con algún criterio, a cada uno de los sitios seleccionados se les pide que nominen a otros miembros de la población. Como convención, diremos que una persona es nominada por un sitio si cualquiera de los miembros de ese sitio la nomina.

Denotaremos por $\tau$ el tamaño de $U$, por $\tau_1$ el de $U_1$, por $\tau_2 = \tau - \tau_1$ el de $U_2 = U - U_1$, y por $M_i$ el número de personas en $A_i$. Obsérvese que $\tau_1 = \sum_{i=1}^{N} M_i$ y que $M = \sum_{i=1}^{n} M_i$ es el número de individuos en $S_0 = \{\text{individuos en sitios } A_i \in S_A\}$. Los conjuntos de variables $\{X_{ij}^{(1)}\}$ y $\{X_{ij}^{(2)}\}$ indicarán el proceso de nominación. Así, $X_{ij}^{(1)} = 1$ si la persona $u_j \in U_1 - A_i$ es nominada por el sitio $A_i$, y $X_{ij}^{(1)} = 0$ en otro caso. Similarmente, $X_{ij}^{(2)} = 1$ si la persona $u_j \in U_2$ es nominada por el sitio $A_i$, y $X_{ij}^{(2)} = 0$ en otro caso.

Como en Félix-Medina *et al.* (2009), supondremos que las $M_i$'s son variables aleatorias independientes con distribución Poisson, y que por carecer de información adicional supondremos que tienen media común $\lambda_1$. Por tanto, dado que $\sum_1^N M_i = \tau_1$, la distribución condicional conjunta de $(M_1, \ldots, M_n, \tau_1 - M)$ es multinomial con parámetro de tamaño $\tau_1$ y vector de probabilidades $(1/N, \ldots, 1/N, 1 - n/N)$. Asimismo, supondremos que dado $M_i$, la distribución condicional de $X_{ij}^{(k)}$ es Bernoulli con probabilidad $p_{ij}^{(k)} = \Pr[X_{ij}^{(k)} = 1 | M_i] = \exp(\alpha_i^{(k)} + \beta_j^{(k)})/[1 + \exp(\alpha_i^{(k)} + \beta_j^{(k)})]$, $i = 1, \ldots, n$; $j = 1, \ldots, \tau_k$, con $u_j \notin A_i$, y $k = 1, 2$. Este modelo se conoce como modelo de Rasch. El parámetro $\alpha_i^{(k)}$ es el efecto del potencial que tiene el sitio $A_i$ de nominar individuos en $U_k$ y $\beta_j^{(k)}$ es el efecto de la suceptibilidad que tiene el individuo $u_j \in U_k - A_i$ de ser nominado. Los efectos $\beta_j^{(k)}$'s se suponen aleatorios con distribución $N(0, \sigma_k^2)$, esto es normal con media cero y varianza $\sigma_k^2$ desconocida.

De los supuestos anteriores se sigue que la probabilidad de que un individuo en $U_k - S_0$ seleccionado al azar sea nominado sólo por los sitios $A_i$'s con $i \in \omega \subseteq \Omega = \{1, \ldots, n\}$ es $\pi_\omega^{(k)}(\sigma_k, \alpha^{(k)}) = \int \prod_{i=1}^{n} \{\exp[x_{\omega i}(\alpha_i^{(k)} + \sigma_k z)]/[1 + \exp(\alpha_i^{(k)} + \sigma_k z)]\} \phi(z) dz$, donde $x_{\omega i} = 1$ si $i \in \omega$, y $x_{\omega i} = 0$ en otro caso, $\alpha^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$ y $\phi(z)$ representa la función de densidad

normal estándar.

## 3. Estimadores máximo verosímiles de los tamaños poblacionales

Félix-Medina *et al.* (2009) propusieron estimar los parámetros $\alpha^{(k)}$ y $\sigma_k$ mediante EMVs condicionales $\hat{\alpha}^{(k)}$ y $\hat{\sigma}_k$ dado el número $R_k$ de individuos en $U_k - S_0$ que fueron nominados, $k = 1, 2$. Los valores de estos estimadores los obtienen maximizando numéricamente la correspondiente función de verosimilitud condicional. Asimismo, sugieren los siguientes EMVs condicionales de $\tau_1$ y $\tau_2$: $\hat{\tau}_1 = (M + R_1)/[1 - (1 - n/N)\hat{\pi}_{\emptyset}^{(1)}(\hat{\sigma}_1, \hat{\alpha}^{(1)})]$ y $\hat{\tau}_2 = R_2/[1 - \hat{\pi}_{\emptyset}^{(2)}(\hat{\sigma}_2, \hat{\alpha}^{(2)})]$, donde $\hat{\pi}_{\emptyset}^{(k)}(\hat{\sigma}_k, \hat{\alpha}^{(k)})$ es un estimador de la probabilidad $\pi_{\emptyset}^{(k)}(\sigma_k, \alpha^{(k)})$ de que un individuo de $U_k - S_0$, seleccionado aleatoriamente, no sea nominado por alguno de los sitios $A_i \in S_A$. Finalmente, proponen estimar $\tau$ mediante $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

## 4. Estimadores de momentos del total y la media poblacional

Supóngase que se desea estimar el total y/o la media de una variable respuesta $y$, tal como gastos médicos mensuales, gastos mensuales en drogas o edad de inicio de consumo de drogas. Sea $y_{kj}$ la variable respuesta asociada con el $j$-ésimo elemento de $U_k$, $j = 1, \ldots, \tau_k$, $k = 1, 2$. Luego, $Y_k = \sum_{j=1}^{\tau_k} y_{kj}$ y $Y = Y_1 + Y_2$ representan el total de $y$ en $U_k$, $k = 1, 2$, y en $U$, respectivamente.

Supóngase que $E(y_{kj}) = \mu_k$, $j = 1, \ldots, \tau_k$, $k = 1, 2$, y que las variables $y_{kj}$ y las probabilidades de nominación $p_{ij}^{(k)}$ son independientes. Note que esto implica que las variables $y_{kj}$ son independientes de diseño muestral. Sea $Y_k^* = E(Y_k) = \tau_k \mu_k$. Asimismo, sea $S_k$ el conjunto de elementos de $U_k - S_0$ que fueron muestreados, $k = 1, 2$. Luego, para los elementos de $U_1$ que fueron muestreados, esto es, en $S_0 \cup S_1$, se tiene que $E\left(\sum_{j \in S_0 \cup S_1} y_{1j}\right) = Y_k^*[1 - (1 - n/N)\pi_{\emptyset}^{(1)}(\sigma_1, \alpha^{(1)})]$. Por tanto $\hat{Y}_1 = \sum_{j \in S_0 \cup S_1} y_{1j}/[1 - (1 - n/N)\hat{\pi}_{\emptyset}^{(1)}(\hat{\sigma}_1, \hat{\alpha}^{(1)})]$ es un estimador de momentos de $Y_1$.

Un razonamiento similar al anterior conduce al siguiente estimador de momentos de $Y_2$: $\hat{Y}_2 = \sum_{j \in S_2} y_{2j}/[1 - \hat{\pi}_{\emptyset}^{(2)}(\hat{\sigma}_2, \hat{\alpha}^{(2)})]$. Luego, un estimador del total $Y$ es $\hat{Y} = \hat{Y}_1 + \hat{Y}_2$.

# 5. Estudio Monte Carlo

Se construyeron cuatro poblaciones de $N = 250$ valores de $M_i$'s. En Poblaciones 1 y 2 los valores se generaron mediante una distribución Poisson con media 7.2, mientras que en Poblaciones 3 y 4 mediante una distribución Binomial negativa con media 7.2 y varianza 21.1. Las probabilidades de nominación $p_{ij}^{(k)}$ se obtuvieron mediante el modelo Rasch descrito en la Sección 2 con $\beta_j^{(k)}$'s generadas mediante una distribución $N(0, 0.66)$ y $\alpha_i^{(k)} = c_k/(M_i^{1/4} + 0.001)$, donde $c_1 = -6.2$ y $c_2 = -6.7$. En las Poblaciones 1 y 3 los valores $y_{kj}$ de la variable respuesta se generaron mediante una distribución ji-cuadrada ($\chi^2$) con 2 grados de libertad y parámetros de no centralidad $\theta_{kj}$, donde $\theta_{1j} = 5 + 30\exp(\beta_j^{(1)})/[1 + \exp(\beta_j^{(1)})]$ y $\theta_{2j} = 5 + 20\exp(\beta_j^{(2)})/[1 + \exp(\beta_j^{(2)})]$, mientras que en Poblaciones 3 y 4 se generaron mediante distribuciones Bernoulli con medias $p_{kj}$, donde $p_{1j} = 0.4\exp(\beta_j^{(1)})/[1 + \exp(\beta_j^{(1)})]$ y $p_{2j} = 0.4\exp(\beta_j^{(2)})/[1 + \exp(\beta_j^{(2)})]$. Note que por la manera en que se generaron las $y_{kj}$'s y las $p_{ij}^{(k)}$'s estos dos conjuntos de variables no son independientes.

El estudio de simulación se realizó seleccionando 1000 muestras de cada una de las poblaciones mediante el diseño propuesto por Félix-Medina y Thompson (2004) y tamaño de muestra inicial $n = 25$. Los estimadores que se consideraron fueron los estimadores $\hat{Y}_k$, $\hat{Y}$, $\hat{\tilde{Y}}_k$ y $\hat{\tilde{Y}}$ propuestos en este trabajo, y los dos conjuntos de estimadores $\{\tilde{Y}_k, \tilde{Y}, \tilde{\tilde{Y}}_k, \tilde{\tilde{Y}}\}$ y $\{\check{Y}_k, \check{Y}, \check{\tilde{Y}}_k, \check{\tilde{Y}}\}$ propuestos por Félix-Medina y Monjardin (2010), derivados bajo el supuesto de probabilidades de nominación homogéneas y basados en los EMV's, propuestos por Félix-Medina y Thompson (2004), y en los estimadores bayesianos, propuestos por Félix-Medina y Monjardin (2006), de los tamaños poblacionales, respectivamente. Los resultados del estudio se presentan en la Tabla 1.

# 6. Conclusiones

Con base en los resultados del estudio de simulación, podemos concluir que si se viola el supuesto de independencia entre la variable respuesta y las probabilidades de nominación, bajo el cual se derivaron los estimadores propuestos, se presentan problemas de sesgo en estos estimadores. Las magnitudes de los sesgos dependen de varios factores. Uno de ellos es el grado de asociación de la variable respuesta y las probabilidades de nominación. Si este es grande, como en el caso de los resultados mostrados, los sesgos pueden ser considerables, mientras que si este es pequeño, los sesgos no son serios (estos resultados no se presentan).

Otro factor es el valor del coeficiente de variación de la variable respuesta. Si este es grande, como en el caso de la distribución Bernoulli considerada en este estudio, los sesgos pueden ser grandes, mientras que si este es pequeño, como en el caso de la distribución Poisson de este estudio, los sesgos no son tan serios. Con respecto a los desempeños de los estimadores derivados bajo el supuesto de probabilidades de nominación homogéneas, los resultados de este estudio indican que estos estimadores tienen en general problemas de sesgo que van de fuertes a moderados.

De los resultados de este estudio no es difícil darse cuenta que un problema de investigación futura es el desarrollo de estimadores de totales y medias poblacionales que tomen en cuenta la asociación entre la variable respuesta y las probabilidades de nominación.

# Bibliografía

Félix-Medina, M. y Thompson, S. (2004), "Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations", *Journal of Official Statistics* **20**, 19–38.

Félix-Medina, M. H., Monjardin, P. E. y Aceves-Castro, A. N. (2009), Link-tracing sampling: estimating the size of a hidden population in presence of heterogeneous nomination probabilities, *in* "Proceedings of the Section on Survey Research Methods of the American Statistical Association", pp. 4020–4033.

Félix-Medina, M. y Monjardin, P. (2006), "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: a bayesian assisted approach", *Survey Methodology* **32**, 187–195.

Félix-Medina, M. y Monjardin, P. (2010), "Combining link-tracing sampling and cluster sampling to estimate totals and means of hidden human populations", *Journal of Official Statistics* **26**, 603–631.

Tabla 1. Sesgos relativos y raíces cuadradas de errores cuadráticos medios relativos de estimadores de totales y medias. Resultados basados en 1000 muestras.

| Esti- | Población 1 Sesgo | | Población 3 Sesgo | | Esti- | Población 1 Sesgo | | Población 3 Sesgo | |
|---|---|---|---|---|---|---|---|---|---|
| mador | rel. | $\sqrt{\text{ecm-}}$ rel. | rel. | $\sqrt{\text{ecm-}}$ rel. | mador | rel. | $\sqrt{\text{ecm-}}$ rel. | rel. | $\sqrt{\text{ecm-}}$ rel. |
| $\hat{Y}_1$ | 0.07 | 0.09 | 0.06 | 0.08 | $\hat{\bar{Y}}_1$ | 0.06 | 0.06 | 0.07 | 0.07 |
| $\hat{Y}_2$ | 0.07 | 0.22 | -0.05 | 0.17 | $\hat{\bar{Y}}_2$ | 0.07 | 0.07 | 0.08 | 0.08 |
| $\hat{Y}$ | 0.07 | 0.10 | 0.04 | 0.07 | $\hat{\bar{Y}}$ | 0.07 | 0.07 | 0.08 | 0.08 |
| $\tilde{Y}_1$ | -0.12 | 0.12 | -0.10 | 0.11 | $\tilde{\bar{Y}}_1$ | 0.06 | 0.06 | 0.07 | 0.07 |
| $\tilde{Y}_2$ | -0.21 | 0.22 | -0.20 | 0.21 | $\tilde{\bar{Y}}_2$ | 0.07 | 0.07 | 0.08 | 0.08 |
| $\tilde{Y}$ | -0.14 | 0.14 | -0.13 | 0.13 | $\tilde{\bar{Y}}$ | 0.07 | 0.07 | 0.08 | 0.08 |
| $\check{Y}_1$ | -0.12 | 0.12 | -0.10 | 0.11 | $\check{\bar{Y}}_1$ | 0.06 | 0.06 | 0.07 | 0.07 |
| $\check{Y}_2$ | -0.21 | 0.22 | -0.19 | 0.20 | $\check{\bar{Y}}_2$ | 0.07 | 0.07 | 0.08 | 0.08 |
| $\check{Y}$ | -0.14 | 0.14 | -0.13 | 0.13 | $\check{\bar{Y}}$ | 0.07 | 0.07 | 0.08 | 0.08 |

| Esti- | Población 2 Sesgo | | Población 4 Sesgo | | Esti- | Población 2 Sesgo | | Población 4 Sesgo | |
|---|---|---|---|---|---|---|---|---|---|
| mador | rel. | $\sqrt{\text{ecm-}}$ rel. | rel. | $\sqrt{\text{ecm-}}$ rel. | mador | rel. | $\sqrt{\text{ecm-}}$ rel. | rel. | $\sqrt{\text{ecm-}}$ rel. |
| $\hat{Y}_1$ | 0.11 | 0.13 | 0.11 | 0.13 | $\hat{\bar{Y}}_1$ | 0.10 | 0.11 | 0.12 | 0.13 |
| $\hat{Y}_2$ | 0.21 | 0.33 | 0.03 | 0.21 | $\hat{\bar{Y}}_2$ | 0.21 | 0.24 | 0.17 | 0.20 |
| $\hat{Y}$ | 0.13 | 0.15 | 0.09 | 0.12 | $\hat{\bar{Y}}$ | 0.12 | 0.13 | 0.14 | 0.14 |
| $\tilde{Y}_1$ | -0.09 | 0.10 | -0.06 | 0.08 | $\tilde{\bar{Y}}_1$ | 0.10 | 0.11 | 0.12 | 0.13 |
| $\tilde{Y}_2$ | -0.11 | 0.15 | -0.13 | 0.17 | $\tilde{\bar{Y}}_2$ | 0.21 | 0.24 | 0.17 | 0.20 |
| $\tilde{Y}$ | -0.09 | 0.10 | -0.08 | 0.09 | $\tilde{\bar{Y}}$ | 0.13 | 0.14 | 0.14 | 0.14 |
| $\check{Y}_1$ | -0.09 | 0.10 | -0.06 | 0.08 | $\check{\bar{Y}}_1$ | 0.10 | 0.11 | 0.12 | 0.13 |
| $\check{Y}_2$ | -0.10 | 0.15 | -0.12 | 0.17 | $\check{\bar{Y}}_2$ | 0.21 | 0.23 | 0.16 | 0.20 |
| $\check{Y}$ | -0.09 | 0.10 | -0.08 | 0.09 | $\check{\bar{Y}}$ | 0.13 | 0.13 | 0.14 | 0.14 |

Notas: Poblaciones 1 y 3: $M_i$'s con distribución Poisson; Poblaciones 2 y 4: $M_i$'s con distribución Binomial negativa; Poblaciones 1 y 2: $y_{kj}$'s con distribución $\chi^2$; Poblaciones 3 y 4: $y_{kj}$'s con distribución Bernoulli.

Memoria FONE XXV (2010), Sección IV, 91–97

# Cotas para la varianza, efecto del diseño y coeficiente de variación de proporciones en el muestreo por conglomerados en dos etapas con tamaños iguales

Alberto Manuel Padilla Terán[a]
*Banco de México*

## 1. Introducción

En el cálculo del tamaño de muestra se usa con frecuencia la fórmula asociada al muestreo aleatorio simple, $MAS$, y, posteriormente, ésta se ajusta por el efecto del diseño, efd, Kish (1965). En el caso del cálculo del tamaño de muestra para la estimación de proporciones y cuando no se cuenta con información de la característica de interés, puede emplearse el valor máximo de la varianza para el estimador de proporciones bajo $MAS$, el cual se alcanza cuando la proporción poblacional adquiere el valor de 0.5, Cochran (1986). Después se aplica un ajuste usando el efecto del diseño, $n = n_{mas} efd(\hat{p})$. En esta expresión $n_{mas} = N'pq/(d^2 + pq/(N-1))$, en la que $N' = N/(N-1)$ y $d^2 = (e_a/t)^2$, donde $e_a$ se refiere al error de estimación absoluto, $t$ al desvío normal y $q = 1 - p$ . Esto conduce al tamaño de muestra más grande para una población, error de estimación absoluto y nivel de confianza dados.

En la literatura se han propuesto algunas cotas para los diseños con elección de elementos o conglomerados de primera etapa con probabilidad proporcional al tamaño, Scott & Smith (1975) y Chauduri & Stenger (2005). Estos esquemas han sido estudiados en la literatura; empero, no proporcionan en general resultados que puedan ser empleados con relativa faci-

[a]ampadilla@banxico.org.mx

lidad en la práctica[1].

El artículo se encuentra organizado de la siguiente manera. En la sección 2 se proporcionan las definiciones, notación y la expresión de varianza para el muestreo bietápico. Las cotas para la varianza, efecto del diseño y coeficiente de variación, junto con varios ejemplos, se encuentran en la sección 3[2].

## 2. Marco teórico

Se supondrá que se trabaja con el enfoque del muestreo probabilístico o del diseño para el caso del muestreo por conglomerados en 2 etapas. Sea $U$ una población finita de $N$ elementos etiquetados como $k = 1, \ldots, N$, $1 < N$. Es usual representar a la población finita por sus etiquetas $k$ como $U = 1, 2, \ldots, k, \ldots, N$.

Los conglomerados se denotarán como $UPM$, unidades primarias de muestreo y a los elementos dentro de conglomerados como $USM$, unidades secundarias de muestreo. $A$ y $B$ representarán al número de $UPM$ en la población y al número de $USM$ dentro de cada $UPM$ respectivamente; en tanto que $a$ y $b$ representarán las respectivas cantidades muestrales. Se supondrá que $A$, $B$, $a$ y $b$ son mayores que uno, $a < A$ y $b < B$. El total de elementos en población y muestra se denotarán como $N = AB$ y $n = ab$, respectivamente. La variable bajo estudio es dicotómica y se representará con $y_{ij}$, en donde $i$ se refiere a la $UPM$ y $j$ a la $USM$. Dicha variable adquirirá el valor de 1 si el j-ésimo elemento de la i-ésima $UPM$ posee la característica de interés y 0 en otro caso. Se trabajará con las proporciones de la i-ésima $UPM$, $p_i = \sum_{j=1}^{B} y_{ij}/B$ y la proporción poblacional $p_U = \sum_{i=1}^{A} p_i/A$.

La varianza del estimador de proporciones en el muestreo bietápico con muestreo aleatorio simple en ambas etapas, (MAS,MAS), puede expresarse como, Cochran (1986): $V(\hat{p}) = \alpha \left( \sum_{i=1}^{A} p_i^2 - A p_U^2 \right) + \beta \left( \sum_{i=1}^{A} p_i(1 - p_i) \right)$. En la expresión anterior, $\alpha = (1 - a/A)/(a(A-1))$ y $\beta = (1 - b/B)B/(abA(B-1))$. La expresión $V(\hat{p})$ puede escribirse como $V(\hat{p}) = \alpha\, V_1(\hat{p}) +$

[1]Cabe mencionar que los teoremas y lemas del presente documento, no han sido publicados en la literatura revisada a la fecha de elaboración del trabajo.

[2]Para un detalle mayor del presente trabajo véase Padilla (2010)

$\beta\, V_2(\hat{p})$, donde $V_1(\hat{p}) = \sum_{i=1}^{A} p_i^2 - A p_U^2$ y $V_2(\hat{p}) = \sum_{i=1}^{A} p_i(1-p_i)$.

# 3. Cotas

A continuación se muestran dos resultados del muestreo por conglomerados en dos etapas para la varianza de proporciones en los que se aprecia el efecto en la varianza cuando algunas de las proporciones de las $UPM$ son 0 y otras toman el valor de 1, o cuando todas son iguales a alguna proporción $p = c$, con $0 < c < 1$.

**Lema 3.1.** *bajo (MAS,MAS), si $p_i = 0$ ó 1 y existen $i$ y $j$, $i \neq j$, tales que $p_i \neq p_j$, entonces $V(\hat{p}) = \alpha\, V_1(\hat{p})$.*

**Lema 3.2.** *bajo (MAS,MAS), si $p_i = p_U$ con $p_U < 1$, $\forall\, i \in \{1, \ldots, A\}$, entonces $V(\hat{p}) = \beta\, V_2(\hat{p})$.*

**Teorema 3.1.** *bajo (MAS,MAS) si $p_i = 0$ ó 1 y existen $i$ y $j$, $i, j \in \{1, \ldots, A\}$, $A \geq 2$ $i \neq j$, tales que $p_i \neq p_j$, $V(\hat{p}) = \alpha\, V_1(\hat{p})$ se maximiza si:*

- *se tienen $A/2$ valores $p_i = 1$, con $A$ par ó*
- *$[A/2] + 1$ valores $p_i = 1$, con $A$ impar ó $[A/2]$ valores $p_i = 1$. En este caso, el valor de $V_1(\hat{p})$ es el mismo para $[A/2]$ y $[A/2] + 1$.*

En esta expresión, $[x]$ denota la parte entera de $x$.

En el siguiente ejemplo se muestran los valores de $V_1(\hat{p})$ para una población con 5 $UPM$ y 3 valores de submuestreo de las $UPM$, para ilustrar el teorema 1.

**Ejemplo 3.1.** *$A = 5$ y cuatro poblaciones con diferentes valores de $p_U$, en los que $p_i = 0$ ó 1; por ejemplo, $p_U = 2/5 = 0.4$ significa $[5/2] = 2$ valores $p_i = 1$ y 3 valores $p_i = 0$.*

| | $a = 2$ | $a = 3$ | $a = 4$ |
|---|---|---|---|
| | $\alpha = 0.075$ | $\alpha = 0.033$ | $\alpha = 0.013$ |
| $p_U = 1/5 = 0.2$ | *0.060* | *0.027* | *0.010* |
| $p_U = 2/5 = 0.4$ | *0.090* | *0.040* | *0.015* |
| $p_U = 3/5 = 0.6$ | *0.090* | *0.040* | *0.015* |
| $p_U = 4/5 = 0.8$ | *0.060* | *0.027* | *0.010* |

*Tabla 1*

*En la Tabla 1 se aprecia, por ejemplo, que si se tiene una población con* $p_U = 2/5 = 0.4$ *y* $a = 2$*, el valor máximo de* $V_{\hat{p}}$ *es* $0.09$ *y ninguna otra configuración de la población con* $p_U = 0.4$ *y el mismo plan de muestreo puede tener una varianza poblacional mayor.*

En los tres siguientes teoremas se exhiben las cotas mínima y máxima para la varianza del muestreo por conglomerados en dos etapas; así como para el efecto del diseño y el coeficiente de variación de la proporción.

**Teorema 3.2.** *bajo (MAS,MAS),* $A \geq 2$ *y para cualquier arreglo* $(p_1, \ldots, p_A)$ *tal que* $\sum_{i=1}^{A} p_i/A = p_U \in (0, 1)$*, el valor de* $V(\hat{p})$ *satisface alguna de las siguientes desigualdades:*

- *si* $\alpha > \beta$, $\beta\, V_2(\hat{p}) < V(\hat{p}) < \alpha\, V_1(\hat{p})$,
- *si* $\alpha < \beta$, $\alpha\, V_1(\hat{p}) < V(\hat{p}) < \beta\, V_2(\hat{p})$,
- *si* $\alpha = \beta$, $V(\hat{p}) = \alpha\, A p_U(1 - p_U)$

**Teorema 3.3.** *bajo (MAS,MAS),* $A \geq 2$ *y para cualquier arreglo* $(p_1, \ldots, p_A)$ *tal que* $\sum_{i=1}^{A} p_i/A = p_U \in (0, 1)$*, el valor del coeficiente de variación,* $cv(\hat{p}) = \sqrt{V(\hat{p})}/\hat{p}$*, satisface alguna de las siguientes desigualdades:*

- *si* $\alpha > \beta$, $cv_2(\hat{p}) < cv(\hat{p}) < cv_1(\hat{p})$,
- *si* $\alpha < \beta$, $cv_1(\hat{p}) < cv(\hat{p}) < cv_2(\hat{p})$,
- *si* $\alpha = \beta$, $cv(\hat{p}) = \alpha\, \sqrt{(}A(1 - p_U)p_U)$,

*donde* $cv_1(\hat{p}) = \alpha\, \sqrt{A(1 - p_U)p_U}$ *y* $cv_2(\hat{p}) = \beta\, \sqrt{A(1 - p_U)p_U}$.

**Teorema 3.4.** *bajo (MAS,MAS), $A \geq 2$ y para cualquier arreglo $(p_1, \ldots, p_A)$ tal que $\sum_{i=1}^{A} p_i/A = p_U \in (0,1)$, el valor del efecto del diseño, $V(\hat{p})/V_{mas}(\hat{p})$, satisface alguna de las siguientes desigualdades:*

- *si $\alpha > \beta$, $efd_2(\hat{p}) < efd(\hat{p}) < efd_1(\hat{p})$,*
- *si $\alpha < \beta$, $efd_1(\hat{p}) < efd(\hat{p}) < efd_2(\hat{p})$,*
- *si $\alpha = \beta$, $efd(\hat{p}) = \alpha\ An(N-1)/(1-f)N$,*

*donde $efd_1(\hat{p}) = \alpha\ An(N-1)/(1-f)N$, $efd_2(\hat{p}) = \beta\ An(N-1)/(1-f)N$ y $V_{mas}(\hat{p}) = (1-f)N' p_U(1-p_U)/n$ y $f = n/N$.*

Por otra parte, con las cotas del teorema 3.2 y la expresión del coeficiente de correlación intraclase, Cochran (1986), en términos de la varianza entre y dentro de las $UPM$, es inmediato determinar las configuraciones de población que conducen a los valores mínimo y máximo de dicho coeficiente.

**Lema 3.3.** *bajo (MAS,MAS), $A \geq 2$ y para cualquier arreglo $(p_1, \ldots, p_A)$ tal que $\sum_{i=1}^{A} p_i/A = p_U \in (0,1)$, el valor del coeficiente de correlación intraclase $\rho$ se expresa como:*
$\rho = \frac{(B-1)V_1(\hat{p})}{(B-1)[V_1(\hat{p})+V_2(\hat{p})]} - \frac{V_2(\hat{p})}{(B-1)[V_1(\hat{p})+V_2(\hat{p})]}$,
*y los valores mínimo, $-1/(B-1)$, y máximo, 1, de $\rho$ se obtienen con $V_1(\hat{p}) = 0$ y $V_2(\hat{p}) = 0$, respectivamente.*

**Ejemplo 3.2.** *A continuación se calculan las cotas para el coeficiente de variación (lím inf cv y sup cv) de la proporción estimada, desviación estándar (lím inf desv y sup desv) y efecto del diseño (lím inf efd y sup efd) para una población con $A = 8$, $a = 2$, $B = 10$, tamaños de submuestreo, $b$, de 2 a 4 $USM$ y $p_U = 0.5$.*

| $b =$ | 2 | 3 | 4 |
|---|---|---|---|
| $\alpha =$ | 0.011 | 0.011 | 0.011 |
| $\beta =$ | 0.011 | 0.006 | 0.004 |
| *lím inf desv=* | *0.146* | *0.114* | *0.091* |
| *lím sup desv=* | *0.149* | *0.146* | *0.146* |
| *lím inf cv=* | *0.293* | *0.228* | *0.183* |
| *lím sup cv=* | *0.298* | *0.293* | *0.293* |
| *lím inf efd=* | *0.967* | *0.945* | *0.878* |
| *lím sup efd=* | *1.003* | *1.563* | *2.257* |

*Tabla 2*

*De la Tabla 2 se aprecia que el límite inferior para la desviación estándar disminuye conforme b se incrementa; empero, la diferencia entre la cota mínima y máxima crece. Para las cotas del coeficiente de variación y el efecto del diseño se observa un comportamiento similar.*

**Ejemplo 3.3.** *Efecto de un diseño muestral en el error de estimación absoluto. Suponga que se tiene una unidad habitacional con* 175 *edificios de departamentos y cada edificio tiene* 8 *departamentos. Se desea calcular el tamaño de muestra para estimar la proporción de departamentos que sufrieron algún robo en el último mes. Supóngase que se cuenta con recursos para visitar a lo más el* 20 % *de los edificios, es decir,* $A = 175x0.20 = 35$ *edificios.*

*Para seleccionar el número de departamentos por edificio en muestra, de la Tabla 2 se desprende que el rango de la varianza disminuye conforme los valores de submuestreo de* b *son cercanos a* 2*, con A y a fijos, por lo cual tomaremos un valor de* $b = 3$ *y como estimación anticipada de* $p_U$ *usaremos* $p_U* = 0.15$*. Las implicaciones de este plan en términos del error de estimación absoluto,* $e_a$*, se pueden evaluar con las cotas del teorema 3.2. Con estos datos,* $N = AB = 1400$*,* $n = 105$ *y usando un desvío normal,* $t_{\alpha/2} = 1.645$*, se tiene que* $\alpha = 0.000131$*,* $\beta = 0.000039$*,* $\alpha > \beta$*, por lo que aplicando la primera cota del teorema 3.2, las cotas inferior y superior de la varianza son,* 0.0009 *y* 0.0030 *respectivamente. De esta manera, el error de estimación absoluto se encontrará entre* 0.05 *y* 0.09*.*

# 4. Conclusiones

Se exhibieron cotas para la varianza, el efecto del diseño y el coeficiente de variación en el caso de la estimación de proporciones para el muestreo por conglomerados en dos etapas con tamaños iguales, suponiendo muestreo aleatorio simple en las dos etapas de selección. Esto facilita el cálculo del tamaño de muestra y también permite evaluar los valores mínimo y máximo posibles de la varianza del estimador de la proporción.

Por otra parte, las fórmulas son sencillas de calcular y únicamente se requieren los elementos de información con los que normalmente se cuenta en la práctica en la etapa de diseño muestral.

Las cotas exhibidas en este artículo no solo se emplean para proporciones calculadas a partir de variables dicotómicas, también se aplican a variables positivas, continuas y acotadas, con cotas conocidas, Scott & Smith (1975).

# Bibliografía

Chaudhuri, A. & Stenger, H. (2005), *Survey Sampling: theory and methods*, 2nd edn, Chapman & Hall/CRC.

Cochran, W. (1986), *Técnicas de Muestreo*, Ed. CECSA, México.

Kish, L. (1965), *Survey Sampling*, New York: Wiley & Sons.

Padilla Terán, A. M. (2000), "Cotas para la varianza, efecto del diseño y coeficiente de variación de proporciones en el muestreo por conglomerados en dos etapas con tamaños iguales", *Memorias electrónicas en extenso de la 3ª Semana Internacional de la Estadística y la Probabilidad. CD ISBN: 978-607-487-162-3.* .

Scott, A.J. & Smith, T. (1975), "Minimax designs for sample surveys", *Biometrika* **62 (2)**, 353–357.

Memoria FONE XXV (2010), Sección IV, 99–104

# Formulación natural del tamaño de muestra para el caso del muestreo por conglomerados en dos etapas*

Javier Suárez Espinosa[a]
*Programa de Estadística, Colegio de Postgraduados*

## 1. Introducción

Un problema esencial en el muestreo probabilístico es obtener el tamaño de muestra con una precisión determinada. El caso del muestreo por conglomerados (MC) en dos etapas es un caso especial, ya que el problema de la obtención del tamaño de muestra involucra dos cálculos: el número de conglomerados que serán muestreados ($n$) y el número de unidades que serán muestreadas dentro de cada conglomerado ($m_i$).

Un enfoque para la obtención de $n$ y las $m_i$´s, en el caso del MC en dos etapas fue propuesto por Cochran (1998, página 346), el cual consiste en minimizar la varianza del estimador de la media poblacional para un costo fijo. Además, si se asume que los tamaños de los conglomerados y los tamaños de muestra de cada conglomerado son iguales; es decir, si $m_i = m$ y $M_i = M$ entonces el tamaño de muestra $n$ se puede obtener como sigue (Lohr, 1999, página 156):

$$n = \frac{C}{c_1 + c_2 m}$$

donde: $C \equiv$ costo total del muestreo, $c_1 \equiv$ costo por conglomerado y $c_2 \equiv$ costo por unidad de muestreo dentro de cada conglomerado (incluyendo costo de la medición)

$$m = \sqrt{\frac{c_1 M(ECMD)}{c_2(ECME - ECMD)}}$$

*Agradecimiento: Este trabajo fue parcialmente financiado por la Linea Prioritaria de Investigación 15 del Colegio de Postgraduados

[a]sjavier@colpos.mx

donde: $ECMD = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\overline{y}_{ip}\right)^2}{N(M-1)}, y \;\; ECME = \frac{\sum_{i=1}^{N} M\left(\overline{y}_{ip}-\overline{y}_{p}\right)^2}{(N-1)}$

$N \equiv$ número de conglomerados en la población

$M \equiv$ número de unidades de muestreo en el i-ésimo conglomerado

$y_{ij} \equiv$ el valor de la característica de interés de la j-ésima unidades de muestreo del i-ésimo conglomerado.

$\overline{y}_p = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{K} \equiv$ media poblacional

$\overline{y}_{ip} = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i} = \frac{t_i}{M_i} \equiv$ media poblacional del i-ésimo conglomerado

Note que la solución depende de cantidades poblacionales que pueden ser estimadas usando una muestra; además, el tamaño de muestra $n$ está en función de $m$ y la fórmula para obtención del tamaño de muestra no involucra criterios de precisión.

Lohr (1999, página 156), bajo la suposición de que los tamaños de conglomerados son iguales e ignorando el factor de corrección, propone obtener el tamaño de muestra con la siguiente fórmula:

$$n = \frac{Z^2_{\alpha/2}\; v}{e^2}$$

$v = \frac{ECME}{M} + (1 - \frac{m}{M})\frac{ECMD}{m}$

$M$, $m$, $ECME$ $y$ $ECMD$ definidos anteriormente, $Z_{\alpha/2} \equiv$ cuantil de la distribución normal estándar que garantiza un cierto nivel de confiabilidad y $e \equiv$ margen de error en la estimación.

Note que para obtener el tamaño de muestra de los conglomerados ($n$) se requiere tener información de las varianzas dentro de los conglomerados; sin embargo, en la práctica dicha información en muchas ocasiones es difícil de obtener. Otros autores que han propuesto la determinación del tamaño de muestra con resultados similares y el mismo inconveniente son: Raj (1968, página 129), Perez (2000, página 429), Sukhatme y Sukhatme(1970, página 237).

La presente investigación pretende encontrar una formulación para obtener el tamaño de

muestra de conglomerados ($n$) sin la necesidad de contar con información de las varianzas dentro de conglomerados, lo cual sería más útil en caso prácticos.

## 2. Propuesta de una formulación natural para la obtención del tamaño de muestra

En esta sección se presenta una propuesta para la solución al problema planteado anteriormente, que hasta donde se tiene conocimiento no ha sido propuesta con anterioridad.

La derivación de la fórmula para el cálculo de tamaño de muestra bajo ciertas consideraciones parte de la siguiente expresión:

$$e^2 = Z^2_{\alpha/2} V\left(\widehat{\theta}\right), \tag{1}$$

donde: $Z^2_{\alpha/2}$ y $e$ fueron definidos previamente y $V\left(\widehat{\theta}\right)$ es la varianza del estimador insesgado de un parámetro de interés (cantidad poblacional de interés).

El estimador de la varianza del estimador de la media poblacional en el MC en dos etapas, se presenta a continuación:

$$\widehat{V\left(\widehat{\overline{y}}_p\right)} = \frac{1}{K^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right], \tag{2}$$

donde:

$s_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{t}_i - \frac{\widehat{t}_p}{N}\right)^2 \equiv$ estimador de la varianza poblacional de los totales de las unidades de muestreo primaria (*ump*)

$s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij}-\overline{y}_i\right)^2 \equiv$ estimador de la varianza poblacional de la i-ésima *ump*

$\overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} \equiv$ media muestral de la i-ésima *ump*

$\widehat{t}_i = M_i\overline{y}_i \equiv$ estimador del total de la i-ésima *ump*

$\widehat{t}_p = \frac{N}{n}\sum_{i=1}^{n}\widehat{t}_i \equiv$ estimador insesgado del total poblacional

Así subtituyendo la ec. 2 en la ec. 1 se tiene la siguiente ecuación:

$\frac{e^2}{Z^2_{\alpha/2}} = \frac{1}{K^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right]$

la cual también puede expresarse como:

$$\frac{K^2e^2}{Z^2_{\alpha/2}} + Ns_t^2 = \frac{1}{n}\left[N^2s_t^2 + N\sum_{i=1}^{n}\frac{M_i^2s_i^2}{m_i} - N\sum_{i=1}^{n}M_is_i^2\right]. \tag{3}$$

Por otro lado, usando los resultados del muestreo aleatorio simple (MAS) se puede mostrar que, para estimar la media poblacional de cada *ump*, el tamaño de muestra $m_i$ tiene la siguiente expresión:

$$m_i = \frac{Z^2_{\alpha/2}s_i^2}{e^2 + \frac{Z^2_{\alpha/2}s_i^2}{M_i}} \tag{4}$$

Ahora bién, bajo la consideración de un MC simple aleatorio en dos etapas; es decir, considerando que la selección de las unidades de muestreo secundarias es usando el muestreo aleatorio simple (MAS) para cada conglomerado en forma independiente y considerando que el nivel de confianza y el margen de error son exactamente los mismos en las ecuaciones 3 y 4, *una formulación natural del tamaño de muestra* puede ser substituyendo la ec. 4 en la ecuación 3, obteniendo el siguiente resultado:

$$n = \frac{N^2s_t^2 + \frac{e^2N}{Z^2_{\alpha/2}}\sum_{i=1}^{n}M_i}{\frac{e^2K^2}{Z^2_{\alpha/2}} + Ns_t^2} \tag{5}$$

Cabe hacer notar lo siguiente:

1. El tamaño de muestra de las *ump* ($n$) no depende de $s_i^2$ ; es decir, no depende de la variabilidad dentro de los conglomerados.

2. La solución de esta ecuación podría resolverse en forma iterativa, aunque su solución es muy difícil ya que las $M_i$ serán obtenidas en forma aleatoria.

3. Si se considera el caso en el que todos los conglomerados son del mismo tamaño; es decir, si $M_i = M$ para todo $i = 1, 2, \ldots N$, entonces se tiene:

$$n = \frac{Z^2_{\alpha/2} s^2_t}{M^2 e^2 \left(1 - \frac{1}{N}\right) + \frac{s^2_t Z^2_{\alpha/2}}{N}} \tag{6}$$

La ec. 6 provee una expresión para el tamaño de muestra para el caso del MC en dos etapas con tamaño de conglomerados iguales. Note que la propuesta para el cálculo del tamaño de muestra sólo tiene una suposición; además mantiene el nivel de confiabilidad deseado.

4. Claramente, si $N$ es suficientemente grande, el límite de la ec. 6 es:

$$n = \frac{Z^2_{\alpha/2} s^2_t}{M^2 e^2} \tag{7}$$

5. Los resultados que se presentan para el caso del estimador la media poblacional, pueden extender fácilmente para el caso del estimador de totales y proporciones.

# 3. Conclusiones

Para el caso del MC en dos etapas con tamaño de conglomerados iguales, el tamaño de muetra, $n$, no depende de otras supocisiones y puede obtenerse con un nivel de precisión deseado. Además, debido a que $n$ no depende de la variabilidad dentro de los conglomerados ($s^2_i$) el cálculo del tamaño de muestra se puede obtener de manera más fácil, lo cual amplia su posibilidad de aplicación en la práctica. Por lo anterior, la propuesta parece tener ventajas sobre investigaciones previas. El resultado puede extenderse fácilmente para la estimación de totales y proporciones. Finalmente, la propuesta puede extenderse a otros esquemas de muestreo.

# Bibliografía

Cochran, W. G. (1998), *Técnicas de muestreo*, CECSA, México.

Groves, R. M. y et al (2004), *Survey Methodology*, John-Wiley and Sons, Inc., New Jersey, USA.

Pérez-López, C. (2000), *Técnicas de Muestreo Estadístico*, Madrid, Alfaomega, España.

Raj, D. (1968), *Sampling theory*, McGraw-Hill, USA.

Sharon, L. (1999), *Sampling: Design and Analysis*, Duxbury Press. USA.

Sukhatme, P. V. y Sukhatme, B. V. (1970), *Sampling theory of surveys with applications*, Iowa State University Press, Ames, Iowa, USA.

Memoria FONE XXV (2010), Sección IV, 105–111

# Caracterización del BLUP de la media poblacional finita $\overline{Y}_j$ en predicción en áreas pequeñas (Small Area Estimation)

Fernando Velasco Luna[a], Mario Miguel Ojeda Ramírez[b]
*Facultad de Estadística e Informática. Universidad Veracruzana*

## 1. Introducción

La teoría de muestreo para poblaciones finitas se encarga de la selección de muestras, de las que se observan y miden características de cada una de las unidades muestreadas; usando estas observaciones la teoría estadística, en este contexto, desarrolla mecanismos para conducir inferencias acerca de ciertas características de la población, como por ejemplo la media poblacional $\overline{Y} = T/N$ Valliant et al. (2000), donde $T$ denota el total poblacional y $N$ denota el número de unidades en la población. Uno de los enfoques de inferencia en la teoría de muestreo de poblaciones finitas para estudiar los procesos de inferencia en el muestreo bietápico es el basado en el Modelo Lineal Mixto. En este enfoque se considera el modelo $\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \mathbf{e}_j$, donde $\mathbf{Y}_j$, $\mathbf{X}_j$ y $\mathbf{Z}_j$ denotan el vector respuesta, y las matrices de diseño (variables explicatorias a nivel 1 y nivel 2), respectivamente, en la $j$-ésima unidad de nivel 2 (área pequeña), la cual cuenta con $N_j$ unidades, $\boldsymbol{\beta}$ denota el vector de parámetros fijos, y $\mathbf{u}_j$ y $\mathbf{e}_j$ los efectos aleatorios de nivel 1 y nivel 2, respectivamente; sea $\mathbf{s}_j$ la muestra de $n_j$ unidades en la $j$-ésima área pequeña, la cual cuenta con $N_j$ unidades en la población, $\mathbf{r}_j$ denotando las unidades en la $j$-ésima área que no están en $\mathbf{s}_j$ y $r_j = N_j - n_j$ el número de unidades no muestreadas. Una vez que la muestra $\mathbf{s}_j$ ha sido obtenida se tiene la descomposición del modelo para la parte observada, que está dado por:

$$\mathbf{Y}_{j\mathbf{s}} = \mathbf{X}_{j\mathbf{s}}\boldsymbol{\beta} + \mathbf{Z}_{j\mathbf{s}}u_j + \mathbf{e}_{j\mathbf{s}} \tag{1}$$

[a] fvelasco@uv.mx
[b] mojeda@uv.mx

y el modelo para la parte no observada, que está dado por:

$$\mathbf{Y}_{j\mathbf{r}} = \mathbf{X}_{j\mathbf{r}}\boldsymbol{\beta} + \mathbf{Z}_{j\mathbf{r}}u_j + \mathbf{e}_{j\mathbf{r}}. \tag{2}$$

Entre todos los predictores, el mejor predictor lineal insesgado ($BLUP$) de la media poblacional finita $\overline{Y}_j = N_j^{-1}\sum_{i=1}^{N_j} Y_{ij}$ en la $j$-ésima área pequeña está dado por $f_j\overline{Y}_{j\mathbf{s}} + (1-f_j)\left[\overline{\mathbf{X}}_{j\mathbf{r}}\overset{\wedge}{\boldsymbol{\beta}}_{\mathbf{s}} + \overline{\mathbf{Z}}_{j\mathbf{r}}\mathbf{G}\mathbf{Z}_{j\mathbf{s}}^t\mathbf{V}_{j\mathbf{ss}}^{-1}\left(\mathbf{Y}_{j\mathbf{s}} - \mathbf{X}_{j\mathbf{s}}\overset{\wedge}{\boldsymbol{\beta}}_{\mathbf{s}}\right)\right]$, donde $f_j = n_j/N_j$, y $\overline{\mathbf{X}}_{j\mathbf{r}}$ y $\overline{\mathbf{Z}}_{j\mathbf{r}}$ son los vectores de medias para las $r_j$ unidades no muestreadas en la $j$-ésima unidad de nivel 2. La media de la población finita $\overline{Y}_j$ se puede descomponer en la media obtenida de la muestra $\overline{Y}_{j\mathbf{s}}$ más la media de las unidades no muestreadas $\overline{Y}_{j\mathbf{r}}$. Para la parte no muestreada se debe de tener una estimación de la media poblacional $\mu_j$ de la $j$-ésima área pequeña, la cual es un efecto mixto. Velasco y Ojeda (2010a) desarrollan la caracterización del $BLUP$ de la media poblacional $\mu_j$ en términos de los operadores proyector, ortogonal $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}\left(\mathbf{Z}^t\mathbf{Z}\right)^{-1}\mathbf{Z}^t$ y oblicuo $\mathbf{P}_{\mathbf{XV}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}$, definidos sobre los subespacios generados por las matrices de diseño. Aunque en la literatura se conocen suficientes resultados acerca de la teoría del álgebra lineal relacionada con la teoría de estimación y prueba de hipótesis en el modelo lineal general (MLG), no existen resultados que caracterizen al $BLUP$ de la media poblacional finita $\overline{Y}_j$ de la $j$-ésima unidad de nivel 2 en términos de las matrices de proyección. En este trabajo se presenta la caracterización del $BLUP$ de la media poblacional finita $\overline{Y}_j$ de la $j$-ésima área pequeña en términos de los operadores proyector $\mathbf{P}_{\mathbf{Z}}$ y $\mathbf{P}_{\mathbf{XV}}$, aplicandose la caraterización obtenida al modelo intercepto aleatorio.

## 2. Efecto mixto

Se considera el modelo dado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

$$\mathbf{u} \sim N\left(\mathbf{0}, \sigma_u^2\mathbf{I}_q\right), \quad \mathbf{e} \sim N\left(\mathbf{0}, \sigma_e^2\mathbf{I}_n\right), \quad Cov\left(\mathbf{e}, \mathbf{u}^{\mathbf{t}}\right) = \mathbf{0}, \tag{3}$$

donde $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X}$ y $\mathbf{Z}$ son matrices de orden $n \times p$ y $n \times q$, respectivamente y $\boldsymbol{\beta} \in \mathbb{R}^p$. En este caso la matriz de varianzas y covarianzas de $\mathbf{Y}$ está dada por $\mathbf{V} = \sigma_u^2\mathbf{Z}\mathbf{Z}^t + \sigma_e^2\mathbf{I}_n$.

Henderson (1975) obtiene el mejor estimador lineal insesgado ($BLUE$) de $\boldsymbol{\beta}$ y el $BLUP$ de $\mathbf{u}$, que están dados por $\overset{\wedge}{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}$ y $\overset{\wedge}{\mathbf{u}} = \mathbf{G}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\overset{\wedge}{\boldsymbol{\beta}}\right)$, respecti-

vamente. Además obtiene el *BLUP* del efecto mixto $\mathbf{k}^t\boldsymbol{\beta} + \mathbf{m}^t\mathbf{u}$ que está dado por:

$$\mathbf{k}^t \overset{\wedge}{\boldsymbol{\beta}} + \mathbf{m}^t \overset{\wedge}{\mathbf{u}} . \tag{4}$$

Velasco y Ojeda (2010b) desarrollan la caracterización del *BLUP* del efecto mixto $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ en términos de los operadores $\mathbf{P_{XV}}$ y $\mathbf{P_Z}$.

**Teorema 2.1.** *Bajo el modelo (3), si se cumple la condición* $n_j\mathbf{P}_{\mathbf{Z}_j} = \mathbf{Z}_j\mathbf{Z}_j^t$, *entonces el BLUP del efecto mixto* $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$ *se expresa en términos de los operadores, proyector oblicuo* $\mathbf{P_{XV}}$ *sobre* $S(\mathbf{X})$ *y proyector ortogonal* $\mathbf{P_Z}$ *sobre* $S(\mathbf{Z})$ *por:*

$$[\mathbf{P_{XV}} + \mathbf{P_Z}\mathbf{B}\mathbf{Q_{XV}}]\,\mathbf{Y}, \tag{5}$$

*donde* $\mathbf{B} = \oplus_{j=1}^{J}\left(b_j\mathbf{I}_{n_j}\right)$ *y* $b_j = n_j\sigma_{u0}^2/\left(n_j\sigma_{u0}^2 + \sigma_e^2\right)$.

*Demostración.* Para una demostración ver Velasco y Ojeda (2010b).

**Observación**. Dada una matriz $\mathbf{Z}_j$ de orden $n_j \times q$, la condición $n_j\mathbf{P}_{\mathbf{Z}_j} = \mathbf{Z}_j\mathbf{Z}_j^t$ se cumple si $\mathbf{Z}_j^t\mathbf{Z}_j = n_j\mathbf{I}_q$, lo cual ocurre si las columnas de la matriz $\mathbf{Z}_j$ son ortogonales y $\sum_{i=1}^{n_j} z_{ij}^2 = n_j$. En términos prácticos, la condición se cumple cuando todos los niveles $j$ de los efectos aleatorios asociados a $\mathbf{Z}$, tienen el mismo número de observaciones (repeticiones).

# 3. Media poblacional $\mu_j$

En está sección se presenta la caracterización del *BLUP* de la media poblacional $\mu_j$, de la $j$-ésima área pequeña, en términos de $\mathbf{P_{XV}}$, $\mathbf{P_Z}$ y $\mathbf{T}_{j\mathbf{s}}$.

Una vez que la muestra $\mathbf{s}$, de tamańo $n$, ha sido obtenida el vector $\mathbf{Y}$, las matrices $\mathbf{X}$ y $\mathbf{V}$, los operadores $\mathbf{Q_{XV}}$ y $\mathbf{P_Z}$, la estimación del parámetro $\boldsymbol{\beta}$ y la predicción del efecto aleatorio $u_j$ se denotarán por medio de $\mathbf{Y_s}$, $\mathbf{X_s}$, $\mathbf{V_s}$, $\mathbf{Q_{X_sV_s}}$, $\mathbf{P_{Z_s}}$, $\overset{\wedge}{\boldsymbol{\beta}}_{\mathbf{s}}$ y $\overset{\wedge}{u}_{j\mathbf{s}}$, respectivamente.

**Teorema 3.1.** $\mathbf{T}_{j\mathbf{s}}$ *dada por* $\mathbf{T}_{j\mathbf{s}} = \mathbf{X}_j\left(\mathbf{X}_\mathbf{s}^t\mathbf{V}_\mathbf{s}^{-1}\mathbf{X}_\mathbf{s}\right)^{-1}\mathbf{X}_\mathbf{s}^t\mathbf{V}_\mathbf{s}^{-1}$ *define una transformación lineal de* $\mathbb{R}^n$ *a* $\mathbb{R}^{N_j}$.

La media poblacional $\mu_j$ se define como $E\left(\overline{Y}_j\,|u_j\right)$, que bajo el modelo (2) con $\mathbf{Z}_j = \mathbf{1}_{n_j}$ está dada por $\overline{X}_{j\mathbf{r}}^t\boldsymbol{\beta} + u_j$, que es un efecto mixto.

**Teorema 3.2.** *Bajo el modelo (3) si* $n_j\mathbf{P}_{\mathbf{Z}_{j\mathbf{s}}} = \mathbf{Z}_{j\mathbf{s}}\mathbf{Z}_{j\mathbf{s}}^t$*, entonces el BLUP de* $\mu_j$ *se expresa en términos de los operadores proyector* $\mathbf{Q}_{\mathbf{X_sV_s}}$ *y* $\mathbf{P}_{\mathbf{Z_s}}$*, y de la transformación lineal* $\mathbf{T}_{j\mathbf{s}}$ *por:*

$$\left[\frac{\mathbf{1}_{N_j}^{*j\mathbf{r}t}}{r_j}\mathbf{T}_{j\mathbf{s}} \;+\; \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j}\left(\mathbf{P}_{\mathbf{Z_s}}\mathbf{B_s}\mathbf{Q}_{\mathbf{X_sV_s}}\right)\right]\mathbf{Y_s},$$

*donde* $\mathbf{1}_n^{*j\mathbf{s}}$ *es un vector de* $0's$ *en* $\mathbb{R}^n$ *con un* 1 *en las posiciones correspondientes a las unidades que pertenecen a la* $j$*-ésima área pequeña,* $\mathbf{B_s} = \oplus_{j=1}^{J}\left(b_j\mathbf{I}_{n_j}\right)$ *y* $b_j = n_j\sigma_{u0}^2/\left(n_j\sigma_{u0}^2 + \sigma_e^2\right)$.

*Demostración.* Para una demostración ver Velasco y Ojeda (2010a).

# 4. Caracterización del *BLUP* de la media poblacional finita $\overline{Y}_j$

En está sección se presenta la caracterización del *BLUP* de $\overline{Y}_j$ en términos de $\mathbf{P_{XV}}$, $\mathbf{P_Z}$ y $\mathbf{T}_{j\mathbf{s}}$.

$\mathbf{1}_{N_j}^{*j\mathbf{r}}$ es un vector de $0's$ en $\mathbb{R}^{N_j}$ con un 1 en las posiciones correspondientes a las unidades de nivel 1, que pertenecen a la $j$-ésima área pequeña, que no están en la muestra.

**Teorema 4.1.** *Bajo el modelo (3), con* $\mathbf{Z}_j = \mathbf{1}_{n_j}$*, el BLUP de la media poblacional finita* $\overline{Y}_j$ *está dado por:*

$$f_j\overline{Y}_{j\mathbf{s}} + (1-f_j)\left[\overline{\mathbf{X}}_{j\mathbf{r}}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}} + \hat{u}_{j\mathbf{s}}\right]. \tag{6}$$

*Demostración.* El *BLUP* de la media poblacional finita $\overline{Y}_j$ está dado por:

$$f_j\overline{Y}_{j\mathbf{s}} + (1-f_j)\left[\overline{\mathbf{X}}_{j\mathbf{r}}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}} + \overline{\mathbf{Z}}_{j\mathbf{r}}\mathbf{G}\mathbf{Z}_{j\mathbf{s}}^t\mathbf{V}_{j\mathbf{ss}}^{-1}\left(\mathbf{Y}_{j\mathbf{s}} + \mathbf{X}_{j\mathbf{s}}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\right)\right] \tag{7}$$

en (7) interviene el término

$$\overline{\mathbf{Z}}_{j\mathbf{r}}\mathbf{G}\mathbf{Z}_{j\mathbf{s}}^t\mathbf{V}_{j\mathbf{ss}}^{-1}\left(\mathbf{Y}_{j\mathbf{s}} + \mathbf{X}_{j\mathbf{s}}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\right) \tag{8}$$

Bajo el modelo (3), si $\mathbf{Z}_j = \mathbf{1}_{n_j}$, entonces $\overline{\mathbf{Z}}_{j\mathbf{r}} = 1$, por lo que (8) toma la forma:

$$\mathbf{G}\mathbf{Z}_{j\mathbf{s}}^t\mathbf{V}_{j\mathbf{ss}}^{-1}\left(\mathbf{Y}_{j\mathbf{s}} + \mathbf{X}_{j\mathbf{s}}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\right) \tag{9}$$

que corresponde a $\hat{u}_{j\mathbf{s}}$. De lo cual se sigue (6).

**Teorema 4.2.** *Bajo el modelo (3), con* $\mathbf{Z}_j = \mathbf{1}_{n_j}$*, si* $n_j \mathbf{P}_{\mathbf{Z}_{js}} = \mathbf{Z}_{js}\mathbf{Z}_{js}^t$*,* $n_j \in \mathbb{R}$*, entonces el BLUP de la media poblacional finita* $\overline{Y}_j$*, se expresa en términos de los operadores proyector* $\mathbf{Q}_{\mathbf{X_s V_s}}$ *y* $\mathbf{P}_{\mathbf{Z_s}}$*, y de la transformación lineal* $\mathbf{T}_{js}$ *por medio de:*

$$\left( \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{N_j} + \frac{r_j}{N_j} \left[ \frac{\mathbf{1}_{N_j}^{*j\mathbf{r}t}}{r_j} \mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j} \left( \mathbf{P}_{\mathbf{Z_s}} \mathbf{B_s} \mathbf{Q}_{\mathbf{X_s V_s}} \right) \right] \right) \mathbf{Y_s},$$

*donde* $\mathbf{B_s} = \oplus_{j=1}^{J} \left( b_j \mathbf{I}_{n_j} \right)$ *y* $b_j = n_j \sigma_u^2 / \left( n_j \sigma_u^2 + \sigma_e^2 \right)$.

*Demostración.* Por el teorema 4.1, $BLUP\left(\overline{Y}_j\right) = f_j \overline{Y}_{j\mathbf{s}} + (1 - f_j) \left[ \overline{\mathbf{X}}_{j\mathbf{r}} \hat{\boldsymbol{\beta}}_{\mathbf{s}} + \hat{u}_{j\mathbf{s}} \right]$. Además, por el teorema 3.2, $BLUP\left(\mu_j\right) = \left[ \frac{\mathbf{1}_{N_j}^{*j\mathbf{r}t}}{r_j} \mathbf{T}_{j\mathbf{s}} \; + \; \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j} \left( \mathbf{P}_{\mathbf{Z_s}} \mathbf{B_s} \mathbf{Q}_{\mathbf{X_s V_s}} \right) \right] \mathbf{Y_s}$ y recordando que $\mu_j$ está dado por $\overline{X}_{j\mathbf{r}}^t \boldsymbol{\beta} + u_j$, se tiene

$$\begin{aligned}
BLUP\left(\overline{Y}_j\right) &= \left( \frac{n_j}{N_j} \right) \overline{Y}_{j\mathbf{s}} \; + \; \left( \frac{r_j}{N_j} \right) BLUP\left(\mu_j\right) \\
&= \left( \frac{n_j}{N_j} \right) \left( \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j} \right) \mathbf{Y_s} \; + \; \left( \frac{r_j}{N_j} \right) BLUP\left(\mu_j\right) \\
&= \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{N_j} \mathbf{Y_s} + \frac{r_j}{N_j} \left[ \frac{\mathbf{1}_{N_j}^{*j\mathbf{r}t}}{r_j} \mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j} \left( \mathbf{P}_{\mathbf{Z_s}} \mathbf{B_s} \mathbf{Q}_{\mathbf{X_s V_s}} \right) \right] \mathbf{Y_s} \\
&= \left( \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{N_j} + \frac{r_j}{N_j} \left[ \frac{\mathbf{1}_{N_j}^{*j\mathbf{r}t}}{r_j} \mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}_n^{*j\mathbf{s}t}}{n_j} \left( \mathbf{P}_{\mathbf{Z_s}} \mathbf{B_s} \mathbf{Q}_{\mathbf{X_s V_s}} \right) \right] \right) \mathbf{Y_s}.
\end{aligned}$$

# 5. Caracterización en el modelo intercepto aleatorio

Se presenta la caracterización del *BLUP* de la media poblacional finita $\overline{Y}_j$, considerando el caso balanceado, es decir $n_j = d \; \forall j = 1, ..., k$, bajo el modelo intercepto aleatorio, ya que este modelo es ampliamente usado en la teoría de estimación en áreas pequñas. Este modelo está dado para el nivel 1 por:

$$Y_{ij} = \mu_j + e_{ij}, \quad i = 1, ..., d, \quad j = 1, ..., k.$$

y para el nivel 2 por:

$$\mu_j = \mu + u_j, \quad j = 1, ..., k.$$

lo cual lleva al modelo combinado:

$$Y_{ij} = \mu + u_j + e_{ij}, \qquad i = 1, ..., d, \qquad j = 1, ..., k. \tag{10}$$

donde $\mu$ es un parámetro fijo; $u_j$ es el efecto aleatorio; $u_j$ y $e_{ij}$ son independientes, con $u_j \sim N(0, \sigma^2_{u0})$ y $e_{ij} \sim N(0, \sigma^2_e)$. El modelo para la $j$-ésima unidad de nivel 2 tiene la forma $\mathbf{Y}_j = \mathbf{1}_d\mu + \mathbf{1}_d u_j + \mathbf{e}_j, j = 1, ..., k$. En este caso $\mathbf{X} = \mathbf{1}_k \otimes \mathbf{1}_d$ y $\mathbf{Z} = \mathbf{I}_k \otimes \mathbf{1}_d$, por lo que

$$\mathbf{P}_{\mathbf{X}_s\mathbf{V}_s} = \frac{\mathbf{1}_{kd}\mathbf{1}^t_{kd}}{kd} \text{ y } \mathbf{P}_{\mathbf{Z}_s} = \frac{(\mathbf{I}_k \otimes \mathbf{1}_d\mathbf{1}^t_d)}{d}. \tag{11}$$

**Teorema 5.1.** *Bajo el modelo intercepto aleatorio (10). El BLUP de la media poblacional finita $\overline{Y}_j$ está dado por:*

$$\frac{n_j}{N_j}\overline{Y}_{j\mathbf{s}} + \frac{r_j}{N_j}\left[\overline{Y}_{\mathbf{s}} \; + \; \frac{c\,(k-1)}{k}\left[\overline{Y}_{j\mathbf{s}} - \overline{Y}_{(-j)\mathbf{s}}\right]\right] \tag{12}$$

*donde $\overline{Y}_{\mathbf{s}}$, $\overline{Y}_{j\mathbf{s}}$ y $\overline{Y}_{(-j)\mathbf{s}}$ denotan la media muestral, la media muestral de la $j$-ésima unidad de nivel 2, y la media muestral de las unidades de nivel 2 restantes, respectivamente.*

*Demostración.* Por el teorema 4.2, considerando el caso balanceado

$$BLUP\left(\overline{Y}_j\right) = \left(\frac{\mathbf{1}^{*j\mathbf{s}t}_n}{N_j} + \frac{r_j}{N_j}\left[\frac{\mathbf{1}^{*j\mathbf{r}t}_{N_j}}{r_j}\mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}^{*j\mathbf{s}t}_n}{d}\left(c\mathbf{P}_{\mathbf{Z}_\mathbf{s}}\mathbf{Q}_{\mathbf{X}_\mathbf{s}\mathbf{V}_\mathbf{s}}\right)\right]\right)\mathbf{Y}_\mathbf{s}.$$

Además, $\mathbf{T}_{j\mathbf{s}} = \frac{\mathbf{1}_{N_j}\mathbf{1}^t_{kd}}{kd}$ y de (11) $c\mathbf{P}_{\mathbf{Z}_s}\mathbf{Q}_{\mathbf{X}_s\mathbf{V}_s} = \frac{c}{kd}\left(k\left(\oplus\mathbf{1}_d\mathbf{1}^t_d\right) - \mathbf{1}_{kd}\mathbf{1}^t_{kd}\right)$, por lo que

$$\begin{aligned}
BLUP\left(\overline{Y}_j\right) &= \left(\frac{\mathbf{1}^{*j\mathbf{s}t}_n}{N_j} + \frac{r_j}{N_j}\left[\frac{\mathbf{1}^{*j\mathbf{r}t}_{N_j}}{r_j}\mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}^{*j\mathbf{s}t}_n}{d}\left(c\mathbf{P}_{\mathbf{Z}_\mathbf{s}}\mathbf{Q}_{\mathbf{X}_\mathbf{s}\mathbf{V}_\mathbf{s}}\right)\right]\right)\mathbf{Y}_\mathbf{s} \\
&= \frac{\mathbf{1}^{*j\mathbf{s}t}_n}{N_j}\mathbf{Y}_\mathbf{s} + \frac{r_j}{N_j}\left[\frac{\mathbf{1}^{*j\mathbf{r}t}_{N_j}}{r_j}\mathbf{T}_{j\mathbf{s}} + \frac{\mathbf{1}^{*j\mathbf{s}t}_n}{d}\left(c\mathbf{P}_{\mathbf{Z}_\mathbf{s}}\mathbf{Q}_{\mathbf{X}_\mathbf{s}\mathbf{V}_\mathbf{s}}\right)\right]\mathbf{Y}_\mathbf{s} \\
&= \frac{\mathbf{1}^{*j\mathbf{s}t}_n}{N_j}\mathbf{Y}_\mathbf{s} + \frac{r_j}{N_j}\left[\overline{Y}_{\mathbf{s}} \; + \; \frac{c\,(k-1)}{k}\left[\overline{Y}_{j\mathbf{s}} - \overline{Y}_{(-j)\mathbf{s}}\right]\right] \\
&= \frac{n_j}{N_j}\overline{Y}_{j\mathbf{s}} + \frac{r_j}{N_j}\left[\overline{Y}_{\mathbf{s}} \; + \; \frac{c\,(k-1)}{k}\left[\overline{Y}_{j\mathbf{s}} - \overline{Y}_{(-j)\mathbf{s}}\right]\right].
\end{aligned}$$

# 6. Conclusiones

En este trabajo se expresó el *BLUP* de la media poblacional finita $\overline{Y}_j$ de la $j$-ésima área pequeña como la suma ponderada de un elemento en $S(\mathbf{X}_j)$ y un elemento en el espacio $S(\mathbf{Z_s})$. Lo anterior al aplicarlo al modelo intercepto aleatorio sin variables explicatorias, considerando el caso balanceado, permitió expresar el *BLUP* de la media poblacional finita $\overline{Y}_j$ como la suma de múltiplos de $\overline{Y}_{\mathbf{s}}$, $\overline{Y}_{j\mathbf{s}}$ y $\overline{Y}_{(-j)\mathbf{s}}$, que denotan la media muestral, la media muestral de las unidades en la $j$-ésima área pequeña y la media muestral de las unidades que no pertenecen a la $j$-ésima área pequeña, respectivamente. Se espera que esta caracterización en términos de los proyectores permita una mejor comprensión de las propiedades del *BLUP* de $\overline{Y}_j$ tal como sucede en la caracterización del estimador de parámetros $\boldsymbol{\beta}$ en el MLG.

# Bibliografía

Henderson, C. (1975), "Best linear unbiased estimation and prediction under a selection model", *Biometrics* **31**, 423–447.

Valliant, R., Dorfman, A. y Royall, R. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley.

Velasco, L. y Ojeda, R. (2010a), "Caracterización del blup de la media poblacional en el modelo lineal general mixto", *Memorias del XXIV Foro nacional de Estadística, INEGI: México* pp. 81–87.

Velasco, L. y Ojeda, R. (2010b), "Caracterización del blup del efecto mixto $x\beta+zu$", *Editores: Taponar S.F.S. and Cruz S.H.A. and Reyes C.H. and Zacarías F.J.D. Aportaciones y Aplicaciones de la Probabilidad y la Estadística,Puebla, México: Benemérita Universidad Autónoma de Puebla* .

# Sección V

# Aplicaciones

Memoria FONE XXV (2010), Sección V, 115–122

# Análisis de confiabilidad para tiempos de eficacia analgésica en pacientes con cólico renoureteral

Fidel Ulín-Montejo[a], Jorge A. Pérez Chávez

Rosa Ma. Salinas-Hernández
*Universidad Juárez Autónoma de Tabasco - IMSS*

## 1. Introducción

El dolor por cólico renoureteral (CR) se ha comparado con los tipo de dolor de mayor intensidad. Así, este trabajo nace de la necesidad de proporcionar un tratamiento rápido para el control urgente del CR, patología muy frecuente y con gran impacto laboral, social y emocional en la población económicamente activa. Clínicamente se considera importante la valoración comparativa del efecto analgésico en el CR de los métodos terapéuticos *Bloqueo Subcostal con Lidocaina* y *Tratamiento Convencional con Metamizol*, debido a la disponibilidad y uso de estos fármacos en los servicios de urgencias del IMSS[3]. Los resultados de este estudio permitirán un manejo más eficaz, oportuno y adecuado de los pacientes en los servicios de urgencias, así como disminuir los costos del manejo de dicha patología.

## 2. Escala Visual Analógica

Se define al dolor como una experiencia sensorial y emocional displacentera, asociada a daño tisular, ya sea real, potencial o descrita en términos de dicho daño. La Escala Visual Analógica (EVA) consiste en una línea horizontal de 10 centímetros, en cuyos extremos se encuentran las expresiones extremas de un síntoma. En el izquierdo se ubica la ausencia o

[a]fidel.ulin@basicas.ujat.mx

menor intensidad (no dolor) y en el derecho la mayor intensidad (peor dolor imaginable). Se pide al paciente que marque en la línea el punto que indique la intensidad. La distancia desde el punto de *no dolor* a la marcada representa la intensidad del dolor. La EVA es un método simple, de uso universal que ocupa poco tiempo, para uso preciso es necesaria la comprensión y colaboración del paciente, buena coordinación motora y visual [3]. Acorde a la EVA, se proponen las alternativas terapéuticas siguientes: **Dolor leve** (*EVA* de 0 a 4); puede ser tratado satisfactoriamente con analgésicos no opioides del tipo de los antiinflamatorios no esteroideos (AINEs). **Dolor moderado** (*EVA* de 5 a 7); puede ser tratado con analgésicos opioides, ya sea en bolo o en infusión contínua; así mismo, puede utilizarse la combinación de estos analgésicos con AINEs. **Dolor severo** (*EVA* de 8 a 10); el dolor intenso puede ser manejado con opioides potentes (morfina y citrato de fentanilo), ya sea en infusión continua, con técnicas de analgesia controlada o de anestesia regional [3].

# 3. Censura

En estudios de analisis sensorial [5], se define una variable aleatoria $T$ como el tiempo al cual el consumidor rechaza una muestra de alimento, definiéndose la *función de confiabilidad* $S(t)$ como la probabilidad de que un consumidor acepte este alimento mas allá del tiempo $t$, esto es $S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(t)$. Donde $f$ y $F$ son las funciones de densidad y distribución para $T$. Aquí, en forma análoga, se define la variable $T$ como el tiempo al cual el dolor ha disminuido; de modo que $S(t)$ se definirá como la *probabilidad de que el paciente sienta dolor severo*, despues de administrado el tratamiento, *mas allá de un tiempo* $t$.

Para ilustrar el concepto de *censura*, considere que los tiempos de observación y aplicación de la EVA a los pacientes fueron $5, 10, 15, 30$ min, y $15, 30, 60, 90$ min, respectivamente. Debido a que los tiempos son discretos, $T$ nunca será observada exactamente, en vez de ello solo se observará que $T \leq 5$, $15 \leq T \leq 30$ o $T \geq 30$. En confiabilidad y superviviencia, esta información es considerada como datos censurados, incompletos o no-detectados[1,2].

**Censura por la Izquierda**: Si para un paciente, el dolor desapareció a los 5 min de aplicado el tratamiento, el tiempo de disminución del dolor es de $T \leq 5$. Es decir, el dolor desapareción en algún momento entre 0 y 5 min. **Censura por Intervalo**: Si un paciente sufría dolor a los 15 min de aplicado el tratamiento, pero a los 30 min el dolor ha disminuido considerablemente, el tiempo de disminución del dolor es de $15 \leq T \leq 30$ min. Las limita-

ciones de recursos, tiempo o personal, impiden observar a los pacientes a los $16, 17, 18, ..., 28$ y 29 min. **Censura por la Derecha**: Si un paciente no ha sentido disminución del dolor aun despues de los 30 min del tratamiento, entonces el tiempo de desaparición del dolor es de $T \geq 30$ min. Esto es, si el paciente pudiera soportar un tiempo suficientemente largo despues de los 30 min, en algún momento sentiría disminuir el dolor, sin embargo las consideraciones éticas no permiten ésto.

# 4. Metodología

## 4.1. Máxima Verosimilitud

La inferencia por máxima verosimilitud (MV) se basa en ajustar modelos por medio de las combinaciones modelos-parámetros para los cuales la probabilidad de los datos sea alta. MV puede aplicarse a una amplia variedad de modelos con datos censurados [2]. La **función de verosimilitud,** $L(\theta)$, es utilizada para la estimación de parámetros y de la función de supervivencia; $L(\theta)$ se define como la probabilidad conjunta de los datos obtenidos.

$$\begin{aligned} L(\theta) &= \prod_{i\in R} S(r_i) \prod_{i\in L} [1 - S(l_i)] \prod_{i\in I} [S(l_i) - S(r_i)] \\ &= \prod_{i\in R} [1 - F(r_i;\theta)] \prod_{i\in L} F(l_i;\theta) \prod_{i\in I} [F(r_i;\theta) - F(l_i;\theta)] \end{aligned} \tag{1}$$

Donde $R$ es el conjunto de observaciones censuradas por la derecha, $r_i$; $L$ es el conjunto de observaciones censuradas por la izquierda, $l_i$; e $I$ es el conjunto de las observaciones censuradas por intervalo. En (1) se muestra como cada uno de los tipos de censura contribuye de manera diferente a la función de verosimilitud.

## 4.2. Modelos Paramétricos y Estimación

Con base en estudios previos, puede suponerse modelos paramétricos adecuados que proporcionen estimaciones precisas para la función de confiabilidad $S(t)$ y otras cantidades de interés[1,2]. Tiempos a la falla, de vida o de rechazo no siguen distribuciones simétricas (normal), sino distribuciones sesgadas a la derecha. Frecuentemente se eligen modelos de probabilidad con distribución de (log)localización-escala [2]:

$$y_p = \ln(T_p) = \mu + \Phi^{-1}(p)\sigma \tag{2}$$

Para $T$ lognormal, se tiene que $\Phi$ es la distribución normal estándar; si $T$ es Weibull, la distribución de los valores mínimos extremos, $\Phi_{sev}(w) = \exp[\exp(w)]$ [2,5]. Si se eligen los modelos lognormal y Weibull, sus funciones de confiabilidad estarían dadas, respectivamente, por:

$$S(t) = 1 - \Phi[\frac{\ln(t) - \mu}{\sigma}];\ S(t) = \Phi_{sev}[\frac{\ln(t) - \mu}{\sigma}]. \tag{3}$$

Los parámetros de los modelos de (log)localización-escala son obtenidos maximizando la función de verosimilitud (1). Ahora, una vez construida la verosimilitud $L(\theta)$ para un modelo asumido, pueden usarse funciones y programas en $R$ [4] para estimar $\mu$ y $\sigma$, maximizando (1) con la solución numérica y simultánea de:

$$\frac{\partial \log L(\mu, \sigma)}{\partial \mu} = 0; \quad \frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = 0. \tag{4}$$

## 4.3. Mediana y Tiempo Medio

En análisis de confiabilidad (superviviencia), el tiempo medio a la falla (de vida)[1,2], se define como: $E(T) = \int_0^\infty S(t)dt$. En este estudio de medición de dolor a traves de una escala sensorial, $M(T)$ y $E(T)$ representan la *mediana* y el *tiempo medio*, respectivamente, al cual los pacientes consideran la disminución o desaparicion de dolor. Para las distribuciones lognormal y Weibull estarían dadas por:

$$M_{lognormal}(T) = exp(\mu), \qquad E_{lognormal}(T) = exp(\mu + \sigma^2/2)$$
$$M_{Weibull}(T) = exp(\mu)[log(2)]^\sigma, \qquad E_{Weibull}(T) = exp(\mu)\Gamma(1 + \sigma)$$

## 4.4. Prueba de Razón de Verosimilitud

Esta prueba puede usarse para elegir el modelo que mejor ajuste los datos, comparando las verosimilitudes estimadas $L(\hat{\theta})$. Así, ambos modelos ajustarían bien los datos si,

$$X^2 = -2ln\left[\frac{L(\hat{\theta}_{Modelo1})}{L(\hat{\theta}_{Modelo2})}\right] \sim \mathcal{X}^2_{(1)}. \tag{5}$$

Análogamente y con las estimaciones obtenidas, puede probarse la hipotesis $H : E(T_{Bloqueo}) = E(T_{Tradicional})$, considerandose éstas como funciones paramétricas del modelo elegido [2,5].

# 5. Grupos Experimentales

Se trataron 90 pacientes con cuadro clínico de cólico renoureteral sin tratamiento en las 8 horas previas al llegar a Urgencias del Hospital General IMSS-Tabasco No. 46; de septiembre 2007 a febrero 2009. Los grupos experimentales fueron definidos como sigue: **Bloqueo** para 50 pacientes, Bloqueo del 12º nervio subcostal con infiltración de Lidocaína simple con técnica de abanico. **Tratamiento Convencional** para 40 pacientes, Metamizol 30 mg/Kg. diluido en 250 ml. de solución salina a pasar en 30 min. Los tratamientos se aplicaron aleatoriamente a los pacientes al tiempo 0, con seguimiento y evaluación analgésica usando EVA, a 5, 10, 15, 30 min. y $15, 30, 60, 90$ min. respectivamante:

**Tabla 1. Datos de los Pacientes con Bloqueo**

| Tiempo a la Disminución del Dolor para el Bloqueo | | | | | | |
|---|---|---|---|---|---|---|
| Frec. | 0 | 5 | 10 | 15 | 30 min | Censura |
| 28 | S | N | N | N | N | Izquierda: $< 5$ |
| 10 | S | S | N | N | N | Intervalo: 5 - 10 |
| 2 | S | S | S | N | N | Intervalo: 10 - 15 |
| 7 | S | S | S | S | N | Intervalo: 15 - 30 |
| 3 | S | S | S | S | S | Derecha: $> 30$ |
| Tiempo a la Disminución del Dolor para el Metamizol | | | | | | |
| Frec. | 0 | 15 | 30 | 60 | 90 min | Censura |
| 1 | S | N | N | N | N | Izquierda: $< 15$ |
| 20 | S | S | N | N | N | Intervalo: 15 - 30 |
| 13 | S | S | S | N | N | Intervalo: 30 - 60 |
| 6 | S | S | S | S | N | Intervalo: 60 - 90 |

# 6. Resultados

## 6.1. Estimaciones para el Grupo con Bloqueo

De los pacientes con bloqueo del 12º nervio subcostal utilizando lidocaína, 28 refirieron un tiempo de desaparición del dolor menor a 5 min, 10 entre 5 y 10 min, 2 pacientes entre 10 y 15 min, 7 entre 15 y 30 min; y en 3 no hubo mejoría a los 30 min.

$R$ [4] provee las estimaciones de los parámetros y de la log-verosimilitud, de donde se eligen los modelos lognormal y Weibull, por buen ajuste y porque han sido utilizada en estudios anteriores. Obteniéndose así, las respectivas medianas $M$ y tiempos medios $E$, para los tiempos de disminución del dolor:

$$Lognormal: \ \mu = 1.41, \ es_\mu = 0.25; \ \ \sigma = 1.34; \ \ -2logL = 124;$$
$$Weibull: \ \mu = 1.89, \ es_\mu = 0.25; \ \ \sigma = 1.48; \ \ -2logL = 127.$$

$$M_{Lognormal}(T) = 4.10, \qquad E_{Lognormal}(T) = 10.05;$$
$$M_{Weibulll}(T) = 3.85, \qquad E_{Weibull}(T) = 8.70.$$

Una prueba de razon de verosimilitud, $P(\mathcal{X}^2_{(1)} > X^2) = 0.47$, asevera que ambos modelos ofrecen resultados similares.

## 6.2. Estimaciones para el Grupo con Tratamiento Convencional

Del grupo de 40 pacientes tratados con analgesia convencional, 1 refirió inicio de la mejoría del dolor a los 15 min, 20 a los 30 min, 13 a los 60 min y los restantes 6 a los 90 min.

$R$ [4] provee las estimaciones de parametros y de la log-verosimilitud, eligiéndose los modelos lognormal y Weibull, por buen ajuste y uso en estudios previos[1,2,7]. Obteniendose las respectivas medianas $M$ y tiempos medios $E$, para los tiempos de disminución del dolor:

$$Lognormal: \ \mu = 3.46, \ es_\mu = 0.081; \ \ \sigma = 0.473; \ \ -2logL = 94.1;$$
$$Weibull: \ \mu = 3.71, \ es_\mu = 0.083; \ \ \sigma = 0.464; \ \ -2logL = 99.7.$$

$$M_{Lognormal}(T) = 31.82, \quad E_{Lognormal}(T) = 35.58;$$
$$M_{Weibulll}(T) = 34.46, \quad E_{Weibull}(T) = 38.23.$$

Una prueba de razon de verosimilitud, $P(\mathcal{X}^2_{(1)} > X^2) = 0.29$, asevera que ambos modelos ofrecen resultados similares.

## 6.3. Comparación de Tratamientos

De las estimaciones con el modelo lognormal para el grupo Bloqueo, la mediana y el tiempo medio de disminución de dolor fueron: $M_B = 4.10$ y $E_B = 10.85$ min; y para el grupo con

Tratamiento Convencional $M_{TC} = 31.82$ y $E_{TC} = 35.58$ min. Obervandose una diferencia marcada entre estos medidas, corroborandose con una prueba de razon de verosimilitud: $P(\mathcal{X}^2_{(2)} > X^2) = 0.023$; estableciendose así, a la luz de los datos, que el tratamiento por Bloqueo es más eficiente.

# 7. Conclusiones

La infiltración de Lidocaína simple en el 12º nervio subcostal es un recurso eficaz para el control rápido del cólico renoureteral de intensidad severa, comparado con la administración de analgesia convencional.

Los resultados de la analgesia convencional utilizando fármacos intravenosos del tipo Metamizol para el control del cólico renal son poco eficaces comparados con el bloqueo subcostal, pues tarda más de 30 min en iniciar el efecto analgésico.

La infiltración subcostal es un método fácil, reproducible en cualquier lugar que reúna las condiciones necesarias, de bajo costo y sin efectos adversos a la dosis administrada.

# Bibliografía

Klein, J. P. y Moeschberger, M. L. (2005), *Survival analysis Techniques for Censored and Truncated Data*, New York: Springer.

Meeker, W. Q. y Escobar, L. A. (1998), *Statistical methods for reliability data*, New York: Wiley.

*Memoria del XXIII Foro Nacional de Estadística* (2009).

Pérez-Chávez, J. A. (2009), "Eficacia analgésica del bloqueo del xii nervio subcostal contra metamizol en pacientes con cólico renoureteral en el servicio de urgencias del hospital general de zona no. 46 imss, villahermosa, tabasco".

Team, R. D. C. (2006), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *www.R-project.org)

Ulín-Montejo, F. y Salinas-Hernández, R. M. (2008), *Análisis de Confiabilidad para la Predicción de Vida Útil de Alimentos.*

Memoria FONE XXV (2010), Sección V, 123–128

# Modelación de fenómenos atmosféricos usando procesos de Poisson no homogéneos

Nahun Israel Loya Monares[a], Hortensia J. Reyes Cervantes
*FCFM, Benemérita Universidad Autónoma de Puebla*

Francisco J. Ariza Hernández
*UAM, Universidad Autónoma de Guerrero*

## 1. Introducción

La contaminación atmosférica en particular por Ozono ($O_3$) es un problema serio en las ciudades industrializadas y con alto índice demográfico, la introducción de sustancias dañinas al medio ambiente provocan desequilibro a los ecosistemas Harrison (2003), en México este fenómeno se presenta en algunas ciudades como: Guadalajara, Ciudad de México y Monterrey Barrientos (2005). La calidad del aire de la Zona Metropolitana del Valle de México (ZMVM) se ha considerado por varias décadas como una de las más contaminadas del mundo debido a las emisiones provenientes de los sectores: transporte, industria, servicios, etc.

El modelar $O_3$ cuando éste supera determinado umbral permitido es importante, ya que permite estimar el comportamiento del mismo, así como observar que covariables pueden aportar información relevante para el desarrollo de este fenómeno. El $O_3$ ejerce su acción a través de varios mecanismos por ejemplo: reacciones con grupos sulfhidrílos, aldehídos y amino de bajo peso molecular; el contaminante provoca efectos nocivos en el tracto respiratorio de las personas, se han hecho estudios que demuestran la estrecha relación entre los niveles de $O_3$ y el grado de mortalidad de las personas que lo respiran Gobierno del D.F. (2010). En esta investigación se analiza el número de incidencias de los niveles $O_3$ que superan la Norma Oficial Mexicana (NOM-1993) de 0.11 ppm Correa (2004), usando un Procesos de Poisson no-homogéneos (PPNH) con función de intensidad log-lineal introduciendo a dicha función

[a]israel_loya@hotmail.com

las covariables: monóxido de carbono ($CO$), bióxido de nitrógeno ($NO_2$), bióxido de azufre ($SO_2$), temperatura ($TMP$), humedad relativa ($HR$), dirección del viento ($WDR$) y velocidad del viento ($WSP$). Los datos para este trabajo son tomados del Sistema de Monitoreo Atmosférico (SIMAT), particularmente se analizan los datos correspondientes al periodo comprendido de 2007 a 2009 de las estaciones: Merced, Pedregal, Plateros, Tlalnepantla y Xalostoc.

# 2. Marco teórico

El PPNH es un proceso estocástico que también es llamado un proceso de conteo, típicamente denotado por $\{N(t), t \geq 0\}$ donde $N(t)$ representa la variable aleatoria que denota el número total de sucesos de algún fenómeno aleatorio de interés que han ocurrido en el instante $t$.

El PPNH es usado a menudo para modelar fallas en sistemas en los cuales los eventos se presentan de forma rara, la diferencia entre un proceso de Poisson homogéneo del no homogéneo radica en que el primero depende de una función de intensidad constante $\lambda, \lambda > 0$, por otro lado el No Homogéneo tiene una función de intensidad que es función de $t$ denotada por $\lambda(t)$, $t \, \epsilon \, T$ donde $T$ generalmente es el tiempo Kingman (1993).

El proceso $\{N(t) : t > 0\}$, es un PPNH con función de intensidad $\lambda(t)$, $t > 0$ si cumple con lo siguiente Basawa y Prakasa (1980):

1. $N(t), t > 0$, tiene incrementos independientes.

2. $Pr\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h)$

3. $Pr\{N(t+h) - N(t) \geq 2\} = o(h)$,
   donde $o(h)$ es cualquier cantidad que después de dividirla por $h$ tiende a cero, como $h \to 0$

## 2.1. El criterio de información de Akaike (AIC)

Es una medida de bondad de ajuste para modelos estadísticos, propuesto por Hirotsugu Akaike Akaike (1980), también definido como un enfoque para seleccionar un buen modelo

de un conjunto de modelos a escoger definido como:

$$AIC = 2k - 2\ln(L), \tag{1}$$

donde $L$ denota la máxima log-probabilidad del modelo estimado, esto es la verosimilitud evaluada en el estimador y $k$, el número de parámetros estimados en el modelo aproximado Rao et al. (2008).

# 3. Aplicación

En este trabajo se usan los registros de tres años de 2007 a 2009 obtenidos del SIMAT y se modelan las incidencias por mes las cuales representan el valor máximo/día de ozono supera el umbral de 0.11 ppm, propuesto en NOM-1993, para tal efecto se utiliza el software estadístico R, usando la metodología siguiente:

## 3.1. Metodología

- Se obtienen los máximos diarios y se registran los tiempos (en días) en donde se presentan las incidencias (valores de $O_3 > 0.11$ ppm) por cada estación considerada, así como los valores de covariables químicas y meteorológicas.

- Se cuenta el número de las incidencias mensuales, i.e. valores de $O_3$ que superan la norma establecida por cada mes del año.

- Se proponen diferentes modelos tomando en cuenta el tiempo t, y combinaciones de covariables para encontrar el mejor modelo con base en el criterio de información de Akaike. En este proceso se usó la función glm() de R.

## 3.2. Modelo

Sea $Y_i$ la variable aleatoria que representa el número de incidencias de concentración de ozono que superan el umbral permitido en el mes $i$. Si las incidencias de ozono por mes se consideran eventos independientes con distribución Poisson, entonces la función de verosimilitud se

obtiene como:

$$\begin{aligned} L(\mathbf{y};\lambda(t,\mathbf{x})) &= f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i;\lambda_i(t,\mathbf{x})) \\ &= \prod_{i=1}^{n} \exp\{-\lambda_i(t,\mathbf{x}_i)\}\frac{\lambda_i(t,\mathbf{x}_i)^{y_i}}{y_i!} \end{aligned}$$

y la log-verosimilitud es:

$$\begin{aligned} l(\mathbf{y};\lambda(t,\mathbf{x})) &= \log L(\mathbf{y};\lambda(t,\mathbf{x})) \\ &= \sum_{i=1}^{n} \log\{\exp\{-\lambda_i(t,\mathbf{x}_i)\}\frac{\lambda_i(t,\mathbf{x}_i)^{y_i}}{y_i!}\} \\ &= \sum_{i=1}^{n} y_i \log\lambda_i(t,\mathbf{x}_i) - \sum_{i=1}^{n}\lambda_i(t,\mathbf{x}_i) - \sum_{i=1}^{n}\log y_i! \end{aligned} \tag{2}$$

Haciendo $\lambda_i(t,\mathbf{x}_i) = \exp\{\beta_0 + \beta_1 t_i + \cdots + \beta_k x_{ki}\}$, $i = 1, 2, \ldots, n$ y sustituyendo en la ecuación (2), se obtiene

$$l(\mathbf{y};\lambda(t,\mathbf{x})) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 t_i+\cdots+\beta_k x_{ki}) - \sum_{i=1}^{n}\exp\{\beta_0+\beta_1 t_i+\cdots+\beta_k x_{ki}\} - \sum_{i=1}^{n}\log y_i! \tag{3}$$

Finalmente, para obtener los estimadores de máxima verosimilitud de los parámetros del modelo $\beta_0, \beta_1, \ldots, \beta_k$ se maximiza a la función log-verosimilitud en (3).

# 4. Resultados y discusión

Se presentan modelos ajustados que describen en el estudio y el número de las incidencias de $O_3$ para cada estación meteorológica, sus gráficos ajustados se muestran detalladamente en la Figura 1:

- Merced. $log(\lambda(t,\boldsymbol{x})) = 4.112 + 0.049T - 4.219NO_2 + 24.745SO_2 - 0.019TMP$
- Pedregal. $log(\lambda(t,\boldsymbol{x})) = 3.876 + 0.0556T + 44.843SO_2 + 0.011TMP - 0.006HR$
- Plateros. $log(\lambda(t,\boldsymbol{x})) = 4.009 + 0.0564T + 28.923SO_2 + 0.0157TMP$
- Tlalnepantla. $log(\lambda(t,\boldsymbol{x})) = 3.408 + 0.051T - 4.783NO_2$
- Xalostoc. $log(\lambda(t,\boldsymbol{x})) = 1.436 + 0.059T + 19.367SO_2 + 0.0036TMP$

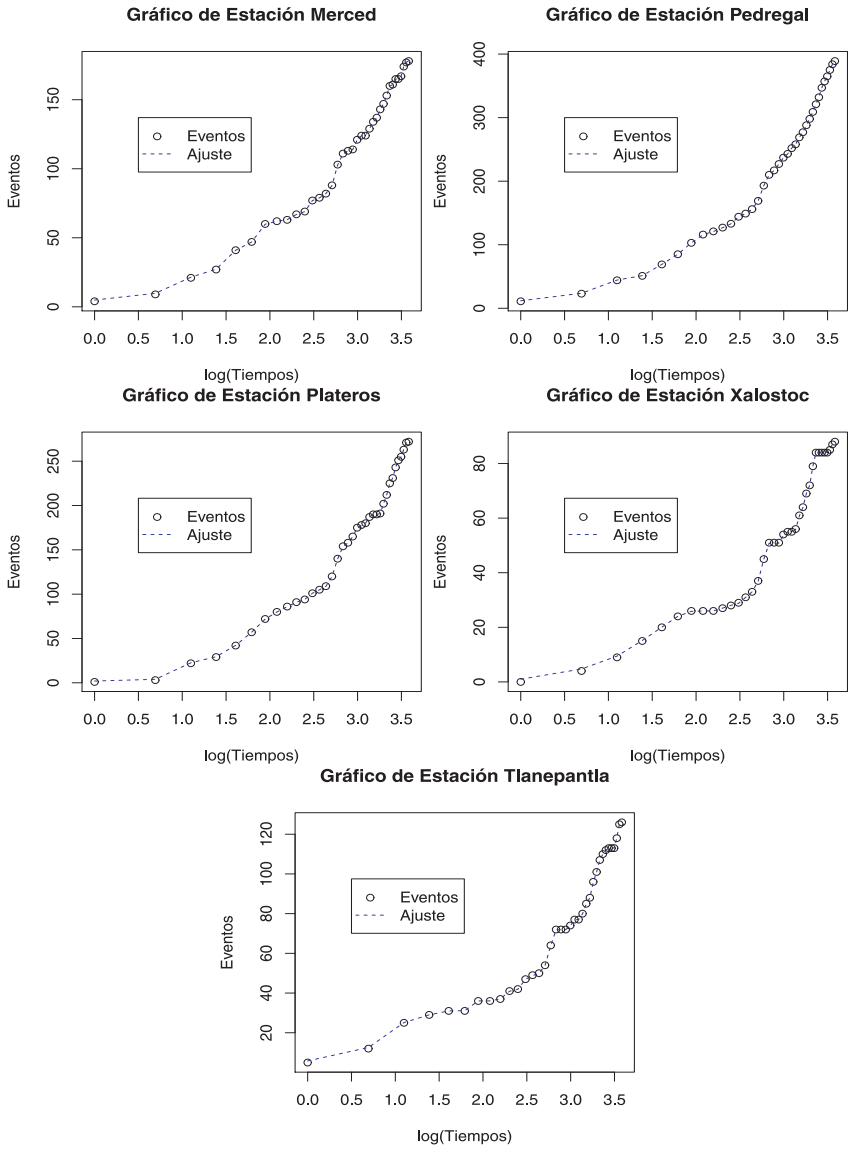



**Figura 1:** Resultados con ajuste. Se presentan los gráficos junto con los modelos ajustados para cada estación meteorológica.

# 5. Conclusiones

Se puede observar que el número de incidencias por mes pueden ser modelados mediante PPNH con función de intensidad loglineal dada en (3) en función del tiempo y ciertas covariables. Para el caso de la estación meteorológica Merced la tasa de ocurrencias está dada por

una constante, así como por el logaritmo del tiempo y se incluyen dos importantes contaminantes que son: $NO_2$ y $SO_2$, a la vez se observa que la TMP influye de forma significativa las incidencias por mes cuando el $O_3$ supera el umbral. En Pedregal el comportamiento esta dado con tasa de ocurrencias representado por una constante relacionada con el logaritmo del tiempo y afectada significativamente por $SO_2$, $TMP$ y $HR$. Plateros es una estación meteorológica en la cual el comportamiento no es semejante a Merced o Pedregal ya que la tasa de ocurrencias está representada por una constante el logaritmo del tiempo y los contaminantes $SO_2$ y la $TMP$. Tlalnepantla es una estación en la cual su comportamiento esta influenciado por una constante, el tiempo, así como el contaminante bióxido de nitrógeno favorece a incrementar las incidencias de ozono. Finalmente Xalostoc estación ubicada al noreste de la Ciudad de México se observa que en su comportamiento incide el tiempo, así como bióxido de azufre y la temperatura. En este estudio, observamos que las covariables que mayormente están presentes para explicar el comportamiento de las incidencias de ozono por mes son el $SO_2$ y $TMP$.

# Bibliografía

Akaike (1980), *Probabilidad y el procedimiento de Bayes, Estadística Bayesiana*, University Press.

Barrientos (2005), *Reporte cuatrimestral de contaminación*, INEGI.

Basawa y Prakasa (1980), *Statistical Inference for Stochastics Processes*, Academy Press.

Correa (2004), *Contaminantes atmosféricos en la zona metropolitana de la ciudad de México*, Instituto Politécnico Nacional.

Gobierno del D.F. (2010), *Sistema de Monitoreo Atmosférico del D.F.*, Gobierno del D.F.

Harrison (2003), *El medio ambiente: Introducción a la química medioambiental y la contaminación*, Acribia.

Kingman (1993), *Poisson Processes*, Oxford Science Publications.

Rao, Toutenburg, Shalabh y Heumann (2008), *Linear Models and Genera-lizations*, Springer.

Memoria FONE XXV (2010), Sección V, 129–135

# Credit Scoring: una aplicación de la estadística*

Soraida Nieto Murillo[a], Blanca Rosa Pérez Salvador
*Universidad Autónoma Metropolitana – Iztapalapa*

José Fernando Soriano Flores
*Subgerente de Riesgo, Bancomer*

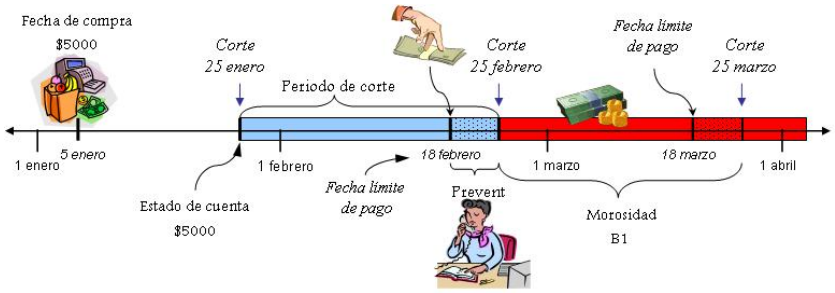
## 1. Introducción

En el país existen aproximadamente 40 millones de tarjetas de crédito, de las cuales 26.5 millones son emitidas por bancos y el resto por otro tipo de establecimientos, como las cadenas comerciales, esto hace que las tarjetas de crédito sean uno de los principales instrumentos de crédito en nuestro país. La cifra de cartera vencida en tarjetas de crédito, es ahora equivalente a 9.9 % del total prestado y esta cifra va en aumento, (Castillo Sánchez (2008)) este dato indica el enorme problema que es para las instituciones de crédito aprobar un préstamo a personas que no van a pagarles. Generalmente las instituciones de crédito compran modelos estadísticos a empresas extranjeras a un alto costo. Por todo lo anterior es prometedor que se desarrollen los modelos de Credit Scoring en México, por la demanda existente.

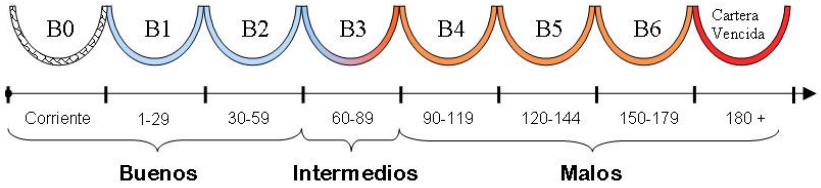
## 2. Tarjeta de crédito: Dinámica para caer en moratoria

Una tarjeta de crédito permite adquirir bienes y servicios hoy y pagarlos después; pero, si por alguna razón el pago no se realiza entonces se entra en mora y la empresa otorgante del crédito comienza a cobrar intereses y comisiones. La dinámica de la mora en una tarjeta de crédito se esquematiza en la siguiente figura.

*Este trabajo se realizó mientras la primera de las autoras era becaria de Conacyt
[a]gussynm@hotmail.com




En este sentido los clientes se clasifican respecto a la mora que presentan sus créditos en diferentes estados o canastas (bucket), esto se esquematiza en la siguiente figura.




El paso de una canasta a otra se puede modelar mediante una matriz de transición cuyos elementos son las probabilidades que en el tiempo $i+1$ se encuentre en la canasta $j$ dado que en el tiempo $i$ está en la canasta $k$.

$$p^{(n)} = P(X_{i+1} = j \mid X_i = k)$$

En el medio financiero a esta matriz de transición se conoce como Roll Rate. La Roll Rate permite identificar la evolución de la cuentas y discriminar a los clientes en dos grupos, los clientes buenos y los clientes malos.

# 3. ¿Qué es el credit scoring?

El credit scoring es una exitosa colección de técnicas estadísticas que se han venido utilizando durante más de 40 años como instrumento para otorgar créditos en la industria del crédito, esto ha permitido el crecimiento del número de consumidores de crédito, crecimiento que ha sido propiciado por el uso de la computadora y las técnicas estadísticas para el manejo de grandes cantidades de datos (Siddiqui (2006)). El credit scoring es una técnica muy rentable, dado que una pequeña mejora en el desempeño puede significar un incremento en las ganancias de los prestamistas.

### 3.1. Tipos de score

Existen tres etapas en el proceso de otorgar un crédito y dependiendo en qué parte del ciclo estemos trabajando, ver figura de abajo, se calcula uno de los siguientes puntajes de score:

- *Aquisition Score* o Score de adquisición. En el departamento de Originación se utiliza este puntaje para la aceptación o rechazo de las solicitudes de crédito con base en variables demográficas y de buró de crédito. Este puntaje estima la probabilidad de incumplimiento de pago de un posible cliente y de esta manera se decide si se acepta o rechaza como posible consumidor de crédito, optimizando la tasa de aprobación de las solicitudes.


- *Behavior score* o Score de comportamiento. Es utilizado en la etapa de administración del ciclo de riesgo. Predice la probabilidad de incumplimiento de los clientes que ya son objeto de crédito en la institución con base en las variables de comportamiento de las cuentas dentro de la propia institución. Permite dar seguimiento al comportamiento de los clientes para que sigan siendo rentables para la empresa.

- *Colection score*. Puntaje que se calcula en la parte de recuperación de cuentas para estimar la probabilidad de recuperar a un cliente. Las variables utilizadas resultan de la combinación de variables de comportamiento y buro de crédito.

## 4. La scorecard

Una scorecard es una tabla que contiene los puntajes asignados a cada atributo de todas las variables. El puntaje determina, por ejemplo, la probabilidad de que un cliente pague la deuda si se le otorgue una tarjeta de crédito. En esta tabla los mayores puntajes corresponden a una mayor probabilidad de pago (Simbaqueba (2004) y Thomas et al (2002)).

Para construir la scorecard se efectua una regresión logística en la cual la variable dependiente $Y$ es de la forma $Y = 0$ para un cliente malo y $Y = 1$ para un cliente bueno. Las variables explicativas $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_p)$ corresponden al cociente de la proporción de

buenos entre la proporción de malos en cada atributo de las diferentes variables demográficas y de buro de crédito. De esta manera se efectúa la regresión

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1^T \mathbf{x}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

donde $p_i = P(Y_i = 1 \mid \mathbf{x}_i)$ para $i = 1,\ 2,\ \ldots,\ n$ y $\beta_1 = (\beta_1,\ \beta_2,\ \ldots,\ \beta_p)$. Los valores en la scorecard para el atributo $j$ de la variable $i$ es igual a $Score_{ij} = Factor\hat{\beta}_i x_{ij}$, además se define un puntaje inicial dado por $Offset + Factor\hat{\beta}_0$, los términos $Offset$ y $Factor$ son dos números que estandarizan en cierto sentido al score. La siguiente tabla corresponde a una Scorecard simple con cinco variables o atributos: edad, estado civil, antigüedad en el empleo, sexo y nivel de estudios

| **Características Edad** | | | | |
|---|---|---|---|---|
| **Atributos** | Menor a 24 años | 25 - 30 años | 31 - 40 años | Mayor a 40 años |
| **Score** | $-40$ | $-28$ | 10 | 30 |
| **Características Estado Civil** | | | | |
| **Atributos** | Casado | Soltero | Otros | |
| **Score** | 12 | 0 | $-60$ | |
| **Características Antigüedad Empleo** | | | | |
| **Atributos** | 0 - 1 años | 2 - 5 años | 6 - 10 años | mayor a 10 años |
| **Score** | $-5$ | 4 | 10 | 15 |
| **Características Sexo** | | | | |
| **Atributos** | Masculino | Femenino | | |
| **Score** | $-10$ | 8 | | |
| **Características Nivel Estudios** | | | | |
| **Atributos** | Medio | Básica | Profesionista | |
| **Score** | 3 | 20 | 27 | |

En esta tabla se puede ver que un individuo con 37 años de edad, soltero, con 5 años de antigüedad en su empleo, masculino y profesionista tendrá un puntaje (score) de $41 = 10 + 0 + 4 - 10 + 27$. La premisa más importante de un score es que, a mayor puntaje de score menos riesgoso es el cliente y viceversa.

El puntaje de separación del escore entre clientes buenos y malos, (puntaje arriba del cual un cliente se hará acreedor de un crédito) o *cut off* es indicado en principio por los analistas de credit score pero se verá influido por las decisiones gerenciales o se basará en las metas corporativas de la propia institución.

# 5. Eficacia de la score card

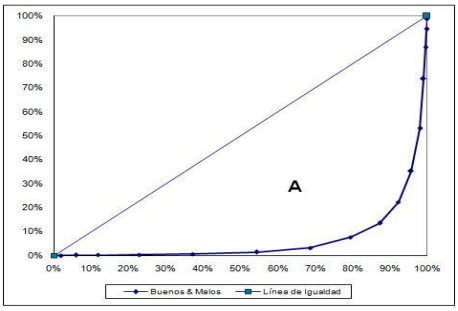
Una vez que se construye la score card debe medirse su eficacia, o el grado de separación de las dos poblaciones dada por el score. Para medir esta eficacia se puede utilizar el coeficiente de Gini, la divergencia o la prueba de Kolmogorov Smirnov que miden la diferencia entre las dos poblaciones (clientes buenos y clientes malos).

## 5.1. Coeficiente de Gini

El coeficiente de Gini se obtiene con la curva de Lorenz de las funciones de distribución asociada al score de los clientes malos $F(x)$, y la distribución asociada a los clientes buenos, $G(x)$. La curva de Lorenz es el conjunto

$$\mathcal{L}(F,G) = \{(u,\ v) \mid u = F(x) \ \text{ y } \ v = G(x), \ \text{ con } \ x \in \Re\}$$

Cuando $F(x) = G(x)$ las dos poblaciones tienen su escore idénticamente distribuidos y por lo tanto el score no separa a las dos poblaciones. Mientras mayor sea la diferencia entre $F(x)$ y $G(x)$ mayor será el grado de separación las dos poblaciones dado por el score.




En vista que la media del score en la población de clientes buenos es mayor que en la población de clientes malos, se satisface la relación $G(x) \leq F(x)$ por lo que, la curva de Lorenz de $F(x)$ y $G(x)$ es una gráfica que está por debajo de la gráfica de la función identidad. La figura a la derecha muestra una curva de Lorenz que separa radicalmente a las dos poblaciones. Por esta razón, el área $A$ que se encuentra entre la identidad y la curva de Lorenz es una medida de la desigualdad entre las distribuciones $F$ y $G$. El coeficiente de Gini resulta de la razón del área $A$ entre el área del triángulo delimitado por la identidad, el eje $x$ y la linea $y = 1$

## 5.2. La divergencia

Cuando construimos un modelo estadístico que clasifica dos poblaciones, se espera que los dos grupos estén estadísticamente bien separados, o que la diferencia entre sus medias sea importante. Esta diferencia se mide con la divergencia dada por expresión:

$$Divergencia = \frac{2(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

## 5.3. Prueba de Kolmogorov Smirnov

Las hipótesis de la prueba de Kolmogorov Smirnov son :

$$H_0\text{: } F(x) = G(x) \quad \text{contra} \quad H_1\text{: } F(x) \neq G(x).$$

La decisión se toma con la estadística $D = \text{máx}_{i \leq i \leq p} |F(x_{i1}) - G(x_{i2})|$ y se rechaza $H_0$ cuando $D$ está por arriba de un valor crítico.

# 6. Conclusiones

Las técnicas de credit scoring son el resultado de la combinación de ciencias y arte. Su implementación es resultado de un trabajo en equipo de los implicados en el proceso de generación e implementación. La aplicación de esta técnica es redituable por los beneficios económicos que produce a las empresas crediticias.

# Bibliografía

E. Castillo Sánchez Mejorada, Presidente de la Asociación de Bancos de México (jueves 11 de diciembre de 2008), "Cada día caen en cartera vencida unos 3 mil 305 préstamos al consumo", *La Jornada. Nacional, sección Economía, nota de Roberto González Amador* . *http://www.jornada.unam.mx/2008/12/11/)

Lilian, S. (2004), *¿Que es el scoring? Una visión práctica de la gestión del riesgo de crédito*, Instituto del Riesgo Financiero.

Lyn, T., Edelman, D. y Crook, J. (2002), *Credit Scoring and its applications*, SIAM, Philadelphia.

Siddiqi, N. (2006), *Credit Risk Scorecards: developing and implementing intelligent credit scoring*, John Wiley & Sons, New Jersey.

Memoria FONE XXV (2010), Sección V, 137–142

# Un modelo de series de tiempo para describir la demanda en grupos escolares con seriación estricta

Blanca Rosa Pérez Salvador[a]
*Universidad Autónoma Metropolitana-Iztapalapa*

## 1. Introducción

La demanda estudiantil en una escuela o facultad cuyas materias se encuentran seriadas se puede pronósticar mediante un modelo autoregresivo de serie de tiempo multivariado. La demanda estudiantil de una materia en un periodo lectivo fijo está en función del número de estudiantes que han acreditado las materias previas, así como del número de estudiantes que han reprobado esa misma materia. En este trabajo se presenta un modelo autoregresivo de series de tiempo multivaluado con el que se describe la demanda para un grupo de materias que presentan seriación estricta.

## 2. El problema

Los planes de estudios universitarios tienen una estructura semejante a la presentada en la figura 1, en esta figura las lineas representan la seriación que siguen las materias.

Considere que se tienen un plan de estudios con $m$ materias, $M_1$, $M_2$, ..., $M_m$; y sea $X_t^k$ la variable que indica el número de estudiantes que se inscriben en la materia $k$ en el periodo lectivo $t$, $k = 1, 2, \ldots, m$; estas variables forman una serie de tiempo autoregresiva formada de la siguiente manera. Cuando $M_i$ es materia de **primer ingreso**, los estudiantes que se inscriben en ella en el periodo $t$ son los que ingresan en ese periodo $t$ más los estudiantes que habiendo reprobado en los periodos $t-1$ a $t-p$, y se inscriben nuevamente a esa materia

[a]psbr@xanum.uam.mx

en el periodo $t$. En el modelo presentado se considera que en todos los periodos hay nuevo ingreso, se puede generalizar a a un modelo cíclico.

$$X_t^i = \mu_0 + \psi_{i1}X_{t-1}^i + \psi_{i2}X_{t-2}^i + \ldots + \psi_{ip}X_{t-p}^i + W_t^i$$

$\mu_0$ corresponde a la matricula de primer ingreso. Dado que la mayoría de los planes de estudio límita a los estudiantes a inscribirse a lo mas dos veces en una misma materia es adecuado considerar que esto determina el valor de $p$.

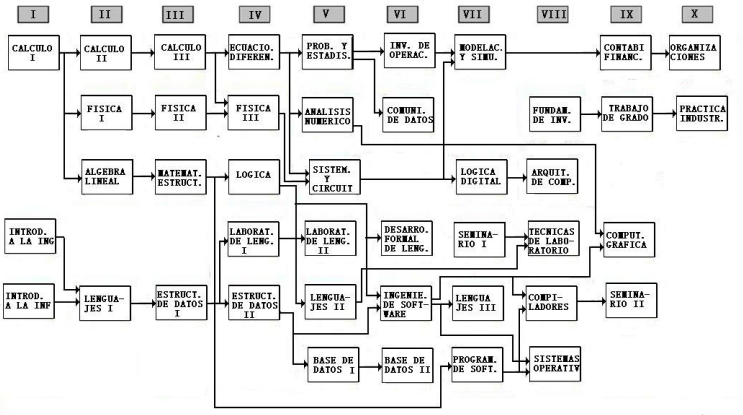



Figura 1. Plan de estudios de una escuela

Por otro lado, si $M_i$ no es una materia, **de primer ingreso**, que tiene como antecedente la materia $M_j$, $j < i$, entonces se satisface la relación

$$X_t^i = \psi_{i1}X_{t-1}^i + \psi_{i2}X_{t-2}^i + \ldots + \psi_{ip}X_{t-p}^i + \varphi_{i1}X_{t-1}^j + \varphi_{i2}X_{t-2}^j + \ldots + \varphi_{ip}X_{t-p}^j + W_t^i$$

Los términos $\psi_{ik}X_{t-k}^i$ corresponden al número de estudiantes que cursaron la materia $M_i$ en el periodo $t-k$, no la aprobaron, y se incriben nuevamente a $M_i$ en el periodo $t$. Los términos $\varphi_{ik}X_{t-k}^j$ corresponden al número de estudiantes que en el periodo $t-k$ cursaron la materia $M_j$ y se inscriben en la materia $M_i$ en el periodo $t$; $W_t^i$ corresponde a la aleatoriedad o ruido blanco de la serie en el periodo $t$ y $p$ corresponde al número de rezagos que aún pueden inscribirse a la materia. Si la materia $M_i$ está seriada con dos o más materias, el modelo se complica un poco más, este caso no lo tratamos en este escrito.

Los parámetros $\varphi_{ik}$ y $\psi_{ik}$ corresponden a una proporción, por lo que son positivos menores que 1, y se esperaría que decrecieran conforme el rezago $k$ crece.

Para un programa escolar de $m$ materias de las cuales $s$ son del primer periodo lectivo, se define el vector dependiente del tiempo $X_t = (X_t^1, X_t^2, X_t^3, \ldots, X_t^m)$, que satisface el modelo autoregresivo de orden $p$

$$X_t = \boldsymbol{\mu}_0 + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots + \Phi_p X_{t-p} + W_t \tag{1}$$

Los elementos de las matrices $\Phi_k = \{\phi_{ij}^k\}$, $k = 1, 2, \ldots, p$, son:

$$\phi_{ij}^k = \begin{cases} \varphi_{ik} & \text{si } M_i \text{ está seriada con } M_j \\ \psi_{ik} & \text{si } i = j \\ 0 & \text{en otro caso} \end{cases}$$

y los elementos del vector $\boldsymbol{\mu}_0 = \{\mu_i^0\}$ son:

$$\mu_i^0 = \begin{cases} \mu_0 & \text{si} \quad i = 1,\ 2,\ \ldots,\ s \\ 0 & \text{en otro caso} \end{cases}$$

El total de parámetros del modelo autoregresivo son $p(2m - s) + 1$; $pm$ de la forma $\psi_{ik}$, $i = 1,\ 2,\ \ldots, m$ y $k = 1,\ 2,\ \ldots, p$; $p(m - s)$ de la forma $\varphi_{ik}$, $i = s + 1,\ s + 2,\ \ldots, m$ y $k = 1,\ 2,\ \ldots, p$, y finalmente el parámetro $\mu_0$, este último parámetro es conocido.

# 3. Supuestos del modelo

Si el número de estudiantes de primer ingreso, $\mu_0$, se mantiene constante a través del tiempo, la serie autoregresiva de orden $p$ se puede suponer estacionaria en media y covarianza, esto es, la media y la covarianza de $X_t$ no depende del tiempo, por lo que

- $E(X_t) = \boldsymbol{\mu}$,
- $V(X_t) = E(X_t - \boldsymbol{\mu})(X_t - \boldsymbol{\mu}) = \Gamma_0$, y
- $Cov(X_t, X_{t+k}) = E(X_t - \boldsymbol{\mu})(X_{t+k} - \boldsymbol{\mu}) = \Gamma_k$. La covarianza solo depende del rezago $k$.

Dado que la multiplicación de matrices no es conmutativa, se tiene que $\Gamma_k$ no necesariamente es simetrica; sin embargo, es fácil probar que $\Gamma_k = \Gamma_{-k}^T$.

De esta manera, la matriz de autocorrelaciones es igual a

$$Corr(X_t, X_{t+k}) = R_k = D^{-1}\Gamma_k D^{-1}$$

donde $D$ es una matriz diagonal con elementos iguales a $var(X_t^i)^{1/2}$

El vector de ruido blanco, $W_t$, satisface las relaciones:

- $E(W_t) = 0$,
- $V(W_t) = \Upsilon$
- $Cov(W_t, W_{t+k}) = 0$
- y $Cov(W_t, X_t) = Cov(X_t, W_t) = \Upsilon$.

# 4. Estimación

Debido a que las matrices $\Gamma_k$ no son simétricas, no es posible utilizar la generalización de las ecuaciones de Yule-Walker para estimar los párametros del modelo, ver (Benkwitz (2000)).

Se puede utilizar, como primera opción, el método de estimación de mínimos cuadrados para obtener los estimadores de las matrices $\Phi_k$, $k = 1, 2, \ldots, p$, aplicándoselo a las ecuaciones de las coordenadas de la serie multivariada, ver Frühwirth-Schnatterv et al, (2004) Liu, et al, 2005.

Una alternativa a los mínimos cuadrados, para estimar las matrices $\Phi_k$ es aprovechar la simetría de la matriz de varianza covarianza, en la relación $\Gamma_0 = E(X_t - \boldsymbol{\mu})(X_t - \boldsymbol{\mu})^T$ cuando se sustituye $X_t$ por $\boldsymbol{\mu}_0 + \Phi_1 X_{t-1} + \ldots + \Phi_p X_{t-1} + W_t$, en el desarrollo de $\Gamma_0$.

$$\Gamma_0 = E(\boldsymbol{\mu}_0 + \Phi_1 X_{t-1} + \ldots + \Phi_p X_{t-1} + W_t - \boldsymbol{\mu})(X_t - \boldsymbol{\mu}) = \Phi_1\Gamma_1 + \ldots + \Phi_p\Gamma_p + \Upsilon,$$

de esta relación, por la simetría de $\Gamma_0$, se obtienen $m(m-1) \div 2$ ecuaciones lineales de los elementos de $\Phi_k$ (de las ecuaciones $\gamma_{ij}^0 = \gamma_{ji}^0$, $i \neq j$).

Si el número de variables en el modelo es menor que el número de ecuaciones $\gamma_{ij}^0 = \gamma_{ji}^0$, se puede encontrar solución para $\Phi_k$ en terminos de los elementos de $\Gamma_k$, esto es posible cuando los números $p$, $m$ y $s$, satisfacen la relación $p(2m - s) < m(m-1) \div 2$.

Los estimadores de $\Phi_k$ se encuentran en función de los estimadores de $\Gamma_k$. Los parámetros $\boldsymbol{\mu}$, $\Gamma_k$ y $R_0$ se estiman de los datos $X_1$, $X_2$, …, $X_T$ usando los momentos muestrales

- $\hat{\boldsymbol{\mu}} = \overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$

- $\hat{\Gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(X_t - \overline{X})(X_{t+k} - \overline{X})^T$
- $\hat{\Gamma}_{-k} = \frac{1}{T}\sum_{t=k+1}^{T}(X_t - \overline{X})(X_{t-k} - \overline{X})^T$
- $\hat{R}_k = \hat{D}^{-1}\hat{\Gamma}_k\hat{D}^{-1}\frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})(X_{t+k} - \overline{X})^T$

## Ejemplo

Suponemos que se tienen tres materias seriadas de la siguiente manera,

$$M_1 \ \rightarrow \ M_2 \ \rightarrow \ M_3$$

y el modelo propuesto es un $AR(1)$, esto es $p = 1$, Las coordenadas del vector de la serie de tiempo son: para $i = 1$

$$X_t^1 = \mu_0 + \psi X_{t-1}^1 + W_t^1$$

Para $i = 2,\ 3$

$$X_t^i = \psi X_{t-1}^i + \varphi X_{t-1}^{i-1} + W_t^i$$

Estas relaciones las podemos escribir matricialmente como

$$\begin{pmatrix} X_t^1 \\ X_t^2 \\ X_t^3 \end{pmatrix} = \begin{pmatrix} \mu_0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & 0 \\ \varphi_2 & \psi_2 & 0 \\ 0 & \varphi_3 & \psi_3 \end{pmatrix} \begin{pmatrix} X_{t-1}^1 \\ X_{t-1}^2 \\ X_{t-1}^3 \end{pmatrix} + \begin{pmatrix} W_t^1 \\ W_t^2 \\ W_t^3 \end{pmatrix}$$

Entonces, se tiene que la matriz de varianza covarianza es

$$\Gamma_0 = \Phi\Gamma_1 + \Upsilon$$

Dado que $\Gamma_0$ y $\Upsilon$ son matrices simétricas, entonces $\Phi\Gamma_1$ también debe ser simétrica. De aquí se sigue que de la matriz

$$\Phi\Gamma_1 = \begin{pmatrix} \psi_1\gamma_{11}(1) & \psi_1\gamma_{12}(1) & \psi_1\gamma_{13}(1) \\ \varphi_2\gamma_{11}(1) + \psi_2\gamma_{21}(1) & \varphi_2\gamma_{12}(1) + \psi_2\gamma_{22}(1) & \varphi_2\gamma_{13}(1) + \psi_2\gamma_{23}(1) \\ \varphi_3\gamma_{21}(1) + \psi_3\gamma_{31}(1) & \varphi_3\gamma_{22}(1) + \psi_3\gamma_{32}(1) & \varphi_3\gamma_{23}(1) + \psi_3\gamma_{33}(1) \end{pmatrix}$$

Se obtienen tres ecuaciones lineales:

$$\psi_1\gamma_{12}(1) = \varphi_2\gamma_{11}(1) + \psi_2\gamma_{21}(1)$$

$$\psi_1\gamma_{13}(1) = \varphi_3\gamma_{21}(1) + \psi_3\gamma_{31}(1)$$
$$\varphi_2\gamma_{13}(1) + \psi_2\gamma_{23}(1) = \varphi_3\gamma_{22}(1) + \psi_3\gamma_{32}(1).$$

Dado que hay 5 variables ($\varphi_1,\ \varphi_2,\ \varphi_3,\ \psi_1,\ \psi_2$) y tres ecuaciones, el sistema no tiene solución única.

Si se tienen 5 materias seriadas de la forma

$$M_1 \ \rightarrow \ M_2 \ \rightarrow \ M_3 \ \rightarrow \ M_4 \ \rightarrow \ M_5$$

con $p = 1$, entonces el número de parámetros del modelo es $5 \times 2 - 1 = 9$ y el número de ecuaciones es $5 \times 4 \div 2 = 10$. Lo que implica que el sistema de ecuaciones tiene solución.

Si $m = 12$, $s = 4$ y $p = 3$ se tienen 60 parámetros y 66 ecuaciones, con lo que es factible obtener una solución para $\hat{\Phi}_k$.

# 5. Conclusiones

El modelo autoregresivo multivariado de orden $p$ es una opción para describir el comportamiento de la matricula a lo largo de la historia académica en una institución educativa, por lo que puede ser una herramienta útil para pronósticar la demanda estudiantil y planificar el número de grupos que se deben abrir en un periodo lectivo determinado. Es importante probar los supuestos considerados y evaluar el modelo, este es un trabajo que se deja para el futuro.

# Bibliografía

Alexander, B. (2000), *Multiple time series analysis*, Sonderforschungsbereich.

Frühwirth-Schnatterv y Sylvia, K. (2004), "Model-based clustering of multiple time-series", *Centre for Economic Policy Research (Great Britain), Euro Area Business Cycle Network for Economic Policy Research* pp. 1–28.

Liu, X., Swift, S., Tucker, A., Cheng, G. y Loizou, G. (2005), *Modelling Multivariate Time Series*, Department of Computer Science,Birkbeck College, University of London,Malet Street, London WC1E 7HX, United Kingdom. *citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.939&rep...)

Memoria FONE XXV (2010), Sección V, 143–148

# Estudio de factores que influyen en la resistencia de los morteros formulados para reparación de vivienda de interés social en la zona costera de Guerrero

Alfredo Cuevas Sandoval[a], Flaviano Godínez Jaimes[b]

Sulpicio Sánchez Tizapa[c]

*Unidad Académica de Ingeniería,Universidad Autónoma de Guerrero*

## 1. Introducción

El mortero es una mezcla de cementante, arena y agua, que tiene la propiedad de fraguar tanto en el aire como en el agua, NMX-C-021 (2004). Esta mezcla se aplica en las construcciones con la finalidad de dar mayor rigidez, evitar deterioro y filtraciones de agentes del medio ambiente. El mortero es el material de construcción de mayor uso desde la antigüedad. Su elaboración ha evolucionado desde el conocimiento empírico al uso de procesos científicos y técnicos considerando los componentes para su elaboración, el proceso de fabricación y la puesta en obra.

A pesar de los adelantos técnicos en la tecnología del mortero para la construcción, es común encontrar problemas relacionados a la exposición con el medio ambiente después de cierto tiempo en servicio como son: daños en la estructura y apariencia del mortero. Dichos daños son fisuras, agrietamientos, manchas, eflorescencias, fragmentación y desgaste en los aplanados, en particular en viviendas de interés social.

[a] acuevas36@hotmail.com
[b] fgodinezj@gmail.com
[c] *ss*tizapa@hotmail.com

En la zona costera del Estado de Guerrero se ha encontrado en viviendas de interés social daños en los aplanados, generados en un lapso de tiempo de 10 a 15 años, en fraccionamientos de las ciudades de Acapulco con 44.3 % de daños en 210 viviendas, Coyuca de Benítez con 48.10 % de daños en 104 viviendas y en Zihuatanejo con 64.6 % de daños en 80 viviendas de interés social Castillo Díaz (2011). Esta situación ha generado preocupación en sus moradores, debido a la posible falla en la seguridad estructural y la pérdida de valor de su bien inmueble.

La rehabilitación y reparación de daños en este tipo de viviendas ha llevado a los encargados a buscar la solución más adecuada en cada lugar donde se ubican las viviendas. La solución abarca desde el diseño de mezcla, proceso de elaboración y manipulación y sistema de puesta en obra.

En este trabajo se analiza una propuesta de elaboración del mortero para la reparación de viviendas utilizando arena del banco más próximo a la ubicación del fraccionamiento de viviendas con daños para que el costo del mortero sea bajo. Acorde con la literatura se considera en la propuesta el uso de fibras artificiales en la mezcla para incrementar su desempeño y resistencia a compresión y de la aplicación de uno de los métodos de curado (hidratación del mortero para que continúen su reacción química el cemento y el agua).

## 2. Materiales y Métodos

Los materiales utilizados para desarrollar el diseño experimental fueron colectados por personal del laboratorio de acuerdo a los criterios de la Norma mexicana, NMX-C-030 (2002). Para la arena se tomaron muestras simples en diferentes puntos del banco. Los bancos muestreados para suministro de las reparaciones fueron: banco Papagayo para viviendas en Acapulco, banco Coyuca para viviendas de Coyuca de Benítez y banco Barrio Nuevo para viviendas en Zihuatanejo. Los tres bancos de arena cumplen la norma mexicana vigente,**?**. Para no introducir variabilidad en el experimento el agua para elaborar la mezcla fue potable y el cementante fue cemento porltland CPC 30 R NMX-C-414 (2004). Los factores estudiados son *tipo de fibra* (sin fibra, de nylon, polipropileno y de vidrio) que se obtuvieron de proveedores de la región y *curado* (sin curado, agua de mar y con agua potable).

La variable respuesta es *resistencia a compresión* del mortero y se determinó de acuerdo

con las normas mexicanas vigentes,**?**. Bajo un diseño experimental de dos factores al azar Montgomery (2005), se elaboraron cubos de mortero de 5X5X5 cm, con dos replicas para las combinaciones de los niveles.

El modelo para estimar el efecto de los factores en cada caso es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk};$$

$$i = 1, 2, 3, 4; \quad j = 1, 2, 3; \quad k = 1, 2$$

donde $y_{ijk}$ es la variable de respuesta (resistencia a compresión del mortero), $\mu$ es la media general, $\alpha_i$ es el efecto del *tipo de fibra*, $\beta_j$ es el efecto de forma de *curado*, $(\alpha\beta)_{ij}$ es el efecto de la interacción tipo de fibra-forma de curado y $\varepsilon_{ijk}$ el error experimental.

# 3. Resultados y Discusión

El análisis se realizó con los paquetes computacionales SPSS v18 y SAS v9. En el modelo ajustado para cada caso (Acapulco, Coyuca de Benítez y Zihuatanejo), donde solo cambio el banco de arena en el modelo, los dos factores, *tipo de fibra* y *curado* son significativos a un $\alpha = .05$ pero no así la interacción *tipo de fibra*curado*. Esto es, el *tipo de fibra* utilizada genera un comportamiento con la misma tendencia en la resistencia del concreto con cada método de *curado*. Los modelos explican 88.0 % para banco Papagayo, 83.2 % para banco Coyuca y 88.9 % para banco Barrio Nuevo, de la varianza de la resistencia del mortero correspondientes a las ciudades de Acapulco, Coyuca de Benítez y Zihuatanejo.

En la Tabla 1A se presentan las comparaciones múltiples del factor *tipo de fibra* para el banco Papagayo, se observan 3 grupos significativamente diferentes, donde la resistencia mayor es de los morteros fabricados con fibra de vidrio y sin fibra. La resistencia más baja corresponde a los morteros hechos con fibra de nylon y fibra de polipropileno.

Para el banco Coyuca, se observan 2 grupos significativamente diferentes (Tabla 1B). La resistencia mayor es de los morteros fabricados con fibra de vidrio, polipropileno y nylon y la resistencia más baja corresponde a los morteros hechos sin fibra.

En la Tabla 1C se presentan las comparaciones múltiples para el banco Barrio Nuevo, se observan 2 grupos estadísticamente diferentes. La resistencia mayor pertenece a morteros fabricados con fibra de vidrio y fibra de polipropileno. La resistencia más baja corresponde a morteros elaborados sin fibra y fibra de nylon.

**Cuadro 1:** Clasificación de la resistencia por tipo de fibra para los tres bancos de suministro.

| | | A | | | B | | | C | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Papagayo | | | Coyuca | | | Barrio Nuevo | | |
| | | Subconjunto | | | Subconjunto | | | Subconjunto | | |
| Tipo de fibra | N | 1 | 2 | 3 | N | 1 | 2 | N | 1 | 2 |
| Sin fibra | 6 | | 158.5 | 158.5 | 6 | 143.9 | | 6 | 129.6 | |
| Nylon | 6 | 137.3 | | | 6 | | 169.3 | 6 | 138.9 | |
| Polipropileno | 6 | 139.5 | 139.5 | | 6 | | 168.6 | 6 | | 161.4 |
| Vidrio | 6 | | | 165.7 | 6 | | 170.6 | 6 | | 162.6 |
| Significación | | .988 | .064 | .715 | | 1.0 | .994 | | .504 | .998 |

Las comparaciones múltiples del factor *curado* para el banco Papagayo muestran 2 grupos significativamente diferentes. La mayor resistencia del mortero es con curado con agua potable y con agua de mar. Las resistencia más baja corresponde al mortero sin curado (Tabla 2A).

En la Tabla 2B se dan las comparaciones múltiples para el factor *curado* para el banco Coyuca en el que se observan dos grupos significativamente diferentes. La mayor resistencia del mortero es la que tuvo curado con agua potable y con agua de mar. Las resistencia más baja corresponde al mortero sin curado.

En la Tabla 2C se presentan las comparaciones múltiples del factor *curado* para el banco Barrio Nuevo, donde se observan 3 grupos estadísticamente diferentes. La resistencia mayor corresponde a mortero curado con agua potable. La resistencia más baja es de morteros no curados.

# 4. Conclusiones

Los factores *tipo de fibra* y *curado* son significativos para explicar la resistencia del mortero.

**Cuadro 2:** Clasificación de la resistencia por tipo de curado para los tres bancos de suministro.

| | A Papagayo Subconjunto | | | B Coyuca Subconjunto | | | C Barrio Nuevo Subconjunto | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Curado | N | 1 | 2 | N | 1 | 2 | N | 1 | 2 | 3 |
| Sin Curado | 8 | 127.8 | | 8 | 144.5 | | 8 | 129.8 | | |
| Con agua de Mar | 8 | | 154.5 | 8 | | 167.6 | 8 | | 146.2 | |
| Con agua potable | 8 | | 168.4 | 8 | | 177.2 | 8 | | | 162.6 |
| Significación | | 1.0 | .083 | | 1.0 | .351 | | 1.0 | 1.0 | 1.0 |

El mortero que tiene mayor resistencia a excepción del elaborado con arena del banco Papagayo es el mortero que contenga algún tipo de fibra. Se observa que dentro de los subconjuntos de los tres bancos de arena la resistencia más alta corresponde al mortero elaborado con fibra de vidrio. Se sugiere que se utilize en los trabajos de reparaciones, además de ser la más económica del mercado.

En cuanto al factor curado, es evidente que es imprescindible aplicar curado para obtener o garantizar la resistencia en el mortero. Se observa que las mayores resistencias se obtienen al aplicar curado con agua potable, por lo que se recomienda este proceso de curado. El agua de mar para aplicar curado muestra que el mortero obtendrá una resistencia adecuada, pero la mayoría de las especificaciones la prohíben, salvo casos especiales.

**Reconocimientos:** Este trabajo fue posible gracias al apoyo de la Arquitecto Urbanista Iris Lizet Castillo Díaz por facilitar los datos experimentales de su trabajo de grado, el Laboratorio de Materiales de la Unidad Académica de Ingeniería de la UAGro., y la ayuda recibida por los estudiantes prestadores de servicio social en la preparación y realización de ensayos.

# Bibliografía

Castillo Díaz, I. L. (2011), *Propuesta de reforzamiento con fibras sintéticas en aplanados de construcción de viviendas de interés social, en zonas costeras del estado de Guerrero.* Tesis de Maestría en proceso.

Cuevas, S., Godínez, J. y Santes, P. A. (2009), *Análisis del efecto de banco de arena, temporada de extracción y edad de prueba, en la resistencia del concreto hidráulico bajo un experimento factorial, Aportaciones y Aplicaciones de la Probabilidad y Estadástica*, Vol. 3, BUAP, México.

GDF-RCDF (2004), *Normas Técnicas Complementaria para diseño y construcción de estructuras de mampostería; Tomo I, Gaceta Oficial del DF*, México.

Kosmatka, Steven, H., Panarese, C. y William, C. (2005), *Diseño y control de mezclas de concreto*, Portland Cement Association, USA.

Montgomery, D. (2005), *Diseño y análisis de experimentos*, 2a edn, Limusa Wiley.

NMX-C-021 (2004), *Norma Mexicana Industria de la construcción-Cemento para albañilería (mortero)-Especificaciones y métodos de prueba*, ONNCCE, México.

NMX-C-030 (2002), *Norma Mexicana Industria de la construcción-Agregados-Muestreo*, ONNCCE, México.

NMX-C-061 (2004), *Norma Mexicana Industria de la construcción-Cemento-Determinación de la resistencia a la compresión de cementantes hidráulicos*, ONNCCE, Secretaría de Economía, México.

NMX-C-085 (2002), *Norma Mexicana Industria de la construcción-Cementos Hidráulicos-Especificaciones y métodos de prueba*, ONNCCE, México.

NMX-C-111 (2004), *Norma Mexicana Industria de la construcción-Agregado para cocnreto-Especificaciones y métodos de prueba*, ONNCCE, México.

NMX-C-414 (2004), *Norma Mexicana Industria de la construcción-Cementos Hidráulicos-Especificaciones y métodos de prueba*, México.

Memoria FONE XXV (2010), Sección V, 149–156

# Relación entre ecuaciones estructurales y correlación canónica en un experimento con guayule

Emilio Padrón Corral, Haydée de la Garza Rodríguez

Armando Muñoz Urbina
*Universidad Autónoma de Coahuila*

Ignacio Méndez Ramírez
*Universidad Nacional Autónoma de México*

## 1. Introducción

Los modelos de ecuaciones estructurales son técnicas principalmente confirmatorias dentro del análisis multivariado con el fin de determinar cuando un cierto modelo es apropiado, esta clase de trabajos son frecuentemente representados gráficamente y dicha técnica implica el uso de las matrices de correlación y de covarianza, y para cada modelo se evalúa el grado de ajuste a los datos. Por lo tanto el objetivo del presente trabajo es obtener un análisis de ecuaciones estructurales que muestre la contribución de las diferentes partes de la planta sobre el rendimiento de resina y hule, así como detectar su relación con el análisis de correlación canónica.

## 2. Metodología

En el presente análisis se muestrearon 35 plantas de guayule de aproximadamente dos años de edad, provenientes de una población silvestre del Ejido Gómez Farias ubicado a 56 $km.$ de Saltillo, Coahuila, México. Arroyo (1999). Estas plantas fueron seccionadas en raíz, corona,

ramas primarias y ramas secundarias; posteriormente se secaron en una estufa para obtener el peso seco, luego cada parte de la planta fue molida. Una muestra de 5 gramos de cada parte de tejido de las plantas fue utilizada para determinar el contenido de resina y hule, utilizando para ello tolueno y acetona como solventes. De las plantas seccionadas se estimaron las variables pesos secos de hule y resina, contenidos de hule y resina, altura, diámetro, además de pesos de hule y resina. Se utilizó el análisis de ecuaciones estructurales y el de correlación canónica, con el fin de estimar los componentes de rendimiento de resina y hule, dentro del análisis de ecuaciones estructurales se obtuvieron los coeficientes de sendero los cuales se estimaron de acuerdo a Cox y Wermuth (1996). Los modelos saturados se ajustaron con el software EQS; Bentler (1995).

# 3. Resultados y Discusión

Al aplicar el análisis de ecuaciones estructurales para analizar los datos sólo se presentan las ecuaciones que contienen las componentes de rendimiento de resina y hule, la ecuación para peso de resina por planta (PR/PL) dada por el paquete estadístico es:

$$\begin{aligned} PR/PL &= 0.227(PRRS) + 0.470(PRRP) \\ &\quad +0.247(PRCOR + 0.212(PRRAI) + 0.002(E_{15}) \qquad ...(1) \end{aligned}$$

Dicha ecuación (1) tuvo un coeficiente de determinación de $R^2 = 0.999996$ el cual contribuye en un 99.9996 porciento de la variación explicada por dicho modelo. En el Cuadro 1 se observa que el efecto directo de PRRP(0.470), contribuye en una mayor cantidad que los efectos directos de PRRS(0.227), PRCOR(0.247) y PRRAI(0.212), con respecto a los efectos indirectos se observa que PRCOR y PRRAI, presentan altos valores a través de PRRP.

En Figura 1, se puede observar que el efecto indirecto de PRCOR a través de PRRP se obtiene del producto de la correlación entre las variables predictoras PRCOR con PRRP(r=0.846), y el efecto directo de la variable PRRP(b = 0.470) sobre PR/PL es decir, (0.846)(0.470) = 0.397. La ecuación para peso de hule por planta (PH/PL) es:

$$\begin{aligned} PH/PL &= 0.187(PHRS) + 0.366(PHRP) \\ &\quad +0.355(PHCOR) + 0.287(PHRAI) + 0.003(E_{15}) \qquad ...(2) \end{aligned}$$

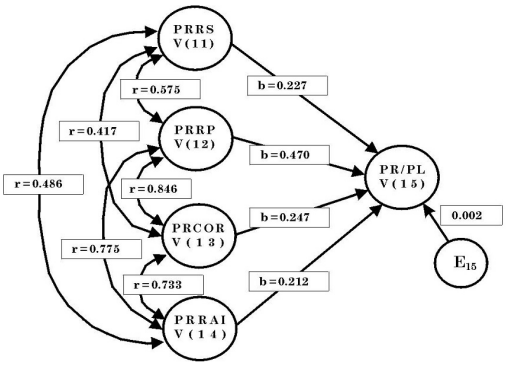



**Figura 1:** Efectos directos $(b)(\rightarrow)$ y correlación $(r)(\leftrightarrow)$.

| | Efectos directos e indirectos | | | | Correlaciones de PRRS(11) |
|---|---|---|---|---|---|
| | PRRS(11) | PRRP(12) | PRCOR(13) | PRRAI(14) | PRRP(12),PRCOR(13) y PRRAI(14) con PR/PL(15) |
| PRRS(11) | **0.227** | 0.270 | 0.103 | 0.103 | $r_{11,15} = 0.6955^{**}$ |
| PRRP(12) | 0.130 | **0.470** | 0.209 | 0.164 | $r_{12,15} = 0.963^{**}$ |
| PRCOR(13) | 0.094 | 0.397 | **0.247** | 0.156 | $r_{13,15} = 0.885^{**}$ |
| PRRAI(14) | 0.110 | 0.364 | 0.181 | **0.212** | $r_{14,15} = 0.858^{**}$ |

Residual=0.002 **Significativo al 1 %

**Cuadro 1:** Efectos directos (negro) e indirectos entre PRRS, PRRP, PRCOR y PRRAI y sus correlaciones con PR/PL

| | Efectos directos e indirectos | | | | Correlaciones de PHRS(16) PHRP(17),PHCOR(18) y PHRAI(19) con PH/PL(20) |
|---|---|---|---|---|---|
| | PHRS(16) | PHRP(17) | PHCOR(18) | PHRAI(19) | |
| PHRS(16) | **0.187** | 0.194 | 0.075 | 0.086 | $r_{16,20} = 0.536^{**}$ |
| PHRP(17) | 0.099 | **0.366** | 0.277 | 0.218 | $r_{17,20} = 0.943^{**}$ |
| PHCOR(18) | 0.039 | 0.286 | **0.355** | 0.239 | $r_{18,20} = 0.897^{**}$ |
| PHRAI(19) | 0.056 | 0.279 | 0.296 | **0.287** | $r_{19,20} = 0.898^{**}$ |

Residual=0.003

**Cuadro 2:** Efectos directos (negro) e indirectos entre PHRS, PHRP, PHCOR y PHRAI y sus correlaciones con PH/PL.

Dicha ecuación (2), tuvo un coeficiente de determinación de $R^2 = 0.999991$ el cual contribuye en un 99.9991 porciento de la variación explicada por dicho modelo. En el Cuadro 2, se observa que los efectos directos de PHRP(0.366), PHCOR(0.355) y PHRAI(0.287) contribuyen en una mayor cantidad que el efecto directo de PHRS(0.187), con respecto a los efectos indirectos se observa que PHRP, PHCOR y PHRAI presentan altos valores a través de PHCOR y PHRAI.

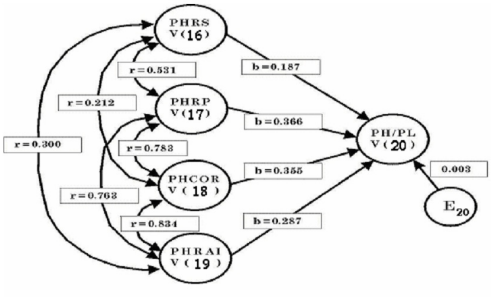



**Figura 2:** Efectos directos $(b)(\rightarrow)$ y correlación $(r)(\leftrightarrow)$

| Raíces | Corr.Can. | Val. Carac. | Ji Cuad. | Gra. Lib. | Prob. |
|---|---|---|---|---|---|
| 1 | 0.999 | 0.998 | 475.3 | 135 | 0.0000** |
| 2 | 0.996 | 0.993 | 341.2 | 112 | 0.0000** |
| 3 | 0.985 | 0.970 | 231.2 | 91 | 0.0000** |
| 4 | 0.980 | 0.961 | 155.6 | 72 | 0.0000** |

** significativo al 0.01 %

**Cuadro 3:** Análisis Estadístico Canónico.

Por lo tanto en Figura 2, se puede observar que el efecto indirecto de PHRAI a través de PHCOR se obtiene del producto de la correlación entre las variables PHRAI y PHCOR(r = 0.834) y el efecto directo de la variable PHCOR(b = 0.355) sobre PH/PL es decir (0.783)(0.355) = 0.296. En lo que respecta a la importancia de la correlación canónica, se puede observar en el Cuadro 3, que las primeras cuatro raíces son altamente significativas ($P < 0.0001$). En Cuadro 4, se observa que las mayores ponderaciones con $U_1$ son PSRS, PSRP, PSCOR, DIAM, PSRAI, PHRS, PHRP, PHCOR, PHRAI y PH/PL. Con respecto al Cuadro 5, las mayores ponderaciones con $V_1$ son, PRRS, PRRP, PRCOR, PRRAI y PR/PL, es decir al incrementar las ponderaciones en las variables de $U_1$ se incrementan las de $V_1$.

| VARIABLES X | $U_1$ | $U_2$ | $U_3$ | $U_4$ |
|---|---|---|---|---|
| PSRS | 0.8632 | 0.4655 | 0.0387 | -0.1369 |
| HRS | 0.0615 | 0.1271 | -0.3243 | - 0.0523 |
| PSRP | 0.8319 | -0.2568 | -0.4338 | -0.1969 |
| APL | 0.3403 | -0.1852 | -0.0172 | -0.2478 |
| HRP | 0.1310 | 0.0476 | -0.1819 | - 0.1645 |
| PSCOR | 0.7855 | -0.6027 | 0.0240 | -0.0704 |
| DIAM | 0.6904 | -0.1556 | -0.1747 | 0.0916 |
| HCOR | 0.2224 | -0.4649 | -0.1261 | - 0.2366 |
| PSRAI | 0.9182 | -0.2894 | -0.0505 | 0.0091 |
| HRAI | 0.1531 | 0.0476 | -0.1819 | - 0.1645 |
| PHRS | 0.6174 | 0.4915 | -0.2461 | -0.0875 |
| PHRP | 0.7749 | -0.2062 | -0.4312 | - 0.2441 |
| PHCOR | 0.7801 | -0.5784 | -0.0241 | -0.0952 |
| PHRAI | 0.8021 | -0.2760 | -0.1187 | 0.2250 |
| PH/PL | 0.8867 | -0.2557 | -0.2451 | -0.0731 |

**Cuadro 4:** Cargas entre las $U_i$ y sus Variables.

| VARIABLES Y | $V_1$ | $V_2$ | $V_3$ | $V_4$ |
|---|---|---|---|---|
| RRS | -0.0124 | -0.0916 | 0.2214 | 0.3996 |
| RRP | 0.2521 | -0.2547 | -0.1462 | 0.2327 |
| RCOR | -0.0006 | -0.3703 | -0.0355 | -0.0118 |
| RRAI | 0.1024 | -0.2734 | -0.2010 | 0.0273 |
| PRRS | 0.8440 | 0.4539 | 0.0935 | -0.0033 |
| PRRP | 0.8454 | -0.3018 | -0.3886 | -0.1144 |
| PRCOR | 0.7676 | -0.5909 | 0.0181 | -0.0594 |
| PRRAI | 0.7795 | -0.2325 | -0.2293 | 0.2941 |
| PR/PL | 0.9334 | -0.2313 | -0.2033 | -0.0065 |

**Cuadro 5:** Cargas entre las $V_i$ y sus Variables.

# 4. Conclusiones

El peso de resina en ramas primarias PRRP presentó la más alta correlación positiva con peso de resina por planta PR/PL. También el PRRP mostró el efecto directo más alto sobre PR/PL, por lo que la selección de genotipos de guayule con mayor número de ramas primarias incrementaría el rendimiento de resina. El peso de hule en ramas primarias PHRP se correlacionó alta y positivamente con peso de hule por planta PH/PL. El efecto directo de PHRP sobre PH/PL fue el de mayor magnitud. por lo tanto, la selección de genotipos con mayor número de ramas primarias también incrementaría el rendimiento de hule. Los efectos indirectos para rendimiento de resina y hule a través de ramas primarias fueron los de mayor magnitud. En los análisis de sendero también se observa que el peso de resina en la corona PRCOR y el peso de hule en la corona PHCOR, fueron los que siguieron en importancia para incrementar los rendimientos de resina y hule respectivamente. El análisis de correlación canónica, mostró que hay asociación entre las variables de hule y resina, en raíces removidas 1,2,3, y 4; no así en el resto donde no hubo significancia. En lo que respecta a los pares de variables canónicas ($U_1$,$V_1$) se observa que a medida que incrementan los pesos de hule, incrementan los pesos de resina; con respecto a los pares de variables ($U_2$,$V_2$),($U_3$,$V_3$),($U_4$,$V_4$), se nota por lo general, que a medida que disminuyen los pesos de hule disminuyen los pesos de resina. Estas conclusiones sugieren que para controlar mejor la producción de guayule procedente de este arbusto, debe tomarse en cuenta para su manejo los pesos de hule y resina por planta, principalmente ramas primarias, corona y raíz.

# Bibliografía

Arroyo, G. (1999), *Evaluación de Arbustos de Guayule (Phartenium argentatum Gray) en una Población Silvestre Regenerada Naturalmente, Tesis Licenciatura en Fitotecnia*, Universidad Autónoma Agraria Antonio Narro, Saltillo, Coahuila, México.

Bentler, M. (1995), "Eqs structural equations program manual multivariate software inc", *49244 Balboa Blvd.# 368. Encino, California USA 91316* .

Cox, D. y Wermuth, N. (1996), *Multivariate Dependencies: Model Analysis and Interpretation*, Chapman and Hall.

Memoria FONE XXV (2010), Sección V, 157–166

# Muestreo por conjuntos ordenados (Ranked Set Sampling) y su aplicación en población de maguey silvestre.

Lorena Alonso[a], Dante Covarrubias[b]
*Universidad Autónoma de Guerrero, Unidad Académica de Matemáticas*

Carlos N. Bouza[c]
*Universidad de la Habana*

## 1. Introducción

Sectores gubernamentales y sociales en México y otros países, han tenido la necesidad de estudiar las relaciones entre los sistemas naturales y la sociedad, esto para la protección del medio ambiente, lo que exige información y conocimiento espacio-temporal, creíble y defendible, Covarrubias y Bouza (2007).

Hasta hace pocos años, la industria del mezcal se desarrolló bajo el supuesto de que disponían de un recurso auto-renovable. Este estudio se hace necesario debido al incremento de la demanda tanto interna, como externa del maguey papalote. Consideramos la recomendabilidad de utilizar la técnica de Muestreo por Conjuntos Ordenados de Rango (MCOR) o como se conoce en ingles Rank Set Sampling (RSS). Esta técnica de muestreo fue motivada por su capacidad para mejorar la precisión de la media muestral, como estimador de la media poblacional. En McIntre (1952), se propone un método, denominado RSS, al trabajar en un problema de forraje. La similitud de los problemas de estimación de la producción de forraje y la asociada al maguey es evidente, dado que con la población de maguey papalote lo que se desea lograr al final es la estimación de la producción.

[a] alonso@uagro.mx
[b] dcova@yahoo.com.mx
[c] bouza@matcom.uh.cu

Mediante MCOR, se utilizara como variable auxiliar la edad del maguey papalote, para la obtención de los estimadores. Se revisará el problema clásico sobre el diseño eficiente de una muestra a través de un muestreo simple basada en información auxiliar, comparado con métodos de muestreo por conjuntos el cual hace un uso apropiado de variables auxiliares.

## 1.1. Planteamiento del problema

El cultivo y producción del Agave cupreata Trel *et* Berger se realiza en Guerrero, en su mayoría en zonas silvestres Gentry (1982), Maradiaga (2004)Alonso et al. (2009). Esta investigación hace referencia a la Región Centro en donde se realizo un inventario "Desarrollo de un Sistema de Inventario y Monitoreo de Maguey Papalote (Agave cupreata Trel. & Berger)", (SIMMP), en el 2004, con el fin de realizar una simulación partiendo de datos auténticos.

Un problema en particular en el Estado de Guerrero es la falta de información registrada sobre maguey papalote, debido a que es una planta que crece de manera natural encontrándose en zonas de difícil acceso, por lo que se hace necesario el uso de una metodología que contenga estrategias eficientes para la recolección de la información y el análisis estadístico. Esto es de suma importancia debido a la creciente extracción del maguey papalote que esta tomando en estos tiempos.

El visible aumento de la demanda de mezcal se debe a la tendencia de mantener las tradiciones y las costumbres, además el gran auge que ha tenido el tequila ha abierto el mercado nacional e internacional, para los diferentes mezcales de México; particularmente la demanda en Guerrero del maguey papalote (Agave cupreata Trel *et* Berger). Uno de los problemas que enfrentan los productores de mezcal en Guerrero, es el desconocimiento a corto y mediano plazo de la disponibilidad de magueyes debido a la carencia de un sistema de inventario.

## 1.2. Objetivo

El objetivo general de la investigación es proponer el muestreo por conjuntos ordenados de rangos, como técnica de muestreo adecuada para implementar un inventario de poblaciones de maguey papalote, que pueda proporcionar periódicamente la información sobre el

estado en que se encuentran y que contribuya a orientar las actividades productivas y de conservación hacia el desarrollo rural sustentable en el Estado de Guerrero.

# 2. Métodos de muestreo

Los estimadores sugeridos son investigados bajo los métodos del muestro simple aleatorio y muestreo por rangos ordenados.

## 2.1. Muestreo Simple Aleatorio

En el Muestreo Aleatrorio Simple (MAS), $m$ unidades de $N$ unidades de una población son elaborados de tal manera que todas las combinaciones posibles de elementos que podrían conformar un tamaño de muestra dado que tienen la misma probabilidad de ser seleccionado. En la práctica habitual, una muestra aleatoria simple se elabora unidad por unidad.

Procedimiento de un MAS:

- Identificar $n$ plantas aleatoriamente.
- Procurar obtener la medición sobre cada planta seleccionada.
- Generar estimadores y varianza (error estándar)

## 2.2. Muestreo por conjuntos ordenados por rangos

En Bouza (2001),se exponen las bondades estadísticas del método, conocido como Muestreo por Conjunto Ordenados por rangos (MCOR). En principio, se realiza utilizando información inicial para obtener una muestra representativa ante lo real. Para poblaciones con distribuciones desiguales que son caras para muestrear el MCOR es recomendable, debido a que puede llevar a una mayor precisión, disminución de toma de muestras, gastos, o ambas cosas.

El procedimiento de MCOR implica la selección de $m$ conjuntos, cada $m$ individuos o unidades muéstrales de la población. Se asume que las unidades dentro de cada conjunto pueden ser rankeadas visualmente sin costo o a un bajo costo. Desde el primer conjunto de $m$ unidades, se selecciona aquella que sea la menor en cuanto a la característica usando

una predicción la característica que se está estudiando, se le asigna un rango, las unidades restantes de la muestra se descartan.

Todo el proceso se puede repetir $r$ veces para obtener una muestra por conjunto ordenado de rangos de tamaño $n = mr$, con su respectivo rango, y desechadas un total de: $rm(m-1)$ observaciones.

# 3. Estimadores de la media usando los métodos de MAS y MCOR

Los estimadores de razón de la media poblacional bajo MAS $\mu_Y$ de la variable de interés $Y$ (ver Cochran (1977)) es:

$$\hat{\mu}_{MAS} = \mu_X(\frac{\bar{Y}_{MAS}}{\bar{X}_{MAS}}) \tag{1}$$

donde

$$\bar{X}_{MAS} = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{2}$$

y

$$\bar{Y}_{MAS} = \frac{1}{m}\sum_{i=1}^{m} Y_i \tag{3}$$

son las medias muéstrales de la variable auxiliar $X$ y la variable de interés $Y$, respectivamente. Las varianzas de $\bar{X}_{MAS}$ y $\bar{Y}_{MAS}$ son:

$$Var(\bar{X}_{MAS}) = \frac{\sigma_X^2}{m} \tag{4}$$

$$Var(\bar{Y}_{MAS}) = \frac{\sigma_Y^2}{m} \tag{5}$$

Los estimadores son estudiados bajo los métodos de MAS y MCOR. En esta investigación se asume que el rango se realiza en la variable $X$ para estimar la media poblacional de la variable $Y$.

Patil Patil (2002), define, cuando los tamaños de las muestras son iguales en cada iteración entonces se dice que es una muestra de conjuntos ordenados balanceada en caso contrario es no balanceada, Bouza (2009), es decir $r_i = r$.

Las propiedades de los estimadores no se alteran: En MAS la media muestral de la variable $X$ es:

$$\hat{\bar{X}}_{MAS} = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} X_{ij} \tag{6}$$

En MCOR la media muestral de la variable $X$ es:

$$\hat{\bar{X}}_{mco} = \frac{1}{r}\sum_{j=1}^{r}\frac{1}{m}\sum_{i=1}^{m} X_{ij} = \frac{1}{m}\sum_{i=1}^{m}\hat{\bar{X}}_i \tag{7}$$

La varianza estimada, $V(\hat{\bar{X}}_{MAS})$ es estimada por:

$$\hat{V}(\hat{\bar{X}}_{mas}) = \frac{1}{mr(mr-1)}\sum_{j=1}^{r}(X_{ij} - \hat{\bar{X}})^2 \tag{8}$$

La varianza estimada para MCOR puede verse en Chen et al. (2004):

$$\hat{V}(\hat{\bar{X}}_{MCOR_j}) = \frac{\hat{\bar{X}}_{MCOR_j}}{r} = \frac{\frac{1}{r-1}\sum_{j=1}^{r}(\hat{\bar{X}}_{MCOR_j} - \hat{\bar{X}}_{MCOR})^2}{r} \tag{9}$$

Un estimador alternativo de varianza:

$$\hat{V}(\hat{\bar{X}}_{mco}) = \frac{1}{mr^2}\sum_{i=1}^{m}\hat{S}_i^2 \tag{10}$$

con

$$\hat{S}_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(X_{ij} - \hat{\bar{X}}_i)^2 \tag{11}$$

Si el diseño no es balanceado los denominadores varían:

$$\hat{V}(\hat{\bar{X}}_{mco}) = \frac{1}{rm^2}\sum_{i=1}^{m}\hat{S}_i^2 \tag{12}$$

Patil (2002), comenta, si las mediciones de las muestras son independientes (aleatorio) e idénticamente distribuidas obtenidas a través del rango perfecto puede conducir a un rendimiento óptimo de MCO; no importando cuánto se desvíen estas características deseables, la eficiencia de muestreo nunca será peor que con MAS, utilizando el mismo número de cuantificaciones. De hecho, cuando la eficiencia se expresa como la precisión relativa (PR): donde $m$ es el tamaño del conjunto. Dado que PR no puede ser menor que uno, el protocoló del MCOR no puede ser peor que el protocoló del MAS.

En McIntre (1952), se sugiere para cuando la función del MCOR decrece como la distribución de la característica fundamental de interés llega a ser cada vez más deformada, que este problema puede subsanarse ubicando unidades de muestra en los rangos en proporción con la desviación típica de cada rango. Esta es la misma aproximación tan usada en el muestreo aleatorio estratificado, conocido como la Asignación de Neyman, y ciertamente sería óptima si hubiera estimaciones a priori confiables de la desviación estándar de los rangos.

$$PR = \frac{\hat{V}(\hat{\bar{X}}_{MAS})}{\hat{V}(\hat{\bar{X}}_{MCOR})}, \tag{13}$$

o el equivalente *ganancia relativa*, propuesta por Takahasi y Wakimoto (1968) tomado de Kaur et al. (1995),

$$GR = 1 - \frac{1}{PR} \tag{14}$$

En Takahasi y Wakimoto (1968) se establecen límites inferior y superior sobre el rendimiento de $\hat{\bar{X}}_{MCOR}$ relativo a MAS. Para la *asignación balanceada*, se tiene:

$$1 \leq PR \leq \frac{m+1}{2} \tag{15}$$

y

$$0 \leq GR \leq \frac{m-1}{m+1} \tag{16}$$

Si tomamos una precisión deseada para nuestro estimador, podemos conseguir que con una muestra pequeña de MCOR ahorrar dinero a través de MCOR, pero tiene costes adicionales, es decir un tamaño de muestra pequeño no necesariamente significa costos más bajos, que con una muestra de MAS.

# 4. Aplicación del estudio

Esta investigación retoma la base de datos proveniente del reporte técnico "Desarrollo de un Sistema de Inventario y Monitoreo de Maguey Papalote (Agave cupreata Trel. & Berger), (SIMMP)"proporcionada por el instituto de investigación Científica área Ciencias Naturales de la Universidad Autónoma de Guerrero. Para examinar la estrategia de muestreo, se procede al siguiente ejercicio: Se realiza en forma computacional 500 repeticiones independientes

del proceso completo que implica la selección de la etapa del maguey, la edad y la cobertura aérea . El procedimiento que se siguió fue el siguiente:

1. Identificar MCOR de $n$ plantas, maguey papalote (conjuntos)

2. Asignar Rango basadas en una estimación rápida.

3. Seleccionar plantas de rango deseado, (costosa) se obtiene con cuidado la medición del maguey papalote.

Se realizan los pasos del 1 al 3 varias veces. formando conjuntos de tamaño: m=5, repitiendo el proceso.

## 4.1. Resultados

El primer resultado que se obtiene se refiere a la cobertura aérea, del parámetro el cual es de 3298.63 m2 aprovechados por el maguey papalote, para el estimador de MAS y MCOR, el tamaño de la muestra fue de 1120, la variable que se empleo como auxiliar fue la edad del maguey papalote y la superficie establecida del maguey papalote, mediante MCOR se necesita en promedio 3.07m2 de superficie para una planta de maguey papalote, para la recuperación del maguey papalote de superficie territorial en las localidades de la zona centro de Guerrero se muestra en la Tabla 1.

| Muestra | Cobertura | Media muestral | Ganancia relativa | Numero de maguey a labrar |
|---|---|---|---|---|
| Parámetro | $3437.64m^2$ | 5876 | 93.21 | 5876 |
| MAS | $4690.568m^2$ | 5511 | 89.30 | 2033 |
| MCO | $4067.761m^2$ | 5866 | 99.45 | 4225 |

**Cuadro 1:** Cobertura aérea especial

Para los resultados de eficiencia y ganancia relativa entre métodos indica que el MCO es más eficiente, que el MAS, esto se observa en la Tabla 1.

# 5. Conclusiones y discusiones

El método de muestreo por conjuntos ordenados por rangos es de fácil aplicación y se puede aplicar tanto en terreno llano, como de pendientes bruscas, combinando muestreo de líneas por transectos. Este muestreo se recomienda para determinar la densidad y cobertura del maguey papalote. Se debe tener en cuenta el patrón espacial del maguey papalote, si la dispersión afecta notablemente al muestreo, por lo que debe ser considerada para realizar un diseño adecuado. El método debe ser escogido de acuerdo con la variabilidad de la población del maguey papalote en estudio, para que los resultados obtenidos sean confiables. Finalmente, para poder clasificar el área de distribución natural del maguey por calidad de sitio, se requiere el establecimiento de parcelas permanentes dentro de las magueyeras, que permita inventariar y regular el establecimiento, velocidad de desarrollo y maduración del maguey papalote.

Otras versiones de este documento se encuentran en:*2do encuentro Iberoamericano de Biometría, (mayo de 2010). "Inventario de maguey papalote de la región centro del Estado de Guerrero". Boca de Río, Veracruz, México.*`www.uv.mx/eib`

*1er Encuentro Internacional del Medio Ambiente, (Noviembre 2010). "Muestreo por Conjuntos Ordenados (Ranked Set Sampling) y su Aplicación en Poblaciones de Maguey silvestre". Puebla, Puebla; México.* `www.eima.todoencomputacion.html`

# Bibliografía

Alonso, L., Covarrubias, D. y Maradiaga, F. (2009), *Estrategias De Muestreo Para El Inventario De Maguey Papalote (Agave Cupreatra Trel & Berger)*, Tesis para obtener el grado de Maestría en Ciencias Área Estadística aplicada, No publicada, Unidad Académica de Matemáticas, UAGro, Guerrero, México.

Bahamondez, C., Lorenz, M., Mery, G. y Varjo, T. (2005), *Forest Assessment for Changing Information Needs*, IUFRO World Series, España.

Bouza, C. (2001), "Model assisted ranked survey sampling: An annotated bibliography", *Biometrical* **43**, 249–259.

Bouza, C. (2009), "El muestreo por conjuntos ordenados por rangos y las perspectivas de su uso en biometría", *Memoria 2do. Encuentro Iberoamericano de Biometrics* **2**, 32–35.

Chen, Z., Bai, Z. y Sinha, K. B. (2004), *Ranked Set Sampling*, Springer, New York.

Cochran, W. G. (1977), *Minería de Datos*, CECSA, México.

Covarrubias, D. y Bouza (2007), *Modelos de Muestreo para la Estimación de Índices de Diversidad*, Tesis para obtener el grado Científico Doctor en Ciencias Matemáticas, No publicada, Facultad de Matemática y Computación, Universidad de la Habana, Cuba.

Gentry, H. (1982), *Agaves of Continental North America*, The University of Arizona Press, Tucson, Arizona.

Gruijter, J., Brus, D., Bierkens, M. y Knotters, M. (2009), *Sampling for Natural Resource Monitoring*, Springer, Berlin.

J.Lawler, J., White, D., Sifneos, J. y Master, L. (2003), "Rare species and the use of indicator groups for conservation planning", *Conservation Biology* **17**(5), 875–882.

Kaur, A., Patil, G. P., Sinha, A. K. y Taillie, C. (1995), "Ranked set sampling: An annotated bibliography.", *Environmental and Ecological Statistics* **2**, 25–54.

Maradiaga, F. (2004), "Desarrollo de un sistema de inventario y monitoreo de maguey papalote (agave cupreata trel. & berger) en el estado de Guerrero", *Fundación PRODUCE Guerrero A.C., Programa de Recursos Biológicos Colectivos (CONABIO) e Instituto de Investigación Científica Área Ciencias Naturales de la UAGro.* .

McIntre, G. A. (1952), "A method for unbiased selective sampling using ranked set", *Journal of Agricultural Research* **15**, 385–390.

Medina, J. C. (2004), *Análisis Comparativo de Técnicas, Metodologías y Herramientas de Ingeniería de Requerimientos.*, Tesis para obtener el grado de Maestría en Ciencias en la especialidad de Ingeniería Eléctrica Opción Computación, Centro de Investigación y de Estudios Avanzados del IPN, México, D.F.

Patil, G. P. (2002), "Ranked set sampling", *Encyclopedia of Environmetrics* **17**, 1684–1690.

SAGARPA (Varios años), "Anuario estadístico de la producción agrícola", *SAGARPA* .

Takahasi, K. y Wakimoto, N. (1968), "On unbiased estimates of population mean based on the sample stratified by means of ordering", *Geoderma* **62**, 233–246.

# Lista de autores

Loya Monares, Nahun I. <israel_loya@hotmail.com>. *FCFM, Benemérita Universidad Autónoma de Puebla*, 123

Mair, Patrick. *WU Vienna University of Economics and Business*, 11

Méndez Ramírez, Ignacio. *Universidad Nacional Autónoma de México*, 149

Muñoz Urbina, Armando. *Universidad Autónoma de Coahuila*, 149

Núñez Antonio, Gabriel <gab.nunezantonio@gmail.com>. *Departamento de Estadística, Universidad Carlos III de Madrid, España*, 31

Nieto Murillo, Soraida <gussynm@hotmail.com>. *Universidad Autónoma Metropolitana – Iztapalapa*, 129

Ojeda Ramírez, Mario Miguel <mojeda@uv.mx>. *Facultad de Estadística e Informática. Universidad Veracruzana*, 105

Padilla Terán, Alberto M. <ampadilla@banxico.org.mx>. *Banco de México*, 91

Padrón Corral, Emilio. *Universidad Autónoma de Coahuila*, 149

Pérez Chávez, Jorge A. *Universidad Juárez Autónoma de Tabasco - IMSS*, 115

Pérez Salvador, Blanca R. <psbr@xanum.uam.mx>. *Universidad Autónoma Metropolitana-Iztapalapa*, 129, 137

Reyes Cervantes, Hortensia J. *FCFM, Benemérita Universidad Autónoma de Puebla*, 123

Romero Mares, Patricia <patricia@sigma.iimas.unam.mx>. *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México*, 3

Rueda Díaz del Campo, Raúl <pinky@sigma.iimas.unam.mx>. *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México*, 3

Russell Noriega, Ma. Guadalupe <mgrussell@uas.uasnet.mx>. *Universidad Autónoma de Sinaloa, Culiacán, Sin.*, 21

Salinas-Hernández, Rosa Ma. *Universidad Juárez Autónoma de Tabasco - IMSS*, 115

Sánchez Tizapa, Sulpicio <*ss*tizapa@hotmail.com>. *Unidad Académica de Ingeniería,Universidad Autónoma de Guerrero*, 143

# Lista de árbitros

El Comité Editorial de la Memoria del XXV Foro Nacional de Estadística agradece la valiosa colaboración de los siguientes árbitros:

1. Aguirre Hernández, Rebeca, *Facultad de Medicina, UNAM.*
2. Alonso Reyes, María del Pilar, *Facultad de Ciencias, UNAM.*
3. Burgueño Ferreira, Juan Andrés, *Colegio de Posgraduados, Campus Montecillo.*
4. Contreras Cristán, Alberto, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*
5. Chávez Cano, Margarita, *Facultad de Ciencias, UNAM.*
6. Chong Rodríguez, Miguel Ángel, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*
7. Domínguez Domínguez, Jorge, *Centro de Investigación en Matemáticas, A.C.*
8. Flores Díaz, José Antonio, *Facultad de Ciencias, UNAM.*
9. Fuentes García, Ruth Selene, *Facultad de Ciencias, UNAM.*
10. Gabriel Pacheco, Carlos, *CINVESTAV-IPN.*
11. Gallardo Franco, Virginia, *Centro Nacional de Control de Energía (CENACE) de Comisión Federal de Electricidad.*
12. García Gómez, Yofre Hernán, *Físico Matemáticas, IPN.*
13. González Farías, Graciela, *Centro de Investigación en Matemáticas, A.C.*

14. Gracia Medrano, Leticia, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*

15. Gutiérrez Peña, Eduardo, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*

16. Hoyos Argüelles, Ricardo, *Banco de México.*

17. Jurado Galicia, Edith, *Instituto Tecnológico de Estudios Superiores de Monterrey campus Santa Fe.*

18. Mena Chávez, Ramsés H., *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*

19. Morales Velasco, Juan, *Universidad Nacional Autónoma de México.*

20. Pérez Salvador, Blanca Rosa, *Universidad Autónoma Metropolitana-Iztapalapa.*

21. Romero Mares, Patricia, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*

22. Romero Martínez Martín, *Instituto Nacional de Salud Pública.*

23. Rueda Díaz del Campo, Raúl, *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM.*

24. Seseña Méndez, Luis Fernando, *Universidad Nacional Autónoma de México.*

25. Villarreal Rodríguez, César Emilio, *Facultad de Ingeniería Mécánica y Eléctrica, UANL.*

26. Villaseñor Alva, José, *Colegio de Postgraduados-Chapingo.*