INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Memoria del XXIV Foro Nacional de Estadística

AME

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Memoria del XXIV Foro Nacional de Estadística

AME

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

310.4 Foro Nacional de Estadística (24° : 2009 : Texcoco, Estado de México).
Memoria del XXIV Foro Nacional de Estadística / Instituto Nacional de Estadística y Geografía, Asociación Mexicana de Estadística. -- México : INEGI, c2010.

176 p. : il.

ISBN 978-607-494-106-7

"Colegio de Postgraduados. Texcoco, Estado de México del 12 al 16 de octubre del 2009"

1. Estadística - Alocuciones, Ensayos, Conferencias. I. Instituto Nacional de Estadística y Geografía. II. Asociación Mexicana de Estadística.

Memoria
del XXIV Foro
Nacional de Estadística

Impreso en México
ISBN 978-607-494-106-7

Esta publicación consta de 1 779 ejemplares y se terminó de imprimir en septiembre de 2010 en los talleres gráficos del
Instituto Nacional de Estadística y Geografía
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso
Fracc. Jardines del Parque, CP 20276
Aguascalientes, Ags.
México

Presentación

En estas memorias publicamos los resúmenes de algunas contribuciones libres presentadas durante el XXIV Foro Nacional de Estadística. La institución sede fue el Colegio de Postgraduados y el evento tuvo lugar en Texcoco, Estado de México, del 12 al 16 de octubre de 2009.

El volumen está integrado por tres secciones:

I. Trabajos de investigación,

II. Aplicaciones,

III. Tesis de licenciatura y maestría.

Los trabajos fueron sometidos a un proceso de arbitraje coordinado por la mesa directiva de la Asociación Mexicana de Estadística. En este proceso, todos los artículos fueron revisados en su forma y contenido; siguiendo, en todo momento, criterios mínimos para evaluar la calidad en sus propuestas, resultados y aplicaciones, con énfasis en la originalidad para los trabajos de la Sección I.

Agradecemos profundamente a todos los autores por su entusiasmo y por la calidad de los trabajos presentados. Agradecemos, además, a todos aquellos colegas que nos apoyaron participando como árbitros, pues con su esfuerzo, contribuyen a la calidad académica de estas memorias. En nombre de la Asociación Mexicana de Estadística expresamos también nuestra gratitud al Colegio de Postgraduados por el apoyo en la realización de este Foro, y al Instituto Nacional de Estadística y Geografía por patrocinar la edición e impresión de esta obra.

El Comité Editorial:

Yolanda Margarita Fernández Ordoñez,

Silvia Ruiz Velasco Acosta.

Índice general

Sección I. Trabajos de Investigación

Sección II. Aplicaciones

Sección III. Tesis de licenciatura y maestría

Sección I

Trabajos de Investigación

Memoria FONE XXIV (2010), Sección I, 5–12

Algunas propiedades de los conjuntos convexos de Barnard y aplicaciones a pruebas de no-inferioridad

Félix Almendra Arao[a]
UPIITA del Instituto Politécnico Nacional

1. Introducción

La necesidad de comparar grupos en muchos campos de la ciencia es universalmente conocida. Particularmente en el campo de ensayos clínicos, en donde recientemente se ha incrementado considerablemente el uso de las pruebas de no-inferioridad en la evaluación de tratamientos nuevos. Las pruebas de no-inferioridad se realizan con el objetivo de mostrar que un tratamiento experimental o nuevo es estadística y clínicamente no inferior a un tratamiento estándar conocido o control activo. El nuevo producto puede ofrecer ventajas de seguridad, tener un método más simple de promover la adherencia, que los costos y las ganancias potenciales sean la razón subyacente o bien que sea de aplicación más fácil. En el cálculo de los niveles de significancia para las pruebas estadísticas de no-inferioridad, las regiones críticas que cumplen la condición de convexidad de Barnard juegan un papel central, ya que debido a un teorema demostrado por Röhmel y Mansmann (1999), cuando la regiones críticas cumplen dicha condición, el nivel de significancia para las pruebas de no-inferioridad puede calcularse de forma mucho más eficiente. En esta investigación, los conjuntos que cumplen la condición de convexidad de Barnard son llamados conjuntos convexos de Barnard, y debido a su antes mencionada relevancia, se estudian sus propiedades. Se realiza un estudio de los conjuntos convexos de Barnard de manera independiente del contexto a partir del cual se originaron. Se obtienen varios resultados, entre ellos, que los conjuntos convexos de Barnard son una geometría convexa, se define el concepto de base para un conjunto convexo de Barnard y

[a]falmendra@ipn.mx

se prueba que todo conjunto convexo de Barnard tiene una base única. Asimismo se proporciona un algoritmo para calcular la cápsula convexa de Barnard de cualquier conjunto. Finalmente, se presentan algunas aplicaciones del concepto de cápsula convexa de Barnard de un conjunto a pruebas de no-inferioridad.

2. Conjuntos convexos de Barnard

Sean n_1 y n_2 dos enteros positivos y $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$.

Definición 2.1. *Un conjunto $C \subseteq \chi$ satisface la condición de convexidad de Barnard si cumple las dos condiciones siguientes:*

1. $(x_1, x_2) \in C \Rightarrow (x_1 - 1, x_2) \in C \ \forall \ 1 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2$
2. $(x_1, x_2) \in C \Rightarrow (x_1, x_2 + 1) \in C \ \forall \ 0 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2 - 1$

Un conjunto $C \subseteq \chi$ que satisface la condición de convexidad de Barnard se llamará conjunto convexo de Barnard.

La colección de todos los subconjuntos de χ que son conjuntos convexos de Barnard, será denotada mediante $\mathcal{C}(\chi)$.

Proposición 2.1. *Se cumplen las afirmaciones siguientes:*

1. $\emptyset, \chi \in \mathcal{C}(\chi)$
2. $A, B \in \mathcal{C}(\chi) \Rightarrow A \cap B \in \mathcal{C}(\chi)$
3. $A, B \in \mathcal{C}(\chi) \Rightarrow A \cup B \in \mathcal{C}(\chi)$

Roy y Stell (2003) presentan la siguiente:

Definición 2.2. *Un espacio de alineamiento es una pareja $A = (X, \mathcal{C})$ donde X es un conjunto y $\mathcal{C}$ es un conjunto de subconjuntos de X tal que se cumplen los axiomas siguientes:*

1. $\emptyset, \chi \in \mathcal{C}$
2. $\forall \ Z \subseteq \mathcal{C}, Z \neq \emptyset \Rightarrow \bigcap Z \in \mathcal{C}$

3. $\forall\ Z \subseteq \mathcal{C}, Z \neq \emptyset \Rightarrow \bigcup Z \in \mathcal{C}$

De la proposición 2.1 se tiene que $A = (X, \mathcal{C}(\chi))$ es un espacio de alineamiento.

Definición 2.3. *Dado un conjunto $A \subseteq \chi$, se define la cápsula de alineamiento de Barnard de A, denotada por [A], como el mínimo conjunto de Barnard que contiene al conjunto A. Se dice que A genera a [A] o que [A] es generado por A.*

Definición 2.4. *Dado un conjunto $C \in \mathcal{C}(\chi)$, se dice que $G \subseteq \chi$ es un conjunto generador de C, si [G]=C.*

Proposición 2.2. *Se cumplen las afirmaciones siguientes:*

1. $[\emptyset] = \emptyset$
2. $A \subseteq [A]\ \forall\ A \subseteq \chi$
3. $A \neq \emptyset \Rightarrow [A] \neq \emptyset$
4. $C \in \mathcal{C}(\chi) \Rightarrow [C] = C$
5. $A \subseteq B \Rightarrow [A] \subseteq [B]$
6. $C \in \mathcal{C}(\chi) \Rightarrow [c] \subseteq C\ \forall\ c \in C$

Proposición 2.3. *Se cumplen las afirmaciones siguientes:*

1. $[A] \cup [B] = [A \cup B]$
2. $[A] \cap [B] \subseteq [A \cap B]$

Es fácil construir un ejemplo donde no se cumpla la contención opuesta a la dada en (2) de la proposición 2.3

Proposición 2.4. *Propiedad de anti-intercambio*
$\forall x, y \in \chi, \forall\ A \subseteq \chi, si\ \forall x \neq y, y \in [A \cup x],\ y \quad \notin [A], entonces\ x \notin [A \cup y]$

Roy y Stell (2003) dan la siguiente:

Definición 2.5. *Una geometría convexa es un espacio de alineamiento que cumple la propiedad de anti-intercambio.*

De la propiedad de anti-intercambio y de que las propiedades (2) y (3) de la proposición 2.1 pueden extenderse a un número finito de conjuntos, se tiene que $A = (\chi, \mathcal{C}(\chi))$ es una geometría convexa. Por tal motivo, en lo que sigue en lugar de referirnos a [A] como la cápsula de alineamiento de Barnard de A, nos referiremos a ella como la cápsula convexa de Barnard de A.

Definición 2.6. *Un conjunto $B \subseteq \chi$ es una base de $C \in \mathcal{C}(\chi)$ si:*

1. *B genera a C*
2. *B es el mínimo conjunto desde el punto de vista de inclusión, que genera a C, es decir, si existe $B' \subseteq \chi$ tal que $[B'] = C$, entonces $B \subseteq B'$.*

Teorema 2.1. *Todo conjunto convexo de Barnard tiene una base única.*

Proposición 2.5. *Si G_1 y G_2 generan a $C \in \mathcal{C}(\chi)$, entonces $[G_1 \cap G_2]$ también genera a C.*

Teorema 2.2. *La intersección de todos los generadores de $C \in \mathcal{C}(\chi)$ es la base de C.*

Proposición 2.6. *Si $(i,j) \in [A]$, entonces existe $(i',j') \in A$ tal que $[(i,j)] \subseteq [(i',j')], [(i'+1, j')] \not\subseteq [A]$ y $[(i', j'-1)] \not\subseteq [A]$. Además, $B = \{(i',j') : (i,j) \in [A]\}$ es la base de [A].*

Para $A \subseteq \chi$, considérense las definiciones siguientes:

Para un valor fijo $x_1 \in \{0, ..., n_1\}$, sea $R_{x_1} = \{x_2 : (x_1, x_2) \in A\}$; si $R_{x_1} \neq \emptyset$, entonces se define $a(x_1) = min R_{x_1}$. Similarmente, para un valor fijo $x_2 \in \{0, ..., n_2\}$, sea $S_{x_2} = \{x_1 : (x_1, x_2) \in A\}$; si $S_{x_2} \neq \emptyset$, entonces se define $b(x_2) = max S_{x_2}$.

Con base en la proposición 2.6 se puede establecer el siguiente algoritmo para construir la cápsula convexa de Barnard de un subconjunto arbitrario A de χ.

1. Tomar $A_1 = A$
2. Definir $j_1 = min\{a(x_1) : (x_1, x_2) \in A\}, i_1 = max\{a(x_1) = j_1\}, A_2 = A_1 - [(i_1, j_1)]$, si $A_2 \neq \emptyset$, entonces repetir el proceso, en caso contrario terminar.

3. Definir $j_2 = min\{a(x_1) : (x_1, x_2) \in A_2\}, i_1 = max\{a(x_1) = j_2\}, A_3 = A_2 - [(i_2, j_2)]$, si $A_3 \neq \emptyset$, entonces repetir el proceso, en caso contrario terminar.

Así este proceso está descrito en forma iterativa y debe terminar en algún paso. Supóngase que el proceso finaliza en el paso k, es decir, que $A_{k+1} = \emptyset$, entonces

$$[A] = \bigcup_{t=1}^{k} [(i_t, j_t)]$$

Las demostraciones de los resultados presentados en esta sección se omitieron por razones de espacio, éstas pueden consultarse en Almendra (2010).

3. Aplicaciones

El propósito de la presente sección es el de ilustrar la utilidad del concepto de cápsula convexa de Barnard, particularmente cuando se aplica dicho concepto a regiones críticas que no son conjuntos convexos de Barnard a fin de calcular de forma práctica los tamaños de prueba para pruebas de no inferioridad, es decir, usando la cápsula convexa de Barnard en lugar de usar la región crítica, la cual, no es un conjunto convexo de Barnard.

Sean X_i variables aleatorias independientes distribuidas binomialmente con parámetros (n_i, p_i), repectivamente, para $i = 1, 2$; donde p_1 y p_2 representan las probabilidades verdaderas de respuesta del tratamiento estándar y nuevo, respectivamente. La hipótesis de no inferioridad es la alternativa en el siguiente juego de hipótesis:

$$H_0 : p_1 - p_2 \geq d_0 \text{ vs } H_a : p_1 - p_2 < d_0$$

donde d_0 es una constante positiva conocida, llamada margen de no inferioridad. El espacio muestral es $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$ y el espacio paramétrico es $\Theta = \{(p_1, p_2) \in [0, 1]^2\}$

Consideremos la prueba estadística de Blackwelder (Blackwelder 1982)

$$T_B(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}_B},$$

donde $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i para $i = 1, 2$ y $\hat{\sigma}_B$ es el estimador de la desviación estándar de $\hat{p}_1 - \hat{p}_2$ dado por $\hat{\sigma}_B = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. La estadística T_B tiene distribución asintótica normal estándar. Así para un nivel de significancia nominal dado α, la región crítica está dada por $R_{T_B} = (x_1, x_2) \in \chi : T_B(X_1, X_2) \leq -z_\alpha$ donde z_α es el percentil superior α de la distribución normal estándar, i. e. $P(Z > z_\alpha) = \alpha$. La función de verosimilitud conjunta está dada por

$$L(p_1, p_2; x_1, x_2) = \prod_{i=1}^{2} \binom{n_i}{x_i} p_i^{x_i}(1 - p_i)^{n_i - x_i},$$

la función potencia es

$$\beta(p_1, p_2) = \sum_{(x_1,x_2)\in R_{T_B}} L(p_1, p_2; x_1, x_2)$$

El tamaño de la prueba está dado por

$$\sup_{(p_1,p_2)\in\Theta_0} \beta_{T_B}(p_1, p_2) \tag{1}$$

donde $\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ es el espacio nulo.

Röhmel y Mansmann (1999) demostraron que cuando la región crítica es un conjunto convexo de Barnard el tamaño de la prueba puede calcularse mediante

$$\begin{array}{c} \text{máx} \\ p_2 = p_1 - d_0 \\ p_1 \in [d_0, 1] \end{array} \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \prod_{i=1}^{2} \binom{n_i}{x_i} p_i^{x_i}(1 - p_i)^{n_i - x_i} I_{[(x_1,x_2)\in R(\alpha)]}. \tag{2}$$

Nótese que por definición, si $R_T \neq \emptyset$ y R_T es un conjunto convexo de Barnard, entonces $(0, n_2) \in R_T$. Por otra parte, $T_B(0, n_2)$ no está definida, entonces $(0, n_2) \notin R_B$ y en consecuencia R_{T_B} no es un conjunto convexo de Barnard, por lo tanto, para calcular el tamaño de prueba no sería posible el uso de la ecuación (2), sino que tendría que hacerse con la fórmula (3) lo cual resulta completamente impráctico debido al tiempo de cómputo que se requeriría. Con el objeto de reducir este tiempo de cómputo se tienen varias alternativas:

1. Considerar el máximo subconjunto convexo de Barnard contenido en la región crítica. Nótese que esta estrategia es aplicable en general, sin embargo, en este caso el máximo subconjunto convexo de Barnard contenido en la región crítica es el conjunto vacío, así, esta estrategia no es recomendable.

2. Redefinir la estadística T_B en aquéllos puntos donde no está definida. Existen varias formas en que se pueden presentar redefiniciones de T_B, una de ellas es la usual, otra se presenta en Almendra (2009) y hay otras posibilidades más, sin embargo en términos generales estas redefiniciones podrían añadir en muchas ocasiones puntos innecesarios a la región crítica, es decir, puntos que si no se agregaran la región crítica también sería un conjunto convexo de Barnard, luego entonces la redefinición de regiones críticas tiene el inconveniente de que en algunos casos podría inflar innecesariamente el tamaño de la prueba.

3. Considerar la cápsula convexa de Barnard de la región crítica. Esta estrategia es óptima para el cálculo de los tamaños de prueba, por tal motivo será preferida por sobre las otras dos.

Existen otras pruebas estadísticas para contrastar no inferioridad tales que como la z-estadística ponderada, la prueba de Hauck y Anderson (1986) y la prueba de Búhning y Viwatgonsen (2005) y algunas combinaciones de ellas; para todas estas pruebas estadísticas, las regiones críticas no son conjuntos convexos de Barnard. Por lo tanto para calcular los tamaños de prueba cuando se usa alguno de dichos procedimientos estadísticos también es recomendable usar la cápsula convexa de Barnard con el objeto de garantizar una región crítica que sea un conjunto convexo de Barnard y que con ello el tamaño de la prueba correspondiente pueda calcularse de forma práctica mediante el teorema de Röhmel y Mansmann.

4. Conclusiones

Se obtuvieron varias propiedades de los conjuntos convexos de Barnard, especialmente aquéllas relacionadas con la cápsula convexa de Barnard. Dichos resultados son interesantes tanto desde el punto de vista teórico como aplicado. También se proporcionaron ejemplos donde se puede aplicar el concepto de cápsula convexa de Barnard y se mostró que este concepto es útil para reducir el tiempo de cómputo para calcular los tamaños de prueba, para pruebas de no inferioridad. Además, se proporcionó un algoritmo para la construcción de la cápsula convexa de Barnard.

Los resultados presentados en este trabajo son útiles para construir regiones críticas que garanticen ser conjuntos convexos de Barnard y en consecuencia que el cálculo de los tamaños de prueba se reduzca considerablemente.

Referencias

Almendra-Arao, F. 2009. "A Study on the Asymptotic Classical Non-inferiority Test for two Binomial Proportions". *Drug Information Journal*, 43 (5): 567-571.

Almendra-Arao, F. 2010. "Barnard Convex Sets". *Communications in Statistics - Theory and Methods*. Aceptado para publicación.

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

Búhning, D. y Viwatwongkasen, C. 2005. "Revisiting proportion estimators". *Statistical methods in medical research*, 14: 1-23.

Hauck, W. y Anderson, S. 1986. "A comparison of large-sample confidence interval methods for the difference of two binomial probabilities". *The American Statistician*, 40: 318-322.

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 ".*Biometrical Journal*, 2: 149-170.

Memoria FONE XXIV (2010), Sección I, 13–18

La prueba de no inferioridad basada en la z-estadística asintótica ponderada

Félix Almendra Arao[a]
UPIITA del Instituto Politécnico Nacional

1. Introducción

El presente trabajo es un resumen de Almendra (2009b). Un problema muy importante en estadística biofarmacéutica es valorar la equivalencia entre dos grupos, muy frecuentemente esto se formula en términos de dos pruebas unilaterales, llamadas pruebas de no inferioridad. En el estudio de pruebas de no inferioridad se consideran varias medidas de discrepancia, una de ellas es la diferencia de proporciones, existen varias pruebas estadísticas que usan esta diferencia, entre ellas sobresalen dos, una basada en la z-estadística con varianza ponderada y otra en la z-estadística clásica con varianza no ponderada.

Para la z-estadística clásica asintótica, Almendra (2009a) calculó los tamaños de prueba, sin corrección por continuidad y con la corrección por continuidad de Hauck-Anderson, para varias configuraciones de tamaños de muestra, niveles de significancia nominales y márgenes de no inferioridad, mostrando que el comportamiento de los tamaños de prueba es demasiado errático y muy alejado de los niveles de significancia nominal.

En este trabajo se calculan los tamaños de prueba para la prueba de no inferioridad basada en la z-estadística asintótica ponderada y se analiza su comportamiento. Se calcularon los tamaños de prueba tanto para la prueba sin corrección por continuidad como para la prueba con cinco correcciones por continuidad. Los tamaños de muestra considerados fueron $30 \leq n_1 = n_2 \leq 100$, los niveles de significancia nominal fueron $\alpha = 0.025, 0.05$, para los márgenes de no inferioridad $d_0 = 0.10$ y 0.15.

[a] falmendra@ipn.mx

El objetivo del trabajo fue examinar el comportamiento de los tamaños de prueba para saber si la prueba asintótica basada en dicha estadística de prueba preserva o no el tamaño de la prueba y en consecuencia poder dar recomendaciones prácticas en relación a su uso.

2. La z-estadística ponderada

Sean X_i variables aleatorias independientes distribuidas binomialmente con parámetros (n_i, p_i), repectivamente, para $i = 1, 2$; donde p_1 y p_2 representan las probabilidades verdaderas de respuesta del tratamiento estándar y nuevo, respectivamente. La hipótesis de no inferioridad es la alternativa en el siguiente juego de hipótesis:

$$H_0 : p_1 - p_2 \geq d_0 \text{ vs } H_a : p_1 - p_2 < d_0 \tag{1}$$

donde d_0 es una constante positiva conocida, llamada margen de no inferioridad. El espacio muestral es $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$ y el espacio paramétrico es $\Theta = \{(p_1, p_2) \in [0, 1]^2\}$

Suissa y Shuster (1985) y Haber (1986) definieron la z-estadística con varianza ponderada como:

$$T_0(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}}, \tag{2}$$

donde $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i para $i = 1, 2$ y $\hat{\sigma}$ es el estimador de la desviación estándar de $\hat{p}_1 - \hat{p}_2$ dado por $\hat{\sigma} = \sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$ donde $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$. Además, se sabe que T_0 en (2) tiene distribución asintótica normal estándar.

Por lo tanto, para un nivel de significancia nominal dado α, la región crítica correspondiente está dada por $R_{T_0} = \{(x_1, x_2) \in \chi : T_0(X_1, X_2) \leq -z_\alpha\}$ donde z_α es el percentil superior α de la distribución normal estándar, i. e., $P(Z > z_\alpha) = \alpha$. En este trabajo se usan las siguientes correcciones por continuidad $C_0 = 0, C_1 = \frac{1}{4min(n_1, n_2)}, C_2 = 2C_1, C_3 = \frac{1}{2n_1} + \frac{1}{2n_2}, C_4 = 6C_1, C_5 = 8C_1$. Así, las pruebas consideradas son $T_i(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0 + C_i}{\hat{\sigma}}$, donde $i = 0, 1, 2, 3, 4, 5$. Nótese que cuando $\hat{p} = 0$, i. e., cuando $\frac{X_1 + X_2}{n_1 + n_2} = 0, 1$, se obtiene $\hat{\sigma} = 0$ y en consecuencia T_i no está definida. Sea $\varphi(x) = \sqrt{x(1 - x)(\frac{1}{n_1} + \frac{1}{n_2})}$, proponemos redefinir

la estadística en este par de puntos de la siguiente forma $\hat{\sigma}(0,0) = \varphi(0.2), \hat{\sigma}(n_1, n_2) = \varphi(n_1 + n_2 - 0.2)$. De la redefinición previa se tiene $\hat{\sigma}(n_1, n_2) = \hat{\sigma}(0,0)$

3. Cálculo de los tamaños de prueba

El principal problema para calcular tamaños de prueba para pruebas de no inferioridad es originado por la existencia de un parámetro perturbador, ésto convierte dicho cálculo en un problema computacionalmente intensivo, nótese además que de acuerdo con Basu (1977) "la eliminación de parámetros perturbadores de un modelo es universalmente reconocido como un problema mayor en estadística". Debido a lo anterior, el cálculo de los tamaños de prueba en este trabajo fue realizado mediante el uso de algunas propiedades que permitieron una considerable reducción en su tiempo de cómputo.

La función de verosimilitud conjunta está dada por

$$L(p_1, p_2; x_1, x_2) = \prod_{i=1}^{2} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i},$$

la función potencia es

$$\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_{T_0}} L(p_1, p_2; x_1, x_2)$$

el tamaño de la prueba está dado por

$$\sup_{(p_1, p_2) \in \Theta_0} \beta_T(p_1, p_2)$$

donde $\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ es el espacio nulo.

Definición 3.1. *Se dirá que una estadística T, o su respectiva región crítica R_T satisface la condición de convexidad de Barnard (C) si cumple las dos condiciones siguientes:*

1. $(x_1, x_2) \in R_T \Rightarrow (x_1 - 1, x_2) \in R_T \ \forall \ 1 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2$

2. $(x_1, x_2) \in R_T \Rightarrow (x_1, x_2 + 1) \in R_T \ \forall \ 0 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2 - 1$

Definición 3.2. *Sean $n_1 = n_2 = n$, se dirá que una estadística T, o su respectiva región crítica R_T satisface la condición de simetría en la misma cola (S) si $(x_1, x_2) \in R_T \Rightarrow (n - x_2, n - x_1) \in R_T \ \forall \ 0 \leq x_1, x_2 \leq n$*

Almendra (2009a) demostró la siguiente:

Proposición 3.1. *Sean $n_1 = n_2 = n$ y $R(\alpha)$ la región crítica para el problema de prueba de hipótesis en (1). Si $R(\alpha)$ cumple las condiciones de convexidad de Barnard y de simetría en la misma cola, entonces el tamaño de la prueba está dado por*

$$\begin{array}{c} \text{máx} \\ p_2 = p_1 - d_0 \\ p_1 \in [d_0, \frac{1+d_0}{2}] \end{array} \quad \sum_{x_1=0}^{n} \sum_{x_2=0}^{n} \prod_{i=1}^{2} \binom{n}{x_i} p_i^{x_i} (1-p_i)^{n-x_i} I_{[(x_1,x_2)\in R(\alpha)]}. \tag{3}$$

La fórmula (3) reduce el tiempo de cómputo de los tamaños de prueba en más de 90 %, en términos generales. Por otra parte, Almendra (2009b) demostró la siguiente:

Proposición 3.2. *Si $n_1 = n_2 = n$, entonces T_i cumple la condición de simetría en la misma cola para $i = 0, 1, 2, 3, 4, 5$.*

Además, Martin y Herranz (2002, 2004a y 2004b) definieron el espacio muestral lícito como $\chi' = \{(x_1, x_2) : \frac{x_1}{n_1} - \frac{x_2}{n_2} < d_0\}$ dichos autores enfatizan la importancia de usar el espacio muestral lícito en lugar del espacio muestral en la construcción de las regiones críticas en pruebas de no inferioridad para la diferencia de dos proporciones a fin de evitar la realización de inferencias incorrectas. Almendra (2009b) demostró la siguiente

Proposición 3.3. *Si T es una estadística de la forma $T(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}}$ donde $\hat{p}_i$ es un estimador de p_i para $i = 1, 2$ y $\hat{\sigma}$ es un estimador de la desviación estándar de $\hat{p}_1 - \hat{p}_2$ y $t_0 < 0$, entonces las regiones críticas correspondientes a los espacio muestral y muestral lícito son iguales.*

De la proposición 3.3 se tiene que en particular las regiones críticas correspondientes al espacio muestral y muestral lícito para la z-estadística asintótica ponderada son iguales. De la proposición 3.2, las regiones críticas para las pruebas que estamos estudiando cumplen la condición (S). Que dichas regiones críticas cumplen la condición (C) fue verificado numéricamente mediante un programa escrito por el autor en S-PLUS. Debido a que, como acabamos de establecer, se cumplen las condiciones de la proposición 3.1, se usó la fórmula 3 para calcular los tamaños de prueba. Para ello se tomó un incremento de 0.001 en p_1.

4. Resultados

En la tabla 1 se muestran los porcentajes de los tamaños de prueba que pertenecen a cada uno de los intervalos especificados. Se consideró que un alto porcentaje de tamaños de prueba en el intervalo $[0, .8\alpha)$ significaba que la prueba era demasiado conservadora, mientras que un alto porcentaje en el intervalo $[.8\alpha, \alpha]$ indicaba una prueba razonablemente conservadora. En constraste, un alto porcentaje de tamaños de prueba en el intervalo $(\alpha, 1.2\alpha]$ indicaba que la prueba era razonablemente liberal y, finalmente, un alto porcentaje de tamaños de prueba en el intervalo $(1.2\alpha, 1]$ indicaba que la prueba era demasiado liberal.

α	d_0	*Intervalo*	T_0	T_1	T_2	T_3	T_4	T_5
		$(.03, 1]$	100.00	100.00	100.00	67.61	32.39	8.45
	0.1	$(.025, .03]$	0.00	0.00	0.00	26.76	15.49	11.27
		$[.02, .025]$	0.00	0.00	0.00	5.63	22.54	15.49
		$[0, 0.02)$	0.00	0.00	0.00	0.00	29.58	69.79
0.025		$(.03, 1]$	100.00	100.00	92.96	42.25	11.27	1.41
	0.15	$(.025, .03]$	0.00	0.00	7.04	33.80	19.72	2.82
		$[.02, .025]$	0.00	0.00	0.00	22.54	23.94	15.49
		$[0, 0.02)$	0.00	0.00	0.00	1.41	45.07	80.28
		$(.06, 1]$	100.00	100.00	100.00	100.00	83.10	32.39
	0.1	$(.05, .06]$	0.00	0.00	0.00	0.00	16.90	36.62
		$[.04, .05]$	0.00	0.00	0.00	0.00	0.00	30.00
		$[0, 0.04)$	0.00	0.00	0.00	0.00	0.00	0.00
0.05		$(.06, 1]$	100.00	100.00	100.00	100.00	78.87	22.54
	0.15	$(.05, .06]$	0.00	0.00	0.00	0.00	19.72	39.44
		$[.04, .05]$	0.00	0.00	0.00	0.00	1.41	38.03
		$[0, 0.04)$	0.00	0.00	0.00	0.00	0.00	0.00

Tabla 1: Porcentajes de tamaños de prueba que pertenecen al intervalo especificado.

5. Conclusiones

Con base en los resultados presentados en la tabla 1, se observa que el comportamiento de los tamaños de prueba para las configuraciones estudiadas de tamaños de muestra, márgenes

de no inferioridad y niveles de significancia nominal no están bien controlados, es decir, son demasiado erráticos. Por lo tanto, para dichas configuraciones el uso de esta prueba, no es recomendable.

Referencias

Almendra Arao, F. 2009a."A Study of the Asymptotic Classical Non-inferiority Test for two Binomial Proportions". *Drug Information Journal*, 43 (5): 567-571.

Almendra Arao, F. 2009b."Behavior of the Asymptotic Pooled z-statistics". *Journal of Biostatistics*, 3 (3):247-256.

Basu, D. 1977. "On the elimination of nuissance parameters". *Journal of the American Statistical Association*, 72:355-366.

Haber, M. 1986."An Exact Unconditional Test for 2x2 Comparative Trial". *Psychological Bulletin*, 99: 129-132.

Martin, A. A. y Herranz, T. I. 2002. "Equivalence Testing for Binomial Random Variables: Which Test to use?", Letter to the editor. *The American Statistician*, 3 : 253-254.

Martin, A. A. y Herranz, T. I. 2004a."Asymptotical test on the equivalence, substantial difference and non-inferiority problems with two proportions". *Biometrical Journal*, 46: 305-319.

Martin, A. A. y Herranz, T. I. 2004b."Exact unconditional non-classics tests on the difference of two proportions". *Computational Statistics and Data Analysis*, 45: 373-388.

Suissa, S. S. y Shuster, J.J. 1985."Exact Unconditional Sample Sizes for 2x2 Binomial Trial". *Journal of the Royal Statistical Society, Ser. A*, 148: 317-327.

Memoria FONE XXIV (2010), Sección I, 19–23

Estimadores ridge en regresión logística cuando hay separación en los datos y colinealidad

Elia Barrera Rodriguez[a], Flaviano Godínez Jaimes[b], Francisco J. Ariza Hernández[c], Ramón Reyes Carreto[d]

Unidad Académica de Matemáticas, Universidad Autónoma de Guerrero

1. Introducción

Regresión logística es uno de los modelos más usados para modelar una variable respuesta dicotómica en función de un conjunto de variables explicatorias. Existen situaciones en que éste modelo presenta problemas, esto ocurre cuando hay separación en los datos y/o cuando hay colinealidad. Cuando hay separación en los datos existe una variable explicatoria o una combinación lineal de las variables explicatorias que predice de manera perfecta la variable respuesta. Geométricamente, hay separación en los datos si existe un hiperplano que separa los éxitos de los fracasos, hay cuasi separación si además, al menos una observación esta en el hiperplano y hay traslape si no existe tal hiperplano. Albert y Anderson (1984) demostraron que el estimador de máxima verosimilitud (EMV) no existe cuando hay separación o cuasi separación y existe y es único cuando hay traslape en los datos. Hay colinealidad en las variables explicatorias si existen fuertes dependencias lineales entre ellas. Ambos problemas causan que el EMV y su varianza estimada tiendan a ser grandes, lo que puede llevar a inferencias erróneas. El objetivo de este trabajo es comparar mediante simulación estimadores que abordan el problema de separación y/o colinealidad en el modelo de regresión logística, en función de su sesgo y error cuadrático medio.

[a]he-lya@hotmail.com
[b]fgodinezj@gmail.com
[c]arizahfj@colpos.mx
[d]rcarreto1@yahoo.com

Sea $\left\{\left(Y_i, \mathbf{x}_i^T\right) : i = 1, ..., n\right\}$ una muestra en regresión binaria, esto es, $Y_i \sim Bernoulli(\pi_i)$ son variables aleatorias independientes y $\mathbf{x}_i^{T*} = (x_{i1}, \ldots, x_{ip})$ son vectores no estocásticos de variables explicatorias de dimensión p $(p < n)$. En regresión logística $\pi_i = e^{\mathbf{x}_i^T \beta} \Big/ \left(1 + e^{\mathbf{x}_i^T \beta}\right)$ y la función de log verosimilitud es:

$$l(\beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\}$$

El EMV, $\hat{\beta}$, se obtiene maximizando $l(\beta)$. Si $\hat{\beta}$ esta en el interior del espacio de parámetros hay que resolver un sistema de ecuaciones no lineal usando el Método de Newton-Raphson (MNR) que en la iteración $k + 1$ establece que:

$$\beta^{k+1} = \beta^k + I\left(\beta^k\right)^{-1} U\left(\beta^k\right),$$

donde $U(\beta) = X^T(Y - \pi)$, $I(\beta) = X^T V X$ $V = diag\{\pi_1(1 - \pi_1), ..., \pi_n(1 - \pi_n)\}$ y X es la matriz diseño de $n \times (p + 1)$ cuyos renglones son $\mathbf{x}_i^T = \left(1\ \mathbf{x}_i^{T*}\right)$.

En presencia de colinealidad en las variables explicatorias, X y $X^T V X$ pueden ser de rango completo pero $\hat{\beta}$ puede ser afectado. Algunos de los efectos que produce la colinealidad son: a) sensibilidad de $\hat{\beta}$ a cambios pequeños en los datos, b) inestabilidad en $\hat{\beta}$, y c) $\hat{V}(\hat{\beta}_i)$ grandes y por tanto intervalos de confianza muy grandes y baja potencia de las pruebas de hipótesis.

Belsley *et al.* (1980) sugirieron usar el número de condición escalado, η_X, para determinar la presencia de colinealidad, el cual se define por $\eta_X = \sqrt{\lambda_1/\lambda_p}$ donde λ_1 y λ_p son el máximo y el mínimo de los valores propios de $X^T X$ después de ser escalada. La colinealidad es un problema de grado, si $\eta_X < 10$ entonces no hay colinealidad, si $10 \leq \eta_X < 30$ entonces hay colinealidad moderada y la inferencia puede ser afectada gravemente, pero si $\eta_X \geq 30$ entonces hay colinealidad severa lo que puede afectar seriamente la estimación y la inferencia.

Enseguida se presentan los estimadores encontrados en la literatura que manejan los problemas de colinealidad y/o separación en los datos.

Heinze y Schemper (2002) encontraron que el estimador de Firth (1993) que se obtiene penalizando la función de verosimilitud con la apriori invariante de Jeffreys existe cuando hay separación en los datos. La función de log verosimilitud penalizada es:

$$l^F(\beta) = l(\beta) + \tfrac{1}{2} \log |I(\beta)|.$$

Así, $\hat{\beta}_F$ es el estimador de β que se obtiene maximizando $l^F(\beta)$ usando el MNR con $U^F(\beta) = U(\beta) + diag(H)\pi(\beta)$, $I^F(\beta) = I(\beta)$ y $H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$. $\hat{\beta}_F$ es afectado por la existencia de colinealidad en las variables explicatorias y no existe en el caso extremo de separación en los datos, esto es, cuando todas las respuestas son del mismo tipo.

Rousseeuw y Christman (2003) introducen el *modelo de regresión logística escondido* en el que las observaciones se transforman en pseudo-observaciones, $\tilde{y}_i = (1 - y_i)\delta_0 + y_i\delta_1$, donde δ_1 y δ_0 son la probabilidad de observar $Y = 1$ dado que el valor verdadero es un éxito y un fracaso respectivamente. Las pseudo-observaciones permiten obtener la función de verosimilitud estimada, $L\left(\tilde{\beta}\,|\tilde{y}_1, ..., \tilde{y}_n\right) = \prod_{i=1}^{n} \pi_i^{\tilde{y}_i}(1 - \pi_i)^{1-\tilde{y}_i}$, y $\hat{\beta}_{RC}$ se obtiene al maximizarla como se describe en el articulo correspondiente. Si $0 < \delta_0 < \delta_1 < 1$ y X es de rango completo entonces $\hat{\beta}_{RC}$ siempre existe y es único. Además, $\hat{\beta}_{RC}$ existe cuando hay separación extrema.

Shen y Gao (2008) para enfrentar ambos problemas, colinealidad y separación, penalizan la función de log verosimilitud doblemente:

$$l^{SG}(\beta) = l(\beta) + \tfrac{1}{2}\log|I(\beta)| - \lambda\|\beta\|^2,$$

y $\hat{\beta}_{SG}$ se obtiene maximizando $l^{SG}(\beta)$ mediante el MNR con $U^{SG}(\beta) = U^F(\beta) - 2\lambda\beta$, $I^{SG}(\beta) = I(\beta)$. El parámetro ridge, λ, controla la reducción en la norma de β y se obtiene minimizando mediante validación cruzada la media del cuadrado del error: $MCE_{VC}(\lambda) \approx \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\pi}_i^{\lambda}\right)^2/(1 - h_{ii})^2$.

El estimador ridge logístico de un paso es dado por:

$$\hat{\beta}_R(k) = \left[X^T\hat{V}X + \lambda\mathbf{I}\right]^{-1}X^T\hat{V}X\hat{\beta} \tag{1}$$

donde λ es el parámetro de ridge, $\mathbf{I}$ la matriz identidad de orden $p + 1$ y $\hat{\beta}$ es el EMV.

El estimador ridge iterativo logístico fue propuesto por le Cessie y van Houwelingen (1992) mediante la penalización de la función de log verosimilitud dada por:

$$l^{CH}(\beta) = l(\beta) - \lambda\|\beta\|^2, \tag{2}$$

y $\hat{\beta}_{RI}$ se obtiene maximizando $l^{CH}(\beta)$ con el MNR con $U(\beta_{RI}) = U(\beta) - 2\lambda\beta$ y $I(\beta_{RI}) = I(\beta) - 2\lambda\mathbf{I}$.

Godínez y Valverde (2005) introducen dos estimadores de tipo $\hat{\beta}_{RC}$, RCA y RCS, donde δ_0 y δ_1 se determinan minimizando $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\pi}_i)^2$ bajo dos situaciones. En la primera δ_0

y $1-\delta_1$ toman valores diferentes y en la segunda toman valores iguales. Los estimadores resultantes se denotan por $\hat{\beta}_{RCA}$ y $\hat{\beta}_{RCS}$. También estudian estimadores ridge logísticos de un paso basados en el estimador de Firth (RF), de Rousseeuw y Christman (RRC, RRCA y RRCS) que se obtienen al usar $\hat{\beta_F}$, $\hat{\beta}_{RC}$, $\hat{\beta}_{RCA}$ y $\hat{\beta}_{RCS}$ en (1) con $\lambda = \lambda_L = (\gamma_1 - 100\gamma_p)/99$, $X = X^*$ es la matriz donde la primer columna es el vector de unos normalizado y las variables se han estandarizado y $\hat{V}$ depende de $\hat{\beta_F}$, $\hat{\beta}_{RC}$, $\hat{\beta}_{RCA}$ y $\hat{\beta}_{RCS}$. El estimador ridge iterativo (RI) se obtiene usando λ_L en (2).

2. Metodología

Se generó una matriz diseño de 3×40 con problemas de colinealidad moderada y severa ($\eta_X =$ 16 y 32) con $X_1 \sim U[0,1]$ y $X_2 = X_1 + cV$; donde c se elige apropiadamente y $V \sim U[0,1]$. El grado de traslape se midió con la función *noverlap* de R generando tres grupos: G0, G1 y G2 que tienen [0, 1], [2, 5] y [6, 9] observaciones traslapadas respectivamente. Se hicieron 2000 repeticiones y los estimadores se compararon en función de su sesgo, $S(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\left(\widetilde{\beta}_r - \beta\right)$, y error cuadrático medio, $ECM(\widetilde{\beta}) = \frac{1}{R}\sum_{r=1}^{R}\left(\widetilde{\beta}_r - \beta\right)^T\left(\widetilde{\beta}_r - \beta\right)$.

3. Resultados

El sesgo de los estimadores ridge es más cercano a cero que el sesgo del EMV. Se observa efecto del porcentaje de observaciones traslapadas pues en el grupo con separación o cuasi separación hay mayor sesgo que en los otros dos grupos. También se observa efecto del grado de colinealidad pues en colinealidad severa hay mayor sesgo de los estimadores. Los estimadores SG y RI tienen sesgo mas pequeño en presencia de colinealidad moderada y separación en los datos. Es mayor el sesgo de β_2 que de β_1. Un comportamiento similar se observa en el ECM en cuanto al efecto del porcentaje de traslape y del grado de colinealidad. El estimador con menor ECM es el Ridge iterativo (Tabla 1).

4. Conclusiones

El estimador de Shen y Gao intenta manejar ambos problemas, pero en los datos que usa no tienen colinealidad pues el $\eta_X < 10$. En este trabajo se encontró que el estimador ridge

iterativo es mejor en sesgo y ECM pero el escenario de simulación es diferente. Es conocido que en presencia de colinealidad el EMV es insesgado y los estimadores ridge son sesgados pero en los resultados no se observa esto tal vez debido al número de repeticiones.

Referencias

Albert, A. y Anderson, J. A. 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 71:1-10.

Belsley, *et. al.* 1980. *Regression diagnostics: Identifying Influential Data and Source of Collinearity*. John Wiley & Sons. New York. 393 p.

Godínez J., F y Ramírez, V.G. 2005. Estimación en el modelo de regresión logística en presencia de datos separados y colinealidad. *Memorias del XIX Foro Nacional de Estadística*. INEGI: México.

Heinze, G. y Schemper, M. 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 21, 2409-2419.

le Cessie, S. y van Houwelingen J.C. 1992. Ridge estimators in logistic regression. *Applied Statistics*. 41(1):191-201.

Rousseeuw, P. J., y Christmann, A. 2003. Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis* 43, 315-332.

Shen J. y Gao, s 2008. A Solution to Separation and Multicollinearity in Multiple Logistic Regression. *Journal of Data Science*, 515-531.

	G	η_x	β	MV	RRC	RRCA	RRCS	SG	RI
		16	β_1	NA	-0.27	2.10	-0.28	-0.19	-0.13
	G0		β_2	NA	-0.63	1.08	-0.63	-0.12	-0.14
		32	β_1	NA	0.50	5.77	-0.35	-0.53	-0.14
			β_2	NA	-0.65	5.11	-0.53	-0.15	-0.14
		16	β_1	0.43	-0.73	-0.75	-0.75	-0.11	-0.21
SESGO	G1		β_2	0.41	-1.00	-1.02	-1.01	-0.32	-0.27
		32	β_1	1.81	-0.70	-0.72	-0.72	-0.28	-0.22
			β_2	-1.28	-0.83	-0.85	-0.84	-0.38	-0.25
		16	β_1	0.20	-0.78	-0.79	-0.79	-0.58	-0.38
	G2		β_2	-0.78	-1.05	-1.05	-1.05	-0.35	-0.51
		32	β_1	0.73	-0.77	-0.77	-0.77	-1.01	-0.41
			β_2	-1.48	-0.89	-0.89	-0.89	-0.22	-0.47
		16	β_1	NA	2.87	289.34	2.38	14.74	0.05
	G0		β_2	NA	2.20	153.33	1.89	13.29	0.05
		32	β_1	NA	3.39	2537.62	3.44	60.91	0.05
			β_2	NA	2.59	1875.25	2.62	59.36	0.05
		16	β_1	72.21	0.57	0.60	0.59	6.05	0.10
ECM	G1		β_2	59.78	1.02	1.05	1.05	5.49	0.13
		32	β_1	323.07	0.53	0.55	0.54	33.41	0.10
			β_2	303.62	0.72	0.74	0.74	32.50	0.11
		16	β_1	18.73	0.62	0.64	0.63	4.69	0.18
	G2		β_2	17.70	1.10	1.12	1.11	3.86	0.30
		32	β_1	91.88	0.60	0.60	0.60	31.40	0.21
			β_2	94.24	0.80	0.81	0.80	28.79	0.26

Tabla 1: Sesgo y ECM para β_1 y β_2 en los diferentes grados de colinealidad y traslape

Memoria FONE XXIV (2010), Sección I, 25–31

Series de tiempo con múltiples puntos de cambio y observaciones censuradas

René Castro Montoya[a]
Universidad Autónoma de Sinaloa

Gabriel A. Rodríguez Yam
Universidad Autónoma Chapingo.

Sergio Pérez Elizalde
Colegio de Postgraduados.

1. Introducción

Debido a factores externos a las variables de interés, una serie de tiempo puede presentar cambios en la estructura del modelo o en algunos de los parámetros y debido a límitaciones en los instrumentos de medición presentar tambien censura en las observaciones. Por ejemplo, cuando se monitorean contaminantes del aire, como pueden ser hidrocarburos aromáticos(PAHs), monóxido de carbono (CO), dióxido de sulfuro (SO_2), etc., las series de tiempo obtenidas pueden tener observaciones censuradas y cambios en la estructura del modelo.

El problema de series de tiempo con puntos de cambios y observaciones censuradas es un problema de inferencia estadística en el que se desconocen los parámetros de el modelo y el número de parámetros. Éste problema se puede formular mediante inferencia conjunta de un indicador r del modelo y el vector de parámetros θ_r, donde el indicador del modelo determina la dimensión n_r.

En este trabajo se propone un modelo bayesiano para series de tiempo con un número desconocido de puntos de cambio y observaciones censuradas, donde cada segmento es un proceso autoregresivo de orden uno. Se consideran iniciales conjugadas para las medias y las

[a]renec@uas.uasnet.mx

varianzas en cada segmento, excepto para los coeficientes autoregresivos, ya que se condiciona para que la serie sea estacionaria en los segmentos.

Para estimar el número y las localizaciones de los puntos de cambio se utiliza el algoritmo de cadenas de Markov Monte Carlo con saltos reversibles (RJMCMC) desarrollado por Green(1995), este algoritmo consiste en crear una cadena de Markov irreducible y aperíodica que alterna saltos entre varios modelos con espacios de parámetros de diferente dimensión, que cumpla la condición de probabilidad de equilibrio, asegurando la convergencia a la distribución final. El problema de censura se resuelve simulando los valores censurados de una distribución normal multivariada de la parte censurada dada la parte observada Jung et. al (2005). Para ilustrar el algoritmo se analiza un conjunto de datos de simulados con distintos porcentajes (0 %,10 %, 40 %) de censura.

2. Modelo bayesiano para el problema de series de tiempo con puntos de cambio y observaciones censuradas

Algunos autores han desarrollado métodos para analizar series de tiempo con observaciones censuradas, e.g., Robinson(1980) trabaja el caso de series de tiempo autoregresivas con observaciones censuradas agrupando las observaciones tal que cada segmento incluya una observación censurada y así se requiera una integral univariada, para simular las observaciones censuradas mediante la esperanza condicional de la parte censurada dada la parte observada. Jung et al(2005) trabaja el caso de series de tiempo autoregresivas con observaciones censuradas simulando las observaciones censuradas mediante un vector de valores de la distribución normal multivariada de la parte censurada dada la parte observada, éste método consiste en los siguientes pasos: i) construir una matriz de permutación, ii) seleccionar valores iniciales para la media y la matriz de covarianzas de la distribución normal multivariada, iii) simular los valores censurados mediante la distribución normal multivariada condicional truncada de la parte censurada dada la parte observada, iv) utilizar éstas para completar el conjunto de observaciones, v) estimar la media y la matriz de covarianzas de la distribución normal multivariada. Ariza et al (2008) utilizan el algoritmo EM para estimar los parámetros de un modelo de espacio de estados con observaciones censuradas. El problema de puntos

de cambio ha sido objeto de estudio de muchos autores, e.g., Davis et al (2006) modelan series de tiempo no estacionarias segmentando la serie en bloques de procesos autoregresivos, se asumen desconocidos el número de puntos de cambio, sus localizaciones y orden de los procesos autoregresivos en cada segmento.

El modelo bayesiano para series de tiempo con puntos de cambio y con observaciones censuradas que se propone en esta sección es como sigue. Sea y_1, y_2, ..., y_n la realización de una serie de tiempo, con k puntos de cambio en las localizaciones $\tau_1, \tau_2, \ldots, \tau_k$, donde k y τ_1, $\tau_2, \ldots, \tau_k$ son desconocidos. Además, se asume que algunas observaciones presentan censura. Por conveniencia se considera censura por la derecha en c_t, $t = 1, 2, \ldots, n$, es decir, en lugar de observar y_t, se tiene $x_t := \text{mín}(y_t, c_t)$. Por conveniencia, se define $\tau_0 := 0$ y $\tau_{k+1} := n$. En cada segmento, condicionado a los parámetros, se asume un proceso autoregresivo de orden 1, es decir,

$$X_t = \mu_i + \phi_i(X_{t-1} - \mu_i) + \epsilon_t, \quad \tau_{i-1} + 1 \leq t \leq \tau_i, \qquad (1)$$
$$i = 1, 2, \ldots, k+1,$$

donde $\epsilon_t \overset{iid}{\sim} N(0, \sigma_i^2)$. Aquí k, τ_1, τ_2, ..., τ_k, $(0 < \tau_1 < \tau_2 < \ldots < \tau_k < n)$, μ_1, μ_2, ..., μ_{k+1}, σ_1^2, σ_2^2, ..., σ_{k+1}^2, ϕ_1, ϕ_2, ..., ϕ_{k+1} son los parámetros del modelo. Las distribuciones iniciales para el número de puntos de cambio y sus localizaciones están dadas por,

$$\begin{aligned} K &\sim U(0, 1, 2, \ldots, k_{max}), \\ f(\tau_i | \tau i - 1, k) &\sim U(\tau_{i-1} + 1, \ldots, n-1), \quad i = 1, 2, \ldots, k, \end{aligned}$$

donde k_{max} es el máximo número de puntos de cambio que se permite en el modelo. Para las medias μ_i y las varianzas σ_i^2, $i = 1, 2, \ldots, k+1$, se consideran distribuciones iniciales conjugadas, i.e.,

$$\mu_i \sim N(\mu_0, \sigma_0^2), \qquad (2)$$
$$\sigma_i^2 \sim Ig(\alpha_0, \beta_0), \qquad (3)$$

donde μ_0, σ_0^2, α_0 y β_0 son hiperparámetros. Para asegurar estacionariedad en cada segmento la distribución inicial que se considera para ϕ_i, $i = 1, 2, \ldots, k+1$, es

$$\phi_i \sim U(-1, 1).$$

Inferencia bayesiana sobre K y $\theta_k = (\tau_1, \tau_2, \ldots, \tau_k, \mu_1, \mu_2, \ldots, \mu_{k+1}, \sigma_1^2, \sigma_2^2, \ldots, \sigma_{k+1}^2, \phi_1, \phi_2, \ldots, \phi_{k+1})$, se basa en la distribución final $f(y_c, \theta_k, k|y_o)$, donde $y_o := \{y_i \,|y_i \leq c_i,\ 1 = 1, 2, \ldots, n\}$ y $y_c := \{c_i \,|y_i > c,\ i = 1, 2, \ldots, n\}$, la cual se puede factorizar como sigue,

$$\begin{aligned} f(y_c, \theta_k, k|y_o) &= f(y_c|y_o, \theta_k, k) f(\theta_k, k|y_o) \\ &\propto f(y_c|y_o, \theta_k, k) f(y_o|\theta_k, k) f(\theta_k, k) \\ &\propto f(y|\theta_k, k) f(\theta_k, k), \end{aligned}$$

3. Análisis de una serie de tiempo con dos puntos de cambio y distintos porcentajes de observaciones censuradas (0 %, 10 %, 40 %)

En esta sección se aplica el algoritmo RJMCMC a un conjunto de datos simulados. El proceso considerado en esta sección tiene 600 observaciones con puntos de cambio en $\tau_1 = 200$ y $\tau_2 = 400$. Los parámetros para de los segmentos autoregresivos son $\mu = (12, 12, 12),\ \sigma = (1.73, 1.22, 1.73)\ y\ \phi = (0.5, 0.79, -0.5)$. Este proceso se expresa como sigue

$$Y_t = \begin{cases} 12 + 0.5(Y_{t-1} - 12) + \epsilon_t, & 1 \leq t \leq 200, \\ 12 + 0.79(Y_{t-1} - 12) + \epsilon_t, & 201 \leq t \leq 400, \\ 12 - 0.5(Y_{t-1} - 12) + \epsilon_t, & 401 \leq t \leq 600, \end{cases} \tag{4}$$

A una realización del modelo en (4) se graficó y censuró 10 % de las observaciones. En el panel (a) de la Figura 1 se muestra una realización de este proceso y en el panel (b) la Figura 1 se muestra la realización con censura.

Los hiperparámetros que se requieren en (2) para las distribuciones iniciales de las medias y las variazas son: $\mu_0 = 12$, $\sigma_0^2 = 3$, $\alpha_0 = 5$, $\beta_0 = 1.2$. El valor inicial para el número de puntos de cambio se fija en $k = 2$. Se implementó el algoritmo RJMCMC (con $50,000$ iteraciones) para generar una muestra de la distribución final.

Con la finalidad de verificar si hay convergencia en el algoritmo RJMCMC se utiliza la prueba de convergencia de Castelloe (1998) implementada en un programa en R, a fin de realizar la prueba de convergencia antes mencionada. En la Figura 2 se muestra las medias del número de puntos de cambio. Como se observa en esta figura, el número de puntos de

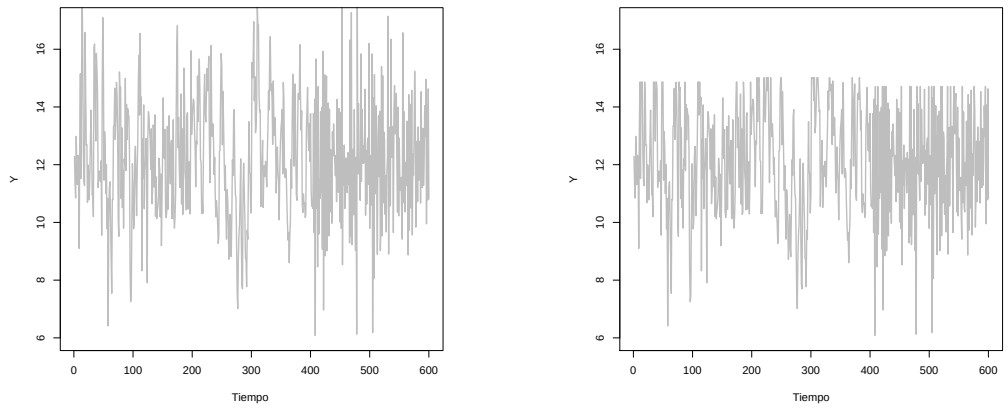


Figura 1: a) Una realización del proceso en (4)., b) Observaciones censuradas (10 %) de la realización del proceso (4).

cambio es $k = 2$. En la Figura 2 se muestra la distribución final de K, el número de puntos de cambio. Como se observa en esta figura, el número de puntos de cambio con mayor probabilidad es $k = 2$. El valor estimado para el número de puntos de cambio es 2.

En el Cuadro 1 se muestran los valores verdaderos, los valores estimados y la desviacion estándar para los puntos de cambio, las localizaciones de los puntos de cambio, así como las medias, las varianzas y los coeficientes autoregresivos en cada segmento que se obtiene al usar la muestra obtenida con el algoritmo RJMCMC.

4. Conclusiones

Se propuso un modelo bayesiano para series de tiempo con puntos de cambio y algunas observaciones censuradas, para las medias y las varianzas en cada segmento se consideran

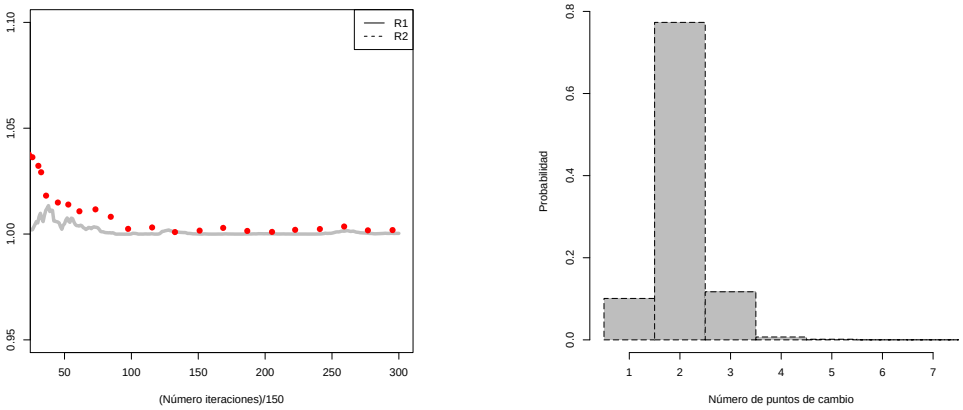


Figura 2: Izquierda) prueba para monitorear convergencia de Castelloe (1998), derecha) Histograma del número de puntos de cambio

distribuciones iniciales conjugadas, ya que el problema de series de tiempo con puntos de cambio y algunas observaciones censuradas implica saltos entre distintos modelos con espacios parametrales de diferente dimensión, los métodos clásicos de MCMC no se pueden implementar para este modelo. Se implementó el algoritmo RJMCMC para generar muestras de la distribución final de $(K,\, \theta_k)$, para resolver el problema de censura se implementó el método de Jung et. al. (2005). En el ejemplo numérico que se presentó se observa que las estimaciones del número de puntos de cambio y de las localizaciones de estos puntos de cambio son razonables. Los segmentos AR(1) se pueden reemplazar por cualquier modelo de serie de tiempo.

Tabla 1: Estimaciones para el caso de dos puntos de cambio.

Parámetros	Sin censura	10 % de censura	40 % de censura
K	2	2	2
	(0.44)	(0.50)	(0.61)
τ_1	209	213	199
	(41.24)	(36.89)	(41.38)
τ_2	391	394	398
	(13.01)	(12.86)	(12.47)
ϕ_1	0.44	0.41	0.34
	(0.06)	(0.06)	(0.079)
ϕ_2	0.77	0.77	0.76
	(0.15)	(0.14)	(0.137)
ϕ_3	-0.52	-0.509	-0.35
	(0.05)	(0.05)	(0.059)

Referencias

Ariza-Hernandez, F. J. and Rodríguez-Yam, G. A.(2008). "Analysis of Time Series with Censured Observations". Asociación Mexicana de Estadística.

Castelloe J.M. (1998).Issues in reversible jump Markov chain Monte Carlo and composite EM analysis, applied to spatial poisson cluster processes",PhD. Thesis University of Iowa.

Davis R. A.,Lee T. C. M., Rodriguez-Yam G. A. (2006). "Structural Breaks Estimation for Non-stationary Time Series Models".*Journal of the American Statistical Association*,101,223-239.

Green P.(1995) "Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination".,*Biometric*,82,711-732.

Jung W. P.,Mark G. G.,Sujit K. G.(2005). "Censored Time Series Analysis with Autoregresive Moving Average Models",*Econometric*,10,234-256.

Robinson P. M.(1980). "Estimation and Forescasting for Time Series Containing Censored or Missing Observations",*Time Series*,167-182. North-Holland Publishing Company.

Memoria FONE XXIV (2010), Sección I, 33–39

Intervalos de confianza para el tamaño de una población de difícil detección en el muestreo por bola de nieve y probabilidades de nominación heterogéneas*

Martín H. Félix Medina[a], Aida N. Aceves Castro y Pedro E. Monjardin
Escuela de Ciencias Físico- Matemáticas de la Universidad Autónoma de Sinaloa

1. Introducción

Existen diferentes métodos para muestrear poblaciones raras [ver Christman (2009) para una revisión actualizada]. Uno de estos es el Muestreo por Bola de Nieve, el cual, 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 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 de muestreo 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 y Monjardin (2008) debilitaron el supuesto de homogeneidad y desarrollaron EMVs del tamaño poblacional bajo el supuesto de probabilidades de nominación heterogéneas. En este trabajo se desarrollan intervalos de confianza tipo Wald y tipo verosimilitud perfil basados en los estimadores propuestos por Félix-Medina y Monjardin (2008).

*Trabajo realizado con apoyos parciales de los proyectos PIFI-2005-25-06 de la SEP y PROFAPI 2008/054 de la UAS.

[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 $S_A = \{A_1, \ldots, A_n\}$ de n sitios, y a las personas de la población de interés que pertenecen 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 τ el tamaño de U, por τ_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, $X_{ij}^{(k)} = 1$ si la persona $u_j \in U_k - A_i$ es nominada por el sitio A_i, y $X_{ij}^{(k)} = 0$ en otro caso, $k = 1, 2$.

Como en Félix-Medina y Monjardin (2008), supondremos que las M_i's son variables aleatorias independientes con distribución Poisson con media λ_1. Así, 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 τ_1 y vector de probabilidades $(1/N, \ldots, 1/N, 1 - n/N)$. Supondremos también 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 Rash. 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 normal con media cero y varianza σ_k^2 $[N(0, \sigma_k^2)]$ desconocida.

3. Estimadores máximo verosímiles

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 $\alpha^{(k)} = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$ y $\phi(z)$ representa la función de densidad normal estándar.

Félix-Medina y Monjardin (2008) demostraron que la función de verosimilitud está dada

por:

$$L(\tau_1, \tau_2, \sigma_1, \sigma_2, \alpha^{(1)}, \alpha^{(2)}) = L_1(\tau_1, \sigma_1, \alpha^{(1)}) L_2(\tau_2, \sigma_2, \alpha^{(2)}),$$

donde:

$$L_1(\tau_1, \sigma_1, \alpha^{(1)}) \propto \frac{\tau_1!}{(\tau_1 - m - r_1)!}(1 - n/N)^{\tau_1 - m} \prod_{\omega \subset \Omega - \emptyset} [\pi_\omega^{(1)}(\sigma_1, \alpha^{(1)})]^{r_\omega^{(1)}} [\pi_\emptyset^{(1)}(\sigma_1, \alpha^{(1)})]^{\tau_1 - m - r_1}$$
$$\times \prod_{i=1}^{n} \prod_{\omega \subset \Omega_i - \emptyset} [\pi_\omega^{(A_i)}(\sigma_1, \alpha_{-i}^{(1)})]^{r_\omega^{(A_i)}} [\pi_\emptyset^{(A_i)}(\sigma_1, \alpha_{-i}^{(1)})]^{m_i - r_\omega^{(A_i)}},$$
$$L_2(\tau_2, \sigma_2, \alpha^{(2)}) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{\omega \subset \Omega - \emptyset} [\pi_\omega^{(2)}(\sigma_2, \alpha^{(2)})]^{r_\omega^{(2)}} [\pi_\emptyset^{(2)}(\sigma_2, \alpha^{(2)})]^{\tau_2 - r_2},$$

$\alpha_{-i}^{(1)} = (\alpha_1^{(1)}, \ldots, \alpha_{i-1}^{(1)}, \alpha_{i+1}^{(1)}, \ldots, \alpha_n^{(1)})$, $\Omega_i = \Omega - \{i\}$ y $\pi_\omega^{(A_i)}(\sigma_1, \alpha_{-i}^{(1)}) = \int \prod_{i' \neq i}^{n} \{\exp[x_{\omega i'}(\alpha_{i'}^{(1)} + \sigma_1 z)]/[1 + \exp(\alpha_{i'}^{(1)} + \sigma_1 z)]\} \phi(z) dz$, con $x_{\omega i'} = 1$ si $i' \in \omega$ y $x_{\omega i'} = 0$ en otro caso, es la probabilidad de que un individuo en A_i que es seleccionado al azar sea nominado sólo por cada uno de los sitios $A_{i'}$'s con $i' \in \omega \subseteq \Omega_i$. Cabe aclarar que en las expresiones de los factores de la función de verosimilitud r_k es el número de distintas personas nominadas en $U_k - S_0$; $r_\omega^{(k)}$ es el número de personas en $U_k - S_0$ que fueron nominadas sólo por cada uno de los sitios A_i con $i \in \omega \subset \Omega$, $k = 1, 2$, y $r_\omega^{(A_i)}$ es el número de personas en A_i que fueron nominadas sólo por cada uno de los sitios $A_{i'}$ con $i' \in \omega \subseteq \Omega_i$.

La maximización numérica de la función de verosimilitud con respecto a τ_1, τ_2, $\alpha^{(1)}$, $\alpha^{(2)}$, σ_1 y σ_2 produce las estimaciones máximo verosímiles $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\alpha}^{(1)}$, $\hat{\alpha}^{(2)}$, $\hat{\sigma}_1$ y $\hat{\sigma}_2$ de estos parámetros. La estimación máximo verosímil de τ es $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

Cabe aclarar que en el proceso de maximización de la función de verosimilitud las probabilidades $\pi_\omega^{(k)}$ y $\pi_\omega^{(A_i)}$ se calculan mediante el método de cuadratura Gaussiana.

4. Intervalos de confianza

4.1. Intervalos tipo Wald

Intervalos tipo Wald de aproximadamente el $100(1-\alpha)\,\%$ de confianza para τ_k y τ están dados por $\hat{\tau}_k \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\tau}_k)}$, $k = 1, 2$, y $\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\tau})}$, donde $z_{1-\alpha/2}$ es el $(1-\alpha/2)$avo cuantil de la distribución $N(0,1)$ y $\hat{V}(\hat{\tau}_k)$ y $\hat{V}(\hat{\tau})$ son estimadores de las varianzas de $\hat{\tau}_k$, $k = 1, 2$, y $\hat{\tau}$. Consideraremos dos tipos de estimadores: basados-en-modelo y basados-en-diseño.

Mediante el enfoque de Sanathanan(1972) se obtiene que si R_k, $k = 1, 2$, y M_i, $i = 1, \ldots, N$, son grandes, entonces estimadores basados-en-modelo de las varianzas de $\hat{\tau}_k$ y $\hat{\tau}$ están dados por:

$$\hat{V}_M(\hat{\tau}_k) = \hat{\tau}_k[\hat{D}_k - \hat{B}_k'\hat{A}_k^{-1}\hat{B}_k]^{-1}, \quad k = 1, 2, \quad \text{y} \quad \hat{V}_M(\hat{\tau}) = \hat{V}_M(\hat{\tau}_1) + \hat{V}_M(\hat{\tau}_2),$$

donde $\hat{A}_k = [\hat{a}_{ij}^{(k)}]_{n+1,n+1}$,

$$\hat{a}_{ij}^{(1)} = \left(1 - \frac{n}{N}\right) \sum_{\omega \subseteq \Omega} \frac{1}{\hat{\pi}_\omega^{(1)}} \left(\frac{\partial \hat{\pi}_\omega^{(1)}}{\partial \hat{\theta}_i^{(1)}}\right) \left(\frac{\partial \hat{\pi}_w^{(1)}}{\partial \hat{\theta}_j^{(1)}}\right) + \frac{1}{N} \sum_{\omega \subseteq \Omega_i} \frac{1}{\hat{\pi}_\omega^{(A_i)}} \left(\frac{\partial \hat{\pi}_\omega^{(A_i)}}{\partial \hat{\theta}_i^{(1)}}\right) \left(\frac{\partial \hat{\pi}_\omega^{(A_i)}}{\partial \hat{\theta}_j^{(1)}}\right),$$

$$\hat{a}_{ij}^{(2)} = \sum_{\omega \subseteq \Omega} \frac{1}{\hat{\pi}_\omega^{(2)}} \left(\frac{\partial \hat{\pi}_\omega^{(2)}}{\partial \hat{\theta}_i^{(2)}}\right) \left(\frac{\partial \hat{\pi}_w^{(2)}}{\partial \hat{\theta}_j^{(2)}}\right),$$

$$\hat{\theta}^{(k)} = (\hat{\theta}_1^{(k)}, \ldots, \hat{\theta}_n^{(k)}, \hat{\theta}_{n+1}^{(k)}) = (\hat{\alpha}_1^{(k)}, \ldots, \hat{\alpha}_n^{(k)}, \hat{\sigma}_k),$$

$$\hat{B}_k = (\hat{b}_1^{(k)}, \ldots, \hat{b}_{n+1}^{(k)})', \quad \hat{b}_j^{(k)} = -(\partial \hat{\pi}_\emptyset^{(k)} / \partial \hat{\theta}_j^{(k)}) / \hat{\pi}_\emptyset^{(k)},\ j = 1, \ldots, n+1,$$

$$\hat{D}_1 = [1 - (1 - n/N)\hat{\pi}_\emptyset^{(1)}] / [(1 - n/N)\hat{\pi}_\emptyset^{(1)}] \quad \text{y} \quad \hat{D}_2 = [1 - \hat{\pi}_\emptyset^{(2)}] / \hat{\pi}_\emptyset^{(2)}.$$

Un estimador basado-en-diseño, $\hat{V}_D(\hat{\tau}_1)$, de la varianza de $\hat{\tau}_1$ está dado por la expresión para $\hat{V}_M(\hat{\tau}_1)$, pero calculando $\hat{D}_1$ por $\hat{D}_1 = \{n/[\hat{\tau}_1(1-n/N)]\}S_M^2 + (1 - \hat{\pi}_\emptyset^{(1)})/[(1-n/N)\hat{\pi}_\emptyset^{(1)}]$, donde $S_M^2 = \sum_1^n (M_i - \bar{M})^2/(n-1)$ y $\bar{M} = \sum_1^n M_i/n$. Un estimador basado-en-diseño de la varianza de $\hat{\tau}_2$ es $\hat{V}_D(\hat{\tau}_2) = \hat{V}_M(\hat{\tau}_2)$ y uno de la varianza de $\hat{\tau}$ es $\hat{V}_D(\hat{\tau}) = \hat{V}_D(\hat{\tau}_1) + \hat{V}_D(\hat{\tau}_2)$.

4.2. Intervalos tipo verosimilitud perfil

Un intervalo de aproximadamente el $100(1-\alpha)\,\%$ de confianza para τ_k es $\{\tau_k : -2\ln[\Lambda(\tau_k)] \leq \chi^2_{1,1-\alpha}\}$, donde $\Lambda(\tau_k) = \max_{\sigma_k, \alpha^{(k)}} L_k(\tau_k, \sigma_k, \alpha^{(k)}) / L_k(\hat{\tau}_k, \hat{\sigma}_k, \hat{\alpha}^{(k)})$, $\hat{\tau}_k$, $\hat{\sigma}_k$ and $\hat{\alpha}^{(k)}$ son los EMV's de τ_k, σ_k y $\alpha^{(k)}$, $k = 1, 2$, y $\chi^2_{1,1-\alpha}$ es el $(1-\alpha)$avo cuantil de la distribución χ^2 con un grado de libertad. Un intervalo para τ se obtiene como el de τ_k, pero reemplazando $L_k(\tau_k, \sigma_k, \alpha^{(k)})$ en la expresión para $\Lambda(\tau_k)$ por $L(\tau, \tau_2, \sigma_1, \sigma_2, \alpha^{(1)}, \alpha^{(2)}) = L_1(\tau - \tau_2, \sigma_1, \alpha^{(1)}) L_2(\tau_2, \sigma_2, \alpha^{(2)})$ y maximizando con respecto a τ_2, σ_1, σ_2, $\alpha^{(1)}$ y $\alpha^{(2)}$. Intervalos ajustados por variación extra Poisson se obtienen como en los casos anteriores, pero se reemplaza $\chi^2_{1,1-\alpha}$ por $(S_M^2/\bar{M})\chi^2_{1,1-\alpha}$.

5. Estudio Monte Carlo

Se construyeron dos poblaciones de $N = 100$ valores M_i's. En la Población I los valores se generaron mediante una distribución Poisson con media 7.2, mientras que en la Población II

mediante una distribución Binomial negativa con media 7.2 y varianza 24.5. Así, se obtuvo que en la Población I: $\tau_1 = 725$, $\tau_2 = 500$ y $\tau = 1225$, y en la Población II: $\tau_1 = 716$, $\tau_2 = 500$ y $\tau = 1216$. Los valores de las probalidades $p_{ij}^{(k)}$ se generaron mediante el modelo Rash que se definió en la Sección 2, con $\alpha_i^{(k)} = 14.0/(M_i^{1/4} + 0.001)$ y $\beta_j^{(k)} \sim N(0, 0.75)$, $k = 1, 2$. El tamaño de la muestra inicial se fijó en $n = 10$. Los resultados para los intervalos tipo Wald se obtuvieron a partir de 500 muestras y, por restricciones de tiempo de cómputo, los resultados de los intervalos verosimilitud perfil a partir de 100 muestras. Los resultados del estudio se muestran en la Tabla 1.

6. Conclusiones

Los resultados del estudio de simulación indican que en presencia de probabilidades de nominación heterogéneas los estimadores de varianzas basados en el supuesto de homogeneidad presentaron sesgos relativos moderadamente grandes (en el intervalo [-0.50,-0.27]), mientras que los basados en el supuesto de heterogeneidad presentaron sesgos relativos de menores tamaños (en el intervalo [-0.03,0.15]), con la excepción de los estimadores de las varianza de $\hat{\tau}_1$ los cuales en la Población II presentaron sesgos relativos grandes (entre -0.55 y -0.35). Cabe aclarar que estos estimadores de varianza presentaron varianzas moderadamente grandes. Los intervalos del 95 % de confianza basados en el supuesto de homogeneidad tuvieron probabilidades de cobertura nulas, mientras que los basados en el supuesto de heterogeneidad probabilidades de cobertura muy variables (en el rango de 0.76 a 0.98).

Tabla 1. Poblaciones I y II. Sesgos relativos y raíces cuadras de errores cuadráticos medios relativos de estimadores de varianzas y probabilidades de cobertura y longitudes relativas de intervalos del 95 % de confianza para τ_1, τ_2 y τ.

	Pob. I		Pob. II			Pob. I		Pob. II	
Estim.	Sesgo		Sesgo		Intervalo	Prob.	Long.	Prob.	Long.
varianza	rel.	$\sqrt{\text{ecm-rel}}$	rel.	$\sqrt{\text{ecm-rel}}$	de confianza	cober.	rel.	cober.	rel.
$\tilde{V}_D(\tilde{\tau}_1)$	-.29	.31	-.50	.51	$\tilde{\tau}_1 \pm 1.96\sqrt{\tilde{V}_D(\tilde{\tau}_1)}$	0.00	.11	0.00	.11
$\tilde{V}_D(\tilde{\tau}_2)$	-.27	.33	-.32	.41	$\tilde{\tau}_2 \pm 1.96\sqrt{\tilde{V}_D(\tilde{\tau}_2)}$	0.00	.13	0.00	.14
$\tilde{V}_D(\tilde{\tau})$	-.36	.38	-.50	.51	$\tilde{\tau} \pm 1.96\sqrt{\tilde{V}_D(\tilde{\tau})}$	0.00	.08	0.00	.09
$\hat{V}_M(\hat{\tau}_1)$	-.03	.17	-.35	.37	$\hat{\tau}_1 \pm 1.96\sqrt{\hat{V}_M(\hat{\tau}_1)}$	0.92	.24	0.89	.24
$\tilde{V}_M(\hat{\tau}_2)$	.08	.94	$.05^4$	1.3^4	$\hat{\tau}_2 \pm 1.96\sqrt{\hat{V}_M(\hat{\tau}_2)}$	0.94	.52	0.92^4	$.63^4$
$\hat{V}_M(\hat{\tau})$	.13	.71	$.02^4$	$.82^4$	$\hat{\tau} \pm 1.96\sqrt{\hat{V}_M(\hat{\tau})}$	0.93	.26	0.95^4	$.30^4$
$\hat{V}_D(\hat{\tau}_1)$	.02	.24	-.55	.56	$\hat{\tau}_1 \pm 1.96\sqrt{\hat{V}_D(\hat{\tau}_1)}$	0.93	.24	0.79	.20
$\tilde{V}_D(\hat{\tau}_2)$	.08	.94	$.05^4$	1.3^4	$\hat{\tau}_2 \pm 1.96\sqrt{\hat{V}_D(\hat{\tau}_2)}$	0.94	.52	0.92^4	$.63^4$
$\hat{V}_D(\hat{\tau})$	.15	.72	$-.01^4$	$.82^4$	$\hat{\tau} \pm 1.96\sqrt{\hat{V}_D(\hat{\tau})}$	0.93	.26	0.92^4	$.29^4$
					ICVP para τ_1	0.90	.21	0.76^2	$.21^2$
					ICVP para τ_2	0.89	.54	0.95^4	$.58^4$
					ICVP para τ	0.84^5	$.26^5$	0.92^{12}	$.33^{12}$
					ICVP ajust. para τ_1			0.93	.39
					ICVP ajust. para τ_2			0.95^4	$.58^4$
					ICVP ajust. para τ			0.98^{12}	$.49^{12}$

Notas: ecm-rel., error cuadrático medio relativo; ICVP, intervalo de confianza verosimilitud perfil. Resultados para intervalos tipo Wald están basados en 500 muestras y para ICVP basados en 100 muestras. Resultados marcados con un superíndice se obtuvieron descartando el número de muestras indicado por el superíndice.

Referencias

Christman, M.C., (2009). Sampling of rare populations . *Handbook of Statistics*, Vol. 29A, Sample Surveys: Design, Methods and Applications, North Holland, C. R. Rao and D. Pfeffermann (eds.), 109-124, Amsterdam.

Félix-Medina, M.H. and Thompson, S.K., (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. and Monjardin, P. E., (2008). Estimación del tamaño de una población de difícil detección en el muestreo por seguimiento de nominaciones y probabilidades de nominación heterogéneas . *Memorias del XXIII Foro Nacional de Estadística*, 49-53.

Sanathanan, L., (1972) Estimating the size of a multinomial population . *Annals of Mathematical Statistics*,43,142-152.

Memoria FONE XXIV (2010), Sección I, 41–52

Curso de Estadística en b-learning basado en los estilos de aprendizaje de los discentes

José Luis García Cué[a]

José Antonio Santizo Rincón[b]

Mercedes Jiménez Velázquez[c]
Colegio de Postgraduados

1. Introducción

Desde el año de 1993 se han ido incorporado las Tecnologías de la Información y la Comunicación (TIC) en cursos de Estadística en el Colegio de Postgraduados (CP) con el objetivo de mejorar los contenidos temáticos, el acceso a diferentes fuentes documentales y el uso de herramientas informáticas que coadyuven al aumento de la calidad educativa en los diferentes postgrados en niveles de Maestría y Doctorado en el CP. A continuación, se hace una lista, en forma cronológica, de algunos de los proyectos donde los autores han utilizado las TIC desde 1993 hasta la fecha:

1. En 1993 se incorporan las TIC a los cursos de servicio de Estadística del CP: Introducción a la Estadística e Introducción a los Diseños Experimentales.

2. De 1994 a 1998 se formaliza la red de computadoras del CP con conexión a Internet y se propone un Modelo de Educación a Distancia para optimizar el aprendizaje mediante el uso TIC.

[a] jlgcue@colpos.mx
[b] jasrg@colpos.mx
[c] mercedes@colpos.mx

3. De 1999 a 2002 se crea el grupo de Educación Participativa para cursos de Estadística vía Internet.

4. En el año 2003 se formaliza el proyecto de Investigación intitulado Materiales Educativos para el Apoyo de la Enseñanza en cursos de Probabilidad y Estadística con la integración de TIC Colegio de Postgraduados-Facultad de Estudios Superiores Zaragoza de la UNAM.

5. De 2004 a 2007 se elabora el proyecto I+D KM-Educa con el propósito de establecer un sistema global de Gestión del Conocimiento para 17 instituciones de educación superior y postgrado de Iberoamérica, en el área de Matemáticas. Se seleccionó la enseñanza de la Estadística en el CP para el proyecto.

6. De los años 2004 al 2010 se utiliza la plataforma educativa Blackboard para la impartición de cursos bajo las modalidades presencial, b-learning y e-learning en el CP.

Además se han ido integrando conceptos tanto didácticos como pedagógicos en los cursos de postgrado a través de las teorías sobre el Conocimiento y del Aprendizaje. A partir del año 2002 se hace una propuesta para incluir los Estilos de Aprendizaje en los cursos de Estadística así como en la construcción de tutoriales y páginas web educativas como apoyo a los cursos presenciales. También, se incorporaron dichos conceptos en la construcción de materiales teóricos, ejemplos y actividades basadas en dichas teorías pare el manejo de plataformas educativas.

Por tal razón, en el presente trabajo, se muestra un ejemplo de la incorporación de las TIC en cursos de Estadística del CP, basados en las preferencias en cuanto a los Estilos de Aprendizaje de los discentes. Para comenzar, se hace referencia a los conceptos de Estilo y Estilos de Aprendizaje. Posteriormente, se explica sobre el uso los Estilos de Aprendizaje en Instituciones educativas y en pedagogía. Después, se hace la presentación del curso de estadística en la modalidad b-learning y se explican los pasos seguidos para elaborar dicho curso en la plataforma educativa Blackboard.

2. Estilos de Aprendizaje

El término Estilo, de acuerdo con Guild y Garger (1998), se comenzó a utilizar por los investigadores a partir del siglo XX en concreto por aquéllos que trabajaron en distinguir las diferencias entre las personas en áreas de la psicología y de la educación. Lozano (2000) después de analizar diversas teorías y reunir múltiples conceptos definió Estilo como "un conjunto de preferencias, tendencias y disposiciones que tiene una persona para hacer algo y que se manifiesta a través de un patrón conductual y de distintas fortalezas que lo hacen distinguirse de los demás".

Los conceptos de Keefe (1982) fueron tomados y adaptados por Catalina Alonso *et al.* (1994) para definir los Estilos de Aprendizaje como "los rasgos cognitivos, afectivos y fisiológicos, que sirven como indicadores relativamente estables, de cómo los discentes perciben, interrelacionan y responden a sus ambientes de aprendizaje". Los rasgos, a los que se refieren pueden diagnosticarse con una serie de instrumentos o cuestionarios ideados por investigadores para distintos colectivos como docentes, discentes, pacientes de sicólogos, directivos, empresarios, trabajadores, administradores, entre otros.

Los instrumentos o cuestionarios han sido sometidos a pruebas de expertos, validez de contenidos y fiabilidad. A lo largo de los años, los autores han presentado resultados de las investigaciones en congresos, foros y se han publicado un gran número de libros y artículos en revistas científicas, algunas disponibles vía Internet (García Cué, 2006). Catalina Alonso (1992) hace una lista de diversos instrumentos utilizados para identificar los Estilos de Aprendizaje. García Cué (2006) complementa la lista de Alonso e identifica 72 diferentes instrumentos. Algunos de ellos, están disponible vía Internet en página web y se pueden contestar de forma gratuita o pagando los derechos (García Cué *et al.*, 2009).

En el año 2004, investigadores como Coffield, Moseley, Hall y Ecclestone analizaron en la Gran Bretaña trece instrumentos que ellos consideran que son los más utilizados en idioma Inglés: Allinson y Hayes; Apter, Dunn y Dunn; Entwistle; Gregorc; Herrmann; Honey y Mumford; Jackson; Kolb; Myers-Briggs; Riding; Sternberg; y Vermunt (Coffield *et al.*, 2004). En idioma español el instrumento más utilizado es el Cuestionario Honey-Alonso de Estilos de Aprendizaje (CHAEA) diseñado por Alonso (1992). El CHAEA ha sido aplicado en diferentes investigaciones y los resultados de estas pesquisas están plasmados en tesis doctorales y en diversos artículos publicados en revistas científicas en idiomas español, inglés

y portugués(García Cué et al., 2009).

Uso de los Estilos de Aprendizaje en Instituciones educativas y en pedagogía

Rita y Kenneth Dunn (1978) enfocaron sus estudios sobre Estilos de Aprendizaje en diferentes niveles educativos en instituciones de Estados Unidos (USA), propusieron un cuestionario de Estilos de Aprendizaje con un modelo de 21 variables que influyen en la manera de aprender. Dichas variables fueron clasificadas en cinco diferentes grupos: *ambiente inmediato* (sonido, luz, temperatura, diseño, forma del medio), *propia emotividad* (motivación, persistencia, responsabilidad, Estructura), *necesidades sociológicas* (trabajo personal, con pareja, dos compañeros, un pequeño grupo y otros adultos), *físicas* (alimentación, tiempo, movilidad, percepción) y *necesidades psicológicas* (analítico-global, reflexivo-impulsivo, dominancia cerebral). La simple enumeración de estas variables aclara la importancia de los Estilos de Aprendizaje. En cada uno de los cinco bloques aparece una repercusión favorable o desfavorable al aprendizaje, en función del Estilo de Aprendizaje de la persona (Gallego y Ongallo, 2003).

Felder y Silverman (1988) propusieron un modelo de Estilos de Aprendizaje para alumnos de enseñanza superior en facultades de Ingeniería de algunas Universidades al oriente de los Estados Unidos (USA). Felder y otros investigadores han obtenido diversos resultados en sus pesquisas desde 1988 hasta la fecha y consideraron que los alumnos tienen diferentes niveles de motivación, distintas actitudes sobre la enseñanza y el aprendizaje, y responden de maneras diversas en algunas prácticas específicas de enseñanza fuera y dentro del aula (Felder y Brent, 2005). También, han demostrado, que un mayor aprendizaje se puede producir cuando los Estilos de Enseñanza coinciden con los Estilos de Aprendizaje. Por tal motivo, consideran necesario que tanto docentes como discentes conozcan sus propias preferencias en cuanto a los Estilos de Aprendizaje, aplicando cualquier instrumento de medición. Asimismo, sugieren que los profesores construyan sus materiales de enseñanza – teoría, ejemplos, ejercicios, actividades, evaluaciones – pensando, en la manera que sea posible, en todos los Estilos de Aprendizaje del instrumento utilizado para medir las preferencias.

Alonso (1992) elaboró el Cuestionario Honey-Alonso de Estilos de Aprendizaje (CHAEA) aprovechando las teorías, aportaciones y experiencias de diversos investigadores, en especial las Keefe, Kolb, Honey-Mumford y adaptó, junto con Domingo Gallego Gil, el Learning Styles Questionnaire (LSQ) de Honey-Mumford al ámbito académico y al idioma español. Después,

desarrolló una investigación con 1371 alumnos de diferentes facultades de las Universidades Complutense y Politécnica de Madrid, España (Alonso *et al.,* 1994). En donde, una de sus aportaciones, fue la identificación de los Estilos de Aprendizaje de los discentes de cada facultad de las dichas universidades. Otra aportación, consistió en la elaboración de una lista con características que determinan el campo de destrezas de cada Estilo:

1. Activo: Animador, Improvisador, Descubridor, Arriesgado, Espontáneo
2. Reflexivo: Ponderado, Concienzudo, Receptivo, Analítico, Exhaustivo
3. Teórico: Metódico, Lógico, Objetivo, Crítico, Estructurado
4. Pragmático: Experimentador, Práctico, Directo, Eficaz, Realista

García Cué (2006) hace una pesquisa - en el Colegio de Postgraduados, México - para determinar la forma en que usan las Tecnologías de la Información y la Comunicación (TIC), tanto los docentes como los discentes de postgrado, de acuerdo con los Estilos de Aprendizaje.
Melaré Vieira *et al.* (2008) hacen una investigación con varios docentes de Instituciones Iberoamenticanas sobre el uso de los Estilos de Aprendizaje para el aprendizaje en la virtualidad basados en las teorías de los Estilos de Aprendizaje de Kolb, Honey-Mumford, Alonso-Gallego y en las de las Tecnologías de la Información y Comunicación (TIC) para definir cuatro Estilos de uso de espacio virtual: Participativo, Buscador e Investigador, Estructurador y planeador, Concreto y Productivo, además, proponen las características de cada uno de ellos.
Lago *et al.* (2008) explican la utilidad pedagógica de la aplicación de teorías de Estilos de Aprendizaje a la hora de seleccionar estrategias de enseñanza-aprendizaje. También, proponen una tipología de actividades polifásicas basadas en modelos educativos y pedagógicos que denominaron Estilos de Aprendizaje y Actividades Polifásicas (EAAP) partiendo de combinaciones de los Estilos Activo, Reflexivo, Teórico y Pragmático propuestos por Alonso *et al.* (1994). La tipología clasifica las actividades en cuatro fases en función del número de estilos que se utilizan simultáneamente:

1. *Actividades monofásicas:* Activo (Rompecabezas, representaciones de teatro); Reflexivo (Exposición narrativa, círculos literarios); Teórico (Resolución de problemas); Pragmático (Trabajo por proyectos).

2. *Actividades bifásicas:* Activo-Reflexivo (Lluvia de ideas); Reflexivo-Teórico (Asistencia a clase magistral); Teórico-Pragmático (Demostraciones científicas); Pragmático-Activo (manualidades).

3. *Actividades trifásicas:* Pragmático-Activo-Reflexivo (Presentación oral del estudiante); Activo-Reflexivo-Teórico (Blogs, Wikis, Webquest); Reflexivo-Teórico-Pragmático (Elaboración de mapas conceptuales); Teórico-Pragmático-Activo (dibujo, fotografía)

4. *Actividades Eclécticas:* Activo-Reflexivo-Teórico-Pragmático (Trabajo por proyectos)

3. Propuesta del curso Estadística

Desde el año 2002 se buscaron diferentes teorías didácticas y pedagógicas que explicaran las razones del porqué los alumnos, que asistían a los cursos de Estadística, aprendían de manera diferente y además, sí los profesores influían en los discentes por su forma de enseñar. Las teorías de Estilos de Aprendizaje despejaron estas dudas, por tal razón, se trabajó en la forma de incorporar dichas teorías en cursos de Estadística apoyados de las Tecnologías de la Información y la Comunicación.

El curso de Estadística en b-learning (modalidad semiprecencial), basado en los Estilos de Aprendizaje, está orientado a discentes de maestría y doctorado del Colegio de Postgraduados; se apoya de TIC en especial el Internet a través de herramientas como: www, software en java, e-mail, foros de discusión, chat, herramientas interactivas web 2.0 y el uso de una plataforma educativa como Blackboard. También, el curso está sustentado en un diseño instruccional que incluye: contenidos temáticos, recursos humanos, administración del CP y sistemas de evaluación (ver figura 1).

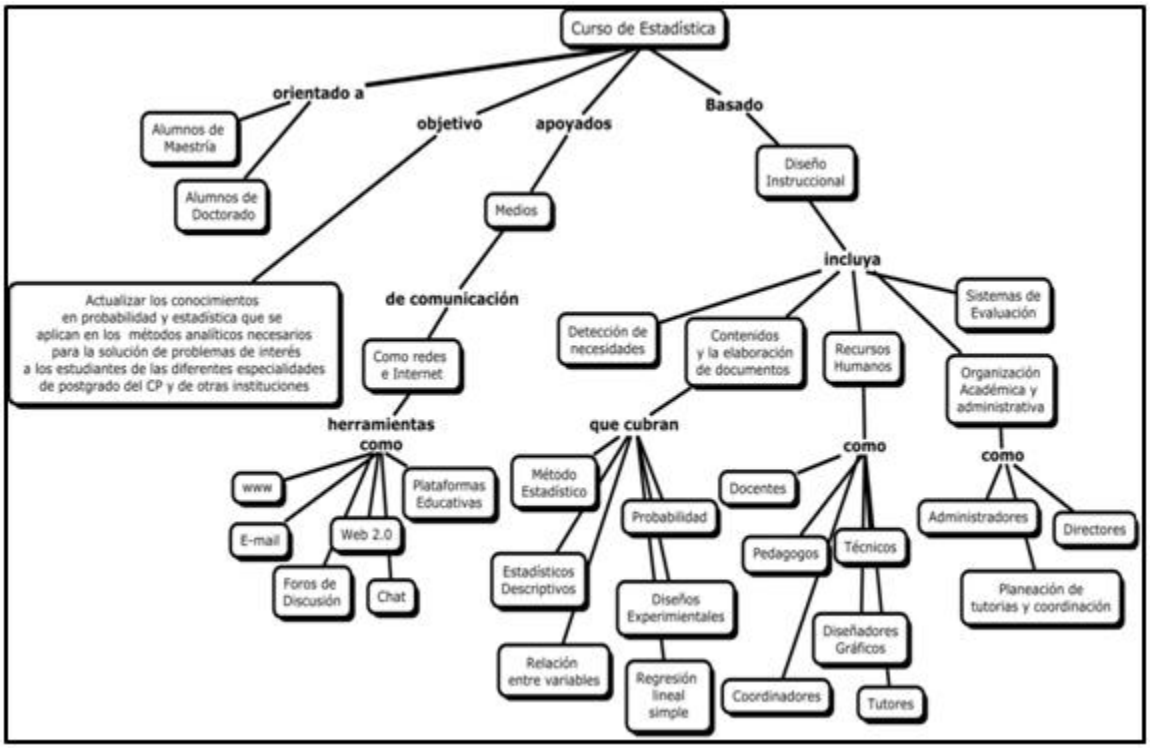


Figura 1 Mapa conceptual del curso de Estadística

Elaboración de materiales en una plataforma educativa

Los materiales del contenido del curso "Estadística en b-learning basado en los Estilos de Aprendizaje" fueron elaborados a través de estructuras didácticas y pedagógicas previamente probadas en otros cursos del CP, siguiendo los pasos siguientes:

a) Propuesta de materiales educativos. Los materiales -teoría, ejercicios y ejemplos – están soportados en una estructura didáctica y de la propuesta teórica de Felder y Brent (2005); esto es, que dichos materiales cubran de manera posible las cuatro preferencias en cuanto a Estilos de Aprendizaje. Por tal motivo, en la parte teórica se plantea, construir materiales que cumplan con los perfiles de cada uno de los Estilos de Aprendizaje: apuntes del profesor, libros de texto, hipervínculos de páginas web, videos de YouTube, libros electrónicos, artículos científicos, documentos, software, entre otros. También se propusieron el número de clases presenciales y a distancia, calendario de actividades, y se nombró el tutor para el curso.

En la parte de actividades se plantean las siguientes estrategias pedagógicas basadas en

cuatro Estilos de Aprendizaje:

1. *Activo*: elaboración de encuestas, liderazgo en trabajo en grupo, propuestas para la solución de problemas.

2. *Reflexivo:* revisión de materiales para el análisis de información y solución de problemas. Revisión de la salida de datos de análisis hechos por los programas informáticos estadísticos SAS y R.

3. *Teórico:* revisión de la teoría, comprobación de fórmulas y desarrollos matemáticos.

4. *Pragmático:* aplicación teórico práctica de la estadística en problemas del área de interés del alumno, uso de programas informáticos estadísticos SAS y R.

b) Captura y elaboración de materiales *Los materiales del curso fueron elaborados utilizando* procesador de textos, presentaciones de diapositivas, archivos PDF, fotografías, videos, imágenes, multimedios, etc. También se elaboraron prácticas para utilizar los paquetes informáticos SAS y R. Asimismo se realizaron páginas web con contenidos disponibles desde la dirección http://colposfesz.galeon.com.

Los materiales elaborados se sujetaron a los formatos didácticos siguientes:

1. *Para las sesiones presenciales:* Nombre del curso; fecha de la sesión; responsable; objetivo de la sesión; contenidos temáticos; material de apoyo para el docente; orden del día; actividades de aprendizaje y fuentes de consulta

2. *Para las sesiones a distancia:* nombre del curso; fecha de la sesión; mediador de la sesión; objetivo de la sesión; contenidos; experiencias de aprendizaje; observaciones; requisitos de entrega; fuentes de consulta

c) Elaboración de Actividades. Las actividades se diseñaron de acuerdo con la propuesta por Lago *et al.*(2008) donde se incluyen los Estilos de Aprendizaje y las estrategias de enseñanza-aprendizaje:

1. *Actividades monofásicas:* Activo (Líder de un grupo de trabajo); Reflexivo (Exposición de un tema en curso); Teórico (Solución de problemas); Pragmático (discusión teórico práctica de un problema).

2. *Actividades bifásicas:* Activo-Reflexivo (Lluvia de ideas); Reflexivo-Teórico (Asistencia a clase en el aula); Teórico-Pragmático (Demostraciones científicas); Pragmático-Activo (diseño de encuestas).

3. *Actividades trifásicas:* Pragmático-Activo-Reflexivo (Presentación oral del discente); Activo-Reflexivo-Teórico (uso de herramientas Web 2.0 como Blogs, Wikis, Webquest, Google Docs); Reflexivo-Teórico-Pragmático (Elaboración de mapas conceptuales); Teórico-Pragmático-Activo (diseño de una investigación apoyado de dibujos y fotografías)

4. *Actividades Eclécticas para todo el grupo:* Elaboración de escritos sobre temas de estadística, manejo del software SAS y del R, trabajo a través de proyectos y presentación al final del curso apoyado de software ofimático.

Las actividades fueron capturadas en el procesador de textos para poder ser accesibles desde la plataforma educativa.

d) Elaboración de pruebas diagnósticas. Dentro de la metodología del curso se planteó evaluar los conocimientos básicos y conceptuales a través de diferentes test con preguntas de opción múltiple, falsa o verdadera, relación de columnas y respuesta corta.

Los test se evalúan de manera independiente a cada una de las actividades, se pueden contestar después de cada unidad temática y están accesibles a los alumnos a través de la plataforma educativa Blackboard.

e) Integración de contenidos. Se organizaron todos los contenidos elaborados - para cubrir el programa del curso: teoría, ejemplos, ejercicios, actividades, pruebas diagnósticas - en la plataforma educativa Blackboard como se muestra en Figura 2.

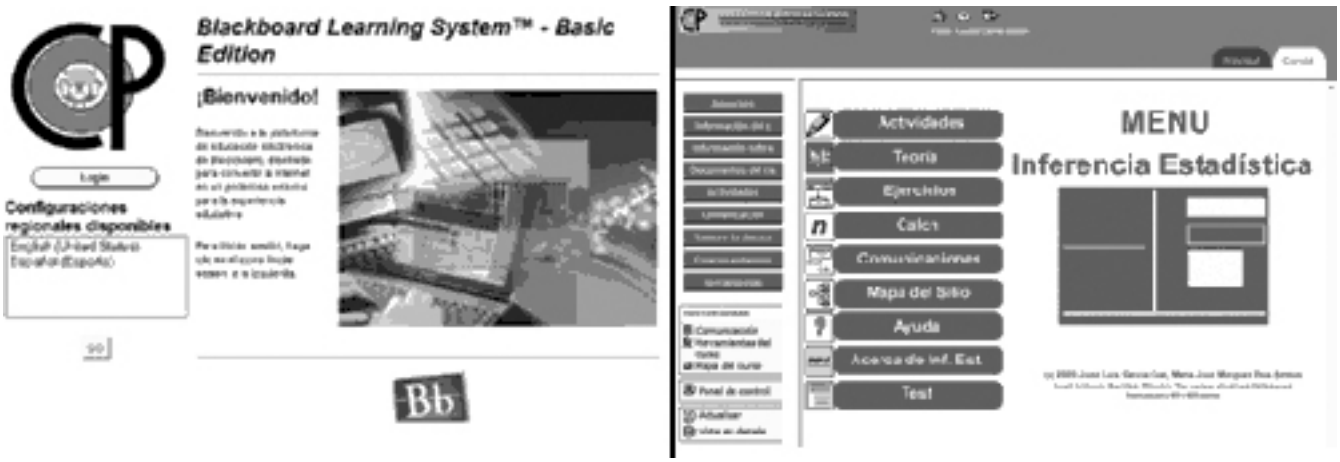


Figura 2. Pantallas del Curso de Estadística en la Plataforma Blackboard

Herramientas de la plataforma educativa

Además del planteamiento del curso y de las diferentes estrategias pedagógicas, se ha trabajado con diferentes recursos y herramientas que tiene la plataforma Blackboard:

1. ***Información de los profesores.-*** Que contiene un breve currículum vitae e hipervínculos de páginas web de los profesores participantes en el curso así como el de los tutores y coordinadores.

2. ***Foros de discusión y el uso de Chat.-*** Con la programación y enlaces a foros y Chat sobre temas de Estadística a otros para la comunicación de dudas o comentarios.

3. ***Herramientas para administrar cursos.-*** Para la administración de matrícula, calificaciones, avisos, mensajes, formación de grupos, glosario de términos, enlaces a otros cursos, entre otras cosas más.

Los resultados de este trabajo evidenciaron que la incorporación de las TIC, los Estilos de Aprendizaje y el empleo de plataformas educativas mejoran los contenidos didácticos y pedagógicos en cursos de Estadística. La modalidad b-learning, además, permite la capacitación de discentes de cualquier *campi* del CP, de investigadores de otras instituciones educativas de postgrado públicas o privadas y de productores agrícolas interesados en el aprendizaje de la estadística.

Referencias

Alonso, C. (1992). *Análisis y Diagnóstico de los Estilos de Aprendizaje en Estudiantes Universitarios*. Tomo I. Madrid: Colección Tesis Doctorales. Editorial de la Universidad Complutense.

Alonso, C.; Gallego D.; Honey, P. (1994). *Los Estilos de Aprendizaje: Procedimientos de diagnóstico y mejora*. Bilbao: Ediciones Mensajero

Coffield, F.; Moseley, D. Hall, E.; Ecclestone, K. (2004). *Learning styles and pedagogy in post-16 learning. A systematic and critical review*. Document in Learning Skills Development Agency. LSDA, PP182. Documento obtenido en la subscripción al LSDA. United kingdom. http://www.lsda.org.uk/files/PDF/1543.pdf

Dunn, R., Dunn, K. (1978). *Teaching Students throught their Individual Learning Styles: A practical aproach*. New Jersey: Prentice Hall.

Felder, R.; Brent, R. (2005). *Understanding Student Differences*. Journal of Engineering Education, 94, 57-72 (2005).

Gallego, D.; Ongallo, C. (2004). *Conocimiento y Gestión*. Madrid: Pearsons Prentice Hall.

García Cué, J.L. (2006). *Los Estilos de Aprendizaje y las Tecnologías de la Información y la Comunicación en la Formación del Profesorado.* Tesis Doctoral. Dirigida por Catalina Alonso García. Madrid: Universidad Nacional de Educación a Distancia.

García Cué, J.L.; Santizo, J.A.; Alonso, C. (2009). *Instrumentos para medir los Estilos de Aprendizaje*. Revista: Learning Styles Review, No.4 Vol. 1. pp 3-21. octubre de 2009. http://www.learningstylesreview.com/

Guild, P.; Garger, S. (1998). *Marching to Different Drummers*. Virginia, USA: ASCD-Association for Supervision and Curriculum Development. 2nd Edition.

Keefe, J. W. (1982). *Assessing student learning styles: An overview*. In Keefe J. W. (ed.), Student Learning Styles and Brain Behavior, Reston, VA: National Association of Secondary School Principals pp. 43-53.

Lago, B.; Colvin, L; Cacheiro, M. (2008). *Estilos de Aprendizaje y Actividades Polifásicas: Modelo EAAP*. Learning Styles Review, No.2 Vol. 1. pp 2-22. Octubre de 2008. http://www.learningstylesreview.com/ el 22/02/2010]

Lozano, A. (2000*). Estilos de Aprendizaje y Enseñanza. Un panorama de la estilística educativa*. ITESM Universidad Virtual - ILCE. México: Trillas.

Melare Vieira, D.; Alonso, C.; Ferreira, S. (1998). *Estilo de uso do espaço virtual*. Learning Styles Review, No.1 Vol. 1. pp 88-108. Abril de 2008.http://www.learningstylesreview.com/

Memoria FONE XXIV (2010), Sección I, 53–59

Puntos de cambio en modelos lineales mixtos

Jésica Hernández Rojano[a]

Posgrado en Ciencias Matemáticas, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

1. Introducción

La detección y estimación de puntos de cambio es un problema presente en muchas áreas de Estadística. Este trabajo presenta algunos métodos para la detección y/o estimación de un punto de cambio en Regresión Lineal (basados en Brown et al (1975) y en Chen y Gupta (2000)) y su extensión al caso de Modelos Lineales Mixtos.

2. Puntos de cambio en regresión lineal

El modelo de Regresión Lineal es:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde $\boldsymbol{x}_i$ es el vector de de las p variables independientes, con primera coordenada igual a 1, para la observación i, $i = 1, \ldots, n$; $\boldsymbol{\beta}$ es el vector de parámetros desconocidos y $\epsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$, son los errores con σ^2 desconocida y no correlacionados entre sí.

El problema de puntos de cambio en regresión lineal múltiple consiste en verificar si existe un cambio en el modelo de regresión en algún punto k y, en caso de existir, estimarlo.

2.1. Métodos

Existen muchos métodos para detectar y estimar puntos de cambio en regresión lineal. A continuación se describirán 3 métodos de detección de puntos de cambio, de los cuales, los últimos dos se basan en las siguientes hipótesis:

[a] jesicahrojano@gmail.com

$$H_0 : \mu_{y_i} = \boldsymbol{x}_i'\boldsymbol{\beta} \quad \text{para} \quad i = 1, \ldots, n,$$

vs.

$$H_1 : \mu_{y_i} = \boldsymbol{x}_i'\boldsymbol{\beta}_1 \quad \text{para} \quad i = 1, \ldots, k, \quad \text{y} \quad \mu_{y_i} = \boldsymbol{x}_i'\boldsymbol{\beta}_2 \quad \text{para} \quad i = k+1, \ldots, n,$$

donde k, $k = p+1, \ldots, n-p$, es la localización del punto de cambio, $\boldsymbol{\beta}$, $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ son vectores que contienen a los parámetros de regresión desconocidos y $\mu_{y_i} = E(y_i)$.

Los métodos explicados se aplicarán a datos de incidencia (cruda) de Diabetes diagnosticada por cada 1000 habitantes en edades de 18 a 79 años en los Estados Unidos de 1980 a 2007 (www.cdc.gov) (Figura 1).

2.1.1. Sumas acumuladas (CUSUMs) de los residuos

Este método utiliza las sumas acumuladas de los residuos, S_r:

$$S_r = \frac{1}{\hat{\sigma}} \sum_{j=p+1}^{r} s_j$$

donde s_j es es el residuo para la j-ésima observación, $\hat{\sigma}^2 = \sum_{j=p+1}^{n} (s_j - \bar{s})^2/(n-p-1)$ y $\bar{s} = \sum_{j=p+1}^{n} s_j/(n-p)$ y r en $r = p+1, \ldots, n-p-1$. Con ayuda del paquete strucchange de R se realizó la gráfica de las CUSUMs de los residuos recursivos. Si hay un punto de cambio la gráfica comenzará a alejarse de la línea del valor medio alrededor de ese punto, como se muesta en la Figura 2, donde la desviación empieza en 1994, así que se considera este año como el punto de cambio.

2.1.2. Técnica del cociente de log-verosimilitudes de Quandt

Esta técnica es apropiada cuando hay un cambio abrupto en el modelo de regresión. Para cada k en $k = p+1, \ldots, n-p-1$, se calcula el valor de λ_k

$$\lambda_k = \ln\left(\frac{\text{max verosimilitud de las observaciones dada } H_0}{\text{max verosimilitud de las observaciones dada } H_1}\right)$$

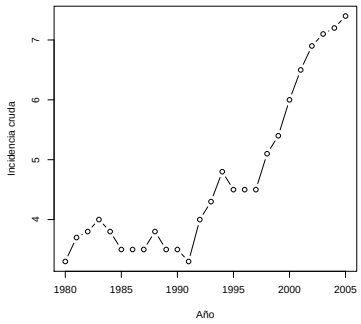


Figura 1: Incidencia de Diabetes

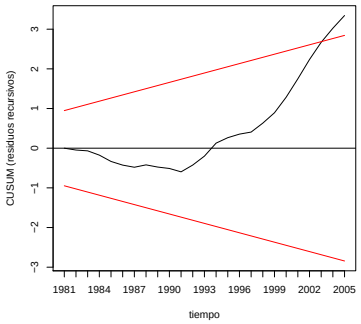


Figura 2: CUSUMs de los residuos recursivos para la Incidencia cruda de Diabetes.

y se considera punto de cambio aquel con λ_k mínimo.

Para la incidencia de diabetes el estimador del punto de cambio es $k = 11$ i.e. 1990 (Figura 3).

2.1.3. Criterio de información de Schwarz (SIC)

Bajo H_0 el criterio de enformación de Schwarz, denotado por $SIC(n)$, se obtiene como:

$$SIC(n) = -2 \ln L_0 \left(\mathbf{b}, \hat{\sigma}^2\right) + (p+2) \ln n$$

y el SIC bajo H_1, $SIC(k)$, es:

$$SIC(k) = -2 \ln L_0 \left(\mathbf{b}_1, \mathbf{b}_2, \hat{\sigma}^2\right) + (2p+3) \ln n$$

Se rechaza H_0 si $SIC(n) > \text{mín}_{p+1 \leq k \leq n-p} SIC(k)$ y el punto de cambio será $\hat{k}$ tal que $SIC(\hat{k}) = \text{mín}_{p+1 \leq k \leq n-p} SIC(k)$.

Para la incidencia cruda $SIC(n)$=58.5582 y el punto de cambio se da en el año 1989 (Figura 4).

3. Puntos de cambio en modelos lineales mixtos

El modelo lineal mixto más simple es:

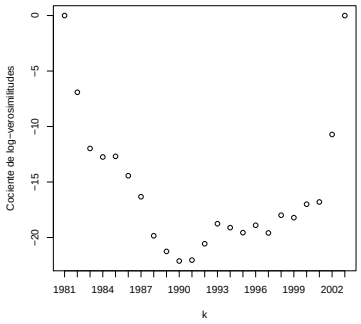


Figura 3: Cociente de log-verosimilitudes de Quandt's vs. Año para la incidencia cruda de diabetes.

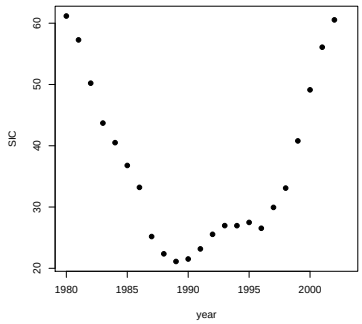


Figura 4: Año vs. SIC para la incidencia cruda de diabetes.

$$y_{ti} = \alpha + \beta x_{ti} + u_{1i} + \epsilon_{ti}$$

para $i = 1, \ldots, m$ observaciones con $t = 1, \ldots, n$ repeticiones. α y β son los parámetros desconocidos, $u_{1i} \sim N(0, \sigma_u^2)$ es el efecto aleatorio, $\epsilon_{ti} \sim N(0, \sigma_\epsilon^2)$ es el error aleatorio con σ_u^2 y σ_ϵ^2 desconocidas.

El caso más fácil es suponer que todos los sujetos tienen el punto de cambio en la misma observación k, de tal manera que la hipótesis que se utilizarán son:

$$H_0 : y_{ti} = \alpha + \beta x_{ti} + u_{1i} + \epsilon_{ti} \quad \text{para} \quad i = 1, \ldots, m, \quad t = 1, \ldots, n$$

vs.

$$\begin{aligned} H_1 : & y_{ti} = \alpha^* + \beta^* x_{ti} + u_{1i}^* + \epsilon_{ti}^* \quad \text{para} \quad t = 1, \ldots, k \quad \text{y} \\ & y_{ti} = \alpha^{**} + \beta^{**} x_{ti} + u_{1i}^{**} + \epsilon_{ti}^{**} \quad \text{para} \quad t = k+1, \ldots, n \end{aligned}$$

para $i = 1, \ldots, m$, donde α^*, α^{**}, β^*, β^{**} son parámetros desconocidos, $u_{1i}^* \sim N(0, \sigma_{u1}^2)$ y $u_{1i}^{**} \sim N(0, \sigma_{u2}^2)$ son los efectos aleatorios y $\epsilon_{ti}^* \sim N(0, \sigma_{\epsilon 1}^2)$ y $\epsilon_{ti}^{**} \sim N(0, \sigma_{\epsilon 2}^2)$ son los errores aleatorios. σ_{u1}^2, σ_{u2}^2, $\sigma_{\epsilon 1}^2$ y $\sigma_{\epsilon 2}^2$ son desconocidos. Para facilitar los cálculos se supondrá que $\sigma_{u1}^2 \neq \sigma_{u2}^2$.

3.1. Métodos

Con los supuestos mencionados arriba los métodos de detección de puntos de cambio para modelos de Regresión Lineal se extienden de manera directa al caso de Modelos Lineales Mixtos. Estos métodos se aplicarán a un conjunto de datos simulados con $n = 15$, $m = 100$, $k = 8$, $\alpha^* = 2$, $\alpha^{**} = 4$, $\beta^* = 11$, $\beta^{**} = 15$, $\sigma^2_{u1} = 4$, $\sigma^2_{u2} = 9$, $\sigma_{\epsilon 1} = 7$ and $\sigma_{\epsilon 2} = 5$.

3.1.1. Prueba CUSUM

La prueba CUSUM para Modelos Lineales Mixtos, Z_r, es la suma acumulada de los residuos de Pearson $r_{ij}, i = 1, \ldots, n, j = 1, \ldots, m$,

$$Z_h = \sum_{i=p+1}^{h} \sum_{j=1}^{m} r_{ij} \quad \text{for} \quad h = p+1, \ldots, n-p-1.$$

La Figura 5 sugiere que el punto de cambio es $k = 8$, que es en realidad el punto de cambio con el que se simularon los datos.

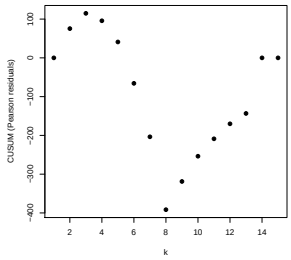


Figura 5: k vs. CUSUM de los residuos de Pearson para los datos simulados.

3.1.2. Cociente de log-verosimilitudes

Este método es una extensión directa del cociente de log-verosimilitudes de Quandt para Regresión Lineal. Como se puede ver en la Figura 6 el valor mínimo del cociente se da en $k = 8$, que es el punto de cambio correcto del ejemplo.

3.1.3. SIC

Este método también es una extensión del método utilizado en Regresión Lineal. Como se muestra en la (Figura 7) el valor mínimo del criterio se toma cuando $k = 8$.

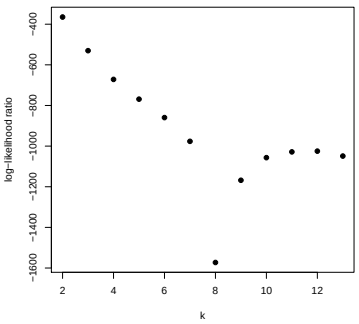


Figura 6: k vs. logaritmo del cociente de verosimilitudes para los datos simulados.

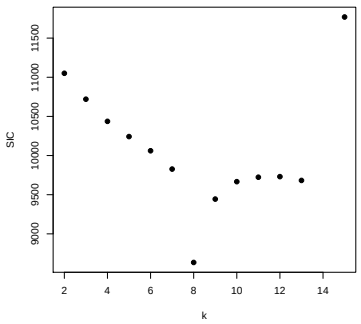


Figura 7: k vs. SIC para los datos simulados.

4. Conclusiones

Suponiendo que todas las observaciones tienen el punto de cambio en la misma repetición y que la desviación estándar de los efectos aleatorios cambia después de ese punto, los métodos propuestos para detección de puntos de cambio encuentran muy bien estos puntos, como se muestra en los ejemplos anteriores. Sin embargo, hay otros problemas a resolver:

- *Estimación del punto de cambio cuando este es diferente en todas las observaciones.* El principal problema en este caso es encontrar un algoritmo óptimo para maximizar la verosimilitud o sumar residuos sobre todas las posibles combinaciones de puntos de cambio.

- *Estimación de las varianzas de los efectos aleatorios.* Se probará la hipótesis de igualdad de varianzas de los efectos aleatorios antes y después del punto de cambio y en caso de rechazarla, se encontrará un estimador apropiado.

- Extensión de los métodos al caso en que la variable de efecto fijo que tiene el punto de cambio tenga también un punto de cambio en el efecto aleatorio.

Referencias

Jie Chen and A. K. Gupta (2000). "Parametric Statistical Change Point ". *Birkhauser Boston*,

R. L. Brown and J. Durbin and J. M. Evans (1975). "Techniques for testing the constancy of regression relationships over time",*Journal of the Royal Statistical Society, Series B (Methodological)*,37,149-192.

Memoria FONE XXIV (2010), Sección I, 61–67

Un estimador insesgado de la varianza del muestreo aleatorio simple usando un diseño mixto aleatorio sistemático

Alberto Manuel Padilla Terán[a]

1. Introducción

El muestreo sistemático es una técnica usada comúnmente en la práctica debido a su simplicidad y facilidad de uso; empero, la principal desventaja consiste en que no existe un estimador insesgado de la varianza del estimador del promedio o del total con una sola muestra sistemática. En la literatura se han propuesto diversas formas de tratar este problema; una de ellas supone que la característica de interés en la población no tiene relación con el orden en el que se encuentra en el marco y se usa entonces la fórmula de estimación de la varianza del muestreo aleatorio simple, MAS, Cochran (1986). Esto es denominado como orden aleatorio, Cochran (1986). Otra forma de estimar consiste en emplear un modelo para la variable de interés y construir una fórmula para la estimación de la varianza, Wolter (1985). Desde la perspectiva del diseño muestral puede permutarse la población previo a la extracción y emplearse la fórmula de estimación de varianza del MAS, Madow & Madow (1944). También puede suplementarse la muestra sistemática con otra muestra sistemática o una aleatoria simple, en lo que se conoce como métodos mixtos, ejemplo Huang (2004). Una comparación de diversas estrategias, ya sea de modelo o diseño se encuentra en Wolter (1985).

En este artículo se propone un diseño mixto aleatorio sistemático en el que primero se selecciona una MAS de tamaño uno y después se extrae una muestra sistemática para completar el tamaño deseado. Bajo este esquema, la media poblacional y la varianza del estimador

[a]ampadilla@banxico.org.mx

de la media, bajo MAS, se estiman insesgadamente con la media muestral y una expresión sencilla para la varianza, Padilla (2009a). La expresión propuesta para la estimación de la varianza no supone que la muestra proviene de una población en orden aleatorio, previniendo al usuario de caer en el error de pretender que se tiene otra cosa, PISE por sus siglas en inglés, un acrónimo acuñado por Valliant (2007), que significa 'pretend it's something else'; tampoco requiere que se aplique una permutación a la población previo a la extracción. Cabe mencionar que no se esperan ganancias en eficiencia comparado con el muestreo sistemático y métodos similares, ya que se estima insesgadamente la varianza del estimador bajo MAS. La comparación del método propuesto se efectúa con el estimador de la varianza usado en el muestreo sistemático bajo el supuesto de orden aleatorio de la población.

En el muestreo sistemático comúnmente se utiliza el supuesto de orden aleatorio y se emplea el estimador de la varianza de la media bajo MAS, $\hat{v}_{oa}(\hat{\bar{y}}) = (1 - n/N)\hat{s}^2_{sis}/n$, donde $\hat{s}^2_{sis}$ se refiere a la varianza entre elementos de la muestra sistemática. Esta es una estrategia adecuada siempre que se cuente con información acerca del orden de los elementos en la población con respecto a la variable de interés. El problema con esta estrategia es que uno puede caer en PISE y trabajar con un estimador sesgado de la varianza o usar dicha fórmula mecánicamente sin cuestionarse acerca del orden previo a la extracción de la muestra de la variable de interés. En Padilla (2009b) se muestra que, con esta estrategia, el sesgo del estimador de esta estrategia $\hat{s}^2_{sis}$ es $\frac{N-1}{N}(1-\rho)S^2_U$, donde ρ es el coeficiente de correlación intraclase, Cochran (1986) y $S^2_U = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, con $\bar{y}_U = \sum_{i=1}^{N} y_i/N$.

En el muestreo sistemático, las estrategias de orden aleatorio y permutación previa a la extracción de la muestra conducen a la varianza de la media muestral bajo MAS, por lo que el uso de alguna de estas estrategias equivale a aceptar a la varianza de la media muestral bajo MAS como la apropiada para la estimación.

En este artículo se exhibe un diseño muestral, junto con un estimador insesgado de la varianza entre elementos, que evita el uso de supuestos, como el de orden aleatorio, el cual es dificílmente verificable en la práctica. Tampoco se requiere efectuar permutaciones previo a la extracción de la muestra sistemática. Esta situación le resultará familiar a aquellos que

han empleado el muestreo sistemático en la práctica y se han enfrentado al problema de estimación de la varianza. El estimador propuesto en este artículo para la estimación de la varianza poblacional entre elementos, está libre de supuestos acerca del orden en la población de la característica de interés por estimar, lo cual es consistente con la teoría del muestreo probabilístico.

Antes de continuar, es necesario mencionar que no se construyó la varianza del estimador del promedio bajo $mmas(1, m)$, ni un estimador de dicha varianza.

2. Marco teórico

Se supondrá que se trabaja con el enfoque del muestreo probabilístico y que se extraerá una muestra de tamaño n de una población U con N elementos, $1 < n < N$.

Definición 2.1. *Diseño mixto aleatorio sistemático es un diseño, en el cual, primero se extrae una muestra de tamaño 1 con MAS de las N unidades de la población y después, de las N-1 unidades restantes, se selecciona una muestra sistemática circular de tamaño* $m = n - 1$*, Murty & Rao (1988).*

Este diseño se denotará como $mmas(1, m)$. Obsérvese que el número de muestras con este diseño es $N(N - 1)$.

Ejemplo 2.1. *Sea U una población con* $N = 7$ *y supóngase que se desea extraer una muestra de tamaño* $n = 3$ *usando un* $mmas(1, 2)$*. En este caso,* $m = 2$ *y hay* $7(7 - 1) = 42$ *muestras. Los índices para las posibles muestras son:*

1 2 5	*2 1 5*	*3 1 5*	*4 1 5*	*5 1 4*	*6 1 4*	*7 1 4*
1 3 6	*2 3 6*	*3 2 6*	*4 2 6*	*5 2 6*	*6 2 5*	*7 2 5*
1 4 7	*2 4 7*	*3 4 7*	*4 3 7*	*5 3 7*	*6 3 7*	*7 3 6*
1 5 2	*2 5 1*	*3 5 1*	*4 5 1*	*5 4 1*	*6 4 1*	*7 4 1*
1 6 3	*2 6 3*	*3 6 2*	*4 6 2*	*5 6 2*	*6 5 2*	*7 5 2*
1 7 4	*2 7 4*	*3 7 4*	*4 7 3*	*5 7 3*	*6 7 3*	*7 6 3*

El primer número de cada entrada se refiere a la selección por MAS, en tanto que los dos números siguientes corresponden a la muestra sistemática circular.

2.1. Estimadores puntuales

En el muestreo mixto aleatorio sistemático, Huang (2004), el estimador Horvitz-Thompson de $\hat{y}_U$, puede ser usado para estimar la media poblacional, siempre que N sea conocido.

Teorema 2.1. *Bajo $mmas(1, m)$ las probabilidades de inclusión de primer orden, π_k, son iguales a n/N para toda k y el estimador Horvitz-Thompson de $\hat{y}_U$ es la media aritmética simple, $\hat{\bar{y}}_{r,s} = \sum_{i=1}^{n} y_i/n$.*

El resultado principal de este artículo se enuncia a continuación.

Teorema 2.2. *Bajo $mmas(1, m)$ un estimador insesgado de la varianza poblacional entre elementos, $S_U^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, es $\hat{s}_{r,s}^2 = \frac{\sum_{i=1}^{m}(y_r - y_{s,i})^2}{2m}$, donde y_r es el valor del elemento seleccionado por el MAS y $y_{s,i}$ son los valores de los elementos seleccionados por la muestra sistemática circular. Un estimador insesgado de la varianza poblacional del estimador del promedio bajo MAS está dado por $\hat{v}_{MAS}(\hat{\bar{y}}_{r,s}) = (1 - n/N)\hat{s}_{r,s}^2/n$.*

Zinger (1980) propuso un estimador insesgado de la varianza entre elementos usando un muestreo parcialmente sistemático, en el que primero se selecciona una muestra sistemática y después se suplementa con una MAS del resto de la población. Infortunadamente, la fórmula propuesta por Zinger es bastante compleja y se requiere especificar valores que no dependen del diseño muestral.

Ejemplo 2.2. *Sea U la población del ejemplo 3.4.2, páginas 80-82, Särndal et al. (1992). Esta población tiene $N = 100$ elementos y la variable y_i toma los valores $1, 2, \ldots, 100$. Las cuatro poblaciones se etiquetarán con las letras A, B, C y D, y presentan la característica siguiente dependiendo del ordenamiento de los elementos:*

Población	*Tipo de ordenamiento de las y_i*	*correlación intraclase ρ*
A	*Tendencia lineal perfecta*	*-0.10*
B	*Varianza mínima para muestreo sistemático*	*-0.11*
C	*Igual varianza dentro de muestras sistemáticas*	*0.989*
D	*Orden aleatorio*	*-0.015*

Usando muestreo sistemático con $n = 10$ *se tienen* $N/n = 10$ *posibles muestras. La media poblacional es* 50.5 *y la varianza poblacional de la media bajo MAS es* $v_{MAS}(\hat{\bar{y}}) = (1 - n/N)S_U^2/n = 75.75$. *Bajo el supuesto de orden aleatorio, oa, los estimadores,* $\hat{s}^2_{sis}$ *y* $\hat{v}_{oa}$, *de cada una de la 10 muestras sistemáticas, se calcularon con las expresiones siguientes:* $\hat{s}^2_{sis,j} = \sum_{i=1}^{10}(y_{ji} - \hat{\bar{y}}_{sis,j})^2/(10-1)$ *y* $\hat{v}_{oa}(\hat{\bar{y}}_{sis,j}) = (1 - 10/100)\hat{s}^2_{sis,j}/10$, *donde j se refiere a la muestra sistemática y* $\hat{\bar{y}}_{sis,j} = \sum_{i=1}^{n} y_{ji}/n$. *Con el fin de comparar la estrategia de estimación de la varianza poblacional entre elementos suponiendo orden aleatorio en la población en el muestreo sistemático con el muestreo mixto aleatorio sistemático, se empleó un diseño* $mmas(1, 9)$ *para las poblaciones A a D del ejemplo en cuestión. En este caso, se generaron para cada población A a D las* $100(100-1) = 9,900$ *posibles muestras y se calcularon diversos valores, que se muestran a continuación.*

Población:	*A*	*B*	*C*	*D*
$S_U^2 =$	*841.7*	*841.7*	*841.7*	*841.7*
$v_{MAS}(\hat{\bar{y}}) =$	*75.75*	*75.75*	*75.75*	*75.75*
Muestreo sistemático:				
Estimador orden aleatorio $\hat{s}^2_{sis} =$	*916.7*	*925.8*	*9.2*	*846.1*
Sesgo relativo $(\bar{s}^2_{sis}) =$	*8.9 %*	*10.0 %*	*98.9 %*	*0.5 %*
Estimador de varianza $\hat{v}_{oa}(\hat{\bar{y}}_{sis}) =$	*82.5*	*83.3*	*0.83*	*76.1*
Coeficiente de variación $(\hat{\bar{y}}_{sis}) =$	*6.0 %*	*0 %*	*60.0 %*	*17.7 %*
Coeficiente de variación $(\hat{s}^2_{sis}) =$	*0 %*	*3.7 %*	*0 %*	*37.7 %*
Muestreo mixto aleatorio sistemático:				
Estimador de varianza $\hat{v}_{MAS}(\hat{\bar{y}}_{r,s}) =$	*75.75*	*75.75*	*75.75*	*75.75*
Coeficiente de variación $(\hat{\bar{y}}_{r,s}) =$	*7.7 %*	*7.7 %*	*7.7 %*	*21.3 %*
Coeficiente de variación $(\hat{s}^2_{r,s}) =$	*46.0 %*	*46.3 %*	*46.6 %*	*60.9 %*

Al comparar los estimadores de varianza $\hat{v}_{oa}(\hat{\bar{y}}_{sis,j})$, $\hat{v}_{MAS}(\hat{\bar{y}}_{r,s})$ *y los coeficientes de variación de los estimadores de la media poblacional y varianza poblaciones entre elementos para ambos diseños, se observa que los estimadores en el muestreo sistemático empleando el supuesto de orden aleatorio, se comportan erráticamente y dependen del orden de la variable en la población. En contraste, en las estimaciones del muestreo mixto aleatorio sistemático se aprecia estabilidad para las poblaciones A, B y C; empero, las distribuciones muestrales de* $\hat{\bar{y}}_{r,s}$ *y* $\hat{s}^2_{r,s}$ *en la población D presentan más variación que su contraparte en el muestreo sistemático. Esto último se debe a la presencia de observaciones influyentes en la distribución de las* $\hat{s}^2_{r,s}$.

3. Conclusiones

Se propuso un estimador insesgado de la varianza poblacional del muestreo aleatorio simple empleando un muestreo mixto aleatorio sistemático. Se mostró que con este diseño no es necesario suponer orden aleatorio en la población o aplicar una permutación antes de extraer una muestra sistemática. También se exhibió el sesgo del estimador comúnmente empleado en el muestreo sistemático, bajo el supuesto de orden aleatorio de la población, el cual depende del coeficiente de correlación intraclase.

Referencias

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

Huang, K. (2004). "Mixed random systematic sampling designs", *Metrika*, 59, pp. 1-11.

Madow, G. W. & Madow, L. H. (1944). "On the theory of systematic sampling", I, *Annals of Mathematical Statistics*, 25, pp. 1-24.

Murthy, M.N. & Rao, T.J. (1988). "Systematic Sampling", *Handbook of Statistics 6: Sampling*, ed. by C.R. Rao, Amsterdam: North Holland.

Padilla Terán, A. M. (2009a). "Un estimador insesgado de la varianza del muestreo aleatorio simple usando un diseño mixto aleatorio-sistemático". *Memorias electrónicas en extenso de la 2ª Semana Internacional de la Estadística y la Probabilidad*, Puebla de Zaragoza, Puebla, México. CD ISBN: 978-607-487-035-0.

Padilla, Alberto (2009b). "An Unbiased Estimator of the Variance of Simple Random Sampling Using Mixed Random-Systematic Sampling", Working Papers 2009-13, *Banco de México*.

Särndal, C.E., Swensson, B. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Valliant, R. (2009). "An Overview of the Pros and Cons of Linearization versus Replication in Establishment Surveys, 2007 International Conference on Establishment Surveys, CD-ROM, Alexandria, VA:" *American Statistical Association*, 929-940.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.

Zinger, A. (1980). "Variance estimation in partially systematic sampling", *Journal of the American Statistical Association*, Vol. 75, No. 369, pp. 206-211.

Memoria FONE XXIV (2010), Sección I, 69–74

El uso de muestras condicionalmente independientes (look alike) en pruebas de bondad de ajuste en modelos lineales generalizados

Silvia Ruiz Velasco Acosta[a], Lizbeth Naranjo Albarrán[b]
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas UNAM

1. Introducción

En el análisis de modelos lineales generalizados es conveniente realizar pruebas de bondad de ajuste de los parámetros desconocidos en las que se busca obtener $p-values$ en el intervalo $[0,1]$ para aceptar o no una hipótesis. Sin embargo, la mayoría de las estadísticas de prueba utilizadas suponen distribuciones asintóticas y por lo tanto producen $p-values$ no exactos.

El método que a continuación se presenta permite generar *p-values* exactos para las estadísticas de prueba cuando ésta admite una estadística suficiente minimal. El procedimiento depende de la construcción de muestras Monte Carlo, llamadas 'look-alike', a partir de la distribución condicional independiente de la muestra, dada la estadística suficiente minimal.

Esta metodología funciona cuando los parámetros desconocidos son de localización y/o escala. Sin embargo, cuando se desconoce el parámetro de forma éste debe estimarse de la muestra y por tanto la situación es diferente; las distribuciones de la estadística de prueba dependerán de los valores verdaderos del parámetro de forma.

2. Modelos lineales generalizados

Los modelos lineales generalizados (MLG) propuestos por Nelder y Wedderburn (1972) están especificados por tres componentes:

[a] silvia@sigma.iimas.unam.mx
[b] lizbeth@sigma.iimas.unam.mx

- El *componente aleatorio*, observaciones independientes con distribución:

$$f(y_i;\theta_i) = \exp\left\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i)\right\}.$$

- El *componente sistemático* o predictor lineal:

$$\eta(\cdot) = \mathbf{X}\beta.$$

- La *función liga*, función monótona y diferenciable que describe la relación entre la media de la i-ésima observación y su predictor lineal:

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \quad i = 1, \ldots, n.$$

La función liga canónica se da cuando se cumple que $\theta = \eta$ y en este caso los parámetros desconocidos de la estructura lineal tienen estadísticas suficientes, dadas por:

$$t_n = \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_{1i}y_i, \ldots, \sum_{i=1}^{n} x_{ki}y_i\right).$$

2.1. Modelo logístico

Si tenemos una variable respuesta que toma valores cero (fracaso) y uno (éxito). En este caso

$$\mathbb{E}(Y_i) = \mu_i = \mathbb{P}(Y_i = 1) = p_i.$$

La distribución para $\mathbf{Y}$ es una binomial, dada por:

$$f(y,p) = \exp\left\{\log\begin{pmatrix} m \\ y \end{pmatrix} + y\log\left(\frac{p}{1-p}\right) + m\log(1-p)\right\},$$

es claro que el parámetro canónico y por lo tanto la liga canónica es $\theta = \log(\frac{p}{1-p})$.

2.2. Modelo Poisson

Si Y es el número de ocurrencias de un evento, su función de distribución puede escribirse como:

$$f(y) = \exp(y\log\mu - \mu)/y!,$$

por lo que $\mathbb{E}(y) = \mu$ y la liga canónica es:

$$\log(\mu_i) = x_i^T\beta.$$

2.3. Bondad de ajuste

Uno de los criterios de bondad de ajuste más usados al ajustar un MLG es la devianza, dada por:

$$\sum 2w_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i) \right\} \phi = D(y;\mu)\phi.$$

Otra medida de discrepancia, que se usa es la X^2 de Pearson generalizada:

$$X^2 = \sum \frac{(y-\hat{\mu})^2}{V(\hat{\mu})}.$$

La distribución asintótica de ambas estadísticas es una χ^2 con n-p grados de libertad. Las distribuciones condicionales de la devianza y la X^2 dadas las estadísticas suficientes, son asintóticamente normales McCullagh (1986).

3. Generación de muestras look-alike

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria de la distribución $F(x;\theta)$, θ un vector de parámetros desconocido. Sea T_n una estadística suficiente minimal para θ. Sea S una estadística. La distribución condicional de la estadística $G(s|t_n)$, no dependerá del valor verdadero de θ. O'Reilly y Gracia (2004), sugieren generar muestras tales que $T_n = t_n$, que ellos llaman 'look-alike', para estimar $G(s|t_n)$.

Lockhart *et al.* (2007) proponen el uso del estimador Rao Blackwell y el muestreador de Gibbs para generar estas muestras. El concepto de la doble transitividad (ver O'Reilly y Quesenberry, 1973) puede usarse para generar muestras look-alike, que cuando se cumple nos dice que si t_m y x_m son conocidas, podemos encontrar t_{m-1} sin conocer explícitamente los valores de $x_1, x_2, \ldots, x_{m-1}$.

Hacemos una transformación de la muestra $x_1, x_2, \ldots, x_n$ y t_n a un nuevo conjunto de n variables, $X_{k+1}, X_{k+2}, \ldots, X_n$, $T_{n1}, T_{n2}, \ldots, T_{nk}$, y se trabaja con la densidad conjunta.

$$f(x_{k+1}, x_{k+2}, \ldots, x_n, t_{n1}, t_{n2}, \ldots, t_{nk}). \tag{1}$$

Una muestra look-alike $(x^*_{k+1}, x^*_{k+2}, \ldots, x^*_n)$ estará generada a partir de esta densidad conjunta en el orden de $x^*_n, x^*_{n-1}, \ldots, x^*_{k+1}$, y cuando estos valores sean conocidos, se encontrarán $x^*_1, x^*_2, \ldots, x^*_k$ resolviendo las k ecuaciones para $t_{n1}, t_{n2}, \ldots, t_{nk}$.

3.1. Método 1: Estimación Rao-Blackwell de $F(x;\theta)$

La estimación de Rao-Blackwell de $F(x;\theta)$, basada en t_n, es $\tilde{F}_n(x|t_n) = P(X_j \leq x|t_n)$, donde x_j es un miembro de una muestra dada.

Sea $t_n^* = t_n$, se genera el primer elemento de la muestra look-alike x_n^*. El segundo elemento de la muestra look-alike, x_{n-1}^*, se genera a partir de la distribución $P(X_{n-1} \leq x|t_n^*, x_n^*)$. Usando la doble transitividad, se puede calcular t_{n-1}^*.

Como x_n^* es independiente de $x_1, x_2, \ldots, x_{n-1}$, y t_{n-1}^*, $P(X_{n-1} \leq x|t_n^*, x_n^*)$ es igual a $\tilde{F}_{n-1}(x|t_{n-1}^*)$. Así x_{n-1}^* se genera a partir de $\tilde{F}_{n-1}(x|t_{n-1}^*)$. Este proceso continúa hasta que se tenga x_{k+1}^*; entonces se calcula $x_1^*, x_2^*, \ldots, x_k^*$ a partir de las ecuaciones $t_{n1}, t_{n2}, \ldots, t_{nk}$ para completar la muestra.

3.2. Método 2: Muestreo de Gibbs

Cuando la estimación de Rao-Blackwell de la distribución no es obtenible, se puede usar el muestreador de Gibbs. Considere la densidad condicional de x_n dado $(x_{k+1}, x_{k+2}, \ldots, x_{n-1}, t_n)$, digamos $f_c(x_n)$.

Considere la densidad conjunta (1), proporcional a $f_c(x_n)$, vista como una función $g_n(x_n)$ de x_n. La muestra look-alike se construye de la siguiente manera:

(a) Usar un procedimiento de aceptación y rechazo con $g_n(x_n)$ para generar un valor x_n^* de $f_c(x_n)$.

(b) Reemplazar x_n por x_n^* en la densidad conjunta, y considerarla como una función de x_{n-1}, $g_{n-1}(x_{n-1})$. La densidad condicional $f_c(x_{n-1}|x_n^*, t_n)$ será proporcional a $g_{n-1}(x_{n-1})$, usar esta función para generar un valor x_{n-1}^* como se hizo para x_n^* en el paso anterior.

(c) Reemplazar x_{n-1} por x_{n-1}^*, conjuntamente con x_n^*, en la densidad conjunta, considerar la nueva función como $g_{n-2}(x_{n-2})$, como una función de x_{n-2}, y generar un nuevo valor x_{n-2}^*.

(d) Repetir el procedimiento hasta obtener x_{k+1}^*; entonces resolver la ecuación de t_n para obtener $x_1^*, x_2^*, \ldots, x_k^*$ y completar la muestra. A esta muestra se le llama $S1$.

Para garantizar convergencia a una muestra con las propiedades deseadas, se genera $S1$, se repiten estos mismos pasos utilizando $S1$ para generar una nueva muestra $S2$; las muestras sucesivas $S1, S2, \ldots$, forman una cadena de Markov que gradualmente convergerá a una muestra que tenga las propiedades deseadas. En este caso, convergen a una muestra look-alike.

4. Muestras *look-alike* en Modelos Lineales Generalizados

4.1. Caso Poisson

Considere el caso en el que se tiene una variable dependiente con distribución Poisson y una variable categórica. Para obtener el estimador de Rao-Blackwell de $F(y;\theta)$ primero se calcula $P(Y_j = y_j|t_j)$ y posteriormente se calcula $P(Y_j \leq y_j|t_j) = \sum_{i=0}^{y_j} P(Y_j = i|t_j)$. En este caso

$$P(Y_m = y_m|t_m) = \sum_{h=1}^{k+1} B\left(t^*_{mh}\ ,\ \frac{1}{1+\sum_{i=1}^{m-1}\prod_{j=1}^{k} x_{ji}^{l_j}(1-x_{ji})^{1-l_j}}\right) 1_{\{h\}}.$$

A partir de esta distribución simulamos.

4.2. Caso binomial

Considere el caso en el que se tiene una variable dependiente con distribución Binomial y una variable categórica. Para obtener la muestra look-alike por el método del muestreador de Gibbs es necesario obtener la densidad condicional $f_c(y_n)$ y la densidad conjunta $g_n(y_n)$.

$$\begin{aligned} g_n(y_n) &= \left[\prod_{i=k+2}^{n}\binom{m}{y_i}\exp\left\{y_i(\beta_0+\beta 1x_{1i}+\cdots+\beta_k x_{ki})\right\}\left(1+e^{\beta_0+\beta 1x_{1i}+\cdots+\beta_k x_{ki}}\right)^{-m}\right] \\ &\times \prod_{h=1}^{k+1}\left[\left(t^*_{nh}-\sum_{i=k+2}^{n}\prod_{j=1}^{k}{}^{m} x_{ji}^{l_j}(1-x_{ji})^{1-l_j}\right)\left(1+e^{\beta_0+\beta 1x_{1i}+\cdots+\beta_k x_{ki}}\right)^{-m}\right. \\ &\times \left.\exp\left\{(\beta_0+\beta 1x_{1i}+\cdots+\beta_k x_{ki})\left(t^*_{nh}-\sum_{i=k+2}^{n}\prod_{j=1}^{k} x_{ji}^{l_j}(1-x_{ji})^{1-l_j}\right)\right\}\right] \end{aligned}$$

$$f_c(y_n) \;=\; \sum_{h=1}^{k+1} 1_{\{h\}} \mathrm{H}\left(2m, t^*_{nh} - \sum_{i=k+2}^{n-1} \prod_{j=1}^{k} x_{ji}^{l_j}(1-x_{ji})^{1-l_j}, m\right).$$

Esto lo utilizamos para simular vía muestreo de Gibbs.

5. Conclusiones

Estamos realizando estudios de simulación para comparar con la distribución asintótica de la devianza condicional y la aproximación sugerida por McCullagh(1986), aunque todavía no los hemos conluído, si hay una diferencia en el nivel de significancia.

Referencias

Lockhart, R.A., O'Reilly, F.J. y Stephens, M.A. (2007). "Use of the Gibbs sampler to obtain conditional tests with applications". *Biometrika* Vol 94, No 4, pp 992-998.

McCullagh, P. (1986). "The Conditional distribution of Goodness of fit statistics for discrete data". *J. Amer. Statist. Assoc.* 81 104-107.

Nelder, J. y Wedderburn, R.W.M. (1972). "Generalized Linear Models". *J. Roy. Statis. Soc.* A135 370-384.

O'Reilly, F. y Gracia-Medrano, L. (2004). "Transformations for Testing the Fit of the Inverse-Gaussian Distribution". *Communications in Statistics (Theory and Methods)*, 33, 919-924.

O'Reilly, F. y Gracia-Medrano, L. (2006). "On the conditonal distribution of goodness-of-fit tests". *Communications in Statistics (Theory and Meth)*, Vol 35, No 3, pp 541-549.

O'Reilly, F. y Quesenberry, C. P. (1973). "The Conditional Probability Integral Transformation and Applications to Obtain Composite Chi-Square Goodness-of-Fit Tests". *Annals of Statistics*, Vol 1, 74-83.

Memoria FONE XXIV (2010), Sección I, 75–81

¿Es la prueba de Blackwelder de no-inferioridad para dos proporciones la mejor prueba disponible?

David Sotres-Ramos[a], Cecilia Ramírez-Figueroa
Colegio de Postgraduados

1. Resumen

En diferentes artículos sobre metodología estadística se recomienda utilizar la prueba de Blackwelder para probar no-inferioridad de dos proporciones independientes, ver por ejemplo Hwang y Morikawa(1999). Sin embargo, actualmente se dispone de pruebas estadísticas mas eficientes que esta prueba de Blackwelder. En este trabajo se calculan y comparan las potencias exactas de las pruebas de Blackwelder y de Farrington-Manning(FM) para diferentes tamaños de muestra (n_1=n_2=n= 45, 60, 70, 75, 80, 90 y 95) y con α=0.05. En todos los casos calculados la potencia de FM resultó superior a la potencia de la prueba de Blackwelder.

2. Introducción

Las pruebas estadísticas de no-inferioridad se utilizan muy frecuentemente en ensayos clínicos. Estas pruebas sirven para demostrar que una terapia nueva (con menores efectos secundarios o menor costo) no es sustancialmente inferior en eficacia a la terapia estándar (Chen et al., 2000). La prueba exacta de Farrington-Manning(FM) para no-inferioridad para dos proporciones independientes ha sido estudiada y recomendada por varios autores, ver por ejemplo Chan(1998) y Röhmel(2005). Ramírez(2008) comparó la prueba exacta de FM con otras 6 pruebas exactas para no-inferioridad. Ramírez(2008) probó que el tamaño de la prueba exacta de FM tiene el mejor comportamiento comparativamente con las otras pruebas estudiadas

[a]sotres.davida@kendle.com

para tamaños de muestra 30< n <100, así como para los tres márgenes de no inferioridad más utilizados en la práctica y para los niveles de significancia nominales 0.01 y 0.05. Sotres-Ramos, et-al(2010) proporcionaron valores críticos y tamaños para la prueba exacta de FM. Por otra parte la prueba de Blackwelder para no-inferioridad es frecuentemente utilizada en la práctica, ver por ejemplo Hwang y Morikawa(1999). El principal objetivo de este trabajo es comprobar que la prueba exacta de FM tiene un nivel de significancia real mucho más cercano al nivel nominal prefijado que el correspondiente de Blackwelder. Adicionalmente se demuestra que la potencia de la prueba de FM es superior a la prueba de Blackwelder para diferentes tamaños de muestra y diferentes combinaciones de parámetros considerados. Se presentan gráficas para sustentar esta afirmación.

3. Pruebas estadísticas consideradas

Sean X_1 y X_2 dos variables aleatorias independientes con distribución binomial y con parámetros (n_1,p_1) y (n_2,p_2) respectivamente, donde p_1 y p_2 representan las probabilidades de respuesta de los tratamientos estándar y nuevo, respectivamente. La hipótesis de interés (hipótesis de no-inferioridad) a ser probada es la alternativa (H_a) en el siguiente juego de hipótesis:

$$H_0 : p_1 - p_2 \geq d_0 \quad vs \quad H_a : p_1 - p_2 < d_0 \tag{1}$$

donde d_0 es el margen de no-inferioridad el cual es una constante positiva y conocida. En el contexto de ensayos clínicos los valores usuales para d_0 son 0.10, 0.15 y 0.20.

La estadística de prueba de FM se define como:

$$T(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}} \tag{2}$$

donde $X_1 = \Sigma X_{1j}$, $X_2 = \Sigma X_{2j}$, y $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i para $i = 1, 2$ y $\hat{\sigma}$ es el estimador de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$; definido por

$$\hat{\sigma} = \left(\frac{\check{p}_1(1 - \check{p}_1)}{n_1} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2} \right)^{1/2}$$

donde $\check{p}_i$ es el estimador de máxima verosimilitud restringida bajo la hipótesis nula de p_i, ver Farrington y Manning (1990). El espacio muestral $\chi = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\}\}$

y el espacio paramétrico es $\Theta = \left\{(p_1, p_2) \in [0,1]^2\right\}$. Para un nivel nominal igual a α, la región de rechazo para la prueba de FM es de la forma:

$$R_T(\alpha) = \{(x_1, x_2) \in \{0, ..., n_1\} \times \{0, ..., n_2\} : T(x_1, x_2) < T(x_1^*, x_2^*)\} \tag{3}$$

la definición de (x_1^*, x_2^*) se establece líneas abajo. A $R_T(\alpha)$ lo denotaremos por R_T o simplemente por R. La función de verosimilitud conjunta es:

$$L\left(p_1, p_2; x_1, x_2\right) = \binom{n_1}{x_1} p_1^{x_1} \left(1 - p_1\right)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} \left(1 - p_2\right)^{n_2 - x_2}$$

y la función de potencia es $\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_T(\alpha)} L(p_1, p_2; x_1, x_2)$

Por lo tanto, el tamaño de la prueba esta dado por $\sup_{(p_1, p_2) \in \Theta_0} \beta_T(p_1, p_2)$ donde $\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$ es el espacio nulo.

El punto (x_1^*, x_2^*) en la ecuación (3) esta definido como

$$T(x_1^*, x_2^*) = \text{máx} \left\{ T(a, b) : \sup_{(p_1, p_2) \in \Theta_0} \left(\sum_{T(x_1, x_2) \leq T(a,b)} L\left(p_1, p_2; x_1, x_2\right) \right) \leq \alpha \right\}$$

La prueba exacta de Blackwelder tiene la misma forma que la prueba de FM definida en (2) y (3) excepto que el estimador de la desviación estandar $\hat{\sigma}$ se define de la siguiente manera

$$\hat{\sigma}_2 = \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}$$

4. Estrategia para calcular el nivel de significancia real

Chan(1998) calculó el nivel de significancia para la prueba de FM tomando el supremo no en todo el espacio nulo (Θ_0) sino calculando el máximo únicamente en $\Theta_0^* = \{(p_1, p_2) \in \Theta : p_1 - p_2 = d_0\}$ el cual es solamente una parte de la frontera del espacio nulo. Computacionalmente ésto representa una inmensa ventaja, pues el tiempo de cómputo se reduce aproximadamente al 0.22 % del tiempo original. Sin embargo, el autor mencionado no justificó formalmente la validez de este argumento. Fue hasta 2005 cuando Röhmel(2005) presenta una prueba formal que justifica el procedimiento utilizado por Chan(1998). En este trabajo se siguió la misma estrategia de Chan (1998).

Definición. *Se dice que una prueba estadística, para el problema en (1), con región de rechazo R_T cumple la* ***condición de convexidad de Barnard*** *(C) si satisface las dos propiedades siguientes:*

a) $(x, y) \in R_T \Longrightarrow (x-1, y) \in R_T \ \ \forall \ \ 1 \leq x \leq n_1,\ 0 \leq y \leq n_2$

b) $(x, y) \in R_T \Longrightarrow (x, y+1) \in R_T \ \ \forall \ \ 0 \leq x \leq n_1,\ 0 \leq y \leq n_2 - 1$

Definición. *Si $n_1 = n_2 = n$, se dice que una región de rechazo R cumple* ***la condición de simetría en la misma cola*** *si $(x, y) \in R \Longrightarrow (n-y,\ n-x) \in R$.*

Proposición 3.1. *Sean $n_1 = n_2 = n$ y $R(\alpha)$ una región crítica para el problema de prueba de hipótesis en (1), si $R(\alpha)$ cumple la condición de convexidad de Barnard y la condición de simetría en la misma cola, entonces el nivel de significancia exacto de la prueba $R(\alpha)$ está dado por*:

$$\alpha^* = \max_{\substack{p_2 = p_1 - d_0 \\ p_1 \in \left[d_0, \frac{1-d_0}{2}\right]}} \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2 - x_2} I_{[(x_1, x_2) \in R(\alpha)]} \qquad (4)$$

Demostración. Se omite por razones de espacio. Proposición 3.2. Las pruebas de FM y de Blackwelder satisfacen la condición de convexidad de Barnard y la condición de simetría en la misma cola. Demostración. Se omite por razones de espacio. Con base en las proposiciones 3.1 y 3.2, el cálculo del nivel de significancia de las pruebas consideradas se hizo aplicando la fórmula en (4) y particionando el intervalo $[d_0, (1-d_0)/2]$ en subintervalos de longitud 0.001. Esto quiere decir que aproximamos el nivel de significancia exacto reemplazando en la fórmula (4) el intervalo continuo $[d_0, (1-d_0)/2]$ por el conjunto por el conjunto finito de puntos (p_1, p_2) tales que $p_1 = \{d_0 + (0.001)\, i : i = 0, 1, 2, ..., 500(1-d_0)\}$ y $p_2 = p_1 - d_0$.

5. Resultados y conclusiones

En el trabajo de Ramírez(2008) se probó que la prueba de FM tiene un nivel de significancia real mucho mas cercano al nivel nominal prefijado que el correspondiente de Blackwelder. Por lo que en lo que resta del escrito nos concentraremos en la comparación de las potencias de las pruebas de FM y de Blackwelder. Para un nivel de significancia α=0.05, d_0=0.10 y tamaños de muestra $n_1 = n_2 = n$= 45, 60, 70, 75, 80, 90 y 95 se calculó la potencia de las pruebas de FM y de Blackwelder. Previo al calculo de las potencias se igualaron

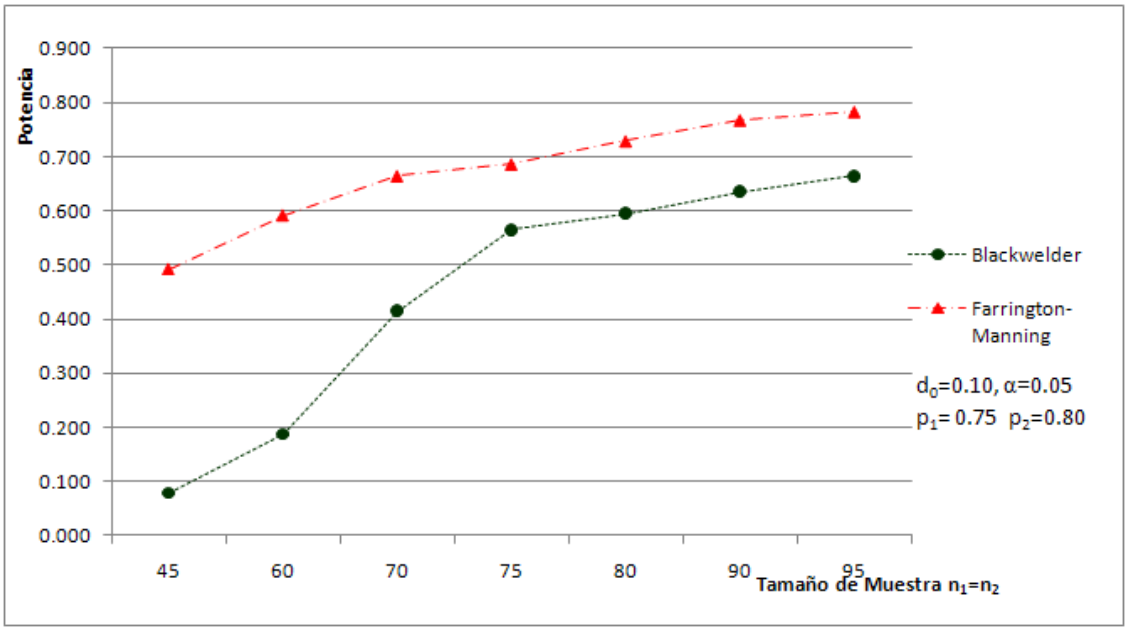


Figura 1: Comparación de potencias uno

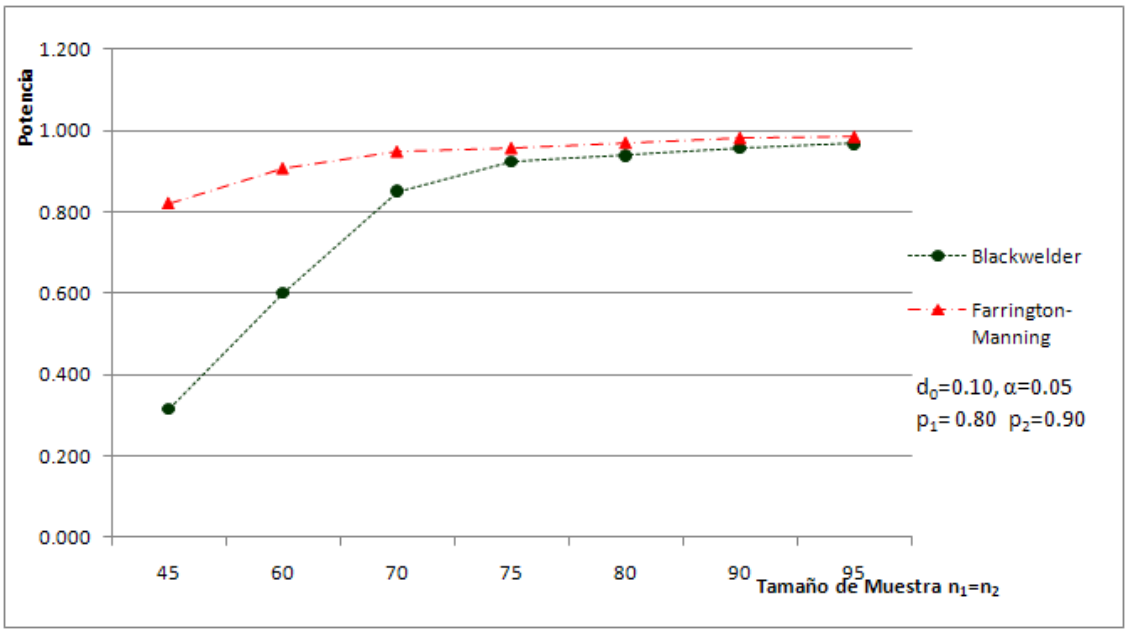


Figura 2: Comparación de potencias dos

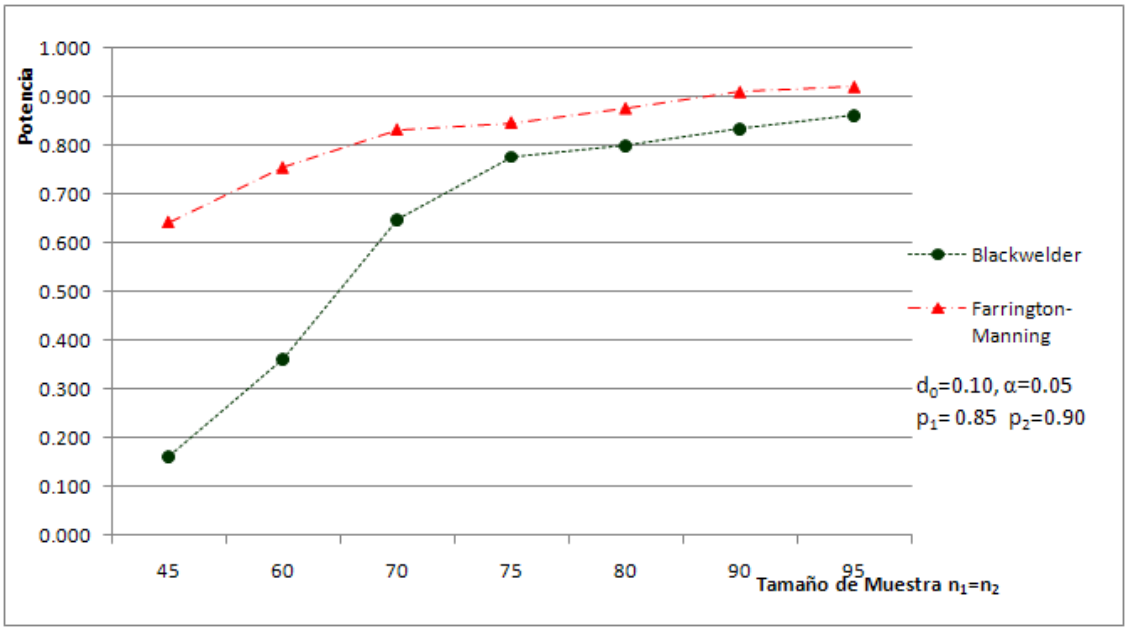


Figura 3: Comparación de potencias tres

los niveles de significancia reales para obtener pruebas del mismo tamaño. Los resultados se presentan en las figuras 1, 2 y 3. Resulta claro de estas figuras que la prueba de FM tiene mayor potencia que Blackwelder uniformemente en todos los tamaños de muestra y para todas las combinaciones de parámetros considerados. Las potencias se calcularon para muchas otras combinaciones de parámetros, pero por brevedad estas no se presentan. Los resultados obtenidos en estas otras comparaciones de potencias son similares a los que se muestran en las figuras 1, 2 y 3.

Referencias

Hwang, I. K., and T.Morikawa. 1999. "Design issues in non-inferiority/equivalence trials". *Drug Information Journal* 33:1205-1218.

Chan, ISF. 1998. "Exact tests of equivalence and efficacy with a non zero lower bound for comparative studies". *Statistics in Medicine* 17:1403-1413.

Chen, J., Tsong, Y. and S. Kang. 2000. "Tests for equivalence or noninferiority between two proportions". *Drug Information Journal* 34:569-578.

Röhmel, J. 2005. "Problems with existing procedures to compute exact unconditional p-values for noninferiority/superiority and confidence intervals for two binomials and how to resolve them". *Biometrical Journal* 47:37-47.

Farrington, C. and G. Manning. 1990. "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.

Ramírez F.C. 2008. "Comparación de Pruebas Exactas y Asintóticas de no inferioridad para dos proporciones independientes". Tesis Doctoral, Especialidad en Estadística, ISEI, Colegio de Postgraduados, México. 89 p.

Sotres-Ramos, D.,Almendra-Arao, F. and C. Ramírez-Figueroa. 2010. "Exact Critical values for Farrington-Manning non-inferiority exact test". *Drug Information Journal* 44:159-164.

Memoria FONE XXIV (2010), Sección I, 83–89

Caracterización del BLUP de la media poblacional en el modelo lineal general mixto

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

1. Introducción

La teoría de espacios vectoriales de dimensión finita proporciona un marco para trabajar en forma didáctica los procesos de inferencia en el modelo lineal general (MLG). Conceptos como subespacio columna y operador proyector, juegan un papel de suma importancia en la estadística teórica, en particular en el estudio de la estimación y predicción en el MLG. La caracterización de estimadores en el MLG por medio del operador proyector permite comprender sus propiedades y plantear generalizaciones de la inferencia. Por otro lado, 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 $\overline{Y} = T/N$ [**?**]. 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 General Mixto (MLGM). En este enfoque se considera el modelo $\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \mathbf{e}_j$, en la j-ésima unidad de nivel 2, la cual cuenta con N_j unidades; 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:

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

La media de la población finita en la j-ésima unidad de nivel 2 $\overline{Y}_j = N_j^{-1}\sum_{i=1}^{N_j} Y_{ij}$ 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 μ_j de la j-ésima unidad de nivel 2, la cual es un efecto mixto. Por la teoría de Henderson (1975) [**?**] un predictor del efecto mixto μ_j está dado por medio de $\overline{\mathbf{X}}_{j\mathbf{r}}\ \hat{\boldsymbol{\beta}} + \overline{\mathbf{Z}}_{j\mathbf{r}}\ \hat{\mathbf{G}}\ \mathbf{Z}_{j\mathbf{s}}^t\mathbf{V}_{j\mathbf{ss}}^{-1}\left(\mathbf{Y}_{j\mathbf{s}} - \mathbf{X}_{j\mathbf{s}}\ \hat{\boldsymbol{\beta}}\right)$ donde $\overline{\mathbf{X}}_{j\mathbf{r}}$ 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. 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 mejor predictor lineal insesgado ($BLUP$) de la media poblacional μ_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 μ_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.

2. Marco teórico

Se considera el modelo dado por:

$$\begin{gathered}\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},\\ \mathbf{u} \sim N\left(\mathbf{0}, \sigma_u^2\mathbf{I}_J\right), \quad \mathbf{e} \sim N\left(\mathbf{0}, \sigma_e^2\mathbf{I}_n\right),\\ Cov\left(\mathbf{e}, \mathbf{u}^{\mathbf{t}}\right) = \mathbf{0},\end{gathered} \tag{3}$$

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

Henderson (1975) obtiene el mejor estimador lineal insesgado ($BLUE$) de $\boldsymbol{\beta}$ y el $BLUP$ de $\mathbf{u}$, que están dados por $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}$ y $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\ \hat{\boldsymbol{\beta}}\right)$, respectivamente. Además del $BLUP$ del efecto mixto $\mathbf{k}^t\boldsymbol{\beta} + \mathbf{m}^t\mathbf{u}$ que está dado por:

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

3. Caracterización del BLUP del efecto mixto $\mathbf{X}\boldsymbol{\beta}+\mathbf{Zu}$

En está sección se presenta 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 3.1. *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. Si se cumple la condición $n_j\mathbf{P}_{\mathbf{Z}_j} = \mathbf{Z}_j\mathbf{Z}_j^t$, entonces la matriz $\mathbf{V}$ se expresa en términos de los operadores proyector $\mathbf{P}_{\mathbf{Z}_j}$ y $\mathbf{Q}_{\mathbf{Z}_j}$ por $\mathbf{V} = \oplus_{j=1}^{J}\left[\left(n_j\sigma_{u0}^2+\sigma_e^2\right)\mathbf{P}_{\mathbf{Z}_j} + \sigma_e^2\mathbf{Q}_{\mathbf{Z}_j}\right]$, y la inversa $\mathbf{V}^{-1}$ se expresa por $\mathbf{V}^{-1} = \oplus_{j=1}^{J}\left[\frac{\mathbf{P}_{\mathbf{Z}_j}}{\left(n_j\sigma_{u0}^2+\sigma_e^2\right)} + \frac{\mathbf{Q}_{\mathbf{Z}_j}}{\sigma_e^2}\right]$, de lo cual y de (4) se tiene:

$$\begin{aligned}
BLUP\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Zu}\right) &= \mathbf{X}\,\hat{\boldsymbol{\beta}} + \mathbf{Z}\,\hat{\mathbf{u}} = \mathbf{P_{XV}}\mathbf{Y} + \sigma_{u0}^2\mathbf{Z}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\,\hat{\beta}\right)\\
&= \mathbf{P_{XV}}\mathbf{Y} + \left(\oplus_{j=1}^{J}\left[\frac{n_j\sigma_{u0}^2\mathbf{P}_{\mathbf{Z}_j}}{\left(n_j\sigma_{u0}^2+\sigma_e^2\right)}\right]\right)\left(\mathbf{Y} - \mathbf{X}\,\hat{\beta}\right)\\
&= \mathbf{P_{XV}}\mathbf{Y} + \mathbf{P_Z}\left(\oplus_{j=1}^{J}\left[\frac{n_j\sigma_{u0}^2\mathbf{I}_{n_j}}{\left(n_j\sigma_{u0}^2+\sigma_e^2\right)}\right]\right)\left(\mathbf{Y} - \mathbf{X}\,\hat{\beta}\right),
\end{aligned}$$

definiendo $b_j = n_j\sigma_{u0}^2/\left(n_j\sigma_{u0}^2+\sigma_e^2\right)$ y $\mathbf{B} = \oplus_{j=1}^{J}\left(b_j\mathbf{I}_{n_j}\right)$ se obtiene (5).

Corolario 3.1. *Bajo el modelo (3), si se cumple* $n_j\mathbf{P}_{\mathbf{Z}_j} = \mathbf{Z}_j\mathbf{Z}_j^t$, *entonces el BLUP del efecto aleatorio* $\mathbf{Zu}$ *se expresa en términos de los operadores proyector, oblicuo* $\mathbf{P_{XV}}$ *sobre* $S(\mathbf{X})$ *y ortogonal* $\mathbf{P_Z}$ *sobre* $S(\mathbf{Z})$, *por medio de* $\mathbf{P_Z}\mathbf{B}\mathbf{Q_{XV}}\mathbf{Y}$, *donde* $\mathbf{B} = \oplus_{j=1}^{J}\left(b_j\mathbf{I}_{n_j}\right)$ *y* $b_j = n_j\sigma_{u0}^2/\left(n_j\sigma_{u0}^2+\sigma_e^2\right)$.

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 además se cumple la condición $\sum_{i=1}^{n_j} z_{ij}^2 = n_j$.

4. Caracterización del *BLUP* de la media poblacional μ_j

En está sección se presenta la caracterización del *BLUP* de la media poblacional μ_j de la j-ésima unidad de nivel 2 en términos de los operadores $\mathbf{P_{XV}}$ y $\mathbf{P_Z}$, y de una transformación lineal definida sobre el subespacio $S(\mathbf{X}_j)$ generado por la matriz de diseño $\mathbf{X}_j$.

Una vez que la muestra $\mathbf{s}$ 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 $\hat{u}_{j\mathbf{s}}$ respectivamente.

Teorema 4.1. $\mathbf{T}_{j\mathbf{s}}$ *dada por* $\mathbf{T}_{j\mathbf{s}} = \mathbf{X}_j \left(\mathbf{X_s^t V_s^{-1} X_s}\right)^{-1} \mathbf{X_s^t V_s^{-1}}$ *define una transformación lineal de* $\mathbb{R}^s$ *a* $\mathbb{R}^j$.

Teorema 4.2. *Bajo el modelo (3) si se cumple la condición* $n_j \mathbf{P}_{\mathbf{Z}_{j\mathbf{s}}} = \mathbf{Z}_{j\mathbf{s}} \mathbf{Z}_{j\mathbf{s}}^t$, *entonces el BLUP del efecto mixto* $\overline{X}_j^t \boldsymbol{\beta} + u_j$ *se expresa en términos de los operadores proyector* $\mathbf{Q_{X_sV_s}}$ *y* $\mathbf{P_{Z_s}}$, *y de la transformación lineal* $\mathbf{T}_{j\mathbf{s}}$ *por:*

$$\left[\frac{\mathbf{1}_{N_j}^t}{N_j}\mathbf{T}_{j\mathbf{s}} \; + \; \frac{\mathbf{1}_{n_s}^{*j\mathbf{s}^t}}{n_{js}} \left(\mathbf{P_{Z_s} B_s Q_{X_sV_s}}\right)\right] \mathbf{Y_s}, \tag{6}$$

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

Demostración. Del corolario 3.1, el *BLUP* de $\mathbf{Zu}$ se expresa en términos de los operadores proyector, oblicuo $\mathbf{P_{X_sV_s}}$ sobre $S(\mathbf{X_s})$ y ortogonal $\mathbf{P_{Z_s}}$ sobre $S(\mathbf{Z_s})$, por:

$$\left(\mathbf{P_{Z_s} B_s} \left(\mathbf{I} - \mathbf{P_{X_sV_s}}\right)\right) \mathbf{Y_s}, \tag{7}$$

donde $\mathbf{B_s} = \oplus_{j=1}^{J} \left(b_j \mathbf{I}_{n_j}\right)$ y $b_j = n_j \sigma_{u0}^2 / \left(n_j \sigma_{u0}^2 + \sigma_e^2\right)$. La relación entre el *BLUP* del vector $\mathbf{Zu}$ y el *BLUP* de u_j está dada por:

$$BLUP\left(u_j\right) = \frac{\mathbf{1}_n^{*j^t}}{n_j} BLUP\left(\mathbf{Zu}\right), \tag{8}$$

donde $\mathbf{1}_n^{*j^t}$ es un vector en $\mathbb{R}^n$ de 0's con un 1 en las posiciones correspondientes a las unidades de nivel 1 que pertenecen a la j-ésima unidad de nivel 2. Así de (7) y (8) el *BLUP* del efecto aleatorio u_j se expresa en términos de los operadores proyector $\mathbf{Q_{X_sV_s}}$ y $\mathbf{P_{Z_s}}$, por:

$$\frac{\mathbf{1}_{n_s}^{*j\mathbf{s}^t}}{n_{js}} \left(\mathbf{P_{Z_s} B_s Q_{X_sV_s}}\right) \mathbf{Y_s}. \tag{9}$$

$\overline{X}_j^t\boldsymbol{\beta}+u_j$ es un efecto mixto, así su *BLUP* está dado por $\overline{X}_j^t\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\ +\ \hat{u}_{j\mathbf{s}}$, y de (9)

$$\begin{aligned} BLUP\left(\overline{X}_j^t\boldsymbol{\beta}+u_j\right) &= \overline{X}_j^t\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\ +\ \hat{u}_{j\mathbf{s}}=\frac{\mathbf{1}_{N_j}^t\mathbf{X}_j}{N_j}\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}\ +\ \hat{u}_{j\mathbf{s}} \\ &= \frac{\mathbf{1}_{N_j}^t}{N_j}\mathbf{X}_j\left(\mathbf{X}_{\mathbf{s}}^t\mathbf{V}_{\mathbf{s}}^{-1}\mathbf{X}_{\mathbf{s}}\right)^{-1}\mathbf{X}_{\mathbf{s}}^t\mathbf{V}_{\mathbf{s}}^{-1}\mathbf{Y}_{\mathbf{s}}+\ \hat{u}_{j\mathbf{s}} \\ &= \frac{\mathbf{1}_{N_j}^t}{N_j}\mathbf{T}_{j\mathbf{s}}\mathbf{Y}_{\mathbf{s}}\ +\ \frac{\mathbf{1}_{n_s}^{*js^t}}{n_{js}}\left(\mathbf{P}_{\mathbf{Z}_{\mathbf{s}}}\mathbf{B}_{\mathbf{s}}\mathbf{Q}_{\mathbf{X}_{\mathbf{s}}\mathbf{V}_{\mathbf{s}}}\right)\mathbf{Y}_{\mathbf{s}}. \end{aligned}$$

La media poblacional μ_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 $\overline{X}_{jr}^t\boldsymbol{\beta}+u_j$. Cuando n_j/N_j es insignificante μ_j toma la forma $\overline{X}_j^t\boldsymbol{\beta}+u_j$.

Teorema 4.3. *Bajo el modelo (3) si se cumple la condición* $n_j\mathbf{P}_{\mathbf{Z}_{js}}=\mathbf{Z}_{j\mathbf{s}}\mathbf{Z}_{j\mathbf{s}}^t$*, y si* n_j/N_j *es insignificante, entonces el BLUP de* μ_j *se expresa en términos de los operadores proyector* $\mathbf{Q}_{\mathbf{X}_{\mathbf{s}}\mathbf{V}_{\mathbf{s}}}$ *y* $\mathbf{P}_{\mathbf{Z}_{\mathbf{s}}}$*, y de la transformación lineal* $\mathbf{T}_{j\mathbf{s}}$ *por:*

$$\left[\frac{\mathbf{1}_{N_j}^t}{N_j}\mathbf{T}_{j\mathbf{s}}\ +\ \frac{\mathbf{1}_{n_s}^{*j\mathbf{s}t}}{n_{js}}\left(\mathbf{P}_{\mathbf{Z}_{\mathbf{s}}}\mathbf{B}_{\mathbf{s}}\mathbf{Q}_{\mathbf{X}_{\mathbf{s}}\mathbf{V}_{\mathbf{s}}}\right)\right]\mathbf{Y}_{\mathbf{s}}, \tag{10}$$

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

Demostración. Como μ_j es el efecto mixto $\overline{X}_j\boldsymbol{\beta}+u_j$, su *BLUP* está dado por $\overline{X}_j^t\,\hat{\boldsymbol{\beta}}_{\mathbf{s}}+\hat{u}_{j\mathbf{s}}$ y del teorema 4.2, el *BLUP* de μ_j se expresa en términos de $\mathbf{Q}_{\mathbf{X}_{\mathbf{s}}\mathbf{V}_{\mathbf{s}}}$, $\mathbf{P}_{\mathbf{Z}_{\mathbf{s}}}$ y $\mathbf{T}_{j\mathbf{s}}$, por:

$$\left[\frac{\mathbf{1}_{N_j}^t}{N_j}\mathbf{T}_{j\mathbf{s}}\ +\ \frac{\mathbf{1}_{n_s}^{*j\mathbf{s}t}}{n_{js}}\left(\mathbf{P}_{\mathbf{Z}_{\mathbf{s}}}\mathbf{B}_{\mathbf{s}}\mathbf{Q}_{\mathbf{X}_{\mathbf{s}}\mathbf{V}_{\mathbf{s}}}\right)\right]\mathbf{Y}_{\mathbf{s}},$$

5. Caracterización en el modelo sólo intercepto

Se presenta la caracterización del *BLUP* de la media poblacional μ_j, considerando el caso balanceado, es decir $n_j=d\ \forall j=1,...,k$, bajo el modelo sólo intercepto sin variables explicatorias, dado por:

$$Y_{ij}=\mu+u_j+e_{ij},\qquad i=1,...,d,\qquad j=1,...,k. \tag{11}$$

donde μ es un parámetro fijo; u_j es el efecto aleatorio; u_j y e_{ij} son independientes, con $u_j\sim N\left(0,\sigma_{u0}^2\right)$ y $e_{ij}\sim N\left(0,\sigma_e^2\right)$. El modelo para la j-ésima unidad de nivel 2 tiene la forma $\mathbf{Y}_j=\mathbf{1}_d\mu+\mathbf{1}_d u_j+\mathbf{e}_j,\qquad j=1,...,k.$

Teorema 5.1. *Bajo el modelo sólo intercepto sin variables explicatorias (11), considerando el caso balanceado y si n_j/N_j es insiginificante, entonces el BLUP de la media poblacional μ_j está dado por:*

$$\overline{Y}_s \ + \ \frac{c\,(k-1)}{k}\left[\overline{Y}_{js} - \overline{Y}_{(-j)s}\right] \tag{12}$$

donde $\overline{Y}_s$, $\overline{Y}_{js}$ y $\overline{Y}_{(-j)s}$ denotan la media muestral, la media muestral de la j-ésima unidad de nivel 2, y la media muestral de las unidades de nivel 1 que no pertenecen a la j-ésima unidad de nivel 2, respectivamente.

Demostración. Definiendo $c = d\sigma^2_{u0}/\left(d\sigma^2_{u0} + \sigma^2_e\right)$, por (10) del teorema 4.3, el *BLUP* de μ_j está dado por $\left[\frac{\mathbf{1}^t_{N_j}}{N_j}\mathbf{T}_{\mathbf{js}} \ + \ \frac{\mathbf{1}^{*\mathbf{js}t}_{n_s}}{d}\left(c\mathbf{P}_{\mathbf{Z_s}}\mathbf{Q}_{\mathbf{X_sV_s}}\right)\right]\mathbf{Y_s}$. Además se cumple para el modelo sólo intercepto (11): a) $\mathbf{P_{XV}} = \frac{\mathbf{1}_{kd}\mathbf{1}^t_{kd}}{kd}$, b) $\mathbf{P_Z P_{XV}} = \mathbf{P_{XV}}$ y c) $\mathbf{T}_{js} = \frac{\mathbf{1}_{N_j}\mathbf{1}^t_{kd}}{kd}$, así:

$$\begin{aligned}
BLUP\left(\mu_j\right) &= \left[\frac{\mathbf{1}^t_{N_j}}{N_j}\frac{\mathbf{1}_{N_j}\mathbf{1}^t_{kd}}{kd} \ + \ \frac{\mathbf{1}^{*js^t}_{n_s}}{d}\left[\frac{c}{kd}\left(k\left(\oplus\mathbf{1}_d\mathbf{1}^t_d\right) - \mathbf{1}_{kd}\mathbf{1}^t_{kd}\right)\right]\right]\mathbf{Y}_s \\
&= \left[\frac{N_j}{N_j}\frac{\mathbf{1}^t_{kd}}{kd} \ + \ \frac{c}{kdd}\left[\mathbf{1}^{*js^t}_{n_s}\left(k\left(\oplus\mathbf{1}_d\mathbf{1}^t_d\right) - \mathbf{1}_{kd}\mathbf{1}^t_{kd}\right)\right]\right]\mathbf{Y}_s \\
&= \frac{\mathbf{1}^t_{kd}}{kd}\mathbf{Y}_s \ + \ \frac{c}{kdd}\left[d\left(k\mathbf{1}^{*js^t}_{n_s} - \mathbf{1}^t_{kd}\right)\right]\mathbf{Y}_s \\
&= \overline{Y}_s \ + \ \frac{c}{kd}\left[(k-1)\,d\overline{Y}_{js} - (kd-d)\,\overline{Y}_{(-j)s}\right] \\
&= \overline{Y}_s \ + \ \frac{c\,(k-1)}{k}\left[\overline{Y}_{js} - \overline{Y}_{(-j)s}\right].
\end{aligned}$$

6. Conclusiones

En este trabajo se expresó el *BLUP* de la media poblacional μ_j como la suma ponderada de un elemento en $S\left(\mathbf{X}_j\right)$ y un elemento en el espacio $S\left(\mathbf{Z_s}\right)$. Lo anterior al aplicarlo al modelo sólo intercepto sin variables explicatorias, considerando el caso balanceado, permitió expresar el *BLUP* de la media poblacional μ_j como la suma de la media muestral, y un múltiplo de la diferencia entre las medias muestrales $\overline{Y}_{js}$ y $\overline{Y}_{(-j)s}$, que denotan la media muestral de las unidades de nivel 1 en la j-ésima unidad de nivel 2 y la media muestral de las unidades de nivel 1 que no pertenecen a la j-ésima unidad de nivel 2, respectivamente. Se espera que esta caracterización en términos de los proyectores permita una mejor comprensión de las propiedades del *BLUP* de μ_j tal como sucede en la caracterización del estimador de

parámetros $\boldsymbol{\beta}$ en el MLG. Al considerar situaciones en las cuales la matriz de diseño $\mathbf{Z}$ involucre a las columnas de la matriz de diseño $\mathbf{X}$, por ejemplo en el modelo de coeficientes aleatorios, se debe cumplir necesariamente que $\mathbf{X}^t\mathbf{X} = d\mathbf{I}$ para que los resultados presentados se puedan aplicar.

Referencias

Henderson, C.R. 1975. Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics* 31:423447.

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

Memoria FONE XXIV (2010), Sección I, 91–102

Construcción de un índice multivariado comparable en el tiempo

José Vences Rivera, Marco Antonio Flores Nájera
Instituto Nacional de Estadística y Geografía, INEGI.

1. Introducción

Los indicadores sintéticos son de gran utilidad para resumir la información contenida en un conjunto, generalmente grande, de variables o indicadores básicos medidos sobre un fenómeno de naturaleza multidimensional, y así orientar a los tomadores de decisiones en los diferentes campos de acción. Un indicador sintético debe ser fácil de calcular, confiable, comparable en el tiempo y en el espacio, sencillo de interpretar y de fácil comunicación para el usuario en general; para su construcción suelen utilizarse técnicas de Análisis Estadístico Multivariado para analizar de manera simultánea tres o más variables. Entre las técnicas más usuales se ubica el Análisis de Componentes Principales (ACP) en sus diferentes variantes.

En la revisión de literatura, no se encontró ningún procedimiento para construir un índice multivariado que cumpliera con las características mencionadas, al menos en el campo de la estadística oficial. Particularmente se analizaron los métodos de Componentes Principales Dinámicas, STATIS y Componentes Principales Comunes. *(Veáse Flury, 1984, 1987; Forni, 2000; Lavit, 1988; Lavit, 1982; y Watson, 1983.)*

El INEGI ha tenido múltiples requerimientos no sólo de información básica, sino también derivada y de consulta. La técnica de ACP ha sido aplicada para dar atención a los siguientes requerimientos:

- Índice de rezago social para reubicar las tiendas de DICONSA.
- Índice para medir el cumplimiento de los derechos humanos.
- Índice para medir el nivel de bienestar de la población.

- Índice para la medición multidimensional de la pobreza.
- Índice de actividad económica.
- Índice para seleccionar a los estudiantes de la Maestría en Ciencias en Estadística Oficial.
- Consulta-crítica del Índice de Marginación.
- Consulta-crítica del Índice de Desarrollo Humano.
- Consulta-crítica del índice de Rezago del Gobierno del D.F.

Lo anterior originó el desarrollo del presente trabajo con el propósito de construir una medida resumen que dé cuenta de la magnitud del fenómeno estudiado.

2. Marco teórico

En el ACP, existen tantas componentes principales independientes como variables de insumo correlacionadas se hayan considerado. La primera componente principal explica la mayor cantidad posible de la varianza conjunta de esas variables para una combinación lineal y expresa el "tamaño promedio" del fenómeno estudiado, es por ello que suele utilizarse como un *índice* cuyo nombre es acorde a la naturaleza del problema, y la calidad del mismo se fortalece cuanto mayor es la varianza explicada. El índice se construye al optimizar una función de varianzas y, es claro, que ningún otro índice-combinación lineal de variables, puede contener tanta información como la primera CP.

Por lo anterior, la técnica de ACP ha sido de las más utilizadas para la construcción de índices socio-económicos que permiten facilitar el diseño de programas gubernamentales y la focalización de recursos, ya que genera un ordenamiento entre las unidades de observación. No obstante, la relevancia teórica del indicador resultante, en la práctica presenta ciertas limitaciones: si bien es cierto que retiene la máxima cantidad de información, es decir, es el que mejor "representa" al conjunto original de variables, y además permite un ordenamiento natural de observaciones, en contraparte, no es de utilidad para medir los cambios en el tiempo; y no es fácil su interpretación y comunicación al usuario en general. Ante esta situación, surge la necesidad de generar un índice multivariado que conserve, en la medida de

lo posible, las propiedades teóricas de la primera CP y supere las limitaciones mencionadas. La metodología utilizada en este trabajo fue la siguiente:

- Se estudió a profundidad la técnica de ACP, *(Jackson, 1991; Johnson, 1982; Peña, 2003; Vences, 1999)*. De ello se desprende que en la combinación lineal de la primera componente principal, el mayor ponderador corresponde a la variable que en promedio está más correlacionada con el resto; el segundo ponderador, en orden decreciente de magnitud, es el que se asocia con la variable que presenta la segunda mayor correlación con las demás; y así sucesivamente.

- Se generaron varios índices alternativos.

- Los diversos índices se compararon con la primera CP en términos de la varianza explicada. Para esto, en principio, se recurrió a la técnica de simulación de Monte Carlo; posteriormente se realizaron aplicaciones con datos reales derivados de los censos y conteos de población y vivienda (INEGI), 1990-2005, y se contrastaron con los resultados esperados.

- El mejor índice sería aquel que en términos de la información retenida fuera similar al obtenido por el ACP, pero de fácil interpretación y simplicidad en su cálculo y, sobre todo, que resultara de utilidad para el usuario.

3. El Índice

Sea p = número de variables correlacionadas: x_1, x_2, ..., x_p.

El índice es una combinación lineal de la forma:

$$I = c_1x_1 + c_2x_2 + \ldots + c_px_p$$

Donde, $c_i = \frac{r_i}{s_i s}$, s_i = desviación estándar de la variable i,

$$s = \frac{r_1}{s_1} + \frac{r_2}{s_2} + \cdots + \frac{r_p}{s_p},$$

y r_i = media cuadrática de las correlaciones entre la variable i y el resto de las variables, dada por

$$r_i = \left(\frac{1}{p-1} \sum_{j=1}^{p} r_{ij}^2 \right)^{1/2}$$

Esto para $i = 1, 2, \ldots, p$; con $i \neq j$.

Así, el índice toma la forma:

$$I = \left(\frac{r_1}{s_1 s} \right) x_1 + \left(\frac{r_2}{s_2 s} \right) x_2 + \cdots + \left(\frac{r_p}{s_p s} \right) x_p = \sum_{i=1}^{p} \frac{r_i}{s_i s} x_i$$

O bien,

$$I = \frac{\left(\frac{r_1}{s_1} \right) x_1 + \left(\frac{r_2}{s_2} \right) x_2 + \cdots + \left(\frac{r_p}{s_p} \right) x_p}{\frac{r_1}{s_1} + \frac{r_2}{s_2} + \cdots + \frac{r_p}{s_p}}$$

El índice constituye un promedio aritmético ponderado, donde la suma de los ponderadores es igual a la unidad. Por tanto, se interpreta en términos de las unidades de las variables originales y es muy fácil de calcular por el usuario en general; aspectos de que adolece la técnica de ACP.

4. Los resultados de la simulación

Se simuló un experimento en que se generaron variables correlacionadas distribuidas aleatoriamente, para diferentes tamaños de población. El experimento se repitió 500 veces, mediante la elaboración de una rutina en lenguaje de programación R.

Cuadro 1. Varianza promedio explicada por I frente a Y_1

Dimensión p	**n=100**			**n=2454**			**n=10000**		
	I	Y_1	Y_1/**I**	**I**	Y_1	Y_1/**I**	**I**	Y_1	Y_1/**I**
3	2.525	2.526	1.00038	2.484	2.485	1.00048	2.475	2.476	1.00045
4	3.277	3.278	1.00020	3.189	3.190	1.00026	3.181	3.182	1.00027
5	3.995	3.995	1.00012	3.869	3.869	1.00016	3.856	3.856	1.00015
6	4.739	4.740	1.00007	4.576	4.577	1.00008	4.551	4.551	1.00010
7	5.481	5.481	1.00004	5.224	5.224	1.00006	5.207	5.207	1.00007
8	6.190	6.190	1.00003	5.891	5.891	1.00004	5.858	5.859	1.00005

I = Nuevo índice, Y_1= Primera Componente Principal.

En el Cuadro 1 se observa que la varianza explicada por el nuevo índice, I, es prácticamente la misma que la correspondiente a la primera componente principal, Y_1. De hecho el cociente de la información retenida por Y_1 respecto al nuevo índice es muy cercano a la unidad.

5. Los resultados en la práctica

Para este análisis, se utilizaron los nueve indicadores de CONAPO aplicados a las 32 entidades federativas del país, para los años 1990, 1995, 2000 y 2005.

Al realizar los cálculos correspondientes, se observó que en principio las correlaciones entre estos indicadores, son en general altas para los cuatro años, por lo que dado el traslape de información, es posible construir índices sintéticos que den cuenta de manera resumida, de la situación multivariada que prevalece en los conjuntos de datos originales.

En el Cuadro 2 se presentan algunas estadísticas descriptivas para los indicadores considerados, en él se observa que en promedio los indicadores de rezago social tienden gradualmente a la baja al transcurrir el tiempo, es decir, experimentan una mejora, pero según el coeficiente de variación, el beneficio del desarrollo entre las entidades no es equitativo; por tanto, el nuevo índice sintético tendrá que reflejar esta situación en términos numéricos.

Por su parte, las mayores dispersiones las presenta el indicador que corresponde a la población que vive en localidades rurales, en tanto que las menores a los que no disponen de energía.

Cuadro 2. Estadísticas descriptivas, 1990 – 2005

Nacional

Estadística	**AÑO**	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9
Promedio	1990	12.54	38.88	22.37	13.12	20.36	57.87	21.07	37.42	62.72
	1995	10.79	24.25	14.35	7.40	14.84	64.91	17.13	34.65	64.26
	2000	9.47	29.48	10.55	5.03	10.49	46.31	14.47	33.55	51.65
	2005	8.36	23.78	5.88	2.72	9.59	41.17	11.11	31.43	46.08
Desviación estándar	1990	6.93	10.14	13.68	7.91	12.61	7.55	12.51	17.62	9.31
	1995	6.04	5.61	12.19	5.27	10.47	7.50	10.58	17.09	11.30
	2000	5.43	8.83	7.70	3.50	8.68	7.92	10.68	16.88	13.29
	2005	5.00	7.80	5.50	1.86	8.12	7.60	9.06	16.57	15.12
Coeficiente de variación	1990	55.26	26.09	61.17	60.29	61.91	13.04	59.40	47.10	14.85
	1995	55.97	23.12	84.90	71.11	70.54	11.56	61.76	49.32	17.58
	2000	57.35	29.94	72.94	69.57	82.70	17.11	73.81	50.33	25.73
	2005	59.80	32.81	93.47	68.44	84.69	18.46	81.57	52.71	32.82

I_1: % de población analfabeta de 15 años o más.

I_2: % de población de 15 años o más sin primaria completa.

I_3: % de ocupantes en viviendas sin drenaje o sin excusado.

I_4: % de ocupantes en viviendas sin energía eléctrica.

I_5: % de ocupantes en viviendas sin agua entubada.

I_6: % de viviendas con hacinamiento.

I_7: % de ocupantes en viviendas con piso de tierra.

I_8: % de pobladores que vive en localidades menores a 5 mil habitantes.

I_9: % de la Población Económicamente Activa que percibe hasta 2 SM.

Eléctrica, por ejemplo, para el año de 2005 la varianza del primero es casi 80 veces más grande que la del segundo, lo cual significa que las varianzas de los indicadores son heterogéneas, a pesar de que están medidos en la misma escala (en %), pero los recorridos son notablemente diferentes. Esto, junto con el hecho de que conceptualmente se establece

que las variables son igualmente importantes conlleva a que el índice compuesto se genere a partir de las variables estandarizadas respecto a su desviación estándar, es decir, que se utilice la matriz de correlaciones, en lugar de la de covarianzas.

En el Cuadro 3 se presentan los resultados obtenidos tanto por la primera componente principal (Y_1) como por el nuevo índice (I), así como los niveles de rezago social basados en I; las observaciones están ordenadas de manera descendente conforme al nuevo índice para 1990. Se destaca que el ordenamiento coincide con el de Y_1, (de hecho las varianzas explicadas por ambos procedimientos también coinciden hasta del orden de centésimas y se ubican cerca del 80 %); sin embargo, obsérvese como sus valores no son fáciles de interpretar, algunos son positivos y otros negativos, en el tiempo suben y bajan, además de que no están en la escala de las variables originales, ni poseen un cero absoluto. En contraste, con el nuevo índice estas áreas de oportunidad son esclarecidas, particularmente se observa que el rezago social disminuye a lo largo del tiempo para todas las entidades federativas.

Cuadro 3. Índices y niveles de rezago social por entidad federativa, 1990-2005
Primera componente principal (Y_1) frente al Nuevo índice (I)

Primera componente principal (Y_1)				**Entidad**	**Nuevo índice (I)**				**Niveles**			
1990	**1995**	**2000**	**2005**		**1990***	**1995**	**2000**	**2005**	**1990**	**1995**	**2000**	**2005**
2.36	2.33	2.25	2.32	**07 Chis**	52.4	42.9	34.1	25.5	6	5	4	3
2.06	1.82	2.08	2.13	**20 Oax**	49.7	39.3	32.9	24.5	5	4	4	3
1.75	1.88	2.12	2.41	**12 Gro**	47.0	39.7	33.2	25.9	5	4	4	3
1.17	0.99	0.88	0.75	**13 Hgo**	41.8	33.2	24.7	17.3	5	4	3	2
1.13	1.12	1.28	1.08	**30 Ver**	41.5	34.2	27.5	19.0	5	4	3	2
0.83	0.79	0.72	0.64	**21 Pue**	38.8	31.8	23.7	16.7	4	4	3	2
0.75	0.75	0.72	0.66	**24 SLP**	38.1	31.5	23.7	16.8	4	4	3	2
0.57	0.59	0.30	0.16	**32 Zac**	36.5	30.3	20.8	14.2	4	4	3	2
0.52	0.66	0.66	0.46	**27 Tab**	36.0	30.9	23.2	15.8	4	4	3	2
0.48	0.77	0.70	0.57	**04 Cam**	35.6	31.6	23.5	16.3	4	4	3	2

* Las observaciones están ordenadas por esta columna, en forma descendente.

Niveles de rezago: <10=1; [10,20)=2; [20,30)=3; [30,40)=4; [40,50)=5; >50=6.

Cuadro 3. Índices y niveles de rezago social por entidad federativa, 1990-2005
Primera componente principal (Y_1) frente al Nuevo índice (I)

. . . Continuación

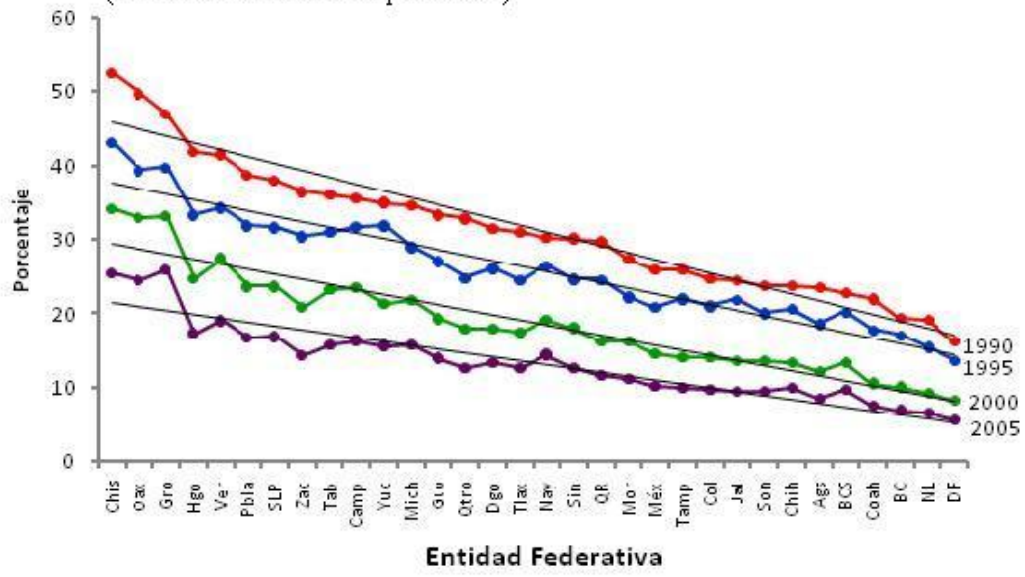
Primera componente principal (Y_1)				**Entidad**	**Nuevo índice (I)**				**Niveles**			
1990	**1995**	**2000**	**2005**		**1990***	**1995**	**2000**	**2005**	**1990**	**1995**	**2000**	**2005**
0.40	0.79	0.38	0.43	**31 Yuc**	34.9	31.8	21.4	15.6	4	4	3	2
0.36	0.39	0.45	0.46	**16 Mich**	34.6	28.9	21.8	15.8	4	3	3	2
0.21	0.13	0.08	0.10	**11 Gto**	33.3	27.0	19.3	13.9	4	3	2	2
0.16	-0.19	-0.11	-0.14	**22 Qtro**	32.8	24.8	18.0	12.6	4	3	2	2
0.01	0.00	-0.11	-0.02	**10 Dgo**	31.5	26.1	18.0	13.3	4	3	2	2
-0.04	-0.23	-0.19	-0.14	**29 Tlax**	31.1	24.4	17.5	12.6	4	3	2	2
-0.13	0.05	0.06	0.19	**18 Nay**	30.2	26.4	19.2	14.4	4	3	2	2
-0.14	-0.21	-0.10	-0.15	**25 Sin**	30.1	24.6	18.1	12.6	4	3	2	2
-0.19	-0.22	-0.36	-0.33	**23 QR**	29.7	24.5	16.3	11.6	3	3	2	2
-0.46	-0.54	-0.36	-0.44	**17 Mor**	27.3	22.2	16.4	11.1	3	3	2	2
-0.60	-0.73	-0.60	-0.62	**15 Mex**	26.0	20.8	14.7	10.1	3	3	2	2
-0.61	-0.57	-0.69	-0.69	**28 Tam**	26.0	22.0	14.1	9.8	3	3	2	1
-0.76	-0.70	-0.69	-0.73	**06 Col**	24.7	21.0	14.1	9.6	3	3	2	1
-0.77	-0.59	-0.76	-0.77	**14 Jal**	24.6	21.8	13.6	9.4	3	3	2	1
-0.86	-0.84	-0.76	-0.75	**26 Son**	23.7	20.1	13.6	9.5	3	3	2	1
-0.87	-0.76	-0.78	-0.68	**08 Chih**	23.7	20.6	13.5	9.8	3	3	2	1
-0.89	-1.04	-0.97	-0.96	**01 Ags**	23.5	18.6	12.2	8.4	3	2	2	1
-0.97	-0.82	-0.80	-0.71	**03 BCS**	22.8	20.1	13.3	9.7	3	3	2	1
-1.05	-1.16	-1.20	-1.14	**05 Coa**	22.0	17.7	10.6	7.4	3	2	2	1
-1.35	-1.25	-1.27	-1.25	**02 BC**	19.4	17.0	10.1	6.8	2	2	2	1
-1.38	-1.47	-1.39	-1.33	**19 NL**	19.1	15.4	9.3	6.5	2	2	1	1
-1.69	-1.71	-1.53	-1.50	**09 DF**	16.3	13.7	8.4	5.5	2	2	1	1

* Las observaciones están ordenadas por esta columna, en forma descendente.

Niveles de rezago: <10=1; [10,20)=2; [20,30)=3; [30,40)=4; [40,50)=5; >50=6.

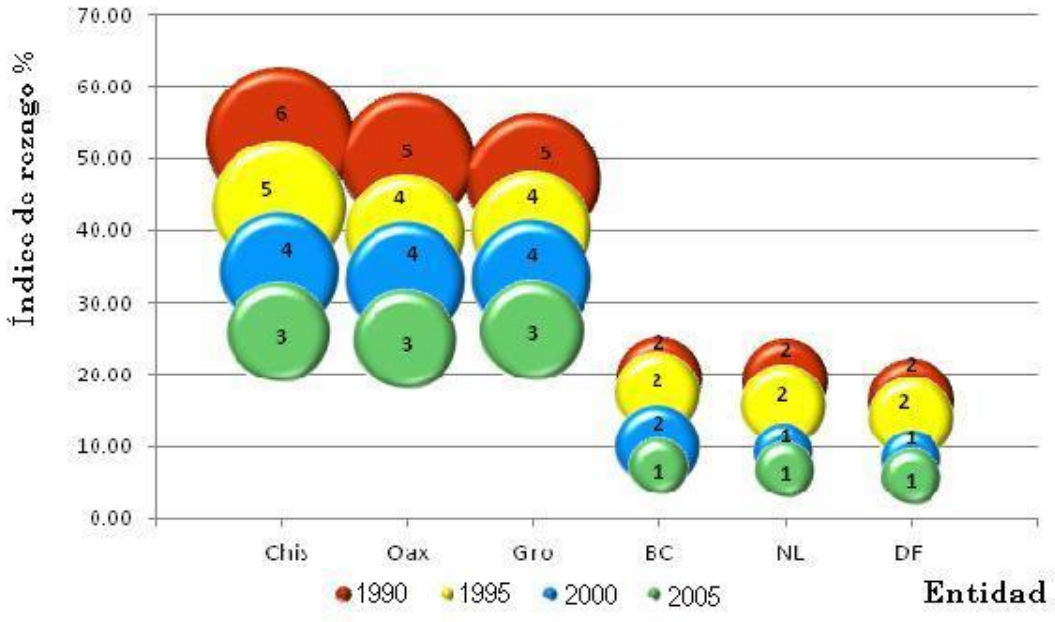
En la Gráfica 1, además de ilustrar la disminución gradual del rezago en el tiempo, se observa que en general la brecha entre las entidades federativas es cada vez más corta.

Gráfica 1. Índices de rezago social por entidad feredativa, 1990 - 2005
(ordenadas conforme a I para 1990)



Mientras tanto, en la Gráfica 2 se destaca que las tres entidades con mayor rezago social son: Chiapas, Oaxaca y Guerrero; en el otro extremo se ubican el Distrito Federal, Nuevo León y Baja California.

Gráfica 2. Índices y niveles de rezago social para entidades extremas 1990 - 2005 (ordenadas conforme a I para 1990)



Una nota

Cuando se utiliza la matriz de varianzas y covarianzas, los coeficientes que se tienen que normalizar son: c_i/S, donde c_i denota la covarianza y se define de manera similar como en el caso de r_i, sólo que ahora el promedio se obtiene sobre p; S es la suma de los c_i. De esta forma,

$$I = \frac{c_1x_1 + c_2x_2 + \cdots + c_px_p}{c_1 + c_2 + \cdots + c_p}$$

$$c_i = \sqrt{\frac{c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2}{p}}\ ,\ para\ i = 1, 2,\ \ldots,\ p$$

6. Conclusiones

El índice refleja la situación que prevalece en las variables originales, está dado en las mismas unidades de medición, permite ordenar las unidades de observación y es comparable en el tiempo. La varianza explicada por el nuevo índice es muy similar a la obtenida por la primera componente principal, con la ventaja adicional de que es muy fácil de calcular y de interpretar.

Puede aplicarse a cualquier conjunto de variables numéricas, correlacionadas y medidas en la misma dirección, es decir, todas deberán ser de bienestar, o bien, todas de rezago. Por tanto, el índice es de uso generalizado en cualquier fenómeno con estas características de medición. Para el caso de variables con menor nivel de medición, pueden hacerse transformaciones en términos porcentuales, por ejemplo.

El nuevo índice sintético es con fines exploratorios y descriptivos, se aplica a datos poblacionales y es de particular importancia para el diseño de programas gubernamentales y la focalización de recursos, en el campo de la estadística oficial. La aplicación del índice con datos muestrales y la consideración de dependencia temporal para realizar inferencia estadística quedan fuera del alcance de este trabajo; serán motivo de discusiones académicas y de estudios posteriores con mayor grado de especificidad.

Referencias

CONAPO. "Índices de Marginación, 1990, 1995, 2000 y 2005". Consejo Nacional de Población.

Flury, B.N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79, 892-898.

Flury, B.N. (1987). Two generalizations of the common principal component model. *Biometrika*, 74, 59-69.

Forni, Mario, Hallin, Marc, Lippi, Marco and Lucrezia Reichlin, (2000). The generalized dynamic factor model : identification and estimation. *The Review of Economics and Statistics*, November.

Jackson, J. Edward. 1991. "A User´s Guide To Principal Components". John Wiley & Sons, Inc. New York.

Johnson Richard A; Dean Wichern. 1982. "Applied Multivariate Statistical Analysis". Prentice-Hall, Inc, New York.

Lavit, CH. (1988): Analyse Conjointe de Tableaux Quantitatifs. Masson, París.

Lavit C y C Roux (1982): Manual del Método Statis. Ponencia en ISUP, París.

Peña, Daniel. 2003. "Análisis de Datos Multivariantes", 1 Ed. McGraw-Hill Interamericana de España.

Vences, José. 1999. "Estadística Multivariada, Análisis de Factores". Instituto de Educación de Aguascalientes.

Watson, Mark W. and Robert F. Engle (1983). Alternative algorithms for the estimation of dynamic factors, MIMIC, and varying coefficient regression models. *Journal of Econometrics* 23, pp.385-400.

Memoria FONE XXIV (2010), Sección I, 103–107

Una prueba por remuestreo para la distribución gamma

José A. Villaseñor Alva[a], Elizabeth González Estrada
Colegio de Postgraduados

1. Introducción

Se dice que la variable aleatoria (v.a.) X tiene distribución gamma con parámetros α y β si su función de densidad está dada por:

$$f_X(x;\alpha,\beta) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \exp\left\{-x/\beta\right\},\ x > 0,$$

donde $\alpha > 0$ y $\beta > 0$ son parámetros de forma y escala. Notación: $X \sim Gamma(\alpha, \beta)$.

La distribución gamma de dos parámetros se ha usado para modelar mediciones no negativas que exhiben unimodalidad y asimetría. Este tipo de datos se obtienen comúnmente en estudios de geología, ecología, economía, meteorología, confiabilidad, medicina y genética (Wilding y Mudholkar, 2008).

En la práctica rara vez se investiga si este modelo es apropiado. Esto se puede deber a que hay pocas pruebas para la hipótesis compuesta de que una muestra aleatoria tiene distribución gamma y a que las pruebas que existen no son fáciles de usar (Marchetti et al., 2002).

Algunas pruebas conocidas para la distribución gamma son:

1. Stephens (1986) sugiere usar la estadística AD de Anderson-Darling para probar la hipótesis de que una muestra aleatoria proviene de una distribución gamma en el caso en que ambos parámetros son desconocidos. Stephens calculó la distribución nula de la estadística AD para diferentes valores del parámetro de forma α y generó tablas para diferentes tamaños de prueba.

[a] jvillasr@colpos.mx

2. H y B de Marchetti et al. (2002). Estas pruebas tratan de explotar la propiedad de independencia entre sumas y razones de variables aleatorias independientes con distribución gamma.

3. Las pruebas condicionales exactas A^2 de Anderson-Darling y W^2 de Cramér-von-Mises sugeridas por Lockhart *et al.* (2007).

En este trabajo se presenta una nueva prueba por remuestreo paramétrico para probar la hipótesis de que una muestra aleatoria sigue una distribución gamma con parámetros desconocidos. También se presentan los resultados de un estudio de potencia por simulación de Monte Carlo para comparar la potencia de la nueva prueba por remuestreo con otras pruebas conocidas usando algunas distribuciones alternativas y tamaños de muestras $n = 30$ y 50.

2. Una prueba de bondad de ajuste por remuestreo

Sea $F(\cdot;\alpha)$ la función de distribución de una v.a. con distribución $Gamma(\alpha, 1)$, la cual es conocida como la distribución gamma estándar, y sea $F^{-1}(\cdot;\alpha)$ la función inversa de $F(\cdot;\alpha)$.

Cuando $X \sim Gamma(\alpha, \beta)$, la función de distribución de X se puede expresar como $F_X(x;\alpha,\beta) = F\left(\frac{x}{\beta};\alpha\right)$. De aquí que:

$$F^{-1}\left(F_X(x;\alpha,\beta);\alpha\right) = \frac{x}{\beta}. \tag{1}$$

Sea $X_1, X_2, ..., X_n$ una m.a. de tamaño n de la distribución $Gamma(\alpha, \beta)$. La función de distribución empírica de $X_1, X_2, ..., X_n$, F_n, se define como $F_n(x) = \frac{\#\text{ de } X_i's<x}{n}$.

Note que un estimador de momentos de α es $\widetilde{\alpha} = \bar{X}^2/s^2$. Por lo tanto, un estimador del lado izquierdo de (1) es $Y = F^{-1}\left(F_n(X), \widetilde{\alpha}\right)$.

Para probar la hipótesis nula H_0 : la m.a. $X_1, X_2, ..., X_n$ tiene distribución $Gamma(\alpha, \beta)$ con α y β desconocidas, se sugiere usar el coeficiente de correlación muestral de las $X_i's$ y $Y_i's$ como la estadística de prueba:

$$R = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \bar{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}},$$

donde $Y_i = F^{-1}\left(F_n(X_i), \widetilde{\alpha}\right), i = 1, 2, ..., n$.

Regla de decisión

Si H_0 es verdadera, por (1) se espera que R esté cerca de 1. Entonces se propone la prueba: rechazar H_0 con un tamaño de prueba γ si $R < k_\gamma$, donde el valor crítico k_γ es tal que $\gamma = P(R < k_\gamma | H_0)$.

2.1. Distribución nula de la estadística de prueba R

La distribución nula de R depende del valor del parámetro α. Entonces se sugiere usar remuestreo paramétrico para obtener el valor crítico k_γ como sigue:

1. Se calcula $\widetilde{\alpha}$ con base en la muestra aleatoria $X_1, X_2, ..., X_n$.
2. Se genera una muestra pseudo aleatoria $X_1^*, X_2^*, ..., X_n^*$ de la distribución $Gamma(\widetilde{\alpha}, 1)$.
3. Se calcula la estadística R con base en $X_1^*, X_2^*, ..., X_n^*$.
4. Se repiten N veces los pasos 2 y 3.
5. Se ordenan ascendentemente los valores calculados de la estadística R, y se denotan como $r_{(1)}, r_{(2)}, ..., r_{(N)}$.
6. La constante crítica k_γ del $\gamma 100\,\%$ se puede aproximar con $r_{(\lfloor \gamma N \rfloor)}$, donde $\lfloor \gamma N \rfloor$ denota el máximo entero que no excede a γN para γ dada.

3. Estudio de potencia

Se hizo simulación de Monte Carlo para estimar la potencia de las pruebas R, AD de Anderson-Darling, basada en estimadores de momentos y en remuestreo paramétrico, H y B de Marchetti et al. (2002). Para hacer los cálculos se hizo un programa en el código R. En cada caso se hicieron 10,000 repeticiones y se hizo remuestreo con $N = 999$.

Para estimar el parámetro de forma α se usó el estimador de momentos porque en un estudio previo de simulación se encontró que al calcular el estimador de máxima verosimilitud (MV) usando métodos numéricos frecuentemente se obtienen estimaciones negativas de α cuando el valor verdadero de α se aproxima a cero (use p.ej. la función $gammafit$ del paquete

R *mhsmm*). Al simular números pseudo aleatorios de algunas distribuciones alternativas frecuentemente se observa que el método numérico para obtener el estimador de MV para α no converge y también produce estimaciones negativas.

Se estimó el tamaño de las pruebas R y AD usando valores del parámetro de forma $\alpha = 0.1, 0.5, 1$ y un tamaño de prueba $\gamma = 0.05$. Los resultados se muestran en la Tabla 1. Se observa que la prueba de Anderson-Darling no preserva el tamaño de prueba cuando el valor verdadero de α es menor o igual que 0.5. Resultados similares fueron presentados en la Tabla 1 de Lockhart *et al.* (2007) para el caso $\alpha = 0.3$. Aparentemente esto se debe a que la distribución nula de la estadística AD aproximada por remuestreo paramétrico no es suficientemente buena para producir una prueba de tamaño γ prefijado.

	$n =$ 30		$n =$ 50	
α	R	AD	R	AD
0.1	0.058	0.21	0.06	0.17
0.5	0.052	0.09	0.051	0.09
1	0.054	0.06	0.055	0.05

Tabla 1: Tamaño estimado de las pruebas R y AD usando $\gamma = 0.05$.

	$n =$ 30				$n =$ 50			
Alternativa	R	AD	H	B	R	AD	H	B
Beta(.5,.5)	0.92	0.99	0.65	0.42	0.99	1.00	0.65	0.42
Uniforme	0.64	0.84	0.39	0.17	0.64	0.98	0.68	0.5
F(1,1)	0.16	0.00	0.29	0.38	0.16	0.09	0.49	0.58
F(30,6)	0.14	0.40	0.2	0.29	0.29	0.63	0.35	0.44
Lnorm	0.11	0.18	0.14	0.21	0.16	0.40	0.23	0.31
Pareto(5)	0.79	0.95	0.78	0.87	0.98	1.00	0.97	0.98
Weibull(.3)	0.11	0.02	0.09	0.14	0.11	0.01	0.13	0.19

Tabla 2: Potencia de cuatro pruebas para la distribución gamma, $\gamma = 0.05$.

4. Conclusiones

De la Tabla 2 se observa que ninguna de las cuatro pruebas es uniformemente más potente. De la Tabla 1 se concluye que para valores del parámetro α menores o iguales que 0.5, la prueba de Anderson-Darling basada en remuestreo paramétrico y en estimadores de momentos no preserva el tamaño fijado debido a la mala aproximación de la distribución nula de la estadística AD dada por el remuestreo.

Referencias

Lockhart, R., O'Reilly, F.; Stephens, M. (2007). Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika*, 97, 4, 992-998.

Marchetti, C.E., Mudholkar, G.S., Wilding, G.E., 2002. Testing goodness-of-fit of the gamma models. In Recent Advances in Statistical Methods (ed. Chaubey, Y.P.). London: World Scientific Publishing Company, Inc.

Stephens, M. (1986). Tests based on EDF statistics. In Goodness-of-Fit Techniques (Eds. D'Agostino and M. A. Stephens), New York: Marcel Dekker.

Wilding, G.E., Mudholkar, G.S. 2008. A gamma goodness-of-fit test based on characteristic independence of the mean and coefficient of variation. *Journal of Statistical Planning and Inference*, 138, 3813 º 3821.

Memoria FONE XXIV (2010), Sección I, 109–115

Log-linear models of categorized variables under distributional assumptions

Alexander von Eye[a]
Michigan State University, University of Vienna

Julian von Eye
Michigan State University

Patrick Mair
WU Vienna University of Economics and Business

1. Introduction

Standard applications of log-linear modeling use categorical variables, typically at the nominal or ordinal scale levels. In many applications, the continuous or ordinal variables are categorized. This is done with the goals of, for instance, (1) reducing the number of categories and thus the size of a cross-classification, (2) comparing individuals with above versus below cutoff scores, or (3) simplifying analysis. After categorization, variables are used as if they were nominal-level, and distributional assumptions are either not made or not taken into account when the data are analyzed.

In this contribution, we pursue two goals. First, we suggest that distributional assumptions be made that reflect data characteristics after categorization. Second, we propose taking these assumptions into account when analyzing categorized data using log-linear models.

2. Log-linear Models

Log-linear models are members of the family of generalized linear models (GLM; McCullagh+Nelder:1989). The GLM can be described by

$$E(Y) = \mu = g^{-1}(X\beta),$$

[a] voneye@msu.edu

where $E(Y)$ is the expectancy of the dependent measure Y, $X\beta$ is the linear predictor, and g is the link function. The dependent measure Y is assumed to follow a distribution from the exponential family. Examples of such functions include the normal, the binomial, and the Poisson distributions. The link function describes the relationship between the linear predictor and the expectancy of the dependent measure. When Y is assumed to follow the normal distribution, the link function typically employed is the identity function. Methods of analysis, in this case, include standard linear regression and ANOVA. Other link functions include the inverse, the squared inverse, the logarithmic, and the logit functions. In the present context, that is, when frequency tables are analyzed, the logarithmic function (base e) is the one of choice. We set $X\beta = \ln(\mu)$, and the model function becomes

$$\mu = \exp(X\lambda)$$

where λ is the vector of model parameters. In standard applications of log-linear modeling, one assumes a GLM with a Poisson random component. In the present context, we also consider the Normal. Parameter interpretation of log-linear models follows (see Mair+vonEye:2007)

$$\lambda = (X'X)^{-1}X' \ln(\mu)$$

The design matrix X contains, in standard applications of log-linear models,

1. main effects of the variables that span the cross-classification under study,
2. interactions,
3. covariates, and
4. special contrasts.

3. Categorizing variables

Categorizing variables involves defining intervals for a variable that is originally continuous, and counting the number of cases that fall inside each interval. Naturally, defining such intervals implies a reduction in the number of possible scores of a scale. In most cases, categorization comes, therefore, with a loss of information. The most popular cutoff points for categorization are the median and the clinical cutoff which is usually located at the 80th

percentile of the distribution. In each of these cases, the number of intervals (categories) after categorization is 2. In other words, the most popular categorization procedure involves dichotomization.

Categorization has been an issue of contention in biostatistics and the social sciences, and researchers carry strong opinions concerning the virtues of categorization (or the lack of it). Critiques of categorization emphasize the loss of information (MacCallum+Zhang:2002) or place statements on their website in which they recommend to "Avoid categorizing continuous variables and predicted values at all costs" (Harrell:2010).

Proponents of categorization note that, in real life, many decisions are binary, such as the decision as to whether a candidate is hired or not, or whether a patient is transferred to a psychotherapist or not. It is argued, then, that statistical analysis should reflect the binary nature of such decisions.

In this contribution, we focus on the distributional characteristics of categorized variables. We note that the loss of information that comes with categorization occurs in two levels, at least. At the first, categorization results in a reduction in the variability of scores. At the second, when categorized variables are analyzed using standard log-linear modeling methods, loss of information occurs when distributional characteristics of the categorized variables are not taken into account. We propose methods that allow the researcher to make information about the distributional characteristics of categorized variables part of a model.

4. Considering distributional information

Before categorization, many variables are supposed to follow specific distributions. In particular in the biological, medical, and social sciences, many variables are considered normally distributed. After categorization, the underlying distribution is unchanged. The categorized variables may still reflect these distributions. For example, when a normally distributed variable is categorized into four categories with two categories covering the scores between the mean and $\pm\,\sigma$. and the other two covering the more extreme scores, in both directions, one would expect the middle two categories to contain 34.1 % of the cases each, and the outer two categories 15.9 % each. This can be applied accordingly to multinormal distributions. When, in log-linear modeling, this information is not taken into account, the resulting model can become unnecessarily complex.

We consider three cases to take this information into account. To introduce these cases, we reformulate the log-linear model as follows:

$$\ln(\mu) = A\lambda + \varepsilon$$

Where ε is the residual, and A is of the form $A = [X|B]$. In this form, X is the usual design matrix, and B contains covariates that reflect distributional information. B is appended to X (appending variables). The three cases differ in the size of B and, therefore, the interpretability and interpretation of the parameters that are estimated for the columns of B.

Case I: B carries one parameter per cell of the cross-tabulation

In Case I, X is of size $[t, m]$, where t is the number of cells in the cross-tabulation under study, that is, the number of rows in X, and m is the number of parameters estimated for the log-linear model of this cross-tabulation. In most models, $m \leq t$. When $m = t$, the model is saturated.

In Case I, B contains one vector for each cell in the cross-classification. Therefore, Matrix B is of size $[t, t]$, and Matrix A is of size $[t, m + t]$. Because m is always $m \geq 1$, A will, in Case I, always have more columns than rows, and the model will always be overidentified. A will assume the form

$$A = \left[\begin{array}{ccc|ccc} x_{11} & \dots & x_{1m} & b_{11} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tm} & 0 & \dots & b_{tt} \end{array}\right].$$

Because of the linear dependencies in A, in Case I, $(A'A)^{-1}$ will not exist, and the effects of the distributional assumptions cannot be estimated independently for each cell.

Case II: B carries less than one parameter per cell of the cross-tabulation

In Case II, X is of size $[t, m]$ again. However, now, B contains less than one vector for each cell in the cross-classification, that is, B is of size $[t, t - k]$ with $0 < k < t$. Therefore, Matrix A is of size $[t, m + t - k]$. In other words, the cells of the cross-tabulation under study are grouped, and one parameter per group is estimated. The number of groups is $t - k$. Because it is always the case that $m \geq 1$, A will, in Case II, be just-identified if $m + t - k = t$. in this case, $(A'A)^{-1}$ will exist and parameters can be estimated. In Case II, A will assume the

form

$$A = \left[\begin{array}{ccc|ccc} x_{11} & \dots & x_{1m} & b_{11} & \dots & b_{1,(k<t)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tm} & b_{t1} & \dots & b_{t,(k<t)} \end{array}\right].$$

The main difference between cases I and II is that, in Case II, the cells are grouped and parameters are estimated for groups of cells. Therefore, whereas B is diagonal in Case I, in Case II, B is rectangular and can contain more than one non-zero element per column. When $m + t - k \leq t$, all parameters of the model will be estimable, in most cases. Note, however, that the estimates will rarely be independent because the columns of A will rarely be orthogonal. In other words, Case II represents the situation in which log-linear model parameters are set equal.

Case III: B is of size $[\mathbf{t}, \mathbf{1}]$

As in Cases I and II, X is of size $[t, m]$. However, now, B is of size $[t, 1]$. Therefore, Matrix A is of size $[t, m + 1]$. In other words, only one parameter is estimated to represent the distributional assumptions. Under this condition, the model will be identified and $(A'A)^{-1}$ will exist when $m + 1 \leq t$. In Case III, A will assume the form

$$A = \left[\begin{array}{ccc|c} x_{11} & \dots & x_{1m} & b_{11} \\ \vdots & \ddots & \vdots & \vdots \\ x_{t1} & \dots & x_{tm} & b_{t1} \end{array}\right].$$

The main difference between cases I and II on one side and Case III on the other is that, in Case III, the effects of the distributional assumptions are expressed in one covariate vector, and only one parameter will be estimated. In Case III, all parameters from Case I are set equal.

5. Data example

For the following application example, we use data from a study on the development of aggression (Finkelstein+vonEye:1994). In 1983, 114 adolescents (69 girls) indicated levels of Physical Aggression Against Peers (PAAP), Verbal Aggression Against Adults (VAAA), and Aggressive Impulses (AI). None of the variables showed significant kurtosis ($p = 0.47$). For the present illustration, each variable was split into three categories (33rd percentiles). Six models are considered:

Model	$LR-X^2$	df	$\Delta LR-X^2$	Δdf	$P(\Delta LR-X^2)$
1. Null model	151.5	26	-	-	-
2. DI only	59.51	21	91.99 (Δ to Model 1)	5	$< .01$
3. Main effect	57.76	20	93.74 (Δ to Model 1)	6	$< .01$
4. Main effect + DI	47.75	19	10.01 (Δ to Model 3)	1	.002
5. All 2-way	12.83	8	44.93 (Δ to Model 3)	12	$< .01$
6. All 2-way + DI	9.19	7	3.64 (Δ to Model 5)	1	.056

Tabla 1: Modeling the structure of adolescent aggression with and without aistributional information (DI; multinormal distribution)

1. Null model ($m = 1$; no matrix B).

2. Distributional information only (normal distribution) (no matrix X); this model involves a first application of Case III.

3. Log-linear main effect model ($m = 6$; no matrix B).

4. Log-linear main effect model plus normal distribution information (probability of pattern estimated under assumption of multivariate normality; see Sommerville:1998, von-Eye+Mair:2008; this model involves to a second application of Case III ($m = 6$; $k = 1$).

5. Log-linear model with all three pair-wise interactions ($m = 17$; no matrix B).

6. Log-linear model with all three pair-wise interactions plus normal distribution information ($m = 17$; $k = 1$); this model involves a third application of Case III.

The results in Table 1 show that, in each case, including information concerning the multivariate distribution of the data improves the model. Only when all two-way interactions are part of the model already, the improvement is non-significant. We conclude that these interactions correlate with some of the distributional characteristics of the data. In other words, these interactions may not only reflect characteristics of the association structure but also characteristics of the multivariate distribution of the three variables that span the table.

6. Discussion

Categorization reduces the number of scores (categories) of variables which may result in loss of information. Additional information is carried by the uni- and multivariate sampling distributions of the variables in a study. If the observed variables follow a sampling distribution before categorization, they will still do so after categorization. This information can be used for modeling. If this information can be expressed in the form of covariates that are included in models, the models may end up being less complex and reflect data characteristics more realistically.

Referencias

J. W. Finkelstein, A. von Eye and M. A. Preece(1994). "The relationship between aggressive behavior and puberty in normal adolescents: A longitudinal study", *Journal of Adolescent Health*,15,319–326.

P. Mair and A. von Eye(2007) "Application scenarios for nonstandard log-linear models",*Psychological Methods*,12,139–156

F. Harrell(2010)http://biostat.mc.vanderbilt.edu/wiki/Main/FrankHarrell

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

P. McCullagh and J. A. Nelder (1989).*Generalized linear models*,Chapman & Hall,London, UK.

P. Sommerville(1998). "A Fortran 90 program for evaluation of multivariate normal and multivariate t integrals over convex regions",*Journal of Statistical Software*,3,4,1–10.

A. von Eye and P. Mair(2008). Evaluating Cluster Solutions with Reference to Data Generation Processes — A Simulation Study, *Memorias del XXII Foro Nacional de Estadistica*,123–131,

Sección II

Aplicaciones

Memoria FONE XXIV (2010), Sección II, 119–124

Efecto de marcas de cemento en la resistencia del concreto

Alfredo Cuevas Sandoval[a]
Unidad Académica de Ingeniería,Universidad Autónoma de Guerrero

Flaviano Godínez Jaimes[b]
Unidad Académica de Matemáticas, Universidad Autónoma de Guerrero

Esteban Rogelio Guinto Herrera[c]
Unidad Académica de Ingeniería,Universidad Autónoma de Guerrero

Roberto Arroyo Matus[d]
Unidad Académica de Ingeniería,Universidad Autónoma de Guerrero

1. Introducción

El cemento es uno de los componentes de mayor costo en una mezcla de concreto hidráulico. Desde su invención hace aproximadamente doscientos años ha habido varios cambios en su composición y aplicaciones. Lo anterior originó también el surgimiento de varias empresas productoras, tipos de cementos y marcas, hasta llegar a una formulación específica a las necesidades, según el proyecto a construir. Para este estudio se consideran cuatro marcas de cemento (las de mayor consumo en la región centro del estado de Guerrero), considerando el mismo tipo, clase resistente y controlando otros factores en laboratorio. Se planeó un experimento para determinar directamente en ensayes de concreto hidráulico cuál tiene un mejor desempeño en la resistencia a compresión y su impacto en el costo en las mezclas (Montgomery, 2005; NMX-C-159, 1988; NMX-C-083, 2007).

[a] acuevas36@hotmail.com
[b] fgodinezj@gmail.com
[c] rguinto2002@yahoo.com.mx
[d] arroyomathus@hotmail.com

El cemento portland es de tipo hidráulico y se obtiene al pulverizar clínkers y sulfato de calcio. Al agregar agua al cemento, sólo o con otros materiales, tiene la propiedad de fraguar y endurecer incluso bajo el agua y ya endurecido conserva su resistencia y estabilidad (NMX-C-414, 2004).

El concreto (hormigón) es mezcla de dos componentes: agregados (arena y grava) y pasta (agua y cemento). La pasta constituye aproximadamente del 25 % al 40 % del volumen total del concreto. El volumen absoluto del cemento varía normalmente entre 7 % y 15 % y el del agua entre 14 % y 21 %. Los agregados constituyen aproximadamente del 60 % al 75 % del volumen total del concreto (Mehta *et al.*, 1998; Kosmatka *et al.*, 2005).

El porcentaje de cemento en una mezcla de concreto es el menor, pero su porcentaje en el costo es el mayor (58 % cemento, 21 % grava, 20 % arena y1 % agua), por esto es importante seleccionar el tipo y la marca de cemento para que alcance la resistencia deseada.

2. Estudio experimental

Este trabajo tiene dos objetivos, por una lado determinar el efecto de los cementos en la resistencia del concreto y otro es verificar si los materiales utilizados en la región centro del Estado de Guerrero cumplen con las sugerencias a nivel nacional. En este documento sólo se presentan resultados del primer objetivo.

Los factores controlados en el experimento son el banco de arena, agua, grava y resistencia de diseño. El banco utilizado fue el "papagayo", el cual previamente había sido verificado que es el mejor en cuanto a sus propiedades físicas y mecánicas (Cuevas *et al.*, 2009). Para el mezclado del concreto se utilizó agua potable. La grava fue del banco "Xocomulco"tipo triturada con tamaño de 3/4". Se elaboraron mezclas de concreto para una resistencia a compresión de diseño con un f'c(f prima c)=250Kg/cm^2 (25 Mpa). Los materiales para la elaboración de concreto cumplen las especificaciones de las normas mexicanas y los reglamentos de construcción vigentes, así como, los procedimientos de ensaye se efectuaron siguiendo los lineamientos indicados en las normas y reglamentos mencionados.

Los factores estudiados son *marca* de cemento (M1, M2, M3, M4) y *edad* de ensaye (3, 7, 14, 21 y 28 días). La variable respuesta es resistencia a compresión del concreto medida de acuerdo con las normas mexicanas (NMX-C-083, 155, 159). El análisis se realizó utilizando

los programas estadísticos SAS, SPSS y JMP.

3. Resultados

Se ajustó un modelo factorial dos factores en completamente el azar.

$$y_{ijk} = \mu + \tau_i + \alpha_i + (\tau\alpha)_{ij} + \varepsilon_{ijk};$$

$$i = 1, 2, 3, 4.; \quad j = 1, 2, ..., 5; \quad k = 1, 2, ..., 9.$$

En el modelo ajustado los dos factores, *edad* y *marca* de cemento son significativos a un nivel de significancia de $\alpha = .05$, pero no así la interacción *marca*edad*. Esto es, la media de la resistencia del concreto de las *marcas* de cemento utilizadas tienen la misma tendencia en cada *edad* de ensaye (Tabla 1).

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	significación
Modelo corregido	358718.620	19	18879.927	11.545	.000
Intersección	10678363.3	1	10678363	6529.785	.000
Edad	308360.829	4	77090.207	47.140	.000
Cemento	30408.420	3	101360140	6.198	.001
Edad*cemento	15555.291	12	1296.274	.793	.658
Error	261653.050	160	1635.332		
Total	11324660.0	180			
Total corregida	620371.670	179			

Tabla 1: Resultados del ajuste del modelo con interacción.

El modelo sin interacción presentó resultados similares para los factores principales, r^2=0.55 y los residuos son independientes y tienen distribución normal (Kolmogorov-Smirnov, valor p=0.125, Cramer-von Mises valor p=0.132). Para el factor *marca* se observan 2 grupos significativamente diferentes (Tabla 2), donde las resistencias mayores de los concretos fabricados son con las marcas de cemento M4 y M2.

Para el factor *edad* se observan 3 grupos significativamente diferentes, donde la mayor resistencia de los concretos fabricados es a la *edad* de 14, 21 y 28 días. Las resistencias más bajas corresponden a los concretos a la *edad* de 3 días (Tabla 3). Este es un comportamiento

		Subconjunto	
Marca de cemento	N	1	2
M3	50	229.7632	
M1	50	238.4060	
M4	40	246.7043	246.7043
M2	40		265.4595
Significación		.202	.131

Tabla 2: Clasificación de la resistencia por *marca* de cemento.

natural, debido a que la resistencia a compresión en el concreto se va incrementando al avanzar la *edad*, hasta llegar a la *edad* de 28 días, donde un concreto de tipo normal alcanza el 100 % de resistencia a compresión.

		Subconjunto		
Edad	N	1	2	3
3 Días	36	167.6578		
7 Días	36		233.2561	
14 Días	36			261.4875
21 Días	36			272.7758
28 Días	36			284.1286
Significación		1.000	1.000	.127

Tabla 3: Clasificación de la resistencia por *edad* de ensaye.

Las comparaciones múltiples para los tratamientos muestran que la resistencia a compresión mayor del concreto incluye combinaciones con todas las *marcas* y *edades* desde los 7 a los 28 días. Las combinaciones estadísticamente iguales y que no cumplen con el f'c de diseño son los grupos C, D, y E (Tabla 4).

4. Conclusiones

A los tres días ninguna *marca* de cemento satisfizo la resistencia a compresión de diseño f'c=250Kg/cm^2 (25 Mpa). En el factor *marca*, no todas las marcas de cemento garantizan alcanzar las resistencia a compresión de un concreto.

La resistencia a compresión de un concreto esta relacionada al factor *edad* y que a partir de los 14 días puede observarse que la resistencia a compresión se cumple con respecto al f'c de proyecto.

Level						Least Sq Mean
21 Días,M2	A					302.62
28 Días,M2	A					301.12
14 Días,M2	A					296.12
28 Días,M1	A	B				287.00
28 Días,M4	A	B				276.00
28 Días,M3	A	B				274.00
21 Días,M1	A	B				271.00
14 Días,M4	A	B				270.12
21 Días,M3	A	B				263.90
21 Días,M4	A	B				256.12
7 Días,M2	A	B				254.62
7 Días,M4	A	B				252.25
14 Días,M3	A	B	C			244.10
14 Días,M1	A	B	C			244.10
7 Días,M1		B	C	D		219.40
7 Días,M3		B	C	D	E	214.80
3 Días,M4			C	D	E	179.00
3 Días,M2				D	E	172.87
3 Días,M1				D	E	170.40
3 Días,M3					E	151.80

Tabla 4: Comparaciones múltiples de las combinaciones *marca*edad*.

Las combinaciones que cumplen y tienen mayor resistencia a compresión con respecto al f'c son las cuatro *marcas de cemento* estudiadas en combinación con la *edad* a partir de los 7 hasta los 28 días.

A los 28 días, que es la edad de garantía, se encontró que las 4 marcas cumplen con la resistencia de diseño, indicando que cualquier *marca* que se utilice en la elaboración de concreto va a producir promedios de resistencias a compresión iguales al tener esta edad.

Como el cemento representa el mayor porcentaje en el costo de un concreto se recomienda utilizar la marca de cemento más económica de las cuatro suministradas en la Región de estudio. Esto permitirá mejorar la calidad de las obras construidas y una reducción importante en los costos a los usuarios.

Referencias

Montgomery, D.C. 2005, *Diseño y análisis de experimentos*, 2ª Edición, Editorial Limusa Wiley, México.

Norma Mexicana, NMX-C-159-1997, Industria de la construcción-Concreto- *Elaboraión y curado en el laboratorio de especimenes de concreto*, México.

Norma Mexicana, NMX-C-083-2002, Industria de la construcción-Concreto-*Determinación de la resistencia a compresión de cilindros de concreto*, México.

Norma Mexicana, NMX-C-414-2004, Industria de la construcción-*Cementos Hidráulicos-Especificaciones y métodos de prueba*, México.

Mehta, K., Monteiro, P., 1998. *Concreto, estructura, propiedades y materiales*, Editorial Imcyc; México.

Kosmatka, S., Panarese, W. 2005. *Diseño y control de mezclas de concreto*, Manual de la Asociación del Cemento Portland, Editorial PCA, USA.

Norma Mexicana, NMX-C-155-2004 Industria de la construcción-Concreto- *Concreto Hidráulico Industrializado*, México.

Cuevas, A., Godínez, F., Santes, 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 la Estadística, Vol. 3, BUAP, México.

Memoria FONE XXIV (2010), Sección II, 125–129

Análisis de patrones espaciales de hongos ectomicorrízicos en el parque nacional Malintzi*

Linares Fleites, G. [a], Marín Castro, M.A., Ticante Roldán, J.A. y Silva Díaz, B

Benemérita Universidad de Puebla Instituto de Ciencias

1. Introducción

Los ecosistemas forestales de México han sido perturbados en diferente medida, por lo que es necesario conocer su funcionamiento para generar alternativas que impidan el avance de procesos de degradación y, por ende, los recursos del bosque puedan ser utilizados bajo un enfoque sustentable. Existen diferentes productos forestales no maderables que pueden ser aprovechados por el hombre, entre los que se encuentran los hongos ectomicorrízicos, que son de extraordinaria importancia para la recuperación y preservación de los bosques. Es de gran interés, económico y social, conocer la distribución de estos hongos en la parte poblana del Parque Nacional Malintzi. Este Parque se localiza en la zona centro-oriental de México formando parte de la Cordillera Neovolcánica y tiene una superficie total de 45 852.45 ha, de las cuales 14 433.81 ha corresponden al estado de Puebla. Para describir la distribución espacial de hongos ectomicorrízicos se ubicaron zonas de muestreo bajo el criterio de localizar áreas de mayor a menor perturbación. Al hacer el recorrido de las áreas seleccionadas, se marcaron los sitios por medio de coordenadas geográficas obtenidas por el sistema de geoposicionamiento geográfico satelital (GPS). La colecta de especímenes se realizó durante los meses de julio a noviembre, detectándose, entre otros, el género Bolletus sp.

*Benemérita Universidad de Puebla Instituto de Ciencias

[a]gladys.linares@icbuap.buap.mx

Perseguimos como objetivo estudiar la distribución espacial de este hongo ectomicorrízico y, en específico, probar la hipótesis de Aleatoriedad Espacial Completa, que es definida como un patrón tal que el número de puntos en cualquier región del plano sigue la distribución de Poisson con media y varianza λ, donde λ es el número medio de organismos por unidad de área ($lambda|A|$ en donde $|A|$ es el área de la región) y todos los puntos en la región se suponen independientes unos de los otros, Diggle(1983). Existen diferentes métodos para estudiar la Aleatoriedad Espacial Completa (en inglés, Complete Spatial Randomness (CSR)), según se tenga en cuenta o no la localización de los puntos. Si no se tiene en cuenta la localización se utilizan pruebas de bondad de ajuste para la distribución de Poisson y si se tiene en cuenta, puede utilizarse alguna medida de autocorrelación espacial, tal como la de Moran o la función $\hat{K}$ (KHAT) de Ripley. El gráfico $\hat{L}$ (LHAT) que se deriva de la función $\hat{K}$ (KHAT) es utilizado en este trabajo para describir la distribución espacial de los hongos del género Bolletus sp encontrados en el Parque Nacional Malintzi. En la siguiente sección se hace un breve recuento de estas técnicas y se muestran los resultados obtenidos. Finalmente, se dan algunas conclusiones.

2. Marco teórico

En la naturaleza, generalmente los organismos no se distribuyen al azar, sino que tienden a estar agrupados o estar espacialmente estructurados. La heterogeneidad espacial medio ambiental, de extraordinaria importancia, no siempre es detectada por las técnicas estadísticas clásicas, por lo que en los últimos años, se han desarrollado nuevas técnicas que permiten reflejar la estructura espacial de los fenómenos ecológicos, Diggle(1983).

En muchas ocasiones, dependiendo de la escala de estudio, los elementos pueden describirse aceptablemente mediante sus coordenadas espaciales (x, y), generándose un conjunto de datos que recibe el nombre de patrón espacial de puntos. La metodología habitual en el estudio de estas estructuras asume que el patrón espacial de puntos de una población, comunidad, etc., es una realización concreta de un proceso espacial de puntos subyacentes. Un proceso de puntos es un proceso estocástico que "genera" patrones de puntos aleatorios que comparten la misma estructura espacial, por ejemplo, patrones de Poisson (distribución completamente al azar), regulares o agrupados, de la Cruz(2006). Bajo la suposición de estacionaridad (el proceso es homogéneo o invariante a la traslación) e isotropía (el proceso es

invariante a la rotación), las características principales de un proceso de puntos pueden ser resumidas por su propiedad de primer orden (o intensidad, que es el número esperado de puntos por unidad de área en cualquier localidad) y por su propiedad de segundo orden, que describe las relaciones entre pares de puntos (por ejemplo, la probabilidad de encontrar un punto en las inmediaciones de otro). En el caso de patrones uniformes o regulares, la probabilidad de encontrar un punto en las inmediaciones de otro es menor de la que tendría un patrón aleatorio, mientras que en los patrones agrupados la probabilidad es mayor. El estimador más popular de las propiedades de segundo orden es la función K de Ripley, que las estima a todas las escalas. La función K se define como $K(r) = \lambda^{-1}$ E [número de puntos en un radio r alrededor de cualquier individuo] donde λ es la densidad de individuos (número de individuos por unidad de área) y E denota el valor esperado o media. La ventaja de la función K es que el valor teórico de K(r) se conoce para varios modelos útiles de procesos de puntos espaciales. Kaluzny et. al.(1997) Por ejemplo, para el proceso de Poisson se cumple que si:

- No hay dependencia espacial: $K(r) = \prod r2$.
- Hay agrupamiento: $K(r) > \prod r2$.
- Hay espaciado regular: $K(r) < \prod r2$.

Dixon(2002) expone los estimadores que se utilizan comúnmente y Venables y Ripley(2002) los desarrollan en lenguaje S. Estos estimadores están programados en el sistema de cómputo S-PLUS 2000 que elabora también gráficos que esclarecen la interpretación. Para la detección de CSR puede utilizarse el gráfico $\hat{L}$ (LHAT) de Ripley. El gráfico $\hat{L}$ (LHAT) representa al estimador de la función $L(r) = \frac{\sqrt{K(r)}}{\pi}$ y es una línea recta para un proceso aleatorio homogéneo de Poisson, lo que corresponde a la hipótesis de aleatoriedad espacial completa. Los puntos por encima de la recta indican la existencia de agrupamientos, mientras que los puntos por debajo indican espaciado regular.

2.1. Detección de CSR.

Contar con las coordenadas geográficas hizo posible estudiar la distribución espacial de dichos hongos. Estadísticos tradicionales para la bondad de ajuste, como la Chi cuadrado,

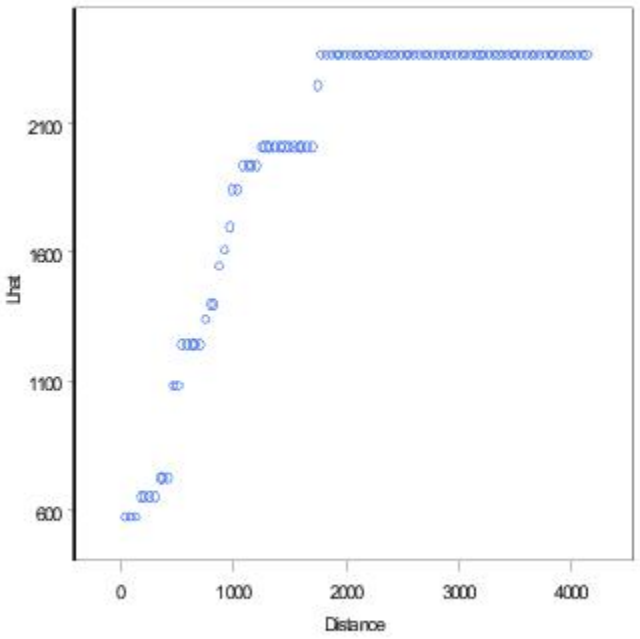


Figura 1: Grafico LHAT para el género *Bolletus sp* de hongos ectomicorrízicos en la parte poblana del Parque Nacional Malintzi, México

no pudieron detectar la heterogeneidad espacial, sin embargo, la medida de autocorrelación espacial de Moran detectó la existencia de agrupaciones de estos hongos. La correlación de Moran fue de 0.4792, con varianza de 0.0636 y error estándar de 0.25222. El estadístico Z fue igual a 2.065, por lo que la hipótesis nula de no correlación espacial fue rechazada con p-empírico (dos colas) de 0.0389. El gráfico $\hat{L}$ (LHAT) (Figura 1) mostró valores positivos, por encima de la recta diagonal del gráfico, lo que indica que el género Bolletus de hongos ectomicorrízicos coleccionados en la parte poblana del Parque Nacional Malintzi no se presenta de manera aleatoria, sino que se ha detectado la existencia de agrupaciones en estos hongos. Sería recomendable obtener bandas de confianza mediante simulación para poder hablar de significancia, lo que se realizará en trabajos futuros.

3. Conclusiones

Los resultados obtenidos muestran la utilidad de los gráficos $\hat{L}$ (LHAT) para esclarecer los patrones espaciales de los hongos ectomicorrízicos en la parte poblana del Parque Nacional Malintzi. En específico, mostró que el género Bolletus de hongos ectomicorrízicos coleccionados en la parte poblana del Parque Nacional Malintzi se presentan agrupados y no de forma aleatoria. Es conocido que dentro de problemática ambiental y el deterioro de los bosques, se encuentra la deforestación y la erosión de los suelos, que se originan, entre otros factores, por el cambio de uso de los suelos, la tala clandestina, incendios forestales, el sobre pastoreo y la falta de una cultura silvícola. Este último factor genera la pérdida irreversible de especies, entre ellas, los hongos ectomicorrízicos. Los árboles, en todo el mundo, aún los que no están siendo cortados, están muriendo por causas diversas, incluyendo enfermedades, contaminación y degradación del suelo. Las amenazas a los bosques no son simples amenazas a los árboles, hay que comprender que cuando desaparezcan los árboles también desaparece todo lo que depende de ellos, desde hongos y microorganismos hasta flora y fauna. El desarrollo de métodos para describir la configuración espacial de un conjunto de puntos y la posible relación con otros fenómenos espaciales es de extraordinaria importancia y es aún un reto para la Estadística como ciencia. En particular, el poder describir cómo está distribuido éste y otros hongos micorrízicos en el Parque permitirá generar una propuesta de reforestación sostenible en el sistema forestal de la Malintzi.

Referencias

De la Cruz Rot, M(2006). "Introducción al análisis de datos mapeados o algunas de las (muchas) cosas que puedo hacer si tengo coordenadas",*Ecosistemas*,3,1-21.

P.J. Diggle (1983),*The statistical analisys of spatial poitn patterns*,Academis Press.

S.P. Kaluzny, S.C. Vega, T.P. Cardoso and A.A. Shelly (1997),*s+Spatial Stats: User's Manual for Windows and Unix*,Seattle, USA, MathSoft, Inc.

P.M. Dixon(2002), "Rippley's K function",*Enciclopedia Of Envirometrics*,3,1796–1803.

W.N. Venables. and B.D. Ripley(2002),*Modern Applied with S*,New York, Springer-Verlag.

Memoria FONE XXIV (2010), Sección II, 131–136

Análisis de conglomerados en el estudio de siete razas de maíz

Emilio Padrón Corral[a], Armando Muñoz Urbina, José Luís de la Riva Canizales, Manuel Antonio Torres Gomar

Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Coahuila, Saltillo, Coahuila, México

Ignacio Méndez Ramírez[b]

Instituto de Investigación en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México

1. Introducción

La evaluación de recursos genéticos es un esfuerzo multidimensional que involucra diversos campos científicos como citogenética, evolución, agronomía y estadística. Una evaluación cuidadosa del germoplasma incrementa grandemente la utilidad de fuentes genéticas en el mejoramiento de este cultivo. El análisis de conglomerados es una técnica multivariada que permite agrupar los objetos en clases de tal manera que sean similares dentro de la misma clase, por lo tanto, en el análisis de conglomerados se pueden definir grupos verdaderos y puede ser útil para la reducción de datos (Manly, 1990). Para la realización del análisis de conglomerados es necesario tener un coeficiente de similitud que mida el parecido entre individuos. La medida de distancia más conocida es la Euclidiana, la cual es adecuada para datos cuantitativos. Crossa et al. (1994) efectuaron la clasificación racial de 80 colectas de maíz Tuxpeño utilizando el análisis de conglomerados.

Acosta et al. (1991) mencionan que el avance del mejoramiento de plantas depende de la disponibilidad de diversidad genética y de su conocimiento. Painting et al (1993) señalan que la variación genética permite el mejoramiento de especies y su expresión es necesaria para

[a] emiliopadron@uadec.edu.mx

[b] imendez@servidor.unam.mx

diferenciar las características deseables de las indeseables. La enorme variabilidad genética del maíz que se mantiene en su descendencia es una herencia que recibimos de las antiguas Civilizaciones Mesoamericanas. El reemplazo de los cultivares primitivos por variedades mejoradas de maíz y de mejor calidad nutritiva es indispensable para una mayor producción de alimentos, pero tal reemplazo no debe significar la pérdida irrecuperable de los cultivares primitivos y de sus parientes silvestres.

2. Materiales y Métodos

Los materiales utilizados en el presente trabajo provienen de una colecta de razas de maíz realizada por el Instituto Mexicano del Maíz y consta de los siguientes genotipos: Criollo Negrito de los Reyes (R1), Ixtlahuaca (R2), Criollo Rojo de San Mateo (R3), Cacahuacintle (R4), Criollo Rosado Pinto Violento (R5), Rosadito Violento de Cashi (R6), Amarillo Criollo de Ixtlahuaca (R7).

Se seleccionaron 10 mazorcas al azar de cada muestra y se tomaron los siguientes caracteres: Longitud de mazorca (LM), Diámetro de mazorca (DM), Número de hileras (NH), Ancho de grano (AG), Espesor de grano (EG), Longitud de grano (LG), Diámetro de olote (DO), Diámetro de raquis (DR). Se midió en centímetros la longitud y el diámetro de la parte media de la mazorca y el diámetro de olote. Las dimensiones de largo, ancho y grueso del grano se determinaron con el vernier: midiendo diez granos por unidad experimental. El análisis de conglomerados se realizo con el paquete computacional Statistica (2000), básicamente lo que el programa hace es una implementación del siguiente algoritmo

1. Examina la matriz de entrada para el par de objetos (i , j) que son más similares (o menos disimilares).

2. Une estos objetos en un nuevo grupo.

3. Usa la matriz para reflejar la supresión del par de objetos, i y j, que fueron unidos y la adición del nuevo objeto correspondiente al nuevo grupo.

4. Regresa al paso 1, si el tamaño de la nueva matriz es mayor de 2 x 2 de otro modo se detiene. Note que dos objetos son suprimidos y uno más es añadido en cada paso, así el algoritmo debe concluir.

Los coeficientes de disimilitud fueron obtenidos utilizando la ecuación de Distancia Euclidiana= $E_{ij} = (\sum_k (xki - xkj)^2)^{\frac{1}{2}}$

3. Resultados y Discusión

En Cuadro 1, se observa que se presentaron diferencias entre las razas para las características evaluadas, las razas 2 (Ixtlahuaca) y 7 (Amarillo Criollo de Ixtlahuaca) mostraron mayor longitud de mazorca, número de hileras y longitud de grano. Se observa también que la raza 4 (Cacahuacintle) fue la de mayor diámetro de mazorca, olote y raquis, presentando además el mayor ancho y espesor de grano, por lo tanto, fue la más divergente con respecto al resto de las razas evaluadas.

Cuadro 1. Promedios de las variables evaluadas en las siete razas de maíz.

Razas	LM (cm)	DM (cm)	NH	AG (mm)	EG (mm)	LG (mm)	DO (cm)	DR (cm)
1	13.6	4.52	17.4	7.122	4.04	13.93	2.39	0.89
2	16.44	4.88	15.77	8.06	3.93	15.63	2.56	1.19
3	13.0	4.26	13.5	7.32	4.03	13.3	2.15	1.10
4	14.64	5.25	13.43	10.66	5.16	14.07	3.21	1.34
5	11.42	4.5	14.0	7.57	4.1	13.37	2.36	0.84
6	10.05	3.9	13.8	7.02	4.21	12.22	1.91	0.74
7	17.67	4.79	15.33	7.78	4.20	14.58	2.32	1.21

En el Cuadro 2, se observa que entre las razas 3 y 5 se detecto el menor coeficiente de disimilitud (1.71) seguidos de las razas 2 y 7. La raza 4 registró los más altos valores de disimilitud con respecto a todo el grupo de genotipos con valores de disimilitud de 4.22 con la raza 3, hasta 6.50 con la raza 6. Las plantas que comparten un conjunto particular de caracteres han recibido un nombre que refleja alguna asociación especial. El nombre puede referirse a la presencia de un carácter o atributo fenotípico pronunciado tal como Cónico por la forma de la mazorca: Reventador por la habilidad de los granos para reventar. Otros pueden indicar el área donde un tipo de maíz llegó a ser predominante. En el grupo de materiales evaluados en este experimento se observó cierta asociación entre los nombres de

las razas y su cercanía, en este caso entre la raza 2 (Ixtlahuaca) y la raza 7 (Amarillo Criollo de Ixtlahuaca) se observó uno de los menores coeficientes de disimilitud con un valor de 1.74.

Cuadro 2. Coeficientes de disimilitud entre razas (Distancia Euclidiana).

Razas	1	2	3	4	5	6	7
1	0.00						
2	3.83	0.00					
3	4.02	4.85	0.00				
4	5.64	4.47	4.22	0.00			
5	4.10	5.82	1.71	4.81	0.00		
6	5.40	7.67	3.20	6.50	2.03	0.00	
7	4.67	1.74	5.23	4.82	6.51	8.22	0.00

En el Cuadro 3, se observan los pasos principales por medio de los cuales se van conformando diferentes grupos de razas según la etapa en que se quiera clasificar a los grupos que contengan la mayor información de la matriz de datos. En el paso 4 la raza 6 (Rosadito Violento de Cashi) se une al grupo de las razas 3 (Criollo Rojo de San Mateo) y 5 (Criollo Rosado Pinto Violento). En el paso 5 se forman los tres grupos principales, cuando la raza 1 (Criollo Negrito de los Reyes) se une al grupo de las razas 2 (Ixtlahuaca) y 7 (Amarillo Criollo de Ixtlahuaca). El proceso continua hasta que en el paso final se unen los grupos de razas (1 2 3 5 6 7) y (4) para conformar un sólo grupo (1 2 3 4 5 6 7) de siete individuos.

Cuadro 3. Valores de distancia entre grupos para la gráfica (fenograma), método de encadenamiento simple.

Paso No.	Distancia	Grupos
1	0.00	(1)(2)(3)(4)(5)(6)(7)
2	1.71	(1)(2)(4)(3 5)(6)(7)
3	1.74	(1)(4)(3 5)(6)(2 7)
4	2.03	(1)(4)(3 5 6)(2 7)
5	3.83	(4)(3 5 6)(1 2 7)
6	4.02	(4)(1 2 3 5 6 7)
7	4.22	(1 2 3 4 5 6 7)

La Figura 1, muestra el grado de asociación entre razas en base a los coeficientes de disimilitud obtenidos de las variables evaluadas. De acuerdo con una distancia de corte

del fenograma de 3.90, se puede observar claramente la formación de tres grupos de razas que corresponden a los tipos: Criollo negrito de los Reyes, Ixtlahuaca y Amarillo criollo de Ixtlahuaca (razas 1, 2 y 7 respectivamente); Criollo rojo de San Mateo, Criollo Rosado Pinto Violento y Rosadito Violento de Cashi (razas 3, 5 y 6 respectivamente); y por último la raza Cacahuacintle (4) que resulto la mas divergente del grupo.

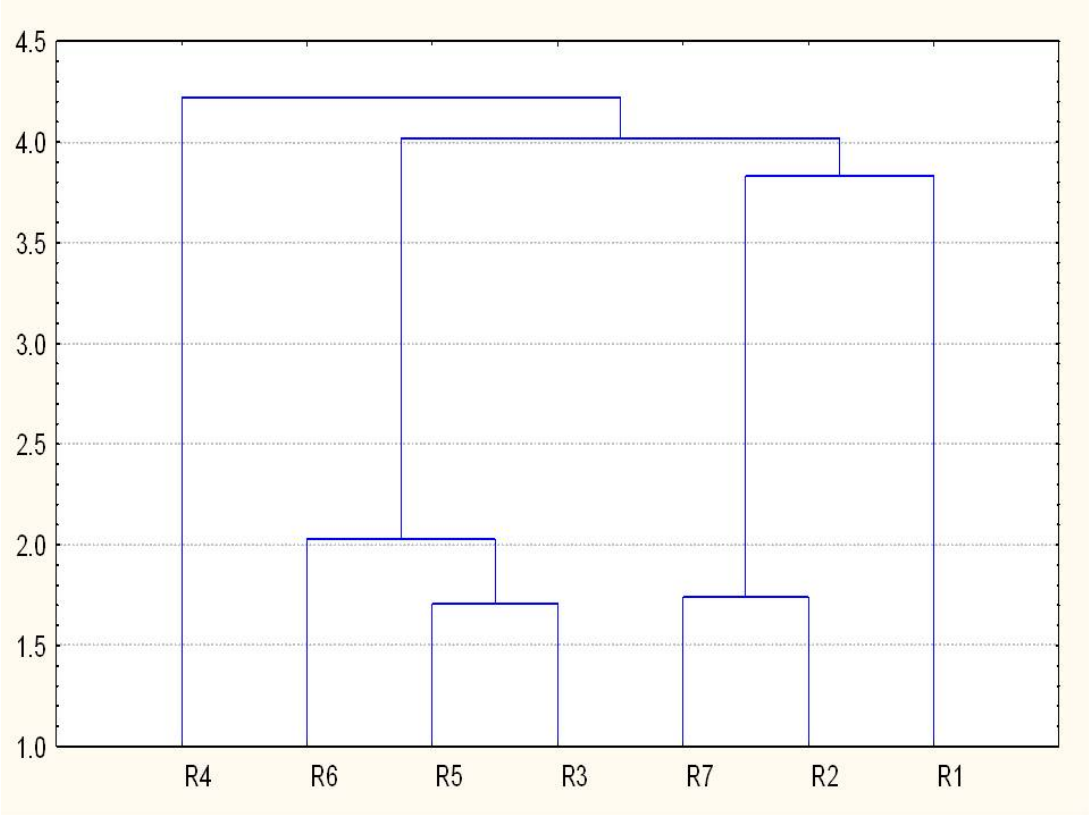


Figura 1. Fenograma para la clasificación de siete razas de maíz en base a ocho características de la mazorca de maíz.

4. Conclusiones

Las razas 2 (Ixtlahuaca) y 7 (Amarillo criollo de Ixtlahuaca) presentaron la mayor longitud de mazorca y longitud de grano, por el contrario las razas 3 (Criollo Rojo de San Mateo), 5 (Criollo Rosado Pinto Violento) y 6 (Rosadito Violento de Cashi) mostraron menor longitud de mazorca, diámetro de mazorca y menores valores para las restantes características. La raza 4 (Cacahuacintle) fue la de mayor diámetro de mazorca, de olote y de raquis presentando

además mayor ancho y espesor de grano.

El análisis de conglomerados produjo una descripción sensible de la relación entre las diferentes razas de maíz. A un nivel de distancia de 3.90 detectó la formación de tres grupos bien definidos formados por las razas (3 5 6); las razas (1 2 7); y la raza 4. Entre las razas 3 y 5 se observó el menor coeficiente de disimilitud (1.71).

Referencias

Acosta, G.J.A., J.S. Muruga y R.F. Cárdenas. 1991. "Utilización y disponibilidad de recursos genéticos de Phseolus en México". *Sociedad Mexicana de Fitogenética*. A.C. Chapingo, México.

Crossa, J., S. Taba, S.A. Eberhart, P. Bretting y R. Vencovsky. 1994. "Practical considerations for maintaining germplasm in maize". *Theor. Appl. Genet*. 89: 89-95.

Manly, B.F.J. 1990. *Multivariante Statistical Methods*. Chapman and Hall, 29 West 35th Street, New York NY 10001.

Painting, K.A., M.C. Perry, R.A. Denning y W.G. Ayad. 1993. "Guía para la documentación de recursos genéticos". Roma, Italia.

Statistica. Kernel release 5.5 A Copyright 1984-2000 by StatSoft, Inic. 2300 East 14 th Street. Tulsa, Ok 74104, USA.

Memoria FONE XXIV (2010), Sección II, 137–142

Efecto de la presencia de datos faltantes en la estimación de componentes de varianza de la interacción genotipo x ambiente

Víctor Prieto Hernández[a]
Facultad de Agronomía. Universidad de la República, Uruguay

Juan Burgueño
Colegio de Postgraduados, México

1. Introducción

En el mejoramiento de cultivares un número extenso de genotipos son normalmente probados sobre un amplio rango de ambientes, que incluye sitios, años, épocas de siembra y prácticas de cultivo, entre otros. Esto se debe a las diferencias de expresión de los genotipos según el ambiente, lo que se conoce como interacción genotipo-ambiente (GEI). Esta respuesta diferencial, medida en valores de caracteres observables (fenotipo) como rendimiento, altura de planta, resistencia a sequía, etc. puede ocasionar incluso el cambio de posición relativa o ranking de cultivares entre diferentes ambientes (Kang and Gauch, 1996; Annichiarichio, 2002)

La consecuencia fundamental para el fitomejorador es que cuanto más se manifieste el componente GEI por sobre el valor genotípico (G), menor heredabilidad para el carácter se obtendrá en el proceso de selección y por ende, mayor es la dificultad de su mejora. Otras implicancias determinantes para un programa de mejoramiento genético son que (a) se dificulta la identificación de materiales superiores, ya que podría existir cambio de rankings de genotipos (lo que se conoce como GEI cruzada) y (b) se incrementan los costos de evaluación, debido a que la prueba debe realizarse en varios lugares representativos de áreas de cultivo claves (Kang, 2002) .

[a]vprieto@fagro.edu.uy

En los programas de mejoramiento genético la metodología de análisis de varianza (ANOVA) ha sido intensamente utilizada para la estimación de componentes de varianza, relacionado a diversas fuentes de variación, incluida la GEI. Ésta metodología es útil para medir la variabilidad genética y estimación de heredabilidad, así como predecir el progreso genético de la selección. También pueden obtenerse estimaciones mediante el método de máxima verosimilitud (ML), aunque se cita en la literatura que éstas son sesgadas (subestimadas). Con el método de máxima verosimilitud restringida (REML) se supera ésta desventaja y produce idénticos estimadores que ANOVA. Una fuerte limitante del método ANOVA es que puede ser inválido para el análisis de experimentos desbalanceados, cosa que no sucede bajo el análisis REML (Freeman, 1973; Crossa, 1990).

En ensayos de múltiple ambiente, las pruebas de cultivares pueden presentar situaciones de desbalance, no necesariamente debido a pérdida no prevista de parcelas. De los diversos motivos, es común que algunos genotipos sean evaluados sólo en algunos sitios o años y que algunos otros ingresen (nuevos materiales a evaluar) o sean descartados. Como consecuencia del proceso selectivo de cultivares se obtiene un conjunto de datos altamente desbalanceado, con un importante número de celdas vacías para los datos de año x lugar x cultivar. Estos datos desbalanceados son comúnmente analizados bajo un enfoque de modelos lineales mixtos REML, que posibilita el tratamiento de observaciones faltantes de forma directa bajo el supuesto de que su naturaleza sea aleatoria (Balzarini, 2002).

El presente trabajo tiene como objetivo cuantificar el cambio en las estimaciones de los componentes de varianza bajo la hipótesis de que la presencia de datos faltantes tiene un efecto importante en el análisis de la GEI.

2. Materiales y Métodos

Para llevar a cabo el trabajo se utilizaron datos de rendimiento en grano (t/ha) de 49 genotipos de trigo tipo semi-árido (ensayos SAWYT) del CIMMYT para 19 localidades de siembra, bajo un diseño experimental de bloques completos al azar con 2 replicaciones por sitio. Esta matriz de datos fue sometida a diferentes pérdidas simuladas de celdas por un proceso de muestro al azar simple, definiéndose nueve tipos (% de pérdida) correspondiente a 5, 10, 15, 20, 25, 30, 35, 40 y 45. De esta manera, se obtuvieron diez muestras para cada tipo de pérdida, que hacen un total de 90 muestras a analizar. El modelo de datos completos

está dado por:

$$\mathrm{Y_{ijk}} = \mu + \mathrm{s_i} + \mathrm{r_{k(i)}} + \mathrm{gs_{ji}} + \mathrm{e_{ijk}}$$

donde Y_{ijk} es la variable de respuesta (fenotipo), μ es la media general, s_i es el efecto del sitio, $r_{k(i)}$ es el efecto aleatorio de la replicación dentro de sitios, gs_{ji} el efecto aleatorio de genotipo y de la interacción genotipo-ambiente y e_{ijk} el error residual. Como el modelo descrito tiene uno o más efectos fijos y aleatorios puede ser escrito como un modelo lineal mixto de la forma convencional:

$$y = X\beta + Zu + \varepsilon$$

donde y es el vector de observaciones, X y Z corresponden a las matrices de incidencia para los efectos fijos y aleatorios respectivamente; β, u y ε son los vectores de efectos fijos, aleatorios y errores residuales, respectivamente. Se asume para los efectos aleatorios que están normalmente distribuidos con una media igual a cero y matriz de varianzas-covarianzas (Σ) con una estructura compuesta de factores analíticos FA(k) donde k=2, tal que:

$$\Sigma = (\Lambda\Lambda' + \Psi)$$

donde Λ es una matriz conteniendo covariables latentes ambientales (loadings ambientales) y Ψ la varianza específica. La comparación entre las matrices de varianzas-covarianzas se basó en el estadístico R de la prueba de Mantel y para el estudio de efectos de escala, traslación y rotación sobre los loadings ambientales se utilizó la técnica de Procrustes. Esta técnica permite evaluar el cambio en la configuración de puntos en un espacio bidimensional debido a los mencionados efectos fundamentales: de escala, de traslación y de rotación. En ambos estudios fue utilizada la librería Vegan implementada en el paquete R (R Development Core Team, 2009).

3. Resultados y Discusión

De la comparación entre la matriz de varianzas-covarianzas del modelo completo y las que surgen de la simulación se observa una consistente pérdida de similaridad, expresada en el coeficiente R (Figura 1). Puede observarse que ésta tendencia también se manifiesta en las matrices de correlaciones para los mismos tratamientos, aunque con mayor grado de dispersión.

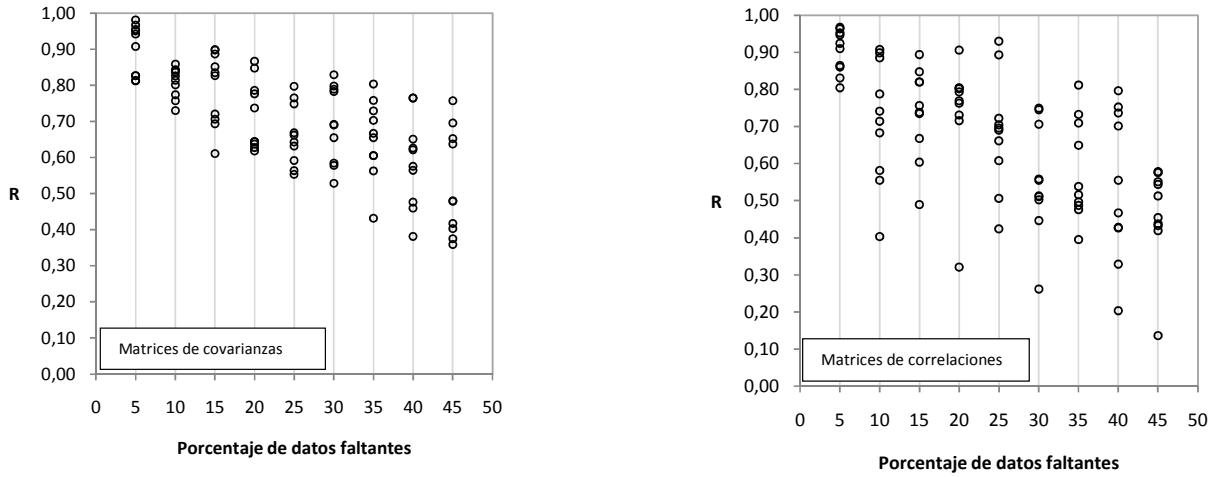


Figura 1: Estadístico de R (test de Mantel) para las matrices de varianzas-covarianzas (izquierda) y las matrices de correlaciones (derecha) según porcentaje de pérdida de datos.

Para evaluar la naturaleza del cambio en los modelos de datos incompletos se utilizaron los loadings ambientales y scores genéticos de las muestras con datos faltantes.

En el caso del efecto de escala se pudo observar un paulatino descenso en los valores, siendo estos muy importantes por encima del 25-30 % en la pérdida de datos. De la misma forma se manifiesta el efecto de traslación en el alejamiento de los puntos al origen del gráfico, y que este es mayor en la medida que sea mayor el porcentaje de pérdida.

Por último, el efecto de rotación se manifiesta en el cambio en los ejes de la configuración de puntos, buscando minimizar la suma de cuadrados de las diferencias entre la matriz original y la que surge de los tratamientos de pérdida. Es así que se generaron 90 gráficos correspondientes a las muestras del análisis, donde cada uno de esos gráficos se observan los ejes de rotación y la dirección y valor de cambio para cada punto (ambiente) en el plano. En la Figura 2 puede observarse la rotación de ejes para el caso de una muestra con 5 % y 45 % de pérdida en las matrices de loadings ambientales.

Claramente, puede observarse que en el caso de una muestra con mayor pérdida, el ángulo de rotación es mucho mayor que si la pérdida es menor, así como también es mayor la magnitud de los residuales para cada punto. Éste mismo efecto se manifiesta en el caso de los scores genéticos.

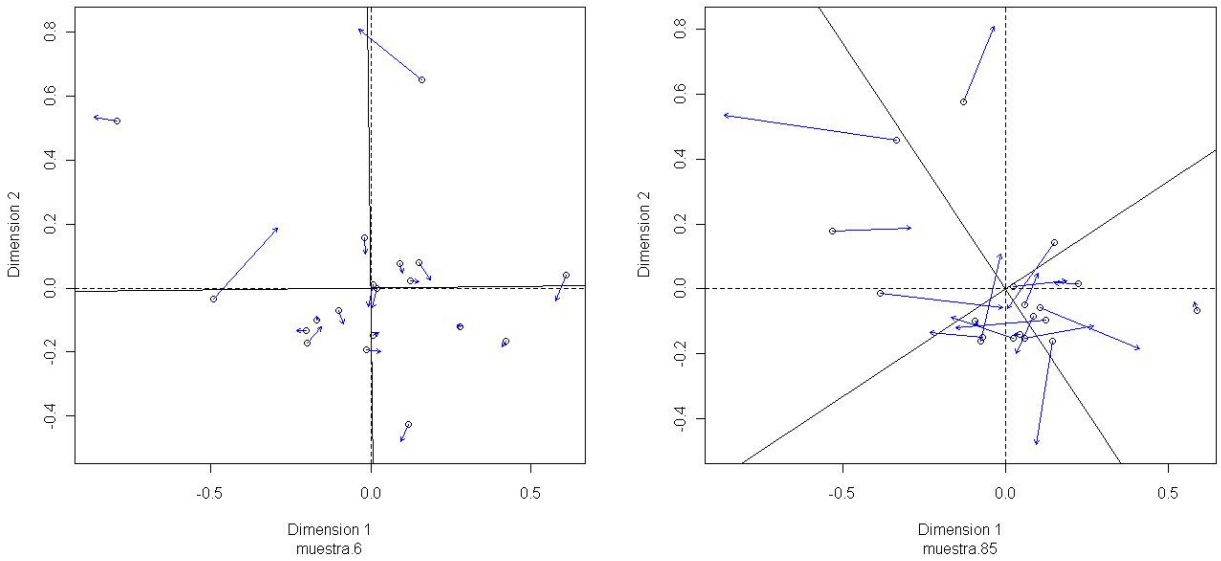


Figura 2: Diagrama de puntos y ejes de rotación correspondientes a los loadings ambientales con un 5 % (izquierda) y 45 % de pérdida de datos (derecha).

4. Conclusiones

Del análisis de los resultados se observa que por encima del 25 % de pérdida, la similaridad entre las matrices de varianzas-covarianzas con respecto a la original cae por debajo de 0.7. De cualquier forma, no permite trazar un punto de corte a partir del cual sería o no recomendable analizar los datos, ya que la caída en la similaridad en función del % de pérdida sigue una tendencia lineal. Teniendo en cuenta ésta tendencia, además del hecho de que la variabilidad entre las muestras con igual % de pérdida es muy similar nos sugiere que su naturaleza es más importante, es decir, no cuantos datos se pierden sino cuales genotipos y en que ambientes se pierden.

Hay un efecto de escala con la pérdida de datos y la tendencia se acentúa cuando la misma supera los 25-30 % de datos faltantes. El efecto de traslación que se produce es más claro cuanto mayor es la pérdida de datos, pero éste fenómeno no se manifiesta para los genotipos. Por último, en cuanto al efecto de rotación no parece tener clara evidencia a través de los porcentajes de pérdida.

El hecho de que la pérdida simulada de datos sea de naturaleza aleatoria nos impide evaluar sus efectos en aquellos casos donde la pérdida no es sistemática, por lo que la evaluación de la GEI es más compleja. Para ello se hace necesario simular patrones de pérdida diferentes, no aleatorios, que se asemejen a situaciones reales de un plan de evaluación de cultivares.

Referencias

Annicchiarico, P. (2002). "Genotype x environment interaction : challenges and opportunities for plant breeding and cultivar recommendations". *FAO plant production and protection paper*, 174. Food and Agriculture Organization of the United Nations.

Balzarini, M. (2002). *Applications of Mixed Models in Plant Breeding*, pp. 353-365. CABI Publishing.

Crossa, J. (1990). "Statistical Analyses of Multilocation Trials", *Advances in Agronomy*, 44,pp. 55-85. Elsevier.

Freeman, G. H. (1973, December). "Statistical methods for the analysis of genotype-environment interactions". *Heredity* 31(3), 339-354.

Kang, M. S. (2002). *Quantitative genetics, genomics, and plant breeding*. CABI Publishing.

Kang, M. S. and H. G. Gauch (1996). *Genotype -by- environment interaction*. CRC Press.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Memoria FONE XXIV (2010), Sección II, 143–149

Modelación de los factores ambientales en niveles altos de ozono

Sara Rodríguez R.[a], Hortensia Reyes C.[b], Gladys Linares F.[c]
Facultad de Ciencias Físico-Matemáticas, Instituto de Ciencias; Benemérita Universidad Autónoma de Puebla

Humberto Vaquera H.[d]
Colegio de Posgraduados

1. Introducción

El ozono que se encuentra en la atmósfera y que varía con la altura, forma parte natural de nuestro planeta. Hay otro ozono que debido a sus concentraciones en la superficie terrestre es de gran interés por los efectos adversos que se producen a la vida Thompson et al (2001), su producción se incrementa por una serie de reacciones químicas y fotoquímicas en el ambiente a partir de algunos precursores, como la intensidad de la luz solar, el mezclado atmosférico, la presencia de lagos y la reactividad de los precursores orgánicos Jazcilevich et al (2002).

Hemos estimado el comportamiento de los niveles altos de ozono de la Zona Metropolitana de la Ciudad de México (ZMCM) en tres estaciones meteorológicas; de la Red Automática de Monitoreo Atmosférico (RAMA) del Sistema de Monitoreo Atmosferico (SIMAT); a través de un modelo de valores extremos generalizados. Se verifica que variables ambientales influyen en la tendencia del ozono y se realiza una comparación en el promedio de las máximas concentraciones de ozono de las estaciones. Se ha obtenido que las tendencia del ozono es decreciente hasta el 2008 y que existe diferencia significativa en los valores promedio de los maximos de las concentraciones de ozono de las tres estaciones.

[a]rguez.sara@gmail.com
[b]hreyes@fcfm.buap.mx
[c]gladyslinares1@yahoo.es
[d]hvaquerah@gmail.com

2. Distribución de valores extremos generalizada

A diferencia de la distribución que surge de usar el teorema del límite central para muestras grandes, la distribución de valor extremo surge del teorema límite de Fisher-Tippet (1928) sobre valores extremos o máximos en muestras de datos, según, dada $X_1, ..., X_n$ una sucesión de variables aleatorias independientes identicamente distribuidas, con distribución $F(x)$ y sea Y_n la estadística de orden máxima con función de densidad $F^n(x)$. Supongamos que existen un par de sucesiones $a_n > 0$ y b_n y una función de distribución $H(x)$ tal que:

$$lim_{n\longrightarrow\infty} P\{(Y_n - b_n)/a_n \leq x\} = lim_{n\longrightarrow\infty} F^n(a_n x + b_n) = H(x). \tag{1}$$

Donde $H(x)$ es una distribución de Valores Extremos (VE), y en adelante la denotaremos como $G_{\mu,\sigma,\varepsilon}(z)$. Las Distribuciones de Valores Extremos son tres diferentes familias de distribución, conocidas como Gumbel, Frechet y Weibull, éstas pueden ser anidadas en una única representación paramétrica conocida como la distribución de Valores Extremos Generalizados (GEV). La distribución GEV estandarizada Reiss et al (2001), incorporando el parámetro de localización μ y el parámetro de escala σ:

$$G_{\mu,\sigma,\varepsilon}(z) \quad = exp{-}[1 + \varepsilon(\tfrac{z-\mu}{\sigma})]^{-\frac{1}{\varepsilon}} \quad \text{si} \quad 1 + \varepsilon(\tfrac{z-\mu}{\sigma}) > 0, \tag{2}$$

$$G_{\mu,\sigma,0}(z) \quad = exp\,(-exp(-\tfrac{z-\mu}{\sigma})) \qquad \text{en el caso de} \quad \varepsilon = 0. \tag{3}$$

con $\theta = (\mu, \sigma, \varepsilon)\epsilon\ \Re\times\Re^{+}\times\Re$. Los parámetros μ y $\sigma > 0$ son los parámetros de localización y de dispersión, respectivamente.

Propone Cox en 1996 el modelo de Regresión de la Distribución Pareto Generalizada (PG) a X^t como el vector de covariables y $\underline{\beta}$ es el vector de parámetros desconocidos, $\sigma = \sigma(\underline{x^t}\ \underline{\beta})$. En el caso de GEV, como $\sigma > 0$, es natural suponer que $\sigma(\underline{x^t}\ \underline{\beta}) = exp(\underline{x^t}\ \underline{\beta})$. En esta expresión se introducen los máximos de las covariables para el periodo de tiempo que se tome. La función de verosimilitud en esta propuesta es $L(Y; \mu, \underline{\beta}, \varepsilon) = \prod_i^n g(y_i; \mu, \underline{\beta}, \varepsilon)$

En forma equivalente se tiene:

$$\exp(-\sum_{i}^{n}[1 + \frac{\varepsilon(y_i - \mu)}{\exp[\mu + \sum_i^k \beta_j x_{ij}]}]^{-\frac{1}{\varepsilon}}) \exp(n\mu + \sum_{i}^{n}\sum_{j}^{k} \beta_j x_{ij}) \tag{4}$$

$$\left\{\prod_{i}^{n}(1 + \frac{\varepsilon(y_i - \mu)}{\exp(\mu + \sum_i^k \beta_{jx_{ij}})})\right\}^{(-\frac{1}{\varepsilon}-1)}. \tag{5}$$

Al obtener los estimadores máximo verosímiles, de la función $-logL(Y;\mu,\underline{\beta},\varepsilon)$, se utiliza el método de Nelder que maneja el módulo ismev de software libre R Stephensaon y Gillei (2004), encontrando los estimadores de los parámetros $\mu, \underline{\beta}$ *y* ε . Para el caso particular, de tener sólo la covariable tiempo para una muestra del modelo anterior, se tendría que μ es el parámetro de localización, β_1 es el parámetro que se está asociando a la tendencia y t_i representa al i-ésimo periodo de tiempo. Es decir, $-logL(X;\mu,exp(\beta_0+\beta_1 t_i),\varepsilon)$, entonces si $\beta_1 = 0$ se tiene que no existe tendencia, si $\beta_1 > 0$ implica que hay tendencia creciente y por último, $\beta_1 < 0$ habrá tendencia decreciente Reyes et al (2007).

Cuando se tiene una muestra de n observaciones y k covariables meteorológicas, sólo se analizará si influye o no en la tendencia. Para analizar la tendencia en la variable tiempo o para probar que alguna de las covariables (j:1,..,k) está contribuyendo en la tendencia de Y, se tiene que probar la hipótesis, $H_0 : \beta_j = 0 \quad vs \quad H_a : \beta_j \neq 0$ con todos los coeficientes de las k variables. Se rechazar H_0 si $|\frac{\widehat{\beta_j}}{\sqrt{v_{jj}}}| > z_{\alpha/2}$ donde α es el nivel de confianza y v_{jj} es el j-ésimo término de la diagonal de la matriz $(E(-\frac{\partial^2 logL}{\partial\beta_i\partial\beta_j}|_{\beta_j=\widehat{\beta_j}}))^{-1}$. En el caso de que se rechace la hipótesis nula, habrá evidencia a un nivel α de significancia, que existe influencia en la tendencia en las observaciones de Y.

2.1. Modelación de la calidad del aire en tres estaciones meteorológicas

Bajo la suposición de condiciones de regularidad en la distribución GEV, para los años de 1990 a 2008, se lleva a cabo el procedimiento usando cuantiles de esta distribución Reyes et al (2009), donde se obtienen los máximos por cada tres días de las variables ambientales de interés para el periodo acordado, esta metodología nos asegura la independencia entre las observaciones.

El paso siguiente consiste en estimar los parámetros de tiempo (t), temperatura (tem), velocidad del viento (wsp), humedad relativa (hr) y dirección del viento (wdr). El modelo correspondiente es el siguiente: $GEV(\widehat{\mu},\widehat{\sigma},\widehat{\varepsilon})$ donde $\widehat{\sigma} = \exp(\widehat{\beta_0} + \widehat{t}\,x_1 + \widehat{tem}\,x_2 + \widehat{wsp}\,x_3 + \widehat{hr}\,x_4 + \widehat{wdr}\,x_5)$.

Presentamos las gráficas de los máximos de cada tres dias de ozono en las estaciones de Plateros, Pedregal y Merced (Figura 1), notando graficamente la disminución en los niveles máximos de concentración de ozono.

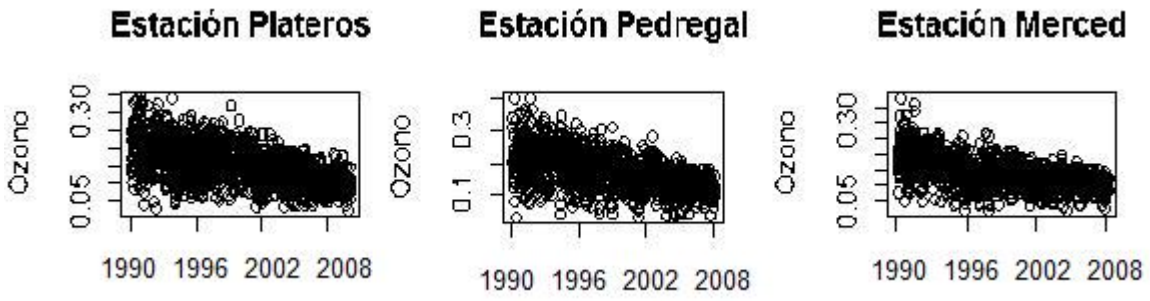


Figura 1: Máximos de Ozono 1990-2008

Sin embargo podemos observar en la Tabla 1, que en las estaciones Plateros y Pedregal más del 75 % de observaciones de concentraciones de ozono sobrepasan la norma mundial de la salud que es de 0.11 ppm. y en la estación Merced más de la mitad de los datos están sobre la norma.

	Plateros	Pedregal	Merced
$Min.$	0.0240	0.027	0.02
$1stQu.$	0.114	0.13	0.1022
$Mediana$	0.148	0.169	0.1330
$Media$	0.1537	0.1722	0.1369
$3rdQu.$	0.19	0.211	0.1660
$Max.$	0.343	0.403	0.377

Tabla 1. Estadísticos descriptivos importantes de las concentraciones de ozono

En la Tabla 2, se presentan los estimadores (emv.) de las variables meteorológicas y su errores estándar(e.s), para las tres estaciones.

Para el análisis de la comparación de tres medias de las concentraciones de ozono, incorporamos al modelo dos variables indicadoras,

$$GEV(\widehat{\mu}, \widehat{\sigma} = \exp(\widehat{\beta_0} + \widehat{\beta_1}\ x_1 + \widehat{\beta_2}\ x_2)\ , \widehat{\varepsilon}\ ), \tag{6}$$

donde $\beta_0 = \mu_3$, $\beta_1 = \mu_1 - \mu_3$ y $\beta_2 = \mu_2 - \mu_3$; μ_1 representa la media de Plateros, μ_2 Pedregal y μ_3 Merced. Definimos las variables indicadoras, $x_1 = 1$ si la observación procede de la estación Plateros o $x_1 = 0$ en cualquier otro caso; $x_2 = 1$ si la observación procede de

Parámetros	Plateros	Pedregal	Merced
	emv. , e.s	emv. , e.s	emv , e.s.
t	-0.00075, 1.995725e-06	-0.00071, 1.996481e-06	-0.00069, 1.996168e-06
tem	-0.06362, 1.997440e-06	-0.07547, 0.002656995	-0.086093, 0.001045413
wsp	-0.03940, 0.01508607	-0.00906, 0.01401765	-0.01057, 0.01913654
hr	-0.006753, 1.997429e-06	-0.00609, 1.997801e-06	
wdr	-0.00121 ,1.997285e-06	-0.00249, 1.997706e-06	-0.00037, 1.997410e-06
ε	-0.23602, 0.00868	-0.25109, 0.01152329	-0.20058, 0.01000607

Tabla 2. Estimadores Máximo Verosímiles para la estaciones

la estación Pedregal o $x_2 = 0$ en cualquier otro caso. En la Tabla 3, se presenta el modelo estadístico que involucra ozono y las variables indicadoras

Parámetros	Esti-M.V.	Var.
β_0	-3.1260242	3.040492e-04
β_1	0.1478137	6.113772e-04
β_2	0.2527483	7.069944e-04

Tabla 3. Comparación de Estaciones

Se encuentra diferencia significativa en los valores promedios de los máximos de las concentraciones de ozono entre las estaciones de Plateros, Pedregal y Merced. Notemos que el hecho de que la estación Merced pertenezca a la región centro de la ZMCM, no implica que tenga en promedio las mayores concentraciones de ozono, esto se debe a las corrientes de aire ya que de acuerdo a la mediana de direcciones de viento que obtenemos, estos predominan en dirección suroeste Jazcilevich et al (2002).

3. Conclusiones

Se observa que los niveles altos de ozono en las estaciones de Plateros, Pedregal y Merced siguen una tendencia decreciente en el periodo 1990-2008. El comportamiento de las variables ambientales para cada estación se explica a continuación.

En la estación Plateros de acuerdo con las pruebas de hipótesis obtenemos que todas las variables influyen en la tendencia decreciente de las concentraciones de ozono.

En la estación Pedregal, concluímos que la velocidad del viento no influye en la tendencia decreciente de las concentraciones de ozono, esto coincide con la información de estudios ambientales los cuales muestran que debido a la magnitud y dirección de vientos, estos arrastran los contaminantes a la parte suroeste dela ZMCM y en esta parte se encuentra ubicada la estación Pedregal. También obtuvimos que las variables restantes si influyen en la tendencia del ozono.

Para la estación Merced la temperatura, velocidad y dirección del viento influyen en la tendencia decreciente de las concentraciones de ozono, sin embargo no obtenemos un estimador para la variable Humedad Relativa.

Al realizar la comparación entre las tres estaciones obtenemos que la media de los niveles altos de ozono en Pedregal son más altos que los de Plateros y Merced, además que la media de Plateros es mayor que la de Merced.

Referencias

Cox, W., and Chu,S.(1996). "Assessment of interannual ozone variation in urban areas from a climatological perspective",*Atmospheric Environment*,30,2615–2625.

Jazcilevich, A.D., Agustín, and Gerardo, R.S. (2002). "A modeling study of air pollution modulation through land-use change in the Valley of Mexico",*Atmospheric environment*,36,2297-2307".

Reiss, R., and Thomas, M. (2001). *Statistical Analysis of extreme values*,Birkhauser Verlag,Germany.

Reyes, H., Vaquera, H., and Villaseñor, J. (2007)."Uso de distribución de valores extremos para investigar tendencias en niveles muy altos de ozono",*Memorias del XXI Foro Nacional de Estadística, INEGI*,107-112.

Reyes, H., Vaquera, H., and Villaseñor, J.,(2009). "Estimation of trends in high ozone levels using the quantiles of the distribution GEV",*Environmetric*

Stephenson and Gillel (2005) "ismev Package:Extreme Values in R",*R Foundation for Statistical Computing*,"http://www.r-project.org",

Thompson, M.L., R., Joel, H.C., Lawrencw, G., Peter, and Paul, D.S.(2001). "A review of statistical methods for the meteorological adjustment of tropospheric ozone",*Atmospheric environment*,35,617-630.

Memoria FONE XXIV (2010), Sección II, 151–157

Estimación de vida útil mediante análisis de datos censurados y pruebas de vida acelerada

Fidel Ulín-Montejo[a], Rosa Ma. Salinas-Hernández y Gustavo A. González Aguilar

Universidad Juárez Autónoma de Tabasco, CIAD - Hermosillo

1. Introducción

La producción de alimentos enfrenta fuertes presiones para desarrollar nuevos productos en tiempos record. En la mayoría de esos productos se espera una vida útil de varias semanas o meses, mientras que el tiempo de prueba se reduce a sólo unos días o semanas. Por ello, las pruebas aceleradas son usadas para obtener información a niveles altos de variables de aceleración (por ej. temperatura), extrapolando esta información para obtener estimaciones sobre condiciones normales de operación o almacenamiento. La ecuación más utilizada para modelar las razones o constantes de reacción en función de la temperatura, como factor de aceleración, es la ecuación de Arrhenius; donde la energía de activación es el parámetro clave para la estimación de vida útil a diferentes temperaturas. Nelson (1990) y Meeker y Escobar (1998) han aplicado análisis de supervivencia y de confiabilidad para pruebas aceleradas de especímenes y materiales industriales tales como componentes eléctricos, lámparas o autopartes. Esos modelos han sido poco utilizados en la estimación de vida útil de alimentos basados en la decisión del consumidor de aceptar o rechazar un producto (Hough *et al*, 2006). Al respecto, una aplicación importante se ilustra en este trabajo.

2. Ecuación de Arrhenius y Temperatura

Suponga que se desea estimar la vida útil de un alimento, basándose en la aceptación/rechazo de su color-apariencia; asumiendo que el oscurecimiento (color café) sigue un modelo cinético

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

de orden cero; esto es $dBC(x,T)/dx = k(T)[BC(x,T)]^n$ con $n = 0$, es decir,

$$BC(x,T) = BC_0 + k(T) \cdot x \tag{1}$$

$BC(x,T)$ es el oscurecimiento al tiempo de almacenamiento x para una temperatura igual a T, BC_0 es el oscurecimiento al tiempo 0 y $k(T)$ es la velocidad de reacción constante a la temperatura T. Ahora, se establece una relación entre la vida útil a la temperatura usual T_U y a la temperatura T basado en la ecuación de Arrhenius

$$k(T) = k_0 \exp\left(-\frac{E_a}{R \cdot T}\right) \tag{2}$$

donde E_a es la energía de activación (cal/mol), $R = 1.98$ cal/mol K es la constante de la ley de los gases y T la emperatura en ^oK. De (2) se define el factor de aceleración (AF)

$$AF = \frac{k(T)}{k(T_U)} = \exp\left[E_R\left(\frac{1}{T_U} - \frac{1}{T}\right)\right] \tag{3}$$

donde T_U se refiere a la temperatura usual y $E_R = E_a/R$. Combinando (1) y (3)

$$BC(x,T) = BC_0 + AF \cdot k(T_U) \cdot x \tag{4}$$

Note que (4) se ha desarrollado en términos de un tiempo de almacenamiento arbitrario x para un determinado producto, y del mismo modo se puede expresar también en términos del tiempo de rechazo $X(T)$ y $X(T_U)$ de un consumidor para una temperatura acelerada (T) y para la temperatura usual (T_U); ésto es,

$$\begin{aligned} BC(X(T),T) &= BC_0 + AF \cdot k(T_U) \cdot X(T) \\ BC(X_U(T_U),T_U) &= BC_0 + k(T_U) \cdot X(T_U) \end{aligned} \tag{5}$$

Se supone que el oscurecimiento al cual el consumidor rechazará el producto es el mismo, independientemente de T; es decir, el consumidor ve el producto y lo rechaza sin conocer T. Por tanto los lados izquierdos de (5) son iguales y se obtiene la ecuación

$$X(T_U) = AF \cdot X(T). \tag{6}$$

Ésto significa que la vida útil en condiciones usuales es igual a la vida útil en condiciones aceleradas multiplicada por AF. Para usar (6) E_a debe ser conocida (estimada). Suponiendo que el oscurecimiento sigue una cinética de primer orden, $dBC(x,T)/dx = k(T)[BC(x,T)]$, de donde $\ln BC(x,T) = \ln BC_0 + k(T) \cdot x$, teniéndose la misma relación entre la vida útil a la temperatura usual y la vida útil a la temperatura T dada en (6).

3. Función de Distribución de Falla Acelerada

La ecuación de probabilidades de falla a distintas temperaturas es la siguiente (Meeker y Escobar, 1998)

$$F(x,T) = F(x \cdot AF, T_U). \tag{7}$$

Si $F(x)$ es una distribución de log-localización-escala, $\ln x_p = \mu + \sigma\Phi^{-1}(p)$, por ejemplo lognormal o Weibull, con parámetros μ y σ, donde Φ es la normal estándar (caso lognormal) o de valores mínimos extremos (caso Weibull). Entonces (7) puede expresarse como:

$$F(x,T) = \Phi\left(\frac{\ln x - \mu_T}{\sigma}\right) = \Phi\left(\frac{\ln(x \cdot AF) - \mu_{T_U}}{\sigma}\right) = \Phi\left(\frac{\ln x - (\mu_{T_U} - \ln AF)}{\sigma}\right). \tag{8}$$

De (8), μ_T puede ser expresada como en los modelos de regresión de log-localización-escala con covariables fijas W; ésto es, $\mu(W) = \beta_0 + \beta_1 g(W)$, donde $g(W)$ puede ser una función lineal, cuadrática o inversa (Meeker y Escobar, 1998). De modo que para este estudio,

$$\mu_T = \beta_0 + \beta_1\left(\frac{1}{T}\right); \quad \beta_0 = \mu_{T_U} - \frac{E_R}{T_U}; \quad \beta_1 = E_R; \quad Z = \frac{1}{T}. \tag{9}$$

En (8) se asume que σ es constante para cada temperatura; acorde a Nelson (1990), donde se explica que σ distintos a diferentes temperaturas resulta físicamente implausible. Sin embargo, Meeker y Escobar (1998) provee una prueba de razón de verosimilitudes para probar este supuesto; osea, la hipótesis $H : \sigma_1 = \sigma_2 = \sigma_3$.

4. Datos Experimentales y Estimación

Muestras de un alimento de origen vegetal fueron almacenadas a distintas temperaturas y extraídos en diferentes tiempos para su análisis sensorial por 60 consumidores habituales:

2°C: 0 - 24 -48 - 96 - 144 - 192 - 240 hrs.
9°C: 0 - 24 - 48 - 72 - 92 - 120 - 144 hrs.
19°C: 0 - 6 -12 - 18 - 24 - 36 - 48 hrs.

Los consumidores contestaron la pregunta ¿Consumiría usted este producto? Las respuestas (si/no), aceptación/rechazo, estuvieron basadas únicamente en la apariencia del producto. En el XXIII Foro Nacional de Estadística, Ulín-Montejo y Salinas-Hernández (2008) presentaron los tipos de datos censurados obtenidos de estudios sensoriales: *Censura por la Izquierda* para un consumidor que rechaza la muestra con 9 h de almacenamiento; es decir, rechaza el producto en algún momento entre 0 y 9 h. *Censura por Intervalo* para un consumidor que acepta muestras almacenadas entre 0 y 9 horas, pero rechaza la muestra almacenada por 18 h. *Censura por la Derecha* para un consumidor que acepta las muestras con cualquier tiempo de almacenamiento, pero que finalmente las rechazara a un tiempo suficientemente largo. A continuación un cuadro ilustrando lo anterior:

Tabla 1. Datos de aceptación/rechazo (S/N) para los consumidores.

Tiempo de almacenamiento a 5°C (en horas)											
Frec.	0	9	18	27	36	45	54	63	72	81	Censura
21	S	S	S	S	S	S	S	S	N	N	Intervalo: 63 - 72
7	S	S	S	S	S	S	S	S	S	S	Derecha: > 81
5	S	N	N	N	N	N	N	N	N	N	Izquierda: < 9

Una vez determinado el tipo de censura se estiman los parámetros y se obtiene el mejor ajuste a través de la función de verosimilitud. La verosimilitud se define como la probabilidad conjunta de los datos obtenidos:

$$L(\theta) = \prod_{i \in R} [1 - F(r_i; \theta)] \prod_{i \in L} [F(l_i; \theta)] \prod_{i \in I} [F(r_i; \theta) - F(l_i; \theta)] \tag{10}$$

Donde R es el conjunto de observaciones censuradas por la derecha, L de las observaciones censuradas por la izquierda e I es el conjunto de las observaciones censuradas por

intervalo. Máxima verosimilitud puede aplicarse a una amplia variedad de modelos con datos censurados y covariables (Meeker y Escobar, 1998)

5. Resultados y Discusión

5.1. Selección del modelo

De acuerdo a la verosimilitud se elige el modelo log-normal. Ahora, usando una prueba de razón de verosimilitud y las log-verosimilitudes estimadas, se compara el ajuste del modelo lognormal para cada condición individual (*modelo completo*), con el ajuste del modelo Arrhenius-lognormal (*modelo reducido*).

La log-verosimilitud del modelo completo, es la suma de las log-verosimilitudes obtenidas para cada temperatura, esto es $\ell_{Completo} = \ell_2 + \ell_9 + \ell_{19} = -75.5 - 72.8 - 78.9 = -227.2$. Luego, para el modelo reducido $\ell_{\beta_0,\beta_1,\sigma} = \ell_{Reducido} = -228.5$. De lo anterior, $Q = -2(\ell_{Reducido} - \ell_{Completo}) = 2.52$, con $Q \sim \chi^2_3$. Entonces, $P(\chi^2_3 > Q) = 0.47$, por lo que no existe evidencia para rechazar el modelo Arrhenius-lognormal.

5.2. Estimación con el Modelo Arrhenius-lognormal

Las estimaciones obtenidas para los parámetros del modelo Arrhenius-lognormal (9), fueron las siguientes ($\pm$ intervalo de confianza del 0.95):

$$\beta_0 = -24 \pm 4; \quad \beta_1 = E_R = 7720 \pm 1351(K); \quad \sigma = 0.70 \pm 0.09 \tag{11}$$

Para estimar la vida útil (junto a un intervalo del 0.95), debe elegirse una probabilidad de rechazo de referencia. Gacula y Singh (1984) mencionan un valor de 0.50, Curia et al (2005) un valor entre 0.25 y 0.50. Para cada temperatura T, la vida útil se obtiene sustituyendo las estimaciones y despejando $x_{0.5}$ de $\ln(x_{0.5}) = \beta_0 + \left(\frac{\beta_1}{T}\right) + \sigma\Phi^{-1}(0.5)$.

Tabla 2. Vida útil estimada para una probabilidad de rechazo del 0.50.

Estimaciones			
Temp °C	Vida Útil	LI hrs	LS hrs
2	88	73	105
9	44	39	49
19	17	14	21

5.3. Estimación por Interpolación

El valor de E_R es importante para estimar la vida útil a temperaturas no experimentadas. Por ejemplo, si la vida útil a 19ºC (292ºK) fue de 17 hrs, para 5ºC se obtendría de (6):

$$X(T_U) = AF \cdot X(T) = exp\left(7720\left(\frac{1}{278} - \frac{1}{292}\right)\right) \cdot 17 = 64.3 \quad hrs \tag{12}$$

De igual modo, con los parámetros estimados en (9) puede obtnerse la función de distribución de rechazos para cualquier temperatura; por ejemplo para 5ºC (278 ºK) se tendría,

$$F(x, 278) = \Phi\left(\frac{\ln(x) - \mu_{278}}{\sigma}\right) = \Phi\left(\frac{\ln(x) - \left(\beta_0 + \frac{E_R}{278}\right)}{\sigma}\right) = \Phi\left(\frac{\ln(x) - 3.76}{0.70}\right) \tag{13}$$

Esta expresión puede ser usada para construir un gráfico de la probabilidad de rechazo por el consumidor contra el tiempo de almacenamiento a 5ºC.

6. Conclusiones

Aquí se ha mostrado que el estudio de análisis sensorial es relativamente simple, pues con sólo responder *si* o *no* a la aceptación de muestras almacenadas a distintos tiempos, los consumidores proveen información muy valiosa, con lo que es posible modelar la probabilidad de rechazo del producto a distintas temperaturas experimentadas o de interés, mediante extrapolación o interpolación. El modelo más adecuado puede ser elegido a través de comparación mediante pruebas de razón de verosimilitudes, requiriéndose al menos tres temperaturas distintas para lograr estimaciones aceptables para los parámetros, tiempos de vida e intervalos de confianza; completandose el proceso de modelación e inferencia estadística.

Referencias

Curia, A., Aguerrido, M., Langohr, K., Hough, G. (2005). "Survival analysis applied to sensory shelf-life of yogurts: I. Argentine formulations". *Journal Food Science*, in press.

Gacula, M. C., Singh, J. (1984). *Statistical methods in food and consumer research*. New York: Academic Press.

Hough, G., Garitta, L., Gómez, G. (2006). "Sensory shelf-life predictions by survival analysis accelerated storage models". *Food Quality and Preference*, **17**, 468-473.

Hough, G., Langohr, K., Gómez, G., Curia, A. (2003). "Survival analysis applied to sensory shelf life of foods". *Journal of Food Science*, **68**, 359-362.

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

Nelson, W. (1990). *Accelerated testing. Statistical models, test plans and data analyses*. New York: Wiley.

Ulín-Montejo, F., Salinas-Hernández, R. M. (2008)"Análisis de Confiabilidad para la Predicción de Vida Útil de Alimentos". *Memoria del XXIII Foro Nacional de Estadística*, 173-180.

Sección III

Tesis de licenciatura y maestría

Memoria FONE XXIV (2010), Sección III, 161–167

A parametric measure of dispersion derived from the generalized mean

Víctor M. Guerrero[a], Claudia Solís-Lemus[b]
Department of Statistics Instituto Tecnológico Autńomo de México (ITAM)

1. Introduction

When carrying out a descriptive statistical analysis of a set of observations from a continuous variable, it is generally agreed that a short and reasonable numerical summary of the data must include both a measure of central tendency and a measure of dispersion. Here, we shall emphasize the idea of measuring central tendency and dispersion in a unified way. We think it is desirable that the two measures be related in some way and that the data lead to choosing them. We present here a new measure of dispersion which is intrinsically linked to a parametric measure of centrality, called generalized mean (see Norris, 1976). Both measures are determined by the same parameter corresponding to the index of the power transformation family. Therefore we argue below that such a parameter can be appropriately chosen by looking for the power transformation that best produces normality (or at least symmetry) in a transformed scale.

Section 2 describes the generalized mean and its properties. In Section 3 we introduce the new parametric measure of dispersion as well as its properties and some interpretations. There, we show that this measure quantifies relative dispersion, and another measure can be obtained from this one if the interest lies in quantifying absolute dispersion. Finally, in Section 4 we show a numerical example in which we establish some comparisons with the coefficient of variation.

[a]guerrero@itam.mx
[b]claudia.sl2904@gmail.com

2. Generalized mean

The present paper requires some knowledge of the generalized mean, that is, the average of transformed data brought back to the original scale by means of retransformation. This can be found in the work by Norris (1976). Let $X > 0$ be a continuous random variable, we will focus only on the sample generalized mean (also called generalized average) and we will define it as $\phi(X, \lambda)$, since it is a function of the data X and the power of the transformation λ. Thus, for a set of n observations $X_1, ..., X_n$ we have

$$\phi(X, \lambda) = \left[\frac{1}{n} \sum_{i=1}^{n} X_i^{\lambda} \right]^{\frac{1}{\lambda}}, \lambda \neq 0 \tag{1}$$

$$\phi(X, 0) = \exp \left[\frac{1}{n} \sum_{i=1}^{n} \ln(X_i) \right]. \tag{2}$$

Some important properties of the generalized average are explained in Norris (1976). Due to the fact that it is an increasing function of λ, we can order the means by evaluating (1) at different values of λ,

$$X_{(1)} = \phi(X, -\infty) \leq \phi(X, -1) \leq \phi(X, 0) \leq \phi(X, 1) \leq \phi(X, \infty) = X_{(n)}.$$

Establishing these inequalities has been the object of various studies and several demonstrations (Norris, 1976) such as those based on the derivative of the generalized mean. This derivative is proven to be positive so that the generalized mean was shown to be an increasing function by many authors among them Shier (1988), Norris (1935) and Berger and Casella (2002).

If λ is chosen in such a way that X^{λ} follows a symmetric distribution then its mean and median are identical and $\frac{1}{n} \sum_{i=1}^{n} X_i^{\lambda}$ is an estimator of the median in the transformed scale. Therefore, since the transformation is monotone, $\phi(X, \lambda)$ becomes an estimator of the median of X in the original scale. The generalized mean can also be written in terms of the Box-Cox transformation (Guerrero, 1982a).

3. A parametric measure of dispersion

The derivative of the generalized mean as a function of λ is

$$\frac{d}{d\lambda}\phi(X,\lambda) = \phi(X,\lambda)\left[\frac{1}{\lambda n}\sum_{i=1}^{n}\left(\frac{X_i}{\phi(X,\lambda)}\right)^{\lambda}\ln\left(\frac{X_i}{\phi(X,\lambda)}\right)\right] \tag{3}$$

and we propose the following function as a parametric measure of dispersion (this measure was first proposed in Guerrero, 1982b)

$$\psi(X,\lambda) = \frac{1}{\lambda n}\sum_{i=1}^{n}\left(\frac{X_i}{\phi(X,\lambda)}\right)^{\lambda}\ln\left(\frac{X_i}{\phi(X,\lambda)}\right). \tag{4}$$

That is, the derivative of the logarithm of the generalized mean $\phi(X,\lambda)$. The proposed measure is relative with respect of the variable in study because it involves division by a measure of central tendency $\phi(X,\lambda)$. This allows a comparison between different groups of data. The derivative of the generalized mean is not a new concept. It has been used in the works of Shier (1988), Norris (1935) and appears in the book by Berger and Casella (2002) to demonstrate that the generalized mean is a monotonous increasing function. In that book, it is also said that the derivative (3) can be interpreted as a measure of entropy, and Shier even points out that the monotone behaviour of the generalized mean is related to the entropy of a probabilistic system.

Nevertheless, the idea of studying it as a measure of dispersion got no further investigation. We will justify the use of the derivative of the generalized mean (divided by the generalized mean) as a relative measure of dispersion. We recall that the derivative can be defined as the instantaneous rate of change with respect of the variable X This means that if the derivative is large, the values of the function will change rapidly near a specific point, while if the derivative is small, then the values of the function will change slowly.

It is important to remember that the objective of this work is to present a measure of dispersion that can describe the behaviour of the data in terms of variability regardless of the possible inference that can be made through the use of the sample measure as an estimator of the population measure. No further research has been made to determine the properties of this estimator and we shall not discuss the advantages or disadvantages regarding its use. That is why we present only the sample expression of the proposed measure.

Based on the results of Berger and Casella (1992), some specific distributions will be analyzed. The most common case results when $\lambda = 1$. The measure of central tendency is the arithmetic mean and the proposed measure of dispersion is given by

$$\psi(X,1) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\phi(X,1)} \ln\left(\frac{X_i}{\phi(X,1)}\right). \tag{5}$$

As

$$\phi(X,1) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

then

$$\psi(X,1) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\bar{X}} \ln\left(\frac{X_i}{\bar{X}}\right). \tag{6}$$

It can be observed that this result is just Theil's entropy (Nguyen, 2008). This means that for $\lambda = 1$ (Normal distribution according to Berger and Casella 1992) we propose the arithmetic sample mean, $\phi(X,1) = \bar{X}$, as a measure of central tendency and $\psi(X,1)$, Theil's entropy, as corresponding measure of relative dispersion.

Another interesting case arises when $\lambda = 0$. The measure of central tendency $\phi(X,0)$ is the geometric sample mean and the measure of dispersion $\psi(X,0)$ can be obtained by calculating the limit of the derivative of its generalized average when $\lambda \to 0$ and then dividing by this generalized mean.(Guerrero, 1982b)

$$\psi(X,0) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n} \left(\ln(X_i) - \ln\phi(X,0)\right)^2\right]. \tag{7}$$

That is, the sum represents the variance of the logarithms of the observations. This result can be corroborated by comparing with Theil (1967), who found an expression for the entropy of the Lognormal distribution (which corresponds to a transformation fo the data with the function $\ln(x)$) and corresponds to the variance of the logarithms of the observations. To sum up, we propose for $\lambda = 0$ (Lognormal distribution) the geometric sample mean as the measure of central tendency (Berger and Casella, 1992) and half the variance of the logarithm of the observations as a measure of dispersion.

The last case to analyse is that of $\lambda = -1$. This transformation is difficult to interpret, however, the generalized average with this parameter results in the harmonic sample mean. The measure of dispersion proposed for this distribution is

$$\psi(X,-1)=\frac{1}{n}\sum_{i=1}^{n}\frac{\phi(X,-1)}{X_i}\ln\left(\frac{\phi(X,-1)}{X_i}\right). \tag{8}$$

So, for the Inverse Gamma distribution (Berger and Casella, 1992), we propose the harmonic sample mean as a measure of central tendency and formula (8) as a measure of dispersion. At the moment the interpretation of this measure of dispersion is unknown, it is important to carry out some further investigation on this aspect. The only thing we can say so far is that it accompanies one of the three most common means, namely arithmetic mean $\phi(X,1)$, geometric mean $\phi(X,0)$ and harmonic mean $\phi(X,-1)$.

4. Numerical examples

To exemplify the use of the proposed measure of dispersion $\psi(X,\lambda)$,we will show comparisons between the coefficient of variation and the measure proposed in an asymmetric model: the Lognormal distribution. We kept the median constant and used three different levels of dispersion represented by the geometric standard deviation (Kirkwood, 1979).for $X\sim$Lognormal(μ,σ^2). We used this measure of dispersion since it is related to the geometric mean so it should be considered a specific measure of dispersion for the Lognormal distribution. Random samples were generated of size $n=100$ and both the coefficient of variation as well as the measure proposed were calculated from each sample. This process was repeated one hundred thousand times and the averages of both measures were calculated. It is important to acknowledge that the number of simulations has no theoretical basis since we only intend to illustrate the behaviour of the proposed measure in comparison with the behaviour of the coefficient of variation. Instead of determining the parameters of the Lognormal distribution, we determined the values of the median and geometric standard deviation

$$median(X)=e^{\mu}\Rightarrow\mu=\ln(median(X))$$

$$GSD(X)=e^{\sigma}\Rightarrow\sigma=\ln(GSD(X))$$

We used $median(X)=100$ with three different geometric standard deviation $GSD(X)=1.5,3,6$. Then we calculated and compared measures of central tendency (population and sample mean, population and sample median, and $\phi(X,0)$), as well as some measures of

dispersion (population and sample standard deviation, population and sample coefficient of variation, population and sample geometric standard deviation, $\Psi(X,0)$ and $\psi(X,0)$, and $\psi(X,0)\phi(X,0)$ as a measure of absolute dispersion), and the coefficient of skewness. The following table summarizes these calculations.

$PopMean$	108.567	182.846	497.886
$SampMean$	108.545	182.589	498.755
$PopMedian$	100.000	100.000	100.000
$SampMedian$	100.040	100.940	102.380
$\phi(X,0)$	100.049	100.490	101.324
$PopSD$	45.893	279.896	2428.390
$SampSD$	45.623	259.303	1568.248
$PopCV$	0.423	1.531	4.877
$SampCV$	0.420	1.390	2.788
$PopGSD$	1.500	3.000	6.000
$SampGSD$	1.497	2.984	5.972
$\Psi(X,0)$	0.082	0.603	1.605
$\psi(X,0)$	0.082	0.598	1.591
$\phi(X,0)\psi(X,0)$	8.155	60.034	161.180
$Skewness$	1.186	3.523	5.406

It can be appreciated that $\phi(X,0)$ looks more alike to the corresponding population measure than the sample median. Also, both the sample standard deviation and coefficient of variation move further away from their population measures as dispersion increases, which is not the case for the geometric standard deviation or $\psi(X,0)$.

5. Conclusions

Here we present a new measure of relative dispersion ($\Psi(X,\lambda)$ as a population measure and $\psi(X,\lambda)$ as a sample measure) that was derived from the generalized mean so it is directly linked to its corresponding measure of central tendency. Both measures are determined by the same parameter which depends on the type of data or model. This paper presents a work still in progress which we consider a preliminary study of this new measure, there is

so much to be written about it since the objetive of our work was merely to present it as another alternative for measuring dispersion. The door remains open for future contributions regarding in depth studies of its characteristics, its advantages and its limitations. Posible future research may include not only random variables with positive values, but all real values, and even discrete or qualitative random variables.

Referencias

Berger, R. y Casella, G. (1992). Deriving generalized means as least squares and maximum likelihood estimates, *The American Statistician*, Vol. 46, No. 4, 279-282.

Berger, R. y Casella, G. (2002). *Statistical inference*, Duxbury Advanced Series, California.

Guerrero, V. M. (1982a). Parametric averages: their selection and use as price indexes, Technical Report No. 684, University of Wisconsin.

Guerrero, V. M. (1982b). Unpublished working paper.

Kirkwood, Thomas. (1979). Geometric means and measures of dispersion, *Biometrics*, Vol. 35, No. 4, 908-909.

Norris, N. (1935). Inequalities among averages, *Annals of Math. Stat.*, Vol. 6, 27-29.

Norris, N. (1976). General means and statistical theory, *The American Statistician*, Vol. 30, No. 1, 8-12.

Nguyen Viet, Cuong. (2008). Do foreign remittances matter to poverty and inequality? Evidence from Vietnam. *Economics Bulletin*, Vol. 15, No. 1, 1-11

Shier D. (1988). The monocity of power means using entropy, *The American Statistician*, Vol. 42, No. 3, 203-204.

Theil, H. (1976). *Economics and Information Theory,* North-Holland Publishing Company.

Memoria FONE XXIV (2010), Sección III, 169–175

Análisis bayesiano del modelo INAR(1)

Lizbeth Naranjo Albarrán[a], Eduardo Gutiérrez Peña[b]
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM

1. Introducción

En este trabajo se discute el análisis Bayesiano de un modelo de series de tiempo para datos categóricos conocido como modelo autoregresivo de valores enteros (INAR por sus siglas en inglés).

2. Modelo multinomial negativo

En el análisis de datos categóricos algunas veces es de interés estudiar la asociación que existe entre las variables de una tabla de contingencia. Para esto Lindley (1964), Good (1967, 1976) y Gutiérrez-Peña (2005) proponen algunas pruebas de independencia. También podemos ajustar un modelo loglineal. Congdon (2001, 2005) describe el desarrollo Bayesiano y utiliza WinBUGS[1] para la estimación y una medida de bondad de ajuste llamada criterio de información de la devianza propuesto por Spiegelhalter *et al.* (2002).

Waller y Zelterman (1997) analizan datos de incidencia de cáncer de 1989 en tres grandes ciudades de Ohio, clasificadas de acuerdo al sitio en el que se encuentra el tumor principal. La Tabla 1 presenta los datos.

La distribución multinomial negativa es una distribución definida para vectores con entradas en los enteros no negativos; cualesquiera dos componentes del vector con esta distribución tienen correlaciones positivas; y cualquier componente presenta varianza mayor a

[a] lizbeth@sigma.iimas.unam.mx
[b] eduardo@sigma.iimas.unam.mx

[1] *Software* de libre acceso en http://www.mrc-bsu.cam.ac.uk/bugs/. Fue diseñado por Spiegelhalter, Thomas y Best y es parte del proyecto BUGS (*Bayesian Inference Using Gibbs Sampling*) que desarrollaron estos investigadores para el análisis Bayesiano de modelos estadísticos a través de métodos de Monte Carlo vía cadenas de Markov.

Ciudad	1	2	3	4	5	6	7	8	9
Cleveland	71	1052	1258	440	488	159	523	169	268
Cincinnati	52	786	988	270	337	133	378	107	160
Columbus	41	517	715	190	212	91	254	77	137

Tabla 1: Muertes de cáncer en tres grandes ciudades de Ohio en 1989 (*National Center for Health Statistics*, 1990). Los sitios principales del tumor son: 1 = cavidad oral; 2 = órganos digestivos y colon; 3 = pulmón; 4 = seno; 5 = genitales; 6 = órganos urinarios; 7 = otros y sitios no especificados; 8 = leucemia; y 9 = tejido linfático.

la media correspondiente:

$$X \sim BN(x|\pi,\alpha) \text{ Verosimilitud} \quad \Rightarrow p(x) = \binom{\alpha+x-1}{x}\pi^{\alpha}(1-\pi)^{x},\ x=0,1,\ldots,\ \alpha>0;$$

$$\pi \sim Beta(\pi|a,b) \text{ Inicial} \qquad \Rightarrow p(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1},\ \pi\in(0,1),\ a,b>0.$$

Una primera prueba para analizar una tabla de dimensión $r \times c$ es la descrita por Gutiérrez-Peña (2005) que considera la distribución final de π sin restricciones y la compara con la distribución final de π bajo la hipótesis nula $\pi = \pi^0 \equiv (\pi^0_{11},\ldots,\pi^0_{rc})'$ donde π_{ij} denota la probabilidad de la celda (i,j) y $\pi^0_{ij} = \pi_{i+}\pi_{+j}$.

El modelo nulo de independencia puede probarse con base en la distribución final de:

$$\delta = \delta(\pi) \equiv \sum_{ij} \log\left(\frac{\pi_{ij}}{\pi^0_{ij}}\right)\log(\pi_{ij}).$$

Esta cantidad puede considerarse como una versión Bayesiana de la devianza; es siempre no negativa y es cero si y sólo si el modelo nulo y el modelo saturado son el mismo, es decir, si y sólo si $\pi^0_{ij} = \pi_{ij}$ para toda $i = 1,\ldots,r$ y $j = 1,\ldots,c$.

Las distribuciones finales de δ concentradas alrededor del cero apoyan el modelo nulo, mientras que las distribuciones finales localizadas lejos del cero conducen a rechazarlo.

En el caso de los datos de la Tabla 1 el valor medio de δ es 6.5 y los cuantiles del 2.5 % y 97.5 % son 5.14 y 8.27, respectivamente. Entonces se puede concluir que las variables no son independientes, por lo que existe una relación entre las incidencias de los tipos de cáncer de las ciudades, es decir, si hay mayor incidencia de una enfermedad en una ciudad, entonces se espera también una mayor incidencia de las otras enfermedades.

3. Modelos para datos de series de tiempo

Los modelos de series de tiempo autoregresivos de valores enteros (INAR) se presentan en Al-Osh *et al.* (1987, 1990) y McKenzie (1988).

3.1. INAR(1)

Sea G una v.a. discreta definida sobre valores enteros no negativos, y sea H_j una sucesión de v.a. binarias i.i.d., independientes de G, tal que $p(H_j = 1) = 1 - p(H_j = 0) = \alpha$ donde $\alpha \in [0, 1]$. El *proceso de adelgazamiento binomial* está definido por:

$$\alpha \circ G = \sum_{j=1}^{G} H_j = B(\alpha, G),$$

donde $B(\alpha, G)$ es una v.a. binomial que se basa en G ensayos con probabilidad de éxito α.

Definimos un proceso INAR(1) $\{Y_t;\ t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ como:

$$Y_t = \alpha \circ Y_{t-1} + I_t$$

donde $\alpha \in [0, 1]$ y I_t es una sucesión de variables aleatorias con valores enteros no negativos no correlacionados, con media μ y varianza σ^2.

Este modelo simplemente establece que los componentes del proceso al tiempo t, Y_t, son: (i) los supervivientes de los elementos del proceso al tiempo $(t-1)$, Y_{t-1}, cada uno con probabilidad de supervivencia α y (ii) elementos que entran al sistema en el intervalo $(t-1, t]$ como términos de innovación (I_t).

3.2. Modelos INAR(1)-Poisson

Un proceso INAR(1) con distribuciones marginales Poisson(λ_t) se obtiene cuando Y_0 y I_t son independientes con distribución Poisson y con parámetros λ_0 y $\iota_t = \lambda_t - \alpha_t \lambda_{t-1}$, respectivamente.

Cuando Y_0 y I_t son variables aleatorias independientes con distribución Poisson con parámetros λ y $\iota_t = \iota = \lambda(1 - \alpha)$, entonces $\{Y_t\}$ es un proceso de Markov estacionario con distribuciones marginales Poisson.

3.3. Modelos INAR(1)-binomial negativa

El proceso INAR(1)-BN para $\{Y_t\}$ se define como $Y_t = \Pi \circ Y_{t-1} + I_t$, donde Π sigue una distribución beta con parámetros $\alpha\theta$ y $(1-\alpha)\theta$, $(0 < \alpha < 1)$, y Y_0 y I_t son independientes con distribución BN(θ,β) y BN($\theta(1-\alpha)$,β), respectivamente.

Una forma más simple de obtener el proceso INAR(1)-BN es a través del proceso INAR(1)-Poisson permitiendo que el parámetro del componente de innovación varíe de acuerdo a una distribución gamma con parámetros $\theta^I = \theta(1-\alpha)$ y β y la probabilidad del proceso de adelgazamiento binomial de acuerdo a una distribución beta con parámetros $\theta^C = \alpha\theta$ y θ^I.

3.4. Modelos INAR(1)-multiomial negativa

La extensión del proceso INAR(1)-BN a R eventos es sencilla cuando los parámetros de cada uno de los eventos varían de acuerdo a la misma distribución gamma con parámetros θ y β_r $(r = 1, \ldots, R)$. En este caso la distribución conjunta de las variables de conteo es una distribución multinomial negativa con R variables (MN), la cual puede factorizarse en una distribución multinomial y una binomial negativa.

4. Ejemplo

Zeger (1988) presenta una lista del número de casos de poliomielitis reportados mensualmente por el *U.S. Centers for Disease Control* para los años de 1970 a 1983 y que fueron publicados en *Morbidity and Mortality Weekly Report Annual Summary.*

Para el ajuste del modelo INAR(1)-BN se requieren los parámetros de la distribución binomial negativa (índice r y probabilidad π). Con este fin, se realizó un primer programa en WinBUGS, que después de 100,000 iteraciones produjo los valores $\hat{r} = 1.211$ y $\hat{\pi} = 0.4705$ como estimadores de estos parámetros. Posteriormente se realizó otro programa en WinBUGS para ajustar el modelo INAR(1)-BN (ver Tabla 2).

Por tanto el modelo queda definido de la siguiente manera:

$$Y_t = \Pi \circ Y_{t-1} + I_t \qquad t = 1, \ldots, 168,$$

donde $Y_0 \sim Poisson(\lambda)$, $I_t \sim Poisson(\lambda^I)$, independientes, y $\Pi \sim Beta(\theta^C, \theta^I)$ y $\lambda^I \sim Gamma(\theta^I, \beta)$, con $\theta^C = \theta\alpha$, $\theta^I = \theta(1-\alpha)$, $\beta = (1-\pi)/\pi$ y $\theta = r$, donde, por los ajustes

node	mean	sd	MC error	2.5 %	median	97.5 %	start	sample
alpha	0.3947	0.06674	4.075E-4	0.2589	0.3964	0.5202	1001	100000
thetaC	0.478	0.08082	4.935E-4	0.3135	0.4801	0.63	1001	100000
thetaI	0.733	0.08082	4.935E-4	0.581	0.7309	0.8975	1001	100000

Tabla 2: Resultados obtenidos por el programa WinBUGS para el ajuste del modelo INAR(1)-BN para los datos de poliomielitis de E.U. de 1970 a 1983.

de la distribución binomial negativa, se tomó $r = 1.211$ y $\pi = 0.4705$.

Debido a que la observación de noviembre de 1972 es muy alta comparada con los demás valores ($y_{15} = 14$), podría considerarse como un valor atípico, por lo que se realizó el mismo ajuste sin esta observación (ver Figura 1).

Las distribuciones finales correspondientes son semejantes, por lo que se concluye que el efecto de la observación de noviembre de 1972 es mínimo.

5. Conclusiones

Los modelos más utilizados en el análisis de datos categóricos se basan en las distribuciones multinomial y Poisson. Sin embargo, la distribución multinomial negativa es una mejor alternativa en situaciones donde los datos presentan correlaciones positivas y/o sobredispersión.

Referencias

Al-Osh, M. A. & Alzaid, A. A. (1987). "First-Order Integer Valued Autoregressive (INAR(1)) Process". *Journal of Time Series Analysis*, 8, 261-275.

Al-Osh, M. A. & Alzaid, A. A. (1990). "An Integer-Valued pth-Order Autoregressive Structure (INAR(p)) Process". *Journal of Applied Probability*, 27, 314-324.

Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley, England.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley, England.

Good, I. J. (1967). "A Bayesian Significance Test for Multinomial Distributions". *Journal of the Royal Statistical Society. Series B (Methodological)*, 29, 339-431.

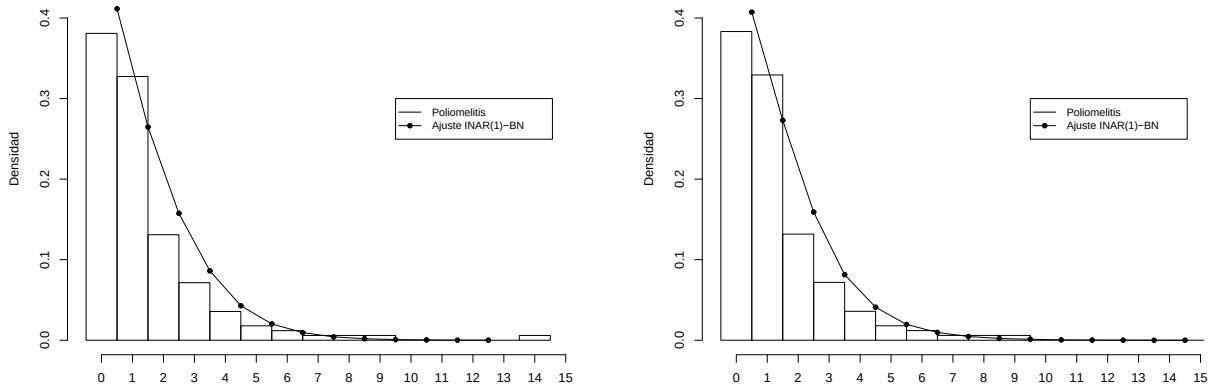


Figura 1: Histograma del número de los casos mensuales de poliomielitis de E.U. de 1970 a 1983 y ajuste INAR(1)-BN con y sin la observación de noviembre de 1972.

Good, I. J. (1976). "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables". *The Annals of Statistics*, 4, 1159-1189.

Gutiérrez-Peña, E. (2005). "Bayesian Methods for Categorical Data". *Encyclopedia of Statistics in Behavioral Science*, 1, 139-146.

Lindley, D. V. (1964). "The Bayesian Analysis of Contingency Tables". *The Annals of Mathematical Statistics*, 35, 1622-1643.

McKenzie, E. (1988). "Some ARMA Models for Dependent Sequences of Poisson Counts". *Advances in Applied Probability*, 20, 822-835.

Spiegelhalter, D., Best, N., Carlin, B. y van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit". *Journal of the Royal Statistical Society*, Series B, 64, 583-639.

Waller, L. A. y Zelterman, D. (1997). "Log-Linear Modeling with the Negative Multinomial Distribution". *Biometrics*, 53, 971-982.

Zeger, S. L. (1988). "A Regression Model for Time Series of Counts". *Biometrika*, Vol. 75, 4, 621-629.

Lista de árbitros

El Comité Editorial de la Memoria del XXIII Foro Nacional de Estadística agradece la valiosa colaboración de los siguientes árbitros:

1. Alegría Hernández, Alejandro. *FLACSO*
2. Barry, Arnold. *University of California, Riverside*
3. Contreras Cristán, Alberto. *IIMAS – UNAM*
4. Díaz Ávalos, Carlos. *IIMAS – UNAM*
5. Eslava Gómez, Guillermina. *Facultad de Ciencias – UNAM*
6. Fuentes García, Ruth S. *Facultad de Ciencias – UNAM*
7. Félix Medina, Martín. *UAS*
8. González Barrios Murguía, José María. *IIMAS – UNAM*
9. Gracia-Medrano Valdelamar, Leticia. *IIMAS – UNAM*
10. Hernández Cid, Rubén. *ITAM*
11. Melaré Vieira Barros, Daniela. *University of Algarve*
12. Méndez Ramírez, Ignacio. *IIMAS – UNAM*
13. Mendoza Ramírez, Manuel. *ITAM*
14. Nakamura Savoy, Miguel. *CIMAT*
15. Nieto Barajas, Luis E. *ITAM*
16. O'Reilly Togno, Federico. *IIMAS – UNAM*

17. Pérrz Salvador, Blanca Rosa. *UAM*
18. Romero Mares, Patrica. *IIMAS –UNAM*
19. Ruiz-Velasco Acosta, Silvia. *IIMAS – UNAM*
20. Soto de la Rosa, Humberto/ *ONU*
21. Téllez Rojo Solís, Martha María *INSP*
22. Villaseñor Alva, José. *Colegio de Postgraduados*