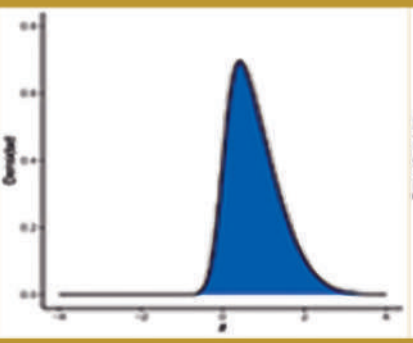
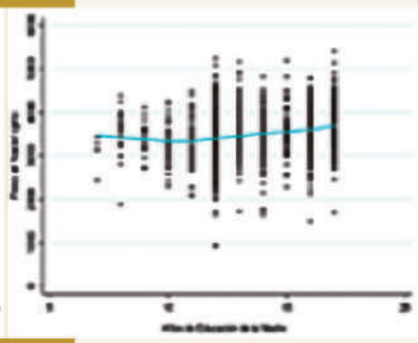
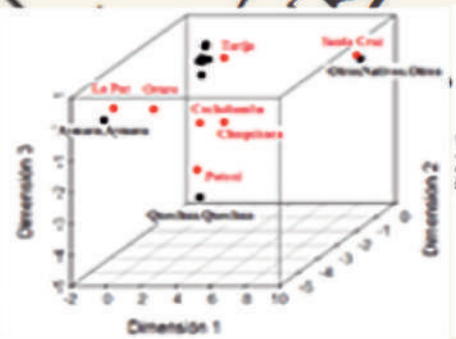
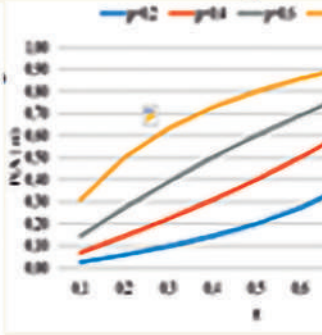
Universidad Mayor
de San Andrés

Varianza

Revista del Instituto de Estadística Teórica y Aplicada

CARRERA DE ESTADISTICA
UMSA - FCPN

UMSA
FCPN
CARRERA
ESTADÍSTICA

IETA
Instituto de Estadística
Teórica y Aplicada

Varianza

Revista de la Carrera de Estadística

Publicación del Instituto de Estadística Teórica y Aplicada

UMSA
FCPN
CARRERA
ESTADÍSTICA

IETA

Instituto de Estadística
Teórica y Aplicada

Número 17

Noviembre, 2020

ISSN 9876-6789
REVISTA VARIANZA
Nº 17 - Noviembre, 2020

PRESENTACIÓN

La Dirección Interina del Instituto de Estadística Teórica y Aplicada (IETA) tiene la satisfacción de presentar a estudiantes y docentes de la Carrera de Estadística, así como a la población en general interesada en temas estadísticos, la REVISTA VARIANZA N° 17.

En esta nueva edición de la revista se exhiben siete artículos científicos, algunos vinculados a la parte teórica de la estadística y otros relacionados con la parte aplicada. Ordenados alfabéticamente, en el primer artículo, luego de adoptar una distribución tipo G para la variación del índice de precios al consumidor, se procede a estimar los parámetros involucrados en dicha distribución mediante procedimientos bayesianos; en el segundo se platea el problema de estimación del parámetro de forma en un modelo normal asimétrico, un problema que surge principalmente en pequeños tamaños de muestra; en el tercero se usa el modelo de regresión logística con interceptos aleatorios para analizar datos de panel, un modelo que resulta ser un caso particular del modelo lineal generalizado de efectos mixtos; en el cuarto se realiza un análisis multivariado de la correspondencia entre un conjunto de variables categóricas usando el método de la descomposición tensorial tucker3, un método alternativo al de análisis de correspondencia clásico; el quinto tiene que ver con la utilidad del cálculo complejo en algunos aspectos de la teoría estadística; en el sexto se comparan las precisiones en las estimaciones de algunos parámetros poblacionales considerando tres planes de muestreo; y en el séptimo artículo se aplica el modelo de respuesta aleatorizada de Warner, un modelo usado cuando las preguntas formuladas a los entrevistados son delicadas.

Se agradece de sobremanera a cada uno de los autores de los artículos por su esfuerzo y contribución para que se concrete esta nueva publicación de la revista varianza. De lo contrario, la continuidad anual de la publicación no hubiese sido posible.

Dr(c) Ramiro Coa Clemente
DIRECTOR a.i.
INSTITUTO DE ESTADISTICA TEÓRICA Y APLICADA

Carrera de Estadística
Instituto de Estadística Teórica y Aplicada (I.E.T.A.)
Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés

La Paz - Bolivia
Edificio Antiguo - Planta Baja
Telefax: 2442100 -2612844
Correos:estadistica@umsa.bo - ieta@umsa.bo

Dedicado a docentes que hacen
investigación científica y a estudiantes
que quieren seguir este camino

Contenido

ESTIMACIÓN DE LOS PARÁMETROS DE LA DISTRIBUCIÓN DE CAMBIOS DE PRECIOS MENSUALES DEL ÍNDICE DE PRECIOS AL CONSUMIDOR DE BOLIVIA

Mgtr. Iván Yony Aliaga Casceres[*] & Dr.(c) Omar Chocotea Poca[**]

✉ *iyaliaga@umsa.bo, powervan@gmail.com*
✉ *ochocotea@umsa.bo*

RESUMEN

La presente investigación estima los parámetros de localización, asimetría y los grados de libertad de una distribución tipo G mediante un método de estimación de libre verosimilitud Bayesiana comparada con el clásico método Metropolis Hastings, utilizando para ello datos de la distribución de la variación mensual del Índice de Precios al Consumidor del Instituto Nacional de Bolivia, periodo ene-1986 a oct-2020, base 2016. Se evidenció que el proceso iterativo algorítmico para el método Metropolis Hastings es más eficiente, sin embargo, el mejor ajuste se encontró con el método ABC-GIBBS.

PALABRAS CLAVE

Analisis Bayesiano, Estimación de libre verosimilitud, Distribución tipo G.

ABSTRACT

The present investigation estimates the location parameters, asymmetry and the degrees of freedom of a G-type distribution using a Bayesian free likelihood estimation method compared to the classic Hastings Metropolis method, using data from the distribution of the monthly variation of the Index of Consumer Prices of the National Institute of Bolivia, period Jan-1986 to Oct-2020, base 2016. It was evidenced that the algorithmic iterative process for the Metropolis Hastings method is more efficient, however, the best fit was found with the ABC method -GIBBS.

KEYWORDS

Bayesian analysis, Free likelihood estimation, Type G distribution.

1. INTRODUCCIÓN

La familia de distribución tipo G fue estudiada por *Marcus* (*1987*), *Andrews et al.* (*1974*), *y Rosinski* (*1991*), es el producto entre dos variables aleatorias independientes, la primera con distribución Normal con varianza constante y la segunda con una deter- minada distribución de probabilidad cuyo espacio soporte está definido en los reales positivos.

La estimación de los parámetros de la familia de distribuciones tipo G es una tarea tediosa, el problema radica en la función de verosimilitud, los métodos de optimización estándar fallan al estimar todos los parámetros y el costo computacional es bastante alto, *Barndorff et al.* (*1981*), *Andrews et al.* (*1974*).

Bajo un enfoque Bayesiano, un parámetro es visto como una variable aleatoria a la que antes de extraer alguna evidencia muestral, se le asigna una distribución previa, con base a un cierto grado de creencia con respecto al comportamiento aleatorio, cuando se obtiene la evidencia muestral surge una actualización en base a la distribución previa y la evidencia muestral, de esta manera se obtiene una nueva distribución llamada distribución posterior.

Existen varios métodos en el ámbito

[*] Carrera de Estadística, Universidad Mayor de San Andrés

[**] Instituto de Estadística, Universidad de Valparaíso

Bayesiano que aproximan la distribución posterior, en este aspecto se presenta los llamados *métodos de verosimilitud libre*, Rubin (1984), Diggle et al. (1984), Tavare et al. (1997), Beaumont et al. (2002), Ratmann et al. (2009), que aproximan de forma iterativa a la distribución posterior teórica resultante.

El término *Cálculo Bayesiano Aproximado*(ABC) fue establecido por los autores: Beaumont et al. (2002), quienes amplían aún más la metodología de los métodos de verosimilitud libre. El ABC es una técnica de aproximación a la distribución posterior de los parámetros de manera iterativa cuando la misma no tiene una función de verosimilitud explícita. Existen diferentes métodos de verosimilitud libre propuestos por diferentes autores: ABC-MCMC, Marjoram et al. (2003), ABC-GIBBS, Turner et al. (2012), Turner & Sederberg (2014). Este último es utilizado en esta investigación.

Los métodos: Metropolis Hastings, Hastings (1970), y ABC-GIBBS, Turner et al. (2012), Turner & Sederberg (2014), son aplicados en la estimación de los parámetros de la distribución de cambios de precios mensuales del Indice de Precios al consumidor de Bolivia desde ene-1986 hasta oct-2020, distribución teórica que presenta asimétria y leptocúrtosis, Blattberg et al. (1974), Kyprianou (2014).

2. MARCO TEÓRICO

2.1. Enfoque Bayesiano

Definición 2.1 ***Teorema de Bayes.*** *Sean* $\boldsymbol{X}$ *y* $\boldsymbol{\theta}$ *variables aleatorias con función de probabilidad* $f(x|\boldsymbol{\theta})$ *y* $\pi(\boldsymbol{\theta})$*, entonces la distribución posterior está dada por:*

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

donde $\boldsymbol{\theta}$ *vector de parámetros y* $\boldsymbol{\Theta}$ *espacio paramétrico.*

Para aproximar $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ se recurre a una serie de métodos aproximados iterativos.

2.2. Método Metropolis Hastings

Este método es una generalización del método Hastings, Metropolis et al. (1953), utiliza una función de densidad propuesta que depende del estado actual de $\boldsymbol{\theta}^{(t)}$. La función de densidad $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ puede ser tan simple como una función de densidad Normal localizada en $\boldsymbol{\theta}^{(t)}$.

Algoritmo 1 Algoritmo Metropolis-Hastings

Entrada: Dar un valor inicial para $\theta^{(0)}$, con $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ función objetivo.
Salida: Distribución posterior de $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta}|\boldsymbol{X})$.

1: **para** $j = 1, 2 \ldots$ **hacer**
2: Generar $\theta^* \sim q(\theta|\theta^{(j)})$
3: Calcular la probabilidad de aceptación.

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}) = min\left\{1, \frac{f(\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}^{(j)})}\frac{q(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j)})}\right\}.$$

4: Generar un número aleatorio $u \sim U(0, 1)$.
5: **si** $u \leq \alpha(\theta^*, \theta^{(j)})$ **entonces**
6: Aceptar θ^*
7: **devolver** $\theta^{(j+1)} = \theta^*$
8: **si no**
9: Rechazar θ^*
10: **devolver** $\theta^{(j+1)} = \theta^{(j)}$
11: **fin si**
12: **fin para**

2.3. Método ABC-GIBBS

El Algortimo ABC-GIBBS fue propuesto por Turner & Zandt (2014), los autores proponen extraer muestras aleatorias directamente de la distribución posterior condicional de los hiperparámetros[1] que no dependan de la función de verosimilitud utilizando el método de Gibbs, Voss (2013), Beaumont et al. (2002), Marjoram et al. (2003), Ratmann et al. (2009), Sisson et al. (2007),Turner &

[1] Los hiperparámetros son parámetros de una distribución previa

Zandt (2014), Gelman et al. (2013).

Definición 2.2 *Sea* $= (\theta_1, \theta_2, \ldots, \theta_p)$ *el vector de parámetros, bajo un nivel de jerarquía de estos parámetros se tiene la siguiente notación para los hiperparámetros, de cada j-ésimo parámetro.*

$$\xi = (\theta_{11}^{+}, \ldots, \theta_{1m}^{+}, \theta_{21}^{+}, \ldots \theta_{2m}^{+}, \ldots, \theta_{p1}^{+}, \ldots, \theta_{pm}^{+}) \quad (1)$$
$$m = 1, 2, 3, \ldots \quad j = 1, 2, \ldots, p.$$

Ambos vectores pueden representarse como:

$$\eta = (\xi, \theta) = (\theta_{11}^{+}, \ldots, \theta_{1m}^{+}, \theta_{21}^{+}, \ldots \theta_{2m}^{+}, \ldots, \theta_{p1}^{+}, \ldots, \theta_{pm}^{+}, \theta_1, \ldots, \theta_p) \quad (2)$$

El vector de hiperparámetros y parámetros sin un elemento k-ésimo y j-ésimo respectivamente se denota por:

$$\xi_{-k} = (\xi_1, \ldots, \xi_{k-1}, \xi_{k+1}, \ldots, \xi_M), \quad M = m * p.$$
$$\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p). \quad (3)$$

Utilizando el muestreo de Gibbs para obtener un gran número de muestras de 3 y es denotado por:

$$\xi_k \sim \pi(\xi|\xi_{-k}, \theta), \theta_j \sim \pi(\theta|\theta_{-j}, \xi), j = 1, \ldots, p \quad (4)$$

La implementación del método ABC-GIBBS se realiza en dos etapas, la primera etapa consiste de hallar la distribución condicional posterior de los hiperparámetros, tomando en cuenta que ξ influye a $\boldsymbol{X}$ solo a través de θ, la distribución condicional posterior de ξ no depende de la verosimilitud π(X,Z|θ, ξ) ya que es constante respecto a ξ.

$$\begin{aligned}\pi(\xi|\boldsymbol{X}, \boldsymbol{Z}, \theta) &\propto \pi(\boldsymbol{X}|\boldsymbol{Z}, \theta, \xi)\pi(\boldsymbol{Z}|\theta, \xi)\pi(\theta|\xi)\pi(\xi) \\ &\propto \pi(\boldsymbol{X}, \boldsymbol{Z}|\theta, \xi)\pi(\theta|\xi)\pi(\xi) \\ &\propto \pi(\theta|\xi)\pi(\xi) \end{aligned} \quad (5)$$

En la segunda etapa del método ABC-GIBBS comprende en hallar la distribución posterior de los parámetros, y debe tomarse en cuenta la dependencia condicional de θ y $\boldsymbol{X}$, la densidad del parámetro ξ es constante con respecto a θ.

$$\begin{aligned}\pi(\theta|\boldsymbol{X}, \boldsymbol{Z}, \xi) &\propto \pi(\boldsymbol{X}|\boldsymbol{Z}, \theta, \xi)\pi(\boldsymbol{Z}|\theta, \xi)\pi(\theta|\xi) \\ &\propto \pi(\boldsymbol{X}, \boldsymbol{Z}|\theta, \xi)\pi(\theta|\xi) \\ &\propto L(\theta, \xi|\boldsymbol{X}, \boldsymbol{Z})\pi(\theta|\xi) \end{aligned} \quad (6)$$

Algoritmo 2 Algoritmo ABC-GIBBS

Entrada: Inicialización para $\theta_{0,j}, \xi_{0k}$, $j = 1, \ldots, p$; $k = 1, \ldots M$ y $\boldsymbol{Z}_0$, datos $\boldsymbol{X} \in \mathbb{R}^n$.
Salida: Distribución posterior de θ: $\pi(\theta|\boldsymbol{X}, \boldsymbol{Z}, \xi)$.

para $i = 1, 2 \ldots$ **hacer**
  para $j = 1, \ldots, p$; $k = 1, 2, \ldots, M$ **hacer**
    Generar ξ^* de $\pi(\xi|\xi_{-k}, \theta)$.
    Generar θ^* de $\pi(\theta|\theta_{-j}, \xi)$.
  fin para
  para $j = 1, \ldots, p$; $k = 1, 2, \ldots, M$ **hacer**
    Generar conjunto de datos $\boldsymbol{X}^* \sim f(\boldsymbol{X}, \boldsymbol{Z}_{i-1}|\theta^*, \xi^*)$.
    Generar una variable aleatoria $\boldsymbol{U} \sim Unif(0, 1)$
    si $\rho(S(\boldsymbol{X}^*), S(\boldsymbol{X})) \leq \epsilon$ y $\boldsymbol{U} < min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta_{i-1,j})} \frac{\pi(\boldsymbol{X}^*, \boldsymbol{Z}_{i-1}|\theta^*, \xi^*)}{\pi(\boldsymbol{X}, \boldsymbol{Z}_{i-1}|\theta^*, \xi^*)}\right\}$ **entonces**
      Generar $\boldsymbol{Z}_i$ de $\pi(\boldsymbol{Z}|\theta^*, \xi^*)$.
      devolver $(\theta_{i,j}, \xi_{i,j}) = (\theta^*, \xi^*)$
    si no
      devolver $(\theta_{i,j}, \xi_{i,j}) = (\theta_{i-1,j}, \xi_{i-1,k})$
    fin si
  fin para
fin para

2.4. Distribuciones tipo G

Definición 2.3 *Una variable aleatoria* $\boldsymbol{X}$ *con función de distribución de probabilidad* $\boldsymbol{F}$ *es infinitamente divisible (i.d.) si para todo n*$\epsilon \mathbb{N}$ *existe n variables aleatorias independientes e idénticamente distribuidos* $X_{n,1}, X_{n,2}, .., X_{n,n}$ *tal que.*

$$\begin{aligned}\boldsymbol{X} &\stackrel{d}{=} \boldsymbol{X}_{n,1} + \boldsymbol{X}_{n,2} + \ldots + \boldsymbol{X}_{n,n} \\ \boldsymbol{X}_{n,j} &\stackrel{d}{=} \boldsymbol{X}_n \quad \forall j = 1, \ldots, n \end{aligned} \quad (7)$$

donde X_n *representa la variable aleatoria con distribución factor común.*

Definición 2.4 *Una variable aleatoria continua* $\boldsymbol{X}$ *tiene distribución de probabilidad tipo G, si está definida como:*

$$\boldsymbol{X} \overset{d}{=} \mu + \beta \boldsymbol{M} + \boldsymbol{M}\boldsymbol{W} \tag{8}$$

donde $\boldsymbol{M}$ *y* $\boldsymbol{W}$ *son variables aleatorias independientes,* $\boldsymbol{W}$ *tiene distribución normal con varianza constante,* $\boldsymbol{M}$ *es una variable aleatoria con una determinada distribución de probabilidad con espacio soporte en los reales positivos y cumple la propiedad de infinita divisibilidad, el parámetro* μ *es el parámetro de localización,* β *es el parámetro de simetría o asimetria.*

La variable aleatoria M tiene distribución Gaussiana Inversa Generalizada y cumple la propiedad infinita divisibilidad, Barndorff et al. (1977), Grosswald (1976).

2.5. Distribución Gaussiana Inversa Generalizada

Definición 2.5 *Una variable aleatoria X tiene distribución Gaussiana Inversa Generalizada (GIG) con parámetros* λ, δ, γ, *si la función de densidad está dado por.*

$$f_X(x) = \frac{(\delta/\gamma)^{\lambda/2}}{2\mathrm{K}_\lambda(\sqrt{\gamma\delta})} x^{\lambda-1} exp\left\{-\frac{1}{2}\left(\delta x + \frac{\gamma}{x}\right)\right\}, \quad x > 0. \tag{9}$$

donde $\mathrm{K}_\lambda(\cdot)$ *es la función Bessel de tercer orden y* $(\gamma,\delta) \in \Theta_\lambda$.

$$\Theta_\lambda = \begin{cases} \{(\gamma,\delta) : \gamma \geq 0, \delta > 0\}, & Si\ \lambda > 0 \\ \{(\gamma,\delta) : \gamma > 0, \delta > 0\}, & Si\ \lambda = 0 \\ \{(\gamma,\delta) : \gamma > 0, \delta \geq 0\}, & Si\ \lambda < 0. \end{cases}$$

Proposición 2.1 *Si* $M \sim GIG(\lambda=-\nu/2, \delta, \gamma=\nu)$, $\nu > 0$ y $\delta \to 0^+$ *entonces* $\boldsymbol{M}$ *tiene distribución Gamma Inversa con parámetros* $(\nu/2, \nu/2)$.

Para obtener la distribución previa y la distribución posterior se identifica la función conjunta de $\boldsymbol{X}$ y $\boldsymbol{M}$, Hichen (2006).

$$f(\boldsymbol{x}|\boldsymbol{M}) = (2\pi)^{-1/2}\boldsymbol{m}^{-1/2} exp\left\{-\frac{1}{2\boldsymbol{z}}(\boldsymbol{x} - (\mu + \beta\boldsymbol{m})^2)\right\}$$
$$\boldsymbol{M} \sim GIG(\lambda, \delta, \gamma) \tag{10}$$

Realizando un cambio de variable en términos de *a* y *b*.

$$a = \frac{\delta}{\gamma};\ b = \gamma\delta;\ \frac{a}{b} = \frac{1}{\delta^2};\ ab = \delta^2;\ \frac{b}{a} = \gamma^2 \tag{11}$$

Se tiene la siguiente la función previa bivariante con a_0, a_1, a_2, a_3, a_4 hiperparámetros.

$$\pi(\mu,\beta) \sim N_2\left(\boldsymbol{m} = \begin{pmatrix} \bar{\mu} \\ \bar{\beta} \end{pmatrix}, \boldsymbol{R} = \begin{pmatrix} \sigma_\mu^2 & \rho\sigma_\mu\sigma_\beta \\ \rho\sigma_\mu\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right) \tag{12}$$

donde:

$$\bar{\mu} = \frac{1}{2(1-\rho^2)a_4}\left(a_2 - \frac{a_0 a_1}{2a_3}\right)$$
$$\bar{\beta} = \frac{1}{2(1-\rho^2)a_3}\left(a_1 - \frac{a_2 a_0}{2a_4}\right)$$
$$\rho = \frac{a_0}{2\sqrt{a_4 a_3}}$$
$$\sigma_\mu^2 = \frac{1}{2(1-\rho^2)a_4}$$
$$\sigma_\beta^2 = \frac{1}{2(1-\rho^2)a_3}$$

Para la distribución previa conjunta $\pi(\lambda, a, b)$ se asume que tiene distribución conjugada a la distribución GIG, esto por razones analíticas.

$$b \sim \Gamma(\zeta, \chi);\ \ a|b \sim GIG\left(-1/2, \sqrt{b\omega\psi}, \sqrt{b\omega/\psi}\right) \tag{13}$$

Para obtener la distribución posterior conjunta de los parámetros μ, β y λ, a, b se realiza de forma separada al ser independientes ambos conjuntos de parámetros por construcción de la distribución tipo G.

$$\pi(\mu,\beta,\lambda,a,b|\boldsymbol{X},\boldsymbol{Z}) \propto \pi(\mu,\beta|\boldsymbol{X},\boldsymbol{Z})\pi(\lambda,a,b|\boldsymbol{Z}) \tag{14}$$

$$\pi(\mu,\beta|\boldsymbol{X},\boldsymbol{Z}) \sim N_2\left(\begin{pmatrix} \mu \\ \beta \end{pmatrix}; \boldsymbol{m}_g, \boldsymbol{D}_g\right) \tag{15}$$

donde:

$$
\begin{aligned}
m_g &= D_g(R_d^{-1}m_d + R_{prior}^{-1}m_{prior}) \\
D_g &= (R_d^{-1} + R_{prior}^{-1})^{-1} \\
m_d &= R_d \begin{pmatrix} \sum x_i/z_i \\ \sum x_i \end{pmatrix} \\
R_d^{-1} &= \begin{pmatrix} \sum z_i^{-1} & n \\ n & \sum z_i \end{pmatrix}
\end{aligned}
$$

$$
\lambda|a,b \propto \frac{\left(a^n \prod_{i=1}^n z_i\right)^\lambda}{K_\lambda(b)^n}
$$

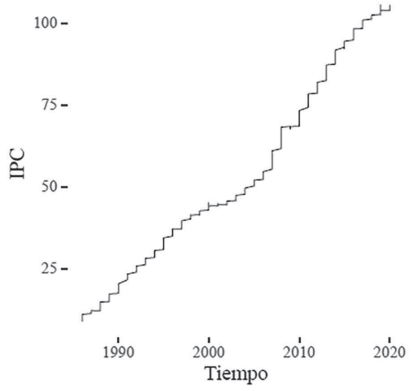
$$
a|\lambda,b \sim GIG\left(n\lambda - 1/2, \sqrt{b\left(\sum_{i=1}^n z_i + \omega\psi\right)}, \sqrt{b\left(\sum_{i=1}^n \frac{1}{z_i} + \frac{\omega}{\psi}\right)}\right)
$$

$$
b|\lambda,a \propto \frac{b^{\zeta-1}}{K_\lambda(b)^n} \exp\left\{-\frac{b}{2}\left(\frac{1}{a}\sum_{i=1}^n \frac{1}{z_i} + a\sum_{i=1}^n z_i + 2\chi\right)\right\} \tag{16}
$$

3. RESULTADOS

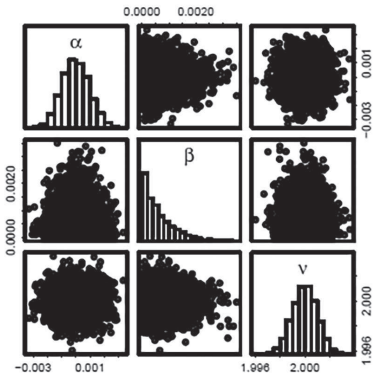
Se aplican los dos métodos anteriormente descritos en la estimación de parámetros de la distribución de cambios de precios (varianciones mensuales) del Índice de Precios al Consumidor con periodicidad desde Ene-1986 hasta Oct-2020, con base 2016=100 del Instituto Nacional de Estadística de Bolivia[2].

Figura 1: Índice de Precios al consumidor, frecuencia mensual, base 2016=100. Fuente: INE-Bolivia



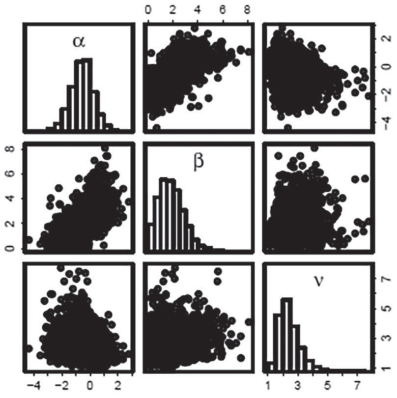
Fuente:Elaboración Propia

Figura 2: Valores de la distribución posterior de los parámetros μ,β yν, bajo el método Metropolis Hastings, sin valores de quemado.



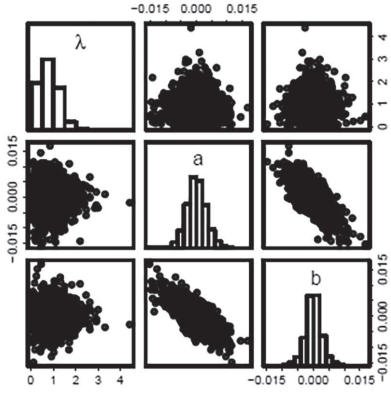
Fuente:Elaboración Propia

Figura 3: Valores de la distribución posterior de los parámetros μ,β,ν, bajo el método ABC-GIBBS, sin valores de quemado



Fuente:Elaboración Propia

Figura 4: Valores de la distribución posterior de los parámetros λ, a y b bajo el método ABC-GIBBS, sin valores de quemado.



Fuente:Elaboración Propia

[2] https://www.ine.gob.bo/index.php/serie-historicaempalmada/

Cuadro 1: Tabla de resultados y diagnóstico de los valores de la distribución posterior de los parámetros

Método/diag.	Estimación					
	$\hat{\mu}$	$\hat{\beta}$	$\hat{\nu}$	$\hat{\lambda}$	$\hat{a}$	$\hat{b}$
METROPOLIS H.	$0{,}12 \cdot 10^{-4}$	$6{,}1 \cdot 10^{-4}$	2			
(4.5m.)/desv.	(9.6)	(5.6)	$(9{,}9 \cdot 10^{-4})$			
AIC	-2975					
BIC	-2964					
Geweke	0.9871	0.8818	0.9820			
Heidelberger	0.6420	0.1830	0.6580			
ABC-GIBBS	$-0{,}55 \cdot 10^{-4}$	$47{,}8 \cdot 10^{-4}$	2.45	$7{,}8 \cdot 10^{-1}$	$-3{,}6 \cdot 10^{-5}$	$3{,}9 \cdot 10^{-4}$
(11.25m.)/desv.	(0.82)	(1.07)	(0.76)	(0.45)	(0.03)	(0.03)
AIC	-3008					
BIC	-2998					
Geweke	0.1999	0.7224	0.1065	0.2642	0.9814	0.8630
Heidelberger	0.2960	0.6910	0.4700	0.1100	0.2160	0.9790

Fuente:Elaboración Propia

4. CONCLUSIÓN

Se aprecia que el método Metropolis Hastings presenta un menor tiempo de procesamiento en la estimación de los parámetros de la distribución de cambios de precios, sin embargo, el método ABCGIBBS presenta un mejor ajuste que el método Metropolis Hastings en los criterios AIC y BIC.

BIBLIOGRAFÍA

1. Andrews, D., & Mallows, C. (1974). Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological), 36 (1), 99–102. Retrieved from http://www.jstor.org/stable/2984774

2. Barndorff, O., & Blaesild, P. (1981). Hyperbolic Distributions and Ramifications: Contributions to Theory and Application (C. Taillie, G. P. Patil, & B. A. Baldessari, Eds.; pp. 19–44). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-009-8549-0_2

3. Barndorff, O., & Halgreen, C. (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete, 38 (4), 309–311. doi: 10.1007/BF00533162

4. Beaumont, M., Zhang, W., & Balding, D. (2002). Approximate Bayesian Computation in Population Genetics. Genetics, 162 (4), 2025–2035.

5. Blattberg, R., & Gonedes, N. (1974). A comparison of the stable and student distributions as statistical models for stock prices. The Journal of Business.

6. Diggle, P., & Gratton, R. (1984). Monte carlo methods of inference for implicit statistical models. Journal of the Royal Statistical Society. Series B (Methodological).

7. Gelman, A., Carlin, J. B., Stern, H. S., & B., R. D. (2013). Bayesian data analysis (A. C. P. Company, Ed.). Chapman & Hall/CRC Texts in Statistical Science. Retrieved from 24000/e2cd 08b709798adecf5e17aba6b1a033

8. Grosswald, E. (1976). The student t-distribution of any degree of freedom is infinitely divisible. Zeitschrift Fur Wahrscheinlichkeitstheorie Und Verwandte Gebiete, 36 (2), 103–109.

9. Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. Biometrika, 57 (1), 97–109. Retrieved from http://www.jstor.org/stable/2334940

10. Hichen, I., S.and Jerome. (2006). Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures. IEEE Transactions on Signal Processing, 54 (9), 3257–3269. doi: 10.1109/TSP.2006.877660

11. Kyprianou, A. E. (2014). Fluctuations of Lévy Processes with Applications: Introductory Lectures (2nd ed.). Springer-Verlag Berlin Heidelberg.

12. Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without Likelihoods. Proceedings of the National Academy of Sciences, 100 (26), 15324–15328.

13. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21 (6), 1087–1092. doi: 10.1063/1.1699114

14. Ratmann, O., Andrieu, C., Wiuf, C., & Richardson, S. (2009). Model criticism based on likelihoodfree inference, with an application to protein network evolution. Proceedings of the National Academy of Sciences, 106 (26), 10576–10581. doi: 10.1073/pnas.0807882106

15. Rosinski, J. (1991). On A Class of Infinitely Divisible Processes Represented as Mixtures of Gaussian Processes (S. Cambanis, G. Samorodnitsky, & M. S. Taqqu, Eds.; pp. 27–41). Boston, MA: Birkhauser Boston. doi: 10.1007/978-1-4684-6778-9_2

16. Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Statist., 12 (4), 1151–1172. doi:10.1214/aos/1176346785

17. Sisson, S. A., Fan, Y., & Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences, 104 (6), 1760–1765. doi:10.1073/pnas.0607208104

18. Tavare, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. Genetics.

19. Turner, B., & Sederberg, P. (2014). A generalized, likelihood-free method for posterior estimation. Psychonomic Bulletin Review, 21 (2), 227–250. doi: 10.3758/s13423-013-0530-0

20. Turner, B., & Zandt, T. (2012). A tutorial on approximate bayesian computation. Journal of

Mathematical Psychology, 56 (2), 69–85. doi: http://dx.doi.org/10.1016/j.jmp.2012.02.005

21. Turner, B., & Zandt, V. (2014). Hierarchical Approximate Bayesian Computation. Psychometrika, 79(2), 185209. Retrieved from http://doi.org/10.1007/s11336-013-9381-x

22. Voss, J. (2013). An Introduction to Statistical Computing: A Simulation-based Approach (1st ed.). Wiley.

UN PROBLEMA EN LA ESTIMACIÓN DEL PARÁMETRO DE FORMA DEL MODELO NORMAL - ASIMÉTRICO

Dr.(c) Omar Chocotea Poca[1] & Mgtr. Iván Yony Aliaga Casceres[2]

✉ *omar.chocotea@postgrado.uv.cl*

RESUMEN

El modelo normal-asimétrico es adecuado cuando la estructura de los datos presenta una moda y asimetría. En esta nota se observa la existencia de un problema de estimación del parámetro de forma utilizando el método de momentos en tamaños de muestra pequeños contemplando algunos aspectos de *Gupta & Gupta* (*2008*) [Test,17, 197–210].

PALABRAS CLAVE

Modelo asimétrico; problema de estimación

ABSTRACT

The normal-asymmetric model is suitable when the data structure presents a mode and asymmetry. This note shows the existence of an estimation problem of the shape parameter using the method of moments in small sample sizes, contemplating some aspects of *Gupta & Gupta* (*2008*) [Test,17, 197–210].

KEYWORDS

Asymmetric model; estimation problem

1. INTRODUCCIÓN

El interés por la clase de distribuciones normalasimétricay relacionadas ha crecido enormemente, el estudio de sus propiedades continúa, varias de sus propiedades permiten tener por ejemplo las propiedades de la distribución normal (ver *Azzalini, 2014; Arellano-Valle et al., 2018*).

La distribución normal-asimétrica de *Azzalini* (*1985*) contempla un parámetro de asimetría/forma: Una variable aleatoria (v.a.) Z tiene una distribución normal-asimétrica estándar con parámetro de forma $\lambda \in \mathbb{R}$, y representaremos por $Z \sim NA(\lambda)$, si su función de densidad de probabilidad (fdp) está dada por

$$\varphi_\lambda(z)=2\varphi(z)\Phi(\lambda z), \qquad (1)$$

donde $\varphi(\cdot)$ y $\Phi(\cdot)$ denotan la fdp y la función de distribución acumulada de la distribución normal estándar (N(0, 1)). Al asumir $\lambda= 0$ se tiene como caso especial a la distribución $N(0,1)$. La fdp de la distribución $NA(\lambda)$ es una densidad unimodal que está sesgada a la izquierda si $\lambda< 0$, y sesgada a la derecha si $\lambda> 0$. La esperanza y la varianza de $Z \sim NA(\lambda)$, están dadas por

$$E[Z] = \sqrt{2/\pi}\,\delta$$

y

$$Var[Z] = 1 - 2\delta^2/\pi,$$

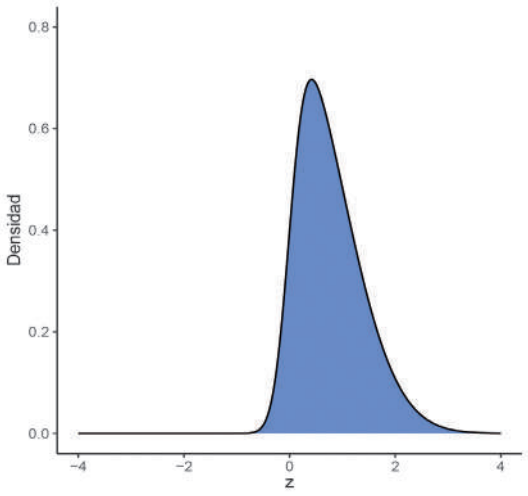
[1] Instituto de Estadística, Universidad de Valparaíso, Chile y Carrera de Estadística, Universidad Mayor de San Andrés, Bolivia

[2] Carrera de Estadística, Universidad Mayor de San Andrés, Bolivia

respectivamente, donde

$$\delta := \delta(\lambda) = \lambda/\sqrt{1+\lambda^2}\ \epsilon\ (-1,1)$$

Figura 1
Fdp de la *NA*(λ) para λ = 4.



Fuente:Elaboración Propia

Si $Z \sim NA(\lambda)$, entonces la v.a. $Y = \xi + \omega Z$ tiene una distribución normal-asimétrica con parámetros de localización $\xi \in \mathbb{R}$, escala $\omega \in \mathbb{R}_+$, y forma $\lambda \in \mathbb{R}$, y representaremos por $Y \sim NA(\xi, \omega, \lambda)$. La fdp de $Y \sim NA(\xi, \omega, \lambda)$ está dada por

$$\varphi_{\xi,\omega,\lambda}(y) = \frac{1}{\omega}\varphi_\lambda((y-\xi)/\omega) \qquad (2)$$

El siguiente resultado proporciona algunas propiedades para la distribución $Z \sim NA(\lambda)$.

Si $Z \sim NA(\lambda)$, entonces las siguientes propiedades son verdaderas: $\varphi_\lambda(0) = \varphi(0)$ para todo λ; $-Z \sim NA(-\lambda)$, equivalente a $\varphi_\lambda(-z) = \varphi_{-\lambda}(z)$ para todo z; y $Z^2 \sim \chi^2_{(1)}$, independientemente de λ.

En la Sección 2, analizamos al estimador de momentos del parámetro de asimetría de la distribución normal-asimétrica, también, en el escenario de una muestra aleatoria extraída de una población normal-asimétrica presentamos la densidad y la función generatriz de momentos (fgm) de la media muestral. En la Sección 3, se efectúa un estudio de simulación para verificar el problema de estimación del parámetro de asimetría vía el método de momento. La Sección 4 ofrece una conclusión.

2. EL ESTIMADOR DE MOMENTO

Cuando $Z \sim NA(\lambda)$, el estimador de momento de λ es la solución de

$$\sqrt{\frac{2}{\pi}}\hat{\delta} = \bar{z} \qquad (3)$$

donde

$$\bar{z} = (1/n)\sum_{i\le n} z_i$$

es la media muestral, y la solución existe si y solamente si

$$|\bar{z}| < \sqrt{2/\pi}$$

Chen et al. (*2004*) incorporan las siguientes proposiciones.

Proposición 1. *Sea $Z_1, Z_2, \ldots, Z_n$ una muestra aleatoria procedente de la NA(λ), entonces la fdp de $\bar{z} = (1/n)\sum_{i\le n} z_i$ está dada por*

$$f(\bar{z}) = 2^n\sqrt{n}\varphi(\sqrt{n}\bar{z}) \times \Phi_n\left(\lambda\bar{z}\left[\frac{1}{1+\lambda^2}\mathbf{I}_n + \frac{1}{n(1+\lambda^2)}\mathbf{1}\mathbf{1}^\top\right]^{1/2}\mathbf{1}\right),$$

donde $\Phi_n(\cdot)$ denota la fda de la $N_n(\mathbf{0}, \mathbf{I}_n)$, $\mathbf{I}_n$ es la matriz identidad, y $\mathbf{1} = (1, \ldots, 1)^\top$.

Proposición 2. *Sea $Z_1, \ldots, Z_n$ una muestra aleatoria procedente de la NA(λ), entonces la fgm de $\bar{z}$ está dada por*

$$M_{\bar{Z}}[t] = 2^n \exp\left(\frac{t^2}{2n}\right)\left[\Phi\left(\delta\frac{t}{n}\right)\right]^n$$

De las Proposiciones 1 y 2 se observa que la distribución de $\bar{z}$ no es conocida y no es fácil de establecer (*ver Gupta & Chen, 2003; Chen et al., 2004*).

Por el Teorema del Límite Central, la distribución asintótica de

$$\sqrt{n}\,\frac{\bar{Z}-\sqrt{2/\pi}\;[\lambda/\sqrt{1+\lambda^2}]}{\sqrt{1-(2/\pi)[\lambda^2/(1-\lambda^2)]}} \qquad (4)$$

es $N(0, 1)$. Por lo tanto, cuando el tamaño de la muestra n es lo suficientemente grande, aunque la distribución exacta de la media muestral estandarizada no es una distribución normal-asimétrica, su distribución asintótica sigue una distribución $N(0, 1)$ (*ver Gupta & Chen, 2003; Chen et al., 2004*).

3. ESTUDIO DE SIMULACIÓN

Basado en $J=1000,\ 5000,\ 10000$ conjuntos de datos, con los valores $n=10,\ 20,\ 30,\ 40,\ 50,\ 100,\ 200$ y $\lambda = 1, 2, 3, 4$, se estima

$$q = Pr[|\bar{Z}| > \sqrt{2/\pi}].$$

Los resultados se muestran en el cuadro siguiente. Las figuras siguientes muestran el historial de las medias muestrales para $J = 10000$ conjuntos de datos y $n = 10, 200,$ donde la banda ploma en el eje vertical inicia en $-\sqrt{2/\pi}$ y termina en $\sqrt{2/\pi}$

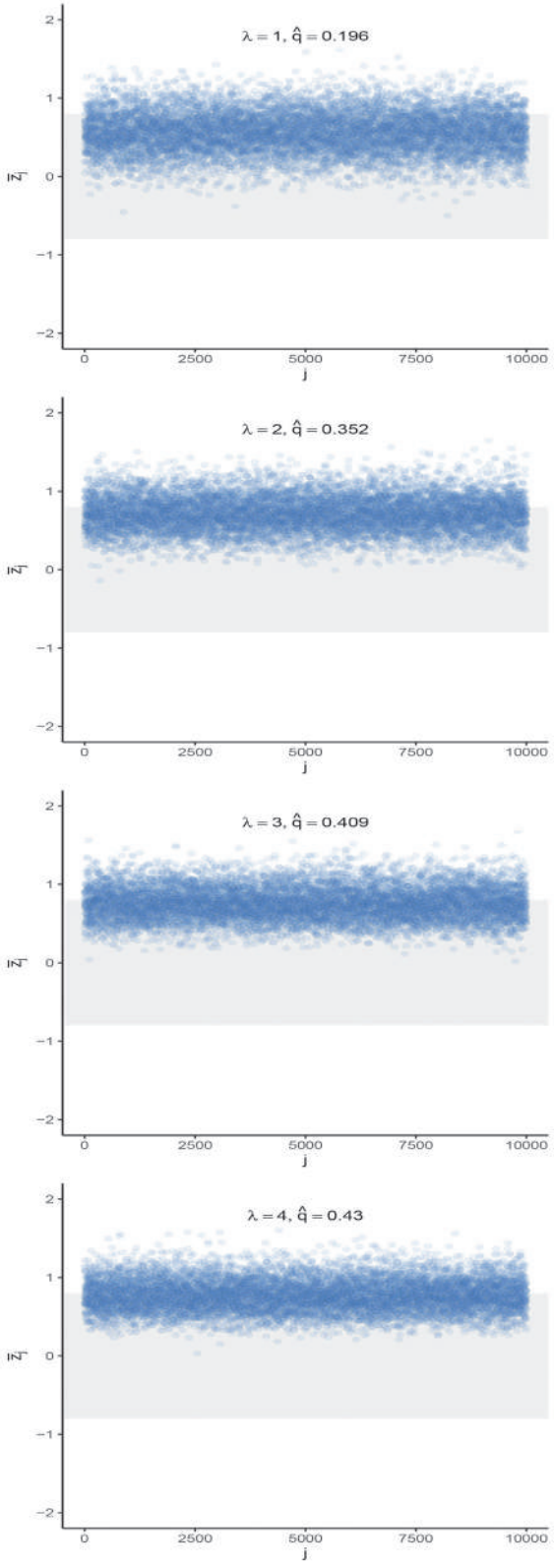
Los resultados indican que: a) si aumenta el tamaño de la muestra disminuye $\hat{q}$, y b) si aumenta parámetro de asimetría aumenta $\hat{q}$.

Tabla 1.
Estimación de $Pr[|\bar{Z}| > \sqrt{2/\pi}]$

J	n \ λ	1	2	3	4
1000	10	0,196	0,352	0,409	0,430
	20	0,098	0,279	0,385	0,440
	30	0,056	0,264	0,385	0.419
	40	0,044	0,229	0,303	0,379
	50	0,016	0,206	0,339	0,379
	100	0,001	0,108	0,278	0,352
	200	0,000	0,047	0,190	0,344
5000	10	0,184	0,350	0,417	0,435
	20	0,097	0,287	0,374	0,447
	30	0,059	0,250	0,356	0,404
	40	0,038	0,209	0,346	0,391
	50	0,019	0,194	0,335	0,395
	100	0,002	0,117	0,270	0,345
	200	0,000	0,047	0,194	0,347
10000	10	0,182	0,349	0,404	0,440
	20	0,100	0,290	0,387	0,416
	30	0,061	0,255	0,355	0,404
	40	0,040	0,222	0,345	0,398
	50	0,022	0,196	0,323	0,389
	100	0,002	0,117	0,260	0,348
	200	0,000	0,049	0,192	0,350

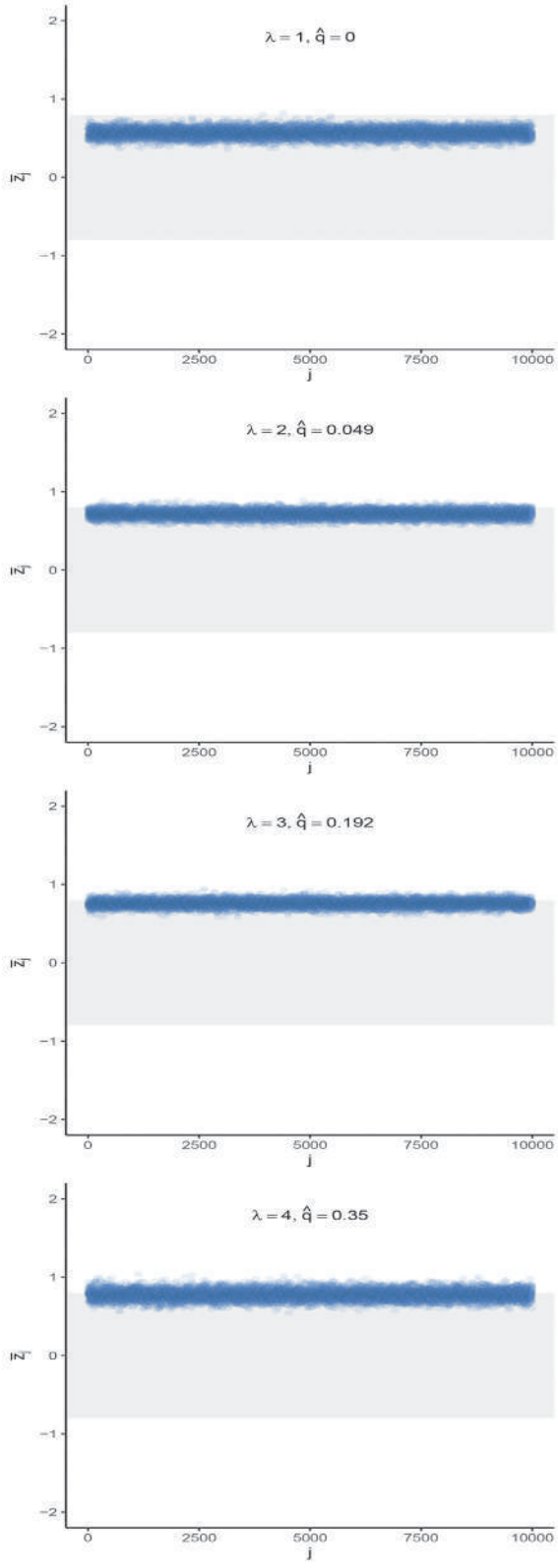
Fuente:Elaboración Propia

Figura 2.
Historial de las medias muestrales para J = 10000 y n = 10.



Fuente: Elaboración Propia

Figura 3.
Historial de las medias muestrales para J = 10000 y n = 200.



Fuente: Elaboración Propia

4. CONCLUSIÓN

En esta nota hemos analizado una opción de estimación del parámetro de asimetría de la distribución normal-asimétrica de *Azzalini* (1985), se observa que se tiene algunos problemas al estimar el parámetro de asimetría en tamaños de muestra moderados. Otros problemas de estimación pueden revisarse en *Pewsey* (*2000*), *Monti* (*2003*), *Liseo & Loperfido* (*2006*), y *Sartori* (*2006*). Algunas alternativas de estimación se ven por ejemplo en *Azzalini* (*2014*) y *Arellano-Valle et al.* (*2018*).

AGRADECIMIENTOS

El primer autor fue parcialmente apoyado por la beca FIB-UV de la Universidad de Valparaíso, Chile. Los autores agradecen al editor por los útiles comentarios

BIBLIOGRAFÍA

1. Arellano-Valle, R. B., Ferreira, C. S., & Genton, M. G. (2018). Scale and Shape Mixtures of Multivariate Skew-Normal Distributions. Journal of Multivariate Analysis, 166, 98–110. 1, 4
2. Azzalini, A. (1985). A Class of Distributions which Includes The Normal Ones. Scandinavian Journal of Statistics, 12(2), 171–178. 1, 4
3. Azzalini, A. (2014). The Skew-Normal and Related Families. Cambridge: Cambridge University Press. 1, 4
4. Chen, J. T., Gupta, A. K., & Nguyen, T. T. (2004). The Density of The Skew Normal Sample Mean and its Applications. Journal of Statistical Computation and Simulation, 74(7), 487–494. 2
5. Gupta, A. & Chen, T. (2003). On The Sample Characterization Criterion for Normal Distributions. Journal of Statistical Computation and Simulation, 73(3), 155–163. 2
6. Gupta, R. D. & Gupta, R. C. (2008). Analyzing Skewed Data by Power Normal Model. Test, 17, 197–210. 1
7. Liseo, B. & Loperfido, N. (2006). A Note on Reference Priors for The Scalar Skew-Normal Distribution. Journal of Statistical Planning and Inference, 136(2), 373–389. 4
8. Monti, A. C. (2003). A Note on The Estimation of The Skew Normal and The Skew Exponential Power Distributions. METRON - International Journal of Statistics, LXI(2), 205–219. 4
9. Pewsey, A. (2000). Problems of Inference for Azzalini 's Skewnormal Distribution. Journal of Applied Statistics, 27(7), 859–870. 4
10. Sartori, N. (2006). Bias Prevention of Maximum Likelihood stimates for Scalar Skew Normal and Skew t Distributions. Journal of Statistical Planning and Inference, 136(12), 4259–4275. 4

REGRESIÓN LOGÍSTICA CON INTERCEPTOS ALEATORIOS. APLICACIÓN A DATOS DE PANEL

Dr(c) Ramiro Coa Clemente *
✉ clementecoa@gmail.com

RESUMEN

En este artículo se presenta sucintamente el modelo lineal generalizado de efectos mixtos, un modelo de mucha utilidad para abordar el análisis estadístico en profundidad, en diferentes campos. Un caso particular de este modelo es el denominado regresión logística con interceptos aleatorios, un modelo alternativo para el análisis de datos de panel. Se ilustra su aplicación en el ámbito de la nutrición. El propósito es determinar si fumar durante el embarazo afecta o no el bajo peso al nacer. Los resultados sugieren un efecto muy significativo del consumo de tabaco durante el embarazo sobre el bajo peso al nacer.

PALABRAS CLAVE

Efectos mixtos, Interceptos aleatorios, Datos de panel

ABSTRACT

This article succinctly presents the generalized linear model of mixed effects, a very useful model to address in-depth statistical analysis in different fields. A particular case of this model is the so-called logistic regression with random intercepts, an alternative model for the analysis of panel data. Its application in the field of nutrition is illustrated. The purpose is to determine whether or not smoking during pregnancy affects low birth weight. The results suggest a very significant effect of tobacco use during pregnancy on low birth weight.

KEYWORDS

Mixed Effects, Random Intercepts, Panel Data

1. EL MODELO LINEAL GENERALIZADO DE EFECTOS MIXTOS

Un Modelo Lineal de Generalizado de Efectos Mixtos (MLGEM) tiene la siguiente forma general:

$$g[E(Y/X,u)] = X\beta + Zu$$

donde Y es un vector de respuestas de dimensión $n*1$ con función de distribución de probabilidad F, X es matriz de covariables $n*p$ asociado al vector de efectos fijos β, Z es una matriz de covariables $n*q$ asociado al vector de efectos aleatorios u, β es vector de efectos fijos $p*1$, u es vector de efectos aleatorios $q*1$, $\eta=X\beta+Zu$ es el predictor lineal, $g(.)$ es la función de enlace para la cual se supone que existe su función inversa $g^{-1}(.)$ de modo que $E(Y/X,\ u) = g^{-1}(X\beta+Zu)=h(\eta)=\mu$, donde μ es el vector de medias de dimensión $n*1$. Al considerar varias definiciones para $g(.)$ y F se tiene una amplia variedad de modelos,

* Ex Director de Investigación de la Unidad de Análisis y Política Social (UDAPSO)

entre los cuales se encuentra el modelo de regresión logística con *interceptos aleatorios*. Generalmente se asume que el vector de efectos aleatorios u tiene una distribución normal multivariada con media 0 y matriz de varianzas-covarianzas Σ de dimensión $q*q$, es decir, $u \sim N_q(0,\Sigma)$. Los efectos aleatorios no son estimados directamente, estos son caracterizados por sus varianzas, denominados comúnmente componentes de varianza. Estos componentes de varianza son elementos de la matriz de varianzas-covarianzas $G = Var(u)$.

El MLGEM permite modelar la correlación dentro de un *cluster* o conglomerado. Esto es, los sujetos dentro de un mismo *cluster* podrían estar correlacionados producto de un intercepto aleatorio compartido, producto de una pendiente aleatoria compartida, o como consecuencia de ambas situaciones.

Cuando se tiene datos *clusterizados*, no es conveniente considerar el total de las n observaciones al mismo tiempo, por el contrario, es ventajoso organizar el modelo mixto como una serie de M *clusters* independientes. La formulación apropiada del modelo es:

$$g[E(Y_j / X_j, u_j)] = X_j \beta + Z_j u_j$$

donde $j=1,\ldots,M$ y el *cluster* j consiste de n_j observaciones. El vector de respuestas Y_j es de dimensión n_j*1 e incluye todas las observaciones correspondientes al j-ésimo *cluster*. Lo mismo para las matrices X_j, Z_j y el vector u_j. Nuevamente se asume que el vector de efectos aleatorios u_j está distribuido normalmente con media 0 y matriz de varianzas-covarianzas Σ de dimensión $q*q$, es decir $u_j \sim N_q(0,\Sigma)$. Este modelo es el propuesto por Laird y Ware (1982) y ofrece dos ventajas importantes. Primero, se puede especificar los términos de los efectos aleatorios con facilidad. Si los *clusters* son escuelas, se puede especificar simplemente un efecto aleatorio al nivel de la escuela. Segundo, el modelo se puede generalizar fácilmente a más de un conjunto de efectos aleatorios. Por ejemplo, si las clases están anidadas dentro de escuelas, el modelo puede ser generalizado para incluir efectos aleatorios a nivel de escuelas y a nivel de clases dentro de escuelas.

La clave para ajustar modelos mixtos cae en la estimación de los componentes de varianza. Existen muchos métodos para tal estimación, uno de ellos es el de máxima verosimilitud (MV). Si $f(Y_j, u_j)$ representa la función de distribución conjunta de Y_j y u_j, la distribución marginal de Y_j es dada por

$$f(Y_j) = \int_{\mathbb{R}^q} f(Y_j, u_j)\, du_j$$

A partir de esta distribución marginal se puede deducir la función de verosimilitud para el *cluster* j, la cual queda expresada como

$$L_j(\beta,\Sigma) = \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma|^{\frac{1}{2}}} \int_{\mathbb{R}^q} e^{\left\{\ln f(Y_j/u_j) - \frac{u_j'\Sigma^{-1}u_j}{2}\right\}}\, du_j.$$

Como se supuso que los M *clusters* son independientes, la función de verosimilitud total para el vector de respuestas Y es dada por

$$L(\beta,\Sigma) = \prod_{j=1}^{M} L_j(\beta,\Sigma).$$

Para aproximar la integral que aparece en $L_j(\beta,\Sigma)$ se recurre a métodos numéricos. Primero, se hace un cambio de variable para

transformar la integral multivariable en un conjunto anidado de integrales univariables; segundo, cada integral univariable puede entonces ser evaluada usando la cuadratura Gauss-Hermite.

2. REGRESIÓN LOGÍSTICA CON INTERCEPTOS ALEATORIOS

Un caso particular del modelo lineal generalizado de efectos mixtos es el modelo de regresión logística con interceptos aleatorios. El modelo es expresado como:

$$P(Y_{ij} = 1)/X_{ij}, u_j) = H(X_{ij}\beta + u_j) = \mathrm{H}(\eta_{ij})$$

donde el efecto aleatorio u_j es una variable unidimensional con distribución normal, es decir, $u_j \sim N(0, \sigma^2)$. Se asume que $u_1,...,u_M$ son independientes.

Para precisar algunas ideas, consideremos el siguiente ejemplo. Asumamos que la variable respuesta representa la ocurrencia o no de cáncer de pulmón y la variable explicativa es la condición de fumador, un factor de riesgo muy importante. Supongamos también que la muestra consiste de M submuestras conducidas en diferentes departamentos del país. Sea j el subíndice asociado al departamento e i a la persona. Entonces la variable binaria Y_{ij} representa la presencia o ausencia de cáncer de pulmón ($Y_{ij} = 1$ = *con cáncer*; $Y_{ij} = 0$ = *sin cáncer*) y X_{ij} representa la condición de fumador de la *i-ésima* persona en el *j-ésimo* departamento ($X_{ij} = 1$ = *fumador*; $X_{ij} = 0$ = *no fumador*). Sea n_j el número de personas encuestadas en el *j-ésimo* departamento. La regresión logística estándar aplicada al conjunto de datos $\{Y_{ij}, X_{ij}\}, j=1,...,M, i=1,...,n_j$, implícitamente asume que la incidencia de cáncer de pulmón es constante para todos los departamentos. Claramente este supuesto puede ser incorrecto porque los departamentos pueden tener diferentes condiciones ambientales, diferentes campañas contra el tabaco, diferentes tradiciones, diferentes políticas de salud y diferente población por edad, entre otros. Estos factores pueden conducir a diferentes incidencias de cáncer entre los departamentos. Por tanto, al asumir que esta incidencia es la misma se puede obtener conclusiones incorrectas con relación al efecto de fumar. No cabe duda que es más coherente y realista asumir que los interceptos difieran de un departamento a otro, por lo que un modelo más apropiado es el expresado anteriormente.

La función de verosimilitud – y consecuentemente la función *log-verosimilitud* – es un caso particular de la anterior función de verosimilitud correspondiente al MLGEM. La función *log-verosimilitud* para la regresión logística con interceptos aleatorios queda expresado como

$$l = l(\beta, \sigma^2)$$

$$= -\frac{M}{2} ln(2\pi\sigma^2) + \beta \sum_{j=1}^{M}\sum_{i=1}^{n_j} Y_{ij}X_{ij}$$

$$+ \sum_{j=1}^{M} ln\left[\int e^{u_j \sum_{i=1}^{n_j} Y_{ij} - \sum_{i=1}^{n_j} ln\left(1+e^{X_{ij}\beta+u_j}\right) - \frac{u_j^2}{2\sigma^2}}\, du_j\right]$$

Para estimar los parámetros β y σ^2 se puede usar el siguiente procedimiento iterativo

$$\hat{\beta}_{s+1} = \hat{\beta}_s + H^{-1}\left(\frac{\partial l}{\partial \beta}|_{\beta=\beta_s}\right)$$

$$\hat{\sigma}^2_{s+1} = \frac{1}{M}\sum_{j=1}^{M}\frac{I_{2j}}{I_{1j}}$$

donde

$$\frac{\partial l}{\partial \beta} = \sum_{j=1}^{M}\sum_{i=1}^{n_j} Y_{ij}X_{ij} - \sum_{j=1}^{M}\frac{\mathbf{I_{3j}}}{I_{1j}}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{M}\frac{I_{2j}}{I_{1j}}$$

y las tres integrales están definidas como

$$I_{1j} = \int_{-\infty}^{\infty} e^{h_j(\beta;u)}\, du$$

$$I_{2j} = \int_{-\infty}^{\infty} u^2\, e^{h_j(\beta;u)}\, du$$

$$\mathbf{I_{3j}} = \int_{-\infty}^{\infty}\left[\sum_{i=1}^{n_j} X_{ij}\,\frac{e^{\beta' X_{ij}+u}}{1+e^{\beta' X_{ij}+u}}\, e^{h_j(\beta;u)}\right] du$$

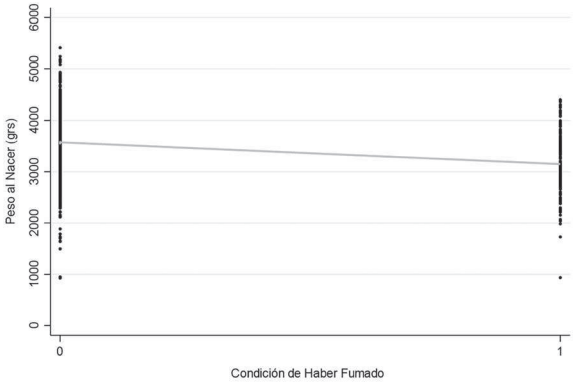
Notar que la integral $\mathbf{I_{3j}}$ es un vector $p*1$ y que H, I_{kj} para $k=1,2,3$, $j=1,\ldots,M$ son calculados en los valores actuales, $\beta = \beta_s$ y $\sigma = \sigma_s$.

3. APLICACIÓN A DATOS DE PANEL

Consideremos los datos de panel donde se tiene 648 *clusters* (mujeres en edad fértil) y en cada *cluster* se observa la condición de peso al nacer para cada uno de tres nacimientos. Luego se tiene un total de 1944 observaciones. La variable respuesta es la condición de bajo peso al nacer. Los nacimientos con bajo peso son los que tuvieron un peso de 2500 gramos o menos al momento de nacer. Una de las variables explicativas consideradas en el análisis es la condición de la madre de haber fumado o no durante cada embarazo. Adicionalmente se incluyeron en el modelo otras variables que, de acuerdo la experiencia, pueden afectar el peso al nacimiento. Estas variables son la edad de la madre, el estado civil de la madre, la educación de la madre, el control prenatal, momento del primer control prenatal, calidad del control prenatal y el sexo del recién nacido. El propósito del análisis es determinar si el haber fumado durante el embarazo tiene un efecto significativo sobre la probabilidad de nacer con bajo peso.

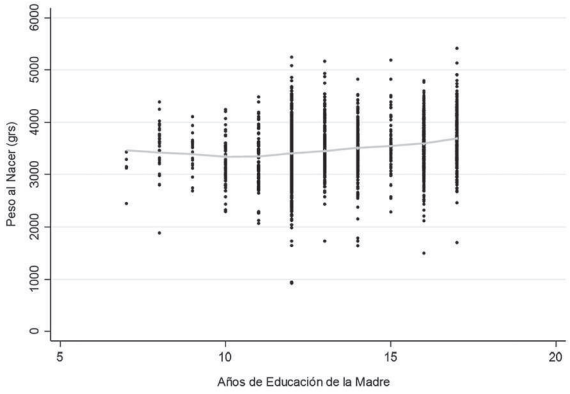
En los siguientes dos gráficos se exhiben las relaciones entre la variable peso al nacer y las variables condición de haber fumado durante el embarazo y educación de la madre. En términos generales se puede apreciar que el peso al nacer de los recién nacidos disminuye cuando la madre fuma durante el embarazo. Al relacionar el peso al nacer con la educación de la madre se observa que desde alrededor de los 12 años de educación comienza a incrementarse suavemente el peso al nacer, sin embargo, previo a este número de años de educación se observa incluso un leve descenso en el peso de los recién nacidos.

Figura 1.
Relación entre el Peso al Nacer y Condición de Fumar



Fuente: Elaboración Propia

Figura 2.
Relación entre el Peso al Nacer y la Educación de la Madre



Fuente: Elaboración Propia

El modelo de regresión logística de *interceptos aleatorios* usado para el análisis de los datos de panel es el siguiente

$$P(Y_{ij} = 1/X_{ij}, u_j) = g^{-1}(u_j + X_{ij}\beta) = \frac{e^{u_j + X_{ij}\beta}}{1 + e^{u_j + X_{ij}\beta}}$$

donde Y_{ij} es la condición de bajo peso al nacer del *i-ésimo* nacimiento para la *j-ésima* madre; X_{ij} es el vector fila de variables explicativas para el *i-ésimo* nacimiento de la *j-ésima* madre; β es el vector de coeficientes de efectos fijos y u_j es el efecto aleatorio de la *i-ésima* madre, que hace que el modelo tenga *interceptos aleatorios*. Los resultados se exhiben en el siguiente cuadro.

Cuadro 1.
Regresión Logística con interceptos aleatorios aplicado a datos de panel

Bajo peso al nacer	Razón de chances	Error estandar robusto	P>|z|	Intervalo de confianza 95%	
Fumó	3,17	1,21	0,003	1,50	6,70
Sexo del recién nacido	0,62	0,20	0,130	0,33	1,15
Edad de la madre	0,95	0,04	0,204	0,88	1,03
Educación de la madre	0,88	0,08	0,178	0,73	1,06
Madre casada	0,49	0,22	0,105	0,20	1,16
Prenatal de calidad intermedia	2,86	1,31	0,022	1,17	7,02
Prenatal de calidad inadecuada	6,19	3,58	0,002	1,99	19,22
Sin control prenatal	0,89	0,78	0,898	0,16	4,97
1er C.prenatal en 2do trimestre	0,65	0,32	0,382	0,24	1,71
1er C.prenatal en 3er trimestre	0,05	0,07	0,034	0,00	0,80
Constante	0,41	0,42	0,382	0,06	3,02

Fuente: Elaboración Propia

Recordemos que el objetivo del ejemplo es principalmente determinar si haber fumado durante el embarazo tiene un efecto significativo sobre la probabilidad de nacer con bajo peso. De los resultados expuestos en el cuadro se puede concluir que el efecto de fumar sobre el bajo peso al nacer es altamente significativo. Cuando la madre fuma, la chance de tener bajo peso al nacer es más de tres veces que cuando la madre no fuma. Por otra parte, como ya se advirtió en el gráfico, la educación de la madre no tiene un efecto significativo sobre el bajo peso al nacer. Sin embargo, la calidad del cuidado prenatal tiene un efecto altamente significativo sobre el bajo peso al nacer. Cuando el cuidado prenatal es de baja calidad, la chance de tener bajo peso al nacer es seis veces más que cuando el cuidado prenatal es adecuado.

4. ALGUNAS CONSIDERACIONES

El modelo lineal generalizado de efectos mixtos se caracteriza por incluir tanto efectos fijos como efectos aleatorios. La introducción de efectos aleatorios permite realizar el análisis de datos con estructura más compleja y, consecuentemente, permite un análisis más próximo de la compleja realidad. En la modelación se puede permitir, por ejemplo, que la probabilidad de nacer con bajo peso varíe de una madre a otra. En cambio, con un modelo lineal generalizado estándar no es posible realizar este tipo de análisis.

Si bien el modelo lineal generalizado de efectos mixtos permite un análisis más profundo de los datos, la maximización de la función *log-verosimilitud* para estimar los coeficientes llega a ser bastante compleja, puesto que involucra la solución de integrales complejas como la integral logística-normal. Para solucionar estas integrales se recurre a métodos de aproximación numérica. Principalmente se recurre al método de cuadratura de Gauss-Hermite.

Un modelo particular y muy importante de la familia de modelos lineales generalizados de efectos mixtos es el modelo de regresión logística con *interceptos aleatorios*. Este modelo es útil para analizar datos de panel.

BIBLIOGRAFÍA

1. Pinheiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS. New York: Springer
2. McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, Second Ed. Chapman and Hall/CRC, London.
3. Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data; Biometrics, Vol. 38, No. 4, pp. 963-974

DESCOMPOSICIÓN TENSORIAL TUCKER3 APLICADO A TABLAS DE CONTINGENCIAS DE TRES VÍAS

M. Sc. Lucy Gabriela Cuarita Ajno*

✉ *lcuarita@fcpn.edu.bo*

RESUMEN

El análisis de datos tensorial se encarga del estudio de datos obtenidos de la medición de más de una variable sobre un conjunto de individuos u objetos, los cuales son ordenados en un tensor de orden superior y donde interesa fundamentalmente la descomposición del tensor en componentes mucho más simples, de tal manera que faciliten la interpretación de los datos. En el campo del Análisis Multivariante, en particular, la técnica del Análisis de Correspondencias Múltiple permite identificar la interacción de los niveles correspondientes a un conjunto de variables de estudio, transformando la tabla de contingencias para luego aplicar la técnica del Análisis de Correspondencias Simple. Por otro lado, el modelo tensorial Tucker3 es un método de descomposición tensorial, que permite modelar la interacción entre las vías de un tensor de tercer orden y de sus componentes, preservando la estructura original de los datos. Hoy en día, los modelos tensoriales son una alternativa en el análisis de datos multivariantes, aunque la mayoría de los trabajos se encuentran en el campo del análisis de datos de tres vías, existen investigaciones que indican que la metodología continuará en ascenso mientras las estructuras de datos sean cada vez más complejas y los investigadores requieran un análisis integral de los datos.

PALABRAS CLAVE

Estadístico de Pearson, Inercia, Interacción, Descomposición Tensorial, Modelo Tucker3, Tensor.

ABSTRACT

Tensor data analysis is responsible for the study of data obtained from the measurement of more than one variable on a set of individuals or objects, which are arranged in a higher order tensor and where the decomposition of the tensor into much more components is of fundamental interest. simple, in such a way that they facilitate the interpretation of the data. In the field of Multivariate Analysis, in particular, the Multiple Correspondence Analysis technique allows identifying the interaction of the levels corresponding to a set of study variables, transforming the contingency table and then applying the Simple Correspondence Analysis technique. On the other hand, the Tucker3 tensor model is a tensor decomposition method that allows modeling the interaction between the pathways of a third order tensor and its components, preserving the original structure of the data. Today, tensor models are an alternative in multivariate data analysis, although most of the work is in the field of three-way data analysis, there is research that indicates that the methodology will continue to rise as long as the structures of data become increasingly complex and researchers require a comprehensive analysis of the data.

KEYWORDS

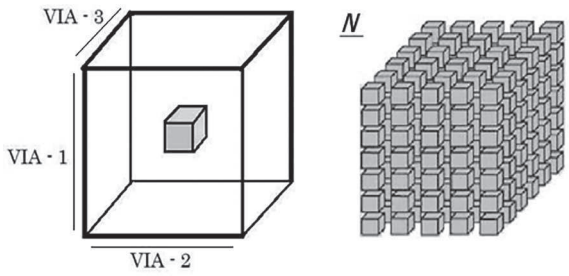
Pearson's statistic, Inertia, Interaction, Tensor Decomposition, Tucker3 Model, Tensor.

* Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés – Bolivia

I. INTRODUCCIÓN

Una tabla de contingencias de tres vías, *N*, es un tensor de orden 3, que tiene una variable fila con *I* niveles (vía-1), una variable columna con *J* niveles (vía-2) y una variable tubo con *K* niveles (vía-3), es decir: $\underline{N} \in \mathbb{N}^{IxJxJ}$ (*Amari S., Cichocki A., Huy A., Zdunek R., 2009*).

Figura N° 1
Tensor de tercer orden $\underline{N} \in \mathbb{N}^{7x5x6}$



El contenido de los cubos en el tensor son generalmente frecuencias absolutas n_{ijk} o frecuencias relativas p_{ijk}. Las diferencias entre las proporciones observadas pueden ser modeladas utilizando como fundamento teórico el modelo de independencia entre variables fila, columna y tubo (*Kroonenberg P.,2008*).

II. MODELO DE INDEPENDENCIA

El modelo de independencia postula que p_{ijk} puede ser expresada como el producto de las proporciones marginales:

$$p_{ijk} = p_{i..}\, p_{.j.}\, p_{..k}$$

donde:

$$p_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\,,\; p_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$$

y

$$p_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}\,.$$

Luego, la contribución del modelo de independencia debe ser sustraído a cada proporción almacenada en los cubos del tensor, es decir: ($p_{ijk} - p_{i..}p_{.j.}p_{..k}$), quedando definidos los cubos con la dependencia entre los niveles de las tres vías.

El análisis de la dependencia implica estandarizar los valores de la dependencia utilizando la raíz cuadrada de los valores esperados, es decir:

$$(p_{ijk} - p_{i..}\, p_{.j.}\, p_{..k}) / \sqrt{p_{i..}\, p_{.j.}\, p_{..k}}$$

A los que se denominan *residuales estandarizados* del modelo de independencia. Luego, la suma de los residuales estandarizados al cuadrado es el *coeficiente de contingencia del promedio al cuadrado de Pearson* más conocida como ***Inercia*** y denotada por Φ^2. Además, si *n* es el total de observaciones entonces el estadístico χ^2 de Pearson se define como: $\chi^2 = \Phi^2\, n$.

III. MEDICIÓN DE LA DEPENDENCIA

Mientras que en tablas de dos vías existe un solo tipo de dependencia, en tablas de tres vías se pueden distinguir: la dependencia total que es la desviación del modelo de independencia de tres vías, la dependencia marginal que es el resultado de la interacción de dos vías y la dependencia de las tres vías la cual es debido a la interacción de las tres vías (*Kroonenberg P.,2008*).

Medición de la dependencia total.

En tablas de contingencias de tres vías con dimensiones I, J y K, la dependencia total es medida por la inercia Φ^2, definida como:

$$\begin{aligned}\Phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[\frac{p_{ijk} - p_{i..}\, p_{.j.}\, p_{..k}}{\sqrt{p_{i..}\, p_{.j.}\, p_{..k}}}\right]^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i..}\, p_{.j.}\, p_{..k}\left[\frac{p_{ijk} - p_{i..}\, p_{.j.}\, p_{..k}}{p_{i..}\, p_{.j.}\, p_{..k}}\right]^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i..}\, p_{.j.}\, p_{..k}\left[\Pi_{ijk}\right]^2 \qquad (1)\end{aligned}$$

donde Π_{ijk} es la medida de la dependencia total en la celda (i,j,k) en la tabla de contingencias de tres vías.

Medición de la dependencia marginal y la dependencia de tres vías.

La dependencia de la celda Π_{ijk} puede ser dividida en contribuciones separadas de las interacciones de dos y tres vías, (*Carlier A. y Kroonenberg P., 1996*). Luego, la descomposición de Π_{ijk} es:

$$\Pi_{ijk} = \frac{p_{ij.} - p_{i..}\,p_{.j.}}{p_{i..}\,p_{.j.}} + \frac{p_{i.k} - p_{i..}\,p_{..k}}{p_{i..}\,p_{..k}} + \frac{p_{.jk} - p_{.j.}\,p_{..k}}{p_{.j.}\,p_{..k}} + \frac{p_{ijk} - \alpha p_{ijk}}{p_{i..}\,p_{.j.}\,p_{..k}} \quad (2)$$

Donde el término que mide el tamaño de la interacción de las tres vías para la celda (i, j, k) es: $\alpha p_{ijk} = p_{ij.}\,p_{..k} + p_{i.k}\,p_{.j.} + p_{.jk}\,p_{i..} - 2p_{i..}\,p_{.j.}\,p_{..k}$. Aplicando la definición de los totales marginales de dos vías, y la última definición para Π_{ijk}, la inercia Φ^2 puede ser particionada como:

$$\begin{aligned}\Phi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i..}\,p_{.j.}\,p_{..k}\,[\Pi_{ijk}]^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i..}\,p_{.j.}\,[\Pi_{ij.}]^2 + \sum_{i=1}^{I}\sum_{k=1}^{K} p_{i..}\,p_{..k}\,[\Pi_{i.k}]^2 + \\ &\quad + \sum_{j=1}^{J}\sum_{k=1}^{K} p_{.j.}\,p_{..k}\,[\Pi_{.jk}]^2 + \\ &\quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{i..}\,p_{.j.}\,p_{..k}\left[\frac{p_{ijk} - \alpha p_{ijk}}{p_{i..}\,p_{.j.}\,p_{..k}}\right]^2 \\ &= \Phi^2_{IJ} + \Phi^2_{IK} + \Phi^2_{JK} + \Phi^2_{IJK} \quad (3)\end{aligned}$$

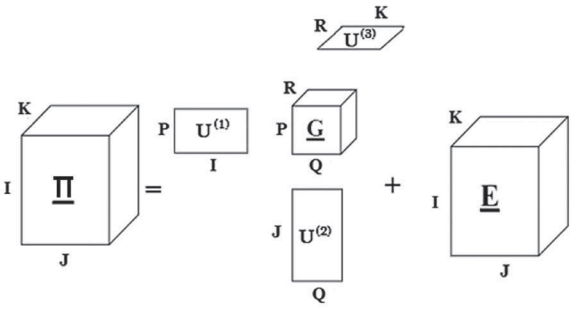
La última relación es una medida de los ajustes para cada interacción, además proporciona las contribuciones de estas interacciones a la dependencia total.

IV. MODELACIÓN DE LA DEPENDENCIA TOTAL

En el caso de tensores de tres vías la modelación implica aplicar una Descomposición en Valores Singulares Generalizada. Al respecto, existen varios candidatos, (*Kroonenberg P. M.; 2008*). En particular, se elige la *Descomposición en Valores Singulares de tres Modos*, más conocida con el nombre de Descomposición Tucker3 y por consiguiente Modelo Tucker3.

La aplicación del modelo Tucker3 en la medida de la dependencia total, implica expresar el tensor $\underline{\Pi}$ de orden tres como:

Figura N° 2
Modelo Tensorial Tucker3 para el tensor Π



Alternativamente, los valores Π_{ijk} se denotan por:

$$\Pi_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\,u^{(1)}_{ip}\,u^{(2)}_{jq}\,u^{(3)}_{kr} + \epsilon_{ijk} \quad (4)$$

V. DIMENSIONALIDAD DEL MODELO TUCKER3

Una característica sobresaliente en la estimación de los parámetros del modelo Tucker3 es la identificación de la dimensionalidad del arreglo central, $\underline{G}$, donde no se consideran todas las combinaciones (P, Q, R), pues no siempre son factibles. Por tal motivo surgen procedimientos algorítmicos como ser: el ***criterio st*** (*Ceulemans E. y Kiers H., 2006*) y el ***diffit*** (*Timmerman M. y Kiers H., 2000*) que permiten determinar las ternas con los mejores porcentajes de ajuste o las menores sumas de residuales al cuadrado como es el caso del ***scree plot multivía***, (*Timmerman M. y Kiers H., 2000*).

VI. ANÁLISIS DE LOS RESIDUALES

El modelo tensorial tucker3 permite detectar características particulares en los datos, como ser puntos atípicos o datos inusuales en los residuales, que pueden revelar características

especiales de algunos datos que no pueden ser modelados y tienen un efecto directo en la salida o estimación del modelo, (*Kroonenberg P. M., 2008*).

Análisis de los residuales estructurados al Cuadrado

La identificación de puntos anómalos en el ajuste, es el resultado de analizar los residuales dentro de la suma de cuadrados para los elementos de cada modo separadamente, utilizando la suma de los residuales al cuadrado y empleando los gráficos de la suma de cuadrados de los residuales relativos-

Análisis de los residuales no estructurados al cuadrado

Los residuales multivía son mucho más complejos que los residuales de dos vías, sin embargo, un estudio no estructurado sería esencialmente el mismo, es decir que se pueden utilizar gráficos bidimensionales de los residuales y de los valores ajustados.

VII. APLICACIÓN

La Encuesta de Hogares del año 2011 cuenta con 33821 personas encuestadas en sus hogares, de donde son de interés personas mayores de edad que dieron respuesta a la pregunta en relación a su *identificación con algún pueblo originario o indígena*, además de su *idioma materno* y el *departamento* de procedencia; realizando las depuraciones correspondientes sólo 28644 personas cumplieron las características especificadas. Posteriormente, se construye una tabla de contingencias donde la vía-1 corresponde a la variable *Identidad* con cinco niveles (Quechua, Aymara, Otros Nativos, Ninguno y NS\NR), la vía-2 tiene asociada a la variable *Idioma Materno* con cuatro niveles (Aymara, Castellano, Otros y Quechua) y a la vía-3 le corresponde la variable *Departamento* con nueve niveles (Chuquisaca, La Paz, Cochabamba, Oruro, Potosí, Tarija, Santa Cruz, Beni y Pando). En general, las especificaciones dadas describen la estructura de un tensor de orden 3, $\underline{N} \in \mathbb{N}^{5x4x9}$.

Al inspeccionar las proporciones marginales (masas de filas, columnas y tubos) de la *Tabla N° 1*, se advierte que el 70% de la población no se identifica como perteneciente a ningún pueblo originario o indígena, es posible que sea debido a que la lengua materna del 73% de la población es el Castellano, sin embargo, un porcentaje significativo de la población (27%) se identifica con el pueblo Quechua o Aymara.

Tabla N° 1. Datos sobre la Identidad: Proporciones Marginales de fila, columna y tubo

Identidad		Idioma Materno		Departamento	
Quechua	0.16	Aymara	0.10	Chuquisaca	0.07
Aymara	0.11	Castellano	0.73	La Paz	0.23
Otros Nativos	0.02	Otros	0.01	Cochabamba	0.18
Ninguno	0.70	Quechua	0.17	Oruro	0.06
NS-NR	0.01			Potosí	0.09
				Tarija	0.07
				Santa Cruz	0.21
				Beni	0.06
				Pando	0.04

Cálculo de la Inercia y el estadístico de Pearson

Utilizando los datos del tensor definido en el apartado anterior y aplicando las relaciones de la Inercia (1) y (3), además de la definición del estadístico χ^2 de Pearson, los resultados de la Tabla N° 2, muestran que existe una correspondencia significativa de la identidad de la población si se toma en cuenta la región de nacimiento y el idioma materno, con un 38% de la inercia, a pesar de que la proporción marginal, Tabla N° 1, mostró que el 70% de la población no tiene una identificación con un pueblo originario.

Además, es bueno señalar que, la correspondencia entre la Identidad y el Idioma Materno es significativa con un 28.77% del total de la inercia. Por lo expuesto, se advierte que las correspondencias identificadas muestran que en nuestro País aún existen grupos humanos

que conservan o se identifican con las raíces culturales de su región o territorio.

Tabla 2. Particionamiento de la dependencia total

Interdependencia	χ^2	Φ^2	Porcentaje
Identidad x IdiomaMat	24891.64	0.869	28.77
Identidad x Departamento	16498.94	0.576	19.06
IdiomaMat x Departamento	12259.63	0.428	14.17
Identidad x IdiomaMat x Departamento	32883.31	1.148	38.00
Total	86533.52	3.021	100.00

Identificación del Modelo

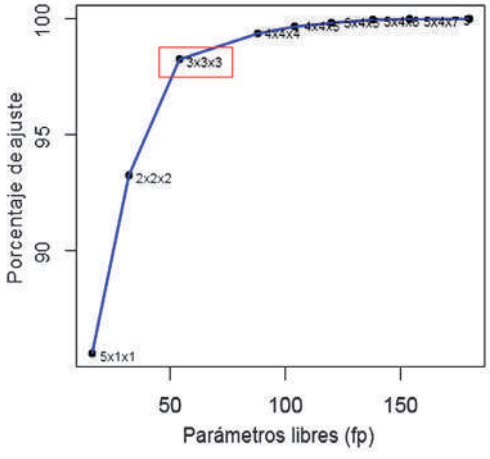
Sea el tensor $\underline{N} \in \mathbb{N}^{5x4x9}$, correspondiente a los datos sobre la Identidad, se construye el tensor de proporciones $\underline{P} \in \mathbb{R}^{5x4x9}$, luego con los valores p_{ijk} de $\underline{P}$ es necesario estimar el tensor $\underline{\Pi} \in \mathbb{R}^{5x4x9}$, cuyas celdas representan las dependencias entre los diferentes niveles de la tres vías. La identificación inicia con la determinación de la dimensión del arreglo central, de la descomposición de Tucker, $\underline{G}$, para tal efecto se utilizan los criterios de dimensionalidad.

Criterio st de Ceulemans y Kiers

La aplicación del criterio *st* indica que la dimensión (P, Q, R) = (3, 3, 3) es adecuada para representar el arreglo central de la Descomposición de Tucker.

Figura N° 3

Criterio st Parámetros libres y el porcentaje de ajuste



Scree plot multivía

En la figura 6.14 (gráfico de la derecha), se observa que a partir del orden (3, 3, 3) el descenso de la Suma de Errores al cuadrado es prácticamente constante, luego se considera que este parámetro es un referente significativo de la dimensionalidad del arreglo central en el modelo Tucker3.

Criterio diffit

En este criterio es relevante analizar la razón del incremento del porcentaje de ajuste resultante del *m-ésimo* componente en relación al ajuste porcentual del *m+1 ésimo* componente, donde el número total de componentes es $S=P+Q+R$. La Tabla N° 3, muestra que la aplicación del criterio diffit determina que la dimensión, del arreglo central, a ser elegida es (P, Q, R) = (3, 3, 3).

Figura N° 4

Scree plot multi vía

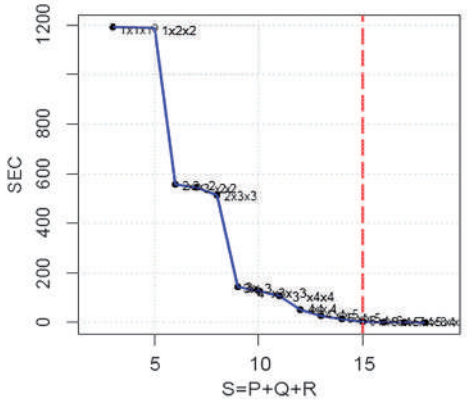


Tabla N° 3. Resultados de la aplicación del Criterio diffit

P	Q	R	S	sec	fit(%)	difs	btm
1	1	1	3	1193.19	85.58	85.58	11.20
1	2	2	5	1189.99	85.62	0.04	-
2	2	2	6	557.92	93.26	7.64	1.71
3	2	2	7	544.98	93.41	0.15	-
2	3	3	8	514.87	93.78	0.37	-
3	**3**	**3**	**9**	**144.21**	**98.26**	**4.48**	**6.59**
4	3	3	10	127.30	98.46	0.20	-
3	4	4	11	108.10	98.69	0.23	-
4	4	4	12	52.08	99.37	0.68	2.27
4	4	5	13	27.70	99.67	0.30	1.88
5	4	5	14	14.37	99.83	0.16	1.23
5	4	6	15	3.31	99.96	0.13	4.33
5	4	7	16	0.90	99.99	0.03	3.00
5	4	8	17	0.42	99.99	0.00	-
5	4	9	18	0.00	100.00	0.01	∞

Contribución de las componentes en cada vía

Los espacios originales de las vías en estudio, debido a la descomposición de Tucker3 empleada, se han reducido a un espacio tridimensional, en donde interesa conocer el porcentaje de la dependencia retenida en cada una de ellas, luego se presentan los siguientes resultados:

Tabla N° 4. Contribución porcentual en cada vía

Vía	Dim.1	Dim.2	Dim.3	Total
Identidad	85.738	8.045	4.474	98.26
Idioma Materno	85.730	8.038	4.490	98.26
Departamento	85.657	7.910	4.691	98.26

Donde 98.26 corresponde al porcentaje total de contribución o dependencia explicada por el modelo Tucker3, en cada una de sus vías.

Estimación de la Inercia y el estadístico de Pearson

El modelo Tucker3, relación (4), permite descomponer el tensor $\underline{\Pi}$ como sigue:

$$\Pi_{ijk} = \sum_{p=1}^{3}\sum_{q=1}^{3}\sum_{r=1}^{3} g_{pqr}u_{ip}^{(1)}u_{jq}^{(2)}u_{kr}^{(3)} + \epsilon_{ijk}$$

de donde $\underline{\widehat{\Pi}}$ se define como:

$$\widehat{\Pi}_{ijk} = \sum_{p=1}^{3}\sum_{q=1}^{3}\sum_{r=1}^{3} \hat{g}_{pqr}\hat{u}_{ip}^{(1)}\hat{u}_{jq}^{(2)}\hat{u}_{kr}^{(3)}$$

Estas estimaciones permiten aproximar los valores calculados de las inercias, expuestos en el Tabla N° 2, de acuerdo a las relaciones (1) y (3).

Tabla N° 5. Porcentaje de ajuste de la inercia con el Modelo Tucker3 seleccionado

Interdependencia	$\hat{\chi}^2$	$\hat{\Phi}^2$	Porcentaje Ajuste
Identidad x IdiomaMat	20451.82	0.714	82.16
Identidad x Departamento	10942.01	0.382	66.32
IdiomaMat x Departamento	8936.93	0.312	72.90
Identidad x IdiomaMat x Departamento	35403.98	1.236	123.03

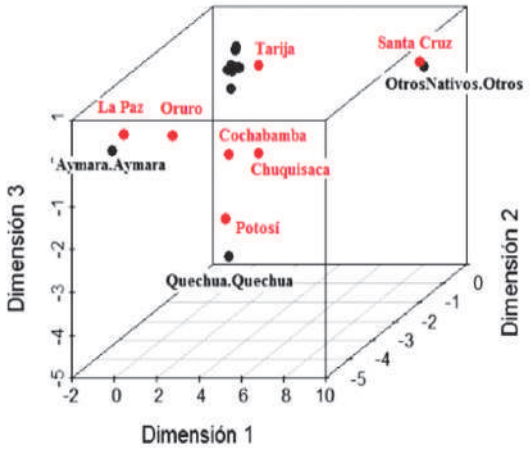
La Tabla N° 5 muestra en su última columna el porcentaje de ajuste de la inercia con el modelo seleccionado, donde la interacción entre el *Idioma Materno* y la *Identidad* tiene el porcentaje más significativo con 82 %. Por otro lado, llama la atención el sobre ajuste identificado para la triple interacción sin embargo no es alarmante si se comparan con las inercias y los valores del estadístico χ^2 de los datos originales y estimados, puesto que se preserva la interpretación de la dependencia dada para cada una de las interacciones.

Representación gráfica e interpretación

La representación gráfica de la dependencia estimada consiste en expresar las coordenadas de las componentes de una vía en un espacio determinado y proyectar sobre el las coordenadas de las componentes de las otras dos vías. Por tal motivo, las coordenadas de los niveles de la vía *Departamento* son transformadas a un espacio en particular, donde luego las coordenadas de los niveles combinados de las vías *Identidad e Idioma materno* son proyectadas en este espacio, como se muestra en la Figura N° 5.

En la Figura N° 5, se pone de manifiesto la dependencia de las tres vías, puesto que el *idioma materno* determina en la mayoría de los casos la identificación de la persona con el pueblo originario o indígena donde se hable ese idioma, por ejemplo si el idioma materno es el Aymara en general su identificación será con el pueblo Aymara que geográficamente e históricamente abarca los departamentos de La Paz y Oruro; un fenómeno similar sucede con las personas Potosinas, Cochabambinas y Chuquisaqueñas cuyo idioma materno es el Quechua su identificación generalmente es con el pueblo Quechua.

Figura N° 5:
Representación de la dependencia estimada de la vía Departamento y las vías combinadas Identidad e Idioma Materno.



En Santa Cruz es donde se tiene una mayor diversidad de pueblos indígenas, luego la identificación con un pueblo indígena tiene una correspondencia directa con su lengua materna. Hay que señalar, que los departamentos de Tarija, Beni y Pando son aquellos donde el idioma materno que impera es el castellano, luego es evidente que su identificación se halla relacionada principalmente con los niveles "Ninguno" y "NS-NR".

Análisis de los Residuales Estructurados

Sea el tensor de residuales, $\underline{E} \in \mathbb{R}^{5x4x9}$ definido por:

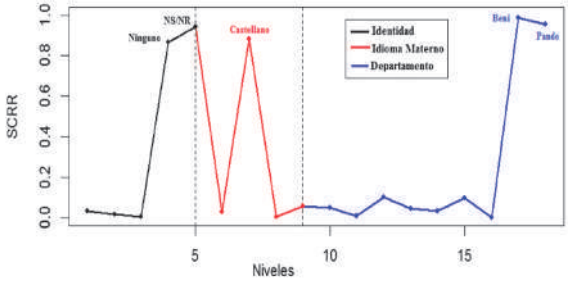
$$E = \Pi - \widehat{\Pi}$$

donde Π es el tensor con las dependencias y $\widehat{\Pi}$ la estimación de las dependencias hallada a través del Modelo Tucker3 de dimensión 3x3x3. El análisis de los residuales estructurados, inicia con el cálculo y representación gráfica de las sumas de cuadrados.

Suma de Cuadrados de los residuales relativos (*SCRR*) *de componentes por vía.*

El gráfico de la Figura N° 6 presenta a las tres vías del tensor de datos: *Identidad*, *IdiomaMat y Departamento* en sus diferentes niveles con las respectivas SCRR. Las representaciones gráficas muestran que los niveles *Ninguno y NS −NR de Identidad; Castellano de Idioma Materno; Beni y Pando de Departamento* no presentan estimaciones razonables (valores cercanos a la unidad) en relación a las restantes componentes en sus vías respectivas.

Figura N° 6.
De izquierda a derecha; Suma de Cuadrados de los Residuales Relativos de cada una de las componentes correspondientes a las vías Identidad, Idioma Materno y Departamento.



La Figura N°6 presenta de manera conjunta la SCRR en donde es evidente la pobre estimación de los niveles *Ninguno, NS.NR, Castellano, Beni y Pando* con el modelo Tucker3 de dimensión 3 × 3 × 3. Las causas pueden ser debido a que:

- La pregunta con respecto a la Identificación con un pueblo originario o indígena, causa que la población en su gran mayoría opte por las opciones *Ninguno, NS.NR*; aspecto que puede asociarse al efecto de la globalización, confusión o carencia de conocimiento sobre las implicaciones del término identidad o identificación con pueblos originarios, entre otros.
- La población en un número significativo asocia su identidad con el idioma materno, es decir que si su idioma materno es el *Castellano* no se identifica con ningún pueblo originario o indígena, sin tomar en cuenta sus raíces y las expresiones

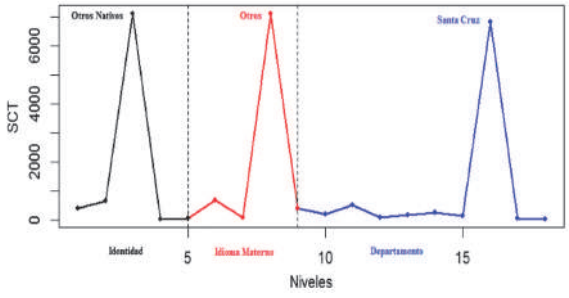
culturales de su región.

- Los departamentos de *Beni y Pando* son regiones afectadas por la migración, en el caso de la migración interna se observa que, si bien el idioma materno es el Castellano, la identificación con los pueblos Aymara, Quechua y otros nativos es significativa, lo que marca una diferencia sobresaliente con relación a los otros departamentos.

Suma de Cuadrados de los totales de componentes por vía (SCT)

La Figura N° 7 muestra la SCT por modo, en donde se identifican tres valores extremadamente altos correspondientes a los niveles *Otrosnativos*, *Otros y SantaCruz* de las vías *Identidad, Idioma Materno y Departamento*, respectivamente.

Figura N° 7. De izquierda a derecha; Suma de Cuadrados de los Totales de las componentes correspondientes a las vías Identidad, Idioma Materno y Departamento.



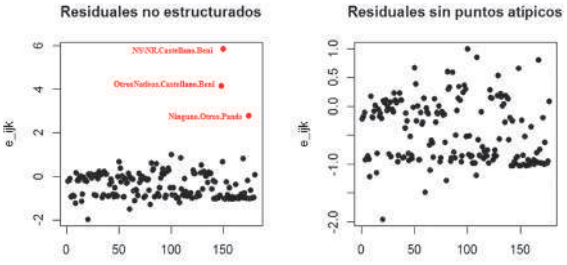
Considerando, que la SCT es calculada del tensor $\underline{\Pi}$ implica que en los componentes identificados existe una interacción significativa, en especial en la interacción de tres vías. Sin embargo, en la Tabla N° 1 se muestra claramente que las proporciones marginales de *Otrosnativos en Identidad y Otros* en *Idioma Materno* corresponden solamente al 2% y 1%, respectivamente, de la población encuestada. Por otro lado, la población encuestada en Santa Cruz es 6105 y si tomamos en cuenta a las personas con los dos niveles anteriores se tiene que es solamente el 1.04 %; luego es posible que en el departamento de Santa Cruz sea significativa esta relación de niveles, sin embargo, puede estar afectando la interacción de las vías en los otros departamentos.

Análisis de los Residuales No Estructurados

El tensor de residuales $\underline{E} \in \mathbb{R}^{5x4x9}$ cuenta con 180 residuales de las diferentes vías, en este apartado interesa solamente el conjunto de datos sin tomar en cuenta las vías y sus componentes.

En la Figura N° 8 se muestra el comportamiento de los 180 residuales, en donde claramente se observa la presencia de tres puntos atípicos que corresponden a las ternas: (NS-NR, Castellano, Beni), (Otros-Nativos, Castellano, Beni) y (Ninguno, Otros, Pando). En general, se conoce que, en Beni o Pando, no existe una identificación directa con algún pueblo originario, luego una alternativa razonable puede ser modelar los datos de la dependencia eliminando la influencia de estos departamentos.

Figura N° 8. De izquierda a derecha; a) Residuales no estructurados donde se identifican tres puntos atípicos; b) residuales no estructurados sin considerar los atípicos.



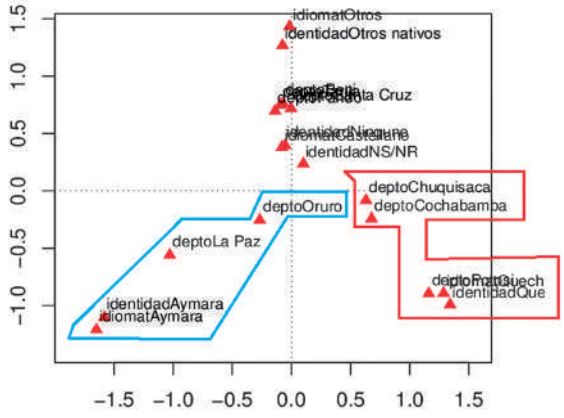
Análisis de Correspondencias Múltiple

El análisis de tablas de contingencias de tres vías se realiza aplicando la Técnica del Análisis de Correspondencias Múltiple, que es la aplicación de un Análisis de Correspondencias Simple en tablas de dos vías donde la tabla de tres vías es transformada en una tabla de dos vías, (*Greenacre M., 2007*).

Transformando los datos en la presente aplicación y utilizando la matriz de Burt, se obtiene los datos de la inercia total y el valor del *estadístico de Pearson*, 2.082948 y 59663,97, respectivamente. Note que el valor de la inercia total está por debajo de la inercia hallada con el modelo Tucker3, además que no es posible calcular las inercias de la interacción de dos y tres vías. Sin embargo, si se observa lo expuesto en la Figura N° 9, se evidencia similitud con las conclusiones emitidas en relación a la Figura N° 5.

Figura N° 9.

Análisis de Correspondencias Múltiple aplicado a la tabla de contingencias de 3 vías, donde se considera las variables Identidad, Idioma Materno y Departamento



VIII. CONCLUSIÓN

La potencialidad del Modelo Tucker3 en la identificación de la interacción o correspondencia entre vías y sus componentes en una tabla de contingencias de 3 vías, se pone de manifiesto si se comparan los resultados obtenidos con métodos convencionales como ser el Análisis de Correspondencias Múltiple. Actualmente, los modelos tensoriales se constituyen en nuevas herramientas en el campo multivariante, los cuales proponen alternativas que pueden robustecer o complementar las conclusiones emitidas con las técnicas convencionales multivariantes.

BIBLIOGRAFÍA

1. Amari S., Cichocki A., Huy A. y Zdunek R. (2009). Non negative matrix and tensor factorizations: Applications to exploratory Multi-Way data Analysis and blind source separation. Editado por John Wiley & Sons Ltd, Reino Unido

2. Carlier A. y Kroonenberg P. (1996). Decompositions and blots in three-way correspondence analysis. Psychometrika, vol. 61, No 2, 355-373.

3. Ceulemans E. y Kiers H. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. British Journal of Mathematical and Statistical Psychology, No 59, 133-150

4. Kroonenberg P. M. (2008). Applied Multiway Data Analysis. Wiley Series in Probability and Statistics. Estados Unidos de Norte América.

5. Timmerman M. y Kiers H. (2000). Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. British Journal of Mathematical and Statistical Psychology, 53, 1-16.

NÚMEROS DE BERNOULLI Y APLICACIONES DEL CÁLCULO COMPLEJO A LA ESTADÍSTICA

Lic. Raúl León Delgado Álvarez

✉ *dea_5@hotmail.com*

RESUMEN

Los trabajos de notables Matemáticos y Estadísticos, como Bernoulli, Cauchy, Hadamard hacen ver que las aplicaciones en el campo de la Estadística son de mucha utilidad y su tratamiento a través del Cálculo Complejo tiene ventajas comparativas que ayudan a la comprensión de sus aplicaciones.

PALABRAS CLAVE

Función Analítica, números de Bernoulli, series de Laurent

ABSTRACT

The works of notable Mathematicians and Statisticians, such as Bernoulli, Cauchy, Hadamard show that the applications in the field of Statistics are very useful and their treatment through Complex Calculus has comparative advantages that help to understand their applications.

KEYWORDS

Analytical function, Bernoulli numbers, Laurent series

1. INTRODUCCIÓN

Son muchas las aplicaciones al campo de la teoría Estadística las que aporta el Cálculo complejo, aquellas en las que no es suficiente el campo de los números reales, como recurso para justificar las características de la las funciones de probabilidad y el soporte contable de las mismas, es por ejemplo importante justificar la convergencia de una suma infinita como:

$$\sum_{n=1}^{\infty} \frac{1}{n^p} \quad , p > 1$$

que para $p = 2$ converge hacia el valor $\pi^2/6$, o equivalentes para hallar características numéricas de las funciones de probabilidad o demostrar éstas que no existen.

El presente artículo mostrará una aplicación del Cálculo Complejo para problemas similares al considerar expansiones dentro de ese campo.

2. DESARROLLO

Sean dos series de potencias:

$$(1) \quad \sum_{i=0}^{\infty} a_i (z-a)^i$$

$$(2) \quad \sum_{i=0}^{\infty} b_i (z-a)^i$$

donde r y g son los radios de convergencia de las series (1) y (2) de coeficientes, números positivos, además $b_0 \neq 0$.

Si $\sigma = min\ (r\ y\ g)$, *si* $r = g = \alpha$

Entonces ambas series serán convergentes en el círculo $|z - a|<\sigma$, si en este círculo hay ceros de la expansión:

$$b_0 + b_1(Z-a) + \cdots + a_n(Z-a)^n \quad (2)$$

Se toma un nuevo circulo de radio menor en cuyo interior, la suma (2) no se anule, existe puesto que el punto a no es cero de la suma (2) debido a la condición de que $b_0 \neq 0$ así entonces existe un circulo $|z - a|<R$.

En el cual ambas series (1) y (2) son convergentes y la suma de la serie (2) carece de ceros.

En el interior de este círculo la relación

$$f_z = \frac{a_0 + a_1(z-a) + \cdots + a_n(z-a)^n + \cdots}{b_0 + b_1(z-a) + \cdots + b_n(z-a)^n + \cdots}$$

Representa una función analítica la cual consecuencia de la regla de derivación del cociente; por lo tanto, existe una serie de potencias

$$c_0 + c_1(z-a) + \cdots + c_n(z-a)^n + \cdots$$

Que expresa a la función $f_{(z)}$ en el entero del círculo $|z - a|<R$ cociente de las series (1) dividendo y (2) divisor, realizando la división por el método de los coeficientes indeterminados (algoritmo de la división)

$$[c_0 + c_1(z-a) + \cdots + c_n(z-a)^n] * [b_0 + b_1(z-a) + \cdots + b_n(z-a) + \cdots + c_n(z-a)^n] * [b_0 + b_1(z-a) + \cdots + b_n(z-a)^n] = a_0 + a_1(z-a) + \cdots + a_n(z-a)^n + \cdots$$

Obsérvese que siendo convergente en el interior del círculo $|z - a|<R$ tiene que ser absolutamente convergente, por eso pueden multiplicar miembro a miembro

Realizando la multiplicación se obtiene:

$$c_0 b_0 + (a_0 b_1 + c_1 b_0)(z-a) +$$
$$(c_0 b_2 c_1 b_1 + c_2 b_0 (z-a)^2 + \cdots)$$
$$c_0 b_n + c_1 b_{n-1} + \cdots + c_n b_0)(z-a)^n + \cdots$$
$$= a_0 + a_1(z-a) + \cdots + a_n(z-a)^n + \cdots$$

Como las sumas de las series de potencias que figuran en el primer y segundo miembro coinciden en el circulo $|z - a|<R$ según el teorema de identidad para los series de potencias los coeficientes de ambos series tienen que ser iguales, de donde de resultan los ecuaciones

$$\begin{cases} c_0 b_0 = a_0 \\ c_0 b_1 + c_1 b_0 = a_1 \\ c_0 b_2 + c_1 b_1 + c_2 b_0 = a_2 \\ \quad \cdots \cdots \\ \quad \cdots \cdots \\ c_0 b_n + c_1 b_{n-1} + \cdots + c_n b_0 = a_n \end{cases}$$

Es un sistema infinito de ecuaciones lineales respecto de los coeficientes desconocidos $c_0, c_1, c_2, \ldots, c_n$ la particularidad de este sistema simplifica su solución y consiste en que para cualquier n, $(n=0,1,2,3,\ldots.)$, las primeras $(n+1)$ ecuaciones contienen las $(n+1)$ incógnitas de las primera ecuación $c_0 = a_0 / b_0$, $(b \neq 0$ según la hipotesis) y sustituyen dos en la segundo se obtiene

$$a_1 = \frac{a_0}{b_0} b_1 + c_1 b_0$$

$$c_1 = a_1 - \frac{a_0}{b_0} b_1$$

$$c_1 = \frac{a_1 b_0 - a_0 b_1}{b^2}$$

Sustituyendo que en la segunda se han hallado las líneas $c_0\ c_1 \ldots c_{n-1}$, sustituyendo en la $(n+1)$ ecuación se obtiene

$$c_n = \frac{a_n - a_0 b_1 c_1 - b_{n-1} - c_n b_1}{{b_0}^{n+1}}$$

Así se puede encontrar el coeficiente de un índice previamente dado, decir c_n en función de $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_n$ en forma de la determinada

$$c_n = \frac{1}{b_0^{n+1}} \begin{vmatrix} b_0 & \cdots & 0 & \ldots & a_0 \\ b_1 b_0 & \ldots & 0 & \ldots & a_1 \\ b_2 b_1 b_0 & \ldots & 0 & \ldots & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_n b_{n-1} b_{n-2} & \cdots & \cdots & \cdots & a_n \end{vmatrix}$$

Ejemplo $F_{(z)} = \frac{z}{e^z - 1}$ esta función analítica en todos los puntos del plano a excepción de los ceros de $e^z - 1$ es decir a excepción de los puntos $0, \pm 2\pi i, \pm 4\pi i, \ldots$

$$e^z - 1 = \frac{z}{1!} + \frac{z^2}{2!} + \cdots \ldots \ldots + \frac{z^n}{n!} + \cdots$$

$$F(z) = \frac{z}{z + \frac{z^2}{2!} + \cdots + \frac{z^n}{n!}}$$

Dividiendo $F(z)$ entre z

$a_0 = 1$ $a_n = 0$, para todo $n \neq 0$

Incluso vale por $z = 0$

$$b_0 = 1 \qquad a_n = \frac{1}{(n+1)!}$$

$$F_{(z)} = \frac{1}{1 + \frac{z}{2!} + \cdots + \frac{z^n}{(n+1)!}} + \cdots$$

La serie del denominador es convergente para cualquier z y tiene los mismos ceros que la función $e^z - 1$, a excepción de un cero en el origen de coordenadas por tanto en el interior del círculo $|z| < 2\pi$ la suma no se anula la primera de las ecuaciones da:

$$c_0(1) = 1 \Rightarrow c_0 = 1$$

Como todos los coeficientes de la serie del dividendo a excepción del coeficiente inicial, son iguales a cero, la $(n+1)$ ecuación tiene la forma:

$$c_0 \frac{1}{(n+1)!} + c_1 \frac{1}{n!} + \cdots + c_{n-1} \frac{1}{2!} + c_n = 0$$

$$n = 1, 2, 3 \ldots$$

$$c_0 b_n + c_1 b_{n-1} + c_2 b_{n-1} + \cdots + c_n b_0 = a_n = 0$$

$$c_0 \frac{1}{(n+1)!} + c_1 \frac{1}{n!} + c_2 \frac{1}{(n-1)!} + \cdots + c_{n-1} \frac{1}{2!} + c_n b_0 = 0$$

También se puede utilizar la formula por el determinante, es decir

$$c_0 = 1$$
$$b_0 = 1, a_0 = 1$$

$$c_n = \begin{vmatrix} 1 & 0 & \ldots & \ldots & 1 \\ \frac{1}{2!} & 1 & \ldots & \ldots & 0 \\ \frac{1}{3!} & \frac{1}{2!} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{(n+1)!} & \frac{1}{n!} & \frac{1}{(n-1)!} & \cdots & 0 \end{vmatrix}$$

$$= (-1)^n \begin{vmatrix} \frac{1}{2!} & 1 & 0 & \ldots & 0 \\ \frac{1}{3!} & \frac{1}{2!} & 1 & \ldots & 0 \\ \frac{1}{4!} & \frac{1}{3!} & \frac{1}{2!} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{(n+1)!} & \frac{1}{n!} & \frac{1}{(n-1)!} & \cdots & \frac{1}{2!} \end{vmatrix}$$

Los números $c_n\, n!$ se denominan números de Bernoulli y designan mediante B_n ; $B_n = C_n!$ Así

$$B_0 = C_0 \quad 0! = 1$$

$$B_n = C_n\, n! = (-1)^n\, n! \begin{vmatrix} \frac{1}{2!} & 1 & 0 & 0 & \ldots & 0 \\ \frac{1}{3!} & \frac{1}{2!} & 1 & 0 & \ldots & 0 \\ \frac{1}{4!} & \frac{1}{3!} & \frac{1}{2!} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{(n+1)!} & \frac{1}{n!} & \frac{1}{(n-1)!} & \frac{1}{(n-2)!} & \cdots & \frac{1}{2!} \end{vmatrix}$$

De donde se obtiene:

$$B_0\frac{1}{0!\,(n+1)!}+B_1\frac{1}{1!\,n!}+\ldots+B_n\frac{1}{n!\,1!}=0$$

Multiplicando ambos lados de la igualdad por *(n+1)!* y observando que $\frac{(n+1)!}{k!(n+1-k)!}$ es el coeficiente binomial $\binom{n+1}{k}$ se tiene

$$B_0\binom{n+1}{0}+B_1\binom{n+1}{1}+\cdots+B_n\binom{n+1}{n}=0 \qquad (n=1,2\ldots)$$

Esta fórmula puede representarse en la siguiente forma

$$(1+B_n)^{(n+1)}-B_n^{(n+1)}=0$$

Como $B_0=1$

$$B_0+2B_1=0\,;\quad B_1=1/2\;;\quad B_0=-1/2$$

$$B_0+3B_1+3B_2=0$$

$$B_2=-1/3\,B_0-B_1=1/6$$

$$B_0+4B_1+6B_2+4B_3=0$$

$$B_3=-1/4\,B_0-B_1-3/2\,B_2=0$$

$$B_3=0$$

$$B_0+5B_1+10B_2+10B_3+5B_4=0$$

$$B_4=-1/5\,B_0-B_1-2B_2+2B_3=-1/30$$

$$B_0+6B_1+15B_2+20B_3+15B_4+6B_5=0$$

$$B_5=0$$

$$B_0+7B_1+21B_2+35B_3+35B_4+21B_3+7B_6=0$$

$$B_6=-1/7-B_0-B_1-3B_2-5B_3-5B_4-3B_5$$

$$B_6=1/42$$

Es decir

$$B_0=1,\ B_1=-1/2,\ B_2=1/6,\ B_3=0,\ B_4=-1/30$$

$$B_5=0,\ B_6=1/42$$

Se demuestra que todos los números de Bernoulli de subíndices impares mayores que la unidad son iguales que $B_{2k+1}=0$, k=1,2,3…

En $\frac{z}{e^z-1}$ sustituyendo *z* por (-*z*) en el desarrollo

$$\frac{z}{e^z-1}=C_0+C_1\,Z+C_2\,Z^2+\cdots+C_nz^n+\cdots$$

$$=B_0+\frac{B_1}{1!}Z+\frac{B_2}{2!}Z^2+\ldots+\frac{B_n}{n!}Z^n+\cdots$$

$$\frac{-z}{e^{-z}-1}=-\frac{ze^z}{(e^{-z}-1)e^z}=\frac{ze^z}{e^z-1}$$

$$=b_0-\frac{b_1}{1!}z+\frac{b_2}{2!}z^2-\frac{b_3}{3!}z^3+\cdots$$

Sumando

$$\frac{z}{e^z-1}-\frac{ze^z}{e^z-1}=\frac{ze^z}{e^z-1}=\frac{z(1-e^z)}{(e^z-1)}=-z$$

$$=2\frac{b_1}{1!}z+\frac{2}{3}b_3Z^3+\cdots+2\frac{B_{2K+1}}{(2K+1)!}Z^{2K+1}+\ldots$$

Basándose en la unicidad del desarrollo de la serie de –z y la última expresión se tiene

$$2B_1=-1\,,\ B_3=B_5=B_{2k+1}=0$$

Solo $B_1=-1/2$ los demas $B_{2k+1}=0$ para todo k=1,2...

De esta manera el cociente solicitado se puede escribir como:

$$\frac{z}{e^z-1}=1-\frac{z}{2}+\sum_{k=1}^{\infty}\frac{B_{2K}}{(2k)!}Z^{2k}$$

Como los puntos singulares de la función $\frac{z}{e^z-1}$ más próximos al origen de coordenadas son $Z_1=2\pi i$, $Z_2=-2\pi i$

En estos puntos la función no está definida y no puede definirse de modo que se conserve la continuidad el radio de convergencia de la serie es 2π.

De esta relación según la fórmula de Cauchy-Hadamard se define que

$$\lim_{n\to\infty}\sqrt[n]{\frac{|B_n|}{n!}}=\lim_{k\to\infty}\sqrt[2k]{\frac{|B_k|}{(2K)!}}=\frac{1}{2\pi}$$

$B_{2K+1}=0$, k=1,2,3, por lo tanto, para cualquier $\epsilon>0$, existe un conjunto infinito de números B_{2k} que satisfacen la desigualdad

$$|B_{2k}|>\frac{2k!}{(2\pi+\epsilon)^{2k}}$$

Es decir son muy grandes en comparación con sus índices 2k.

Del desarrollo anterior se pueden obtener los desarrollos de las funciones de *zcotg z; tag z; zcosec z* al hacer

$$Cotg\ z=\frac{\cos z}{sen\ z}$$

$$=\frac{\frac{e^{iz}+e^{-iz}}{2}}{\frac{e^{iz}-e^{iz}}{2i}}=i\frac{e^{iz}+e^{-iz}}{e^{iz}-e^{iz}}$$

$$Cotg\ Z=i\left[\frac{e^{iz}+\frac{1}{e^{iz}}}{e^{iz}-\frac{1}{e^{iz}}}\right]=i\left[\frac{\frac{e^{2iz}+1}{e^{iz}}}{\frac{e^{2iz}-1}{e^{iz}}}\right]$$

$$=i\left(\frac{e^{2iz}+1}{e^{2iz}-1}\right)$$

Se puede revisar como

$$Cotg\ z=i+\frac{zi}{e^{2iz}-1}$$

De donde

$$z\ Cotg\ z=iz+\frac{2iz}{e^{2iz}-1}$$

La función: $\frac{2iz}{e^{2iz}-1}$ se puede desarrollar de acuerdo a

$$\frac{z}{e^z-1}=1-\frac{z}{2}+\sum_{k=1}^{\infty}\frac{Z_{2k}}{(2k)!}z^{2k}$$

Haciendo *2iz* como variable nueva que converge para $|2iz|<2\pi=>|z|<\pi$, el cociente:

$$\frac{2iz}{e^{2iz}-1}=1-\frac{2iz}{2}+\sum_{k=1}^{\infty}\frac{B_{2k}}{(2k)!}(2iz)^{2k}$$

lo que significa:

$$z\ Cotg\ z=1+\sum_{k=1}^{\infty}(-1)^k2^{2k}\frac{B_{2k}}{(2k)!}z^{2k}$$

Considerando identidades trigonométricas se puede expresar como

$$Tag\ z=\sum_{k=1}^{\infty}(-1)^{k-1}\left(2^{2k}\left(2^{2k}-1\right)\right)\frac{B_{2k}}{(2k)!}z^{2k}$$

Donde $|z|<\pi/2$, también para la

$$Sec\ z=\sum_{k=1}^{\infty}(-1)^k2^{2k}\frac{E_{2k}}{(2k)!}z^{2k}$$

Donde $|z|<\pi/2$, los coeficientes E_{2k} se denominan números de Euler, y se determinan por las ecuaciones:

$$E_0 = 1$$
$$E_0 + \binom{2n}{2} E_2 + \binom{2n}{4} E_4 + \cdots + \binom{2n}{2n-2} E_{2n-2} + E_{2n} = 0, \quad n = 1,2,3 \ldots$$

Algunos términos serán:

$$E_0=1,\ E_2=-1,\ E_4=5,\ E_6=-61,\ E_8=1385.$$

Del desarrollo en serie de Laurent para un punto arbitrario $z \in D$

$$f(z) = \sum_{-\infty}^{\infty} a_n (z - z_0)^n$$

donde los coeficientes

$$a_n = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(t)dt}{(t - z_0)^{n+1}} \qquad n = 0, \pm 1, \pm 2, \ldots$$

donde $\Gamma: |z-z_0|=\lambda,\ r<\lambda<\mathbb{R}$

Donde $\sum a_n e^{-\lambda_n z}$, los coeficientes a_n son complejos, pero los λ_n son números reales no negativos que satisfacen: $\lambda_{n+1} > \lambda_n$ para n=1, 2,3…. El $\lim_{n\to\infty} \lambda_n = 0$

Dicha serie se denomina serie de Dirichlet general y los lambda se llaman exponentes de la serie, de esta manera $\sum_{n=1}^{\infty} \frac{a_n}{n^z}$ se conoce como la serie de Dirichlet ordinaria o clásica.

Escribiendo el desarrollo en serie de Laurent

$$Cotg\ z - \frac{1}{z} = \sum_{k=1}^{\infty} \left(\frac{1}{z - k\pi} + \frac{1}{z + k\pi} \right)$$

integrando término a término a lo largo de la curva arbitraria que pasa por origen de coordenadas y no pasa por $k\pi$ $k=1, 2, 3$…..

$$\int_0^z \left(cotz - \frac{1}{z} \right) dz = \sum_{k=1}^{\infty} ln\left(\frac{k\pi - z}{k\pi} \cdot \frac{k\pi + z}{k\pi} \right) = \sum_{k=1}^{\infty} \ln\left(1 - \frac{z^2}{k^2\pi^2}\right)$$

Interpretando la suma infinita como un límite de una suma, se tiene:

$$ln \frac{senz}{z} = \lim_{n\to\infty} \sum_{k=1}^{\infty} ln\left(1 - \frac{z^2}{k^2 z^2}\right)$$

Donde $\frac{senz}{z} = \lim_{n\to\infty} \prod_{k=1}^{\infty} \left(1 - \frac{z^2}{k^2\pi^2}\right)$

pero:

$$\frac{1}{z - \pi i} + \frac{1}{z + \pi i} = -\sum_{k=0}^{\infty} \frac{z^k}{(i\pi)^{k+1}} + \sum_{k=0}^{\infty} \frac{(-1)^k z^k}{(i\pi)^{k+1}} = -2 \sum_{k=1}^{\infty} \frac{z^{2k-1}}{(i\pi)^{2k}}$$
$$para \quad |z| < \pi i$$

Por lo tanto los coeficientes de las potencias pares de z, en el desarrollo de *cotg z -1/z* son iguales a cero, mientras que los impares se expresan como:

$$-2 \sum_{j=1}^{\infty} \frac{1}{(j\pi)^{2n}} = \frac{2}{\pi^{2n}} \sum_{j=1}^{\infty} \frac{1}{j^{2n}}$$

$$Cotg\ z - \frac{1}{z} = \sum_{m=1}^{\infty} \left[\frac{-2}{\pi^{2m}} \sum_{j=1}^{\infty} \frac{1}{j^{2m}} \right] z^{2m-1}$$

También se mostró el desarrollo:

$$cotgz - \frac{1}{z} = \sum_{m=1}^{\infty} (-1)^m \frac{2^{2m}}{(2m)!} B_{2m} z^{2m-1}$$

Comparando los dos desarrollos se tiene:

$$\frac{-2}{\pi^{2m}} \sum_{j=1}^{\infty} \frac{1}{j^{2m}} = (-1)^m \frac{2^{2m}}{(2m)!} B_{2m}$$

$$\frac{2}{\pi^{2m}}\sum_{j=1}^{\infty}\frac{1}{j^{2m}}=(-1)^{m-1}\frac{2^{2m}B_{2m}}{(2m)!}$$

De dónde despejando:

$$\sum_{j=1}^{\infty}\frac{1}{j^{2m}}=\pi^{2m}2^{2m-1}(-1)^{m-1}\frac{B_{2m}}{(2m)!}$$

Finalmente para distintos valores de m se tienen las sumas siguientes

$$m=1,\qquad \sum_{j=1}^{\infty}\frac{1}{j^{2}}=\frac{\pi^{2}}{2!}(-1)^{0}\frac{2}{6},$$

porque B_2=1/6 , realizando operaciones:

$$\sum_{j=1}^{\infty}\frac{1}{j^{2}}=\frac{\pi^{2}}{6}$$

$$m=2,\qquad \sum_{j=1}^{\infty}\frac{1}{j^{4}}=\frac{\pi^{4}(-1)^{3}2^{3}}{24}B_{4},$$

pero B_4=(-1)/30, realizando operaciones

$$\sum_{j=1}^{\infty}\frac{1}{j^{4}}=\frac{\pi^{4}}{90}$$

m=3, considerando B_6=1/42

$$\sum_{j=1}^{\infty}\frac{1}{j^{6}}=\frac{\pi^{6}}{945}$$

De esta manera se pueden obtener varias sumas.

3. CONCLUSIONES

En el desarrollo del presente artículo, se pudo observar que el tratamiento desde el punto de vista del Calculo Complejo, primero observando las ventajas de operaciones básicas como la división entre polinomios dentro del campo complejo, permite analizar las ventajas de trabajar con funciones analíticas, posteriormente haciendo uso del desarrollo de serie de Laurent, se pueden comparar desarrollos de series que al igualar coeficientes da lugar a sumas infinitas como las que se pudo observar al final del desarrollo.

Aunque esta manera de obtener sumas infinitas no es la única, dado que se pueden comparar con desarrollos de Fourier para obtener los anteriores resultados.

Cuando se hace uso de las fórmulas de Euler Mac Laurin, las mismas son muy laboriosas el método expuesto aquí puede simplificar muchos cálculos innecesarios dado que se puede considerar los números de Bernoulli como elementos ya desarrollados y así obtener más fácilmente las indicadas sumas infinitas.

BIBLIOGRAFÍA

1. Alexei Ivanovich MARKUSHEVICH, Teoría de las funciones de variable Compleja, Editorial MIR. Moscú.
2. Jerrold E. MARSDEN, Michael HOFFMAN, Análisis Básico de Variable Compleja, Editorial Trillas.
3. KRASNOV M.L.KISELIOV, A, I MAKARENKO, Funciones de Variable Compleja Editorial MIR Moscú.

ANÁLISIS DE PRECISIÓN DE ESTIMADORES EN TÉCNICAS DE MUESTREO

Lic. Jaime Tito Pinto Ajhuacho

✉ *titojaime_pinto@yahoo.com*

RESUMEN

La información permite adquirir el conocimiento necesario para la toma de decisiones en diversas áreas donde se lo requiera, existen diversas técnicas para la captura de datos, como las encuestas por muestreo, que se apoyan diseños muestrales con diversidad de técnicas, en ellos es bueno analizar la precisión del estimador, porque interesa conocer o aproximarse al parámetro poblacional.

Analizar y comentar las técnicas sobre la precisión es importante, sobre todo comparar entre ellas y ver cual se aproxima más al parámetro poblacional.

Para este propósito apoyado en una información muestral, se analizó y comparo técnicas como el Muestreo Aleatorio Simple, Muestreo Estratificado y el Método de Postestratificación.

PALABRAS CLAVE

Comparación de precisión de estimadores

ABSTRACT

The information allows to acquire the necessary knowledge for decision-making in various areas where it is required, there are various techniques for data capture, such as sample surveys, which are supported by sample designs with a variety of techniques, in which it is good to analyze the precision of the estimator, because it is of interest to know or approximate the population parameter.

Analyzing and commenting on the precision techniques is important, especially comparing between them and seeing which one is closest to the population parameter.

For this purpose, supported by sample information, techniques such as Simple Random Sampling, Stratified Sampling and the Post-Stratification Method were analyzed and compared.

KEYWORDS

Estimator precision comparison

La información que es un grupo organizado de datos procesados que integran un mensaje sobre un determinado ente o fenómeno; permiten adquirir el conocimiento necesario para la toma de decisiones en diversas áreas donde se lo requiera.

El Dato, que es una expresión que explica las características de algo que se esté analizando o simplemente se quiere conocer, es viabilizado acudiendo a diversas técnicas para su captura de datos, como las encuestas por muestreo, que se apoyan diseños muestrales con diversidad de técnicas, en ellos es bueno analizar la precisión del estimador, porque interesa conocer o aproximarse al parámetro poblacional.

Las encuestas por muestreo, al igual que cualquier investigación profunda, se ven afectadas por el error ajeno al muestreo y el error que influye en los resultados de las investigaciones por muestreo, lo constituye los errores muestrales, los que están estrechamente relacionados con el diseño estadístico utilizado para la selección de la muestra; y que mediante un buen esquema de muestreo y proceso de estimación, es posible reducirlos considerablemente.

Por lo que el soporte básico de los estudios por muestreo, es el de proporcionar a partir de una muestra, resultados o estimaciones.

Analizar y comentar las técnicas sobre la precisión es importante, sobre todo comparar entre ellas cual se aproxima más al parámetro poblacional.

Para este propósito apoyado en una información muestral, se analizó y comparo técnicas como la del Muestreo Aleatorio Simple, Muestreo Estratificado y el Método de Postestratificación.

El Muestreo Estratificado brinda una buena precisión con sus estimadores, pero a veces se da el caso que no se dispone de la información para la realización de los estratos y realizamos la aplicación de otra técnica como por ejemplo el Muestreo Aleatorio Simple, y posteriormente podemos con la información muestral establecer los estratos, porque el muestreo estratificado ofrece una variación relativa menor.

Recordemos que el muestreo estratificado realiza primero una partición de la población en subpoblaciones que se denominan estratos, y dentro de cada estrato se realiza el muestreo de forma independiente.

Las utilidades del muestreo estratificado son:

- Sirve cuando se quiere obtener una precisión distinta para cada subpoblación. De esta forma se puede controlar qué muestra pertenece a cada estrato, y así controlar su precisión.

- Se utiliza también cuando es necesario plantear distintas tácticas de muestreo según las subpoblaciones.

- Si los estratos que se utilizan son más homogéneos que la población, la utilización del muestreo estratificado permite ganar precisión frente al aleatorio simple.

Planteando metodologías que minimicen la varianza del estimador, se puede a posteriori asignar las unidades de una muestra a los estratos, que trabaja el Muestreo Estratificado y de este modo obtener mejores estimaciones; se puede advertir que los tamaños muestrales en cada estrato variaran de tamaño.

Estratificando a "a posterior" se puede estimar la media poblacional $\bar{y}$ mediante una media pos estratificada $\bar{y}_{post}$. Así, el tamaño muestra en el estrato h, n_h es aleatorio antes de seleccionar la muestra y fijo una vez seleccionada; el estimador postestratificado será similar al estratificado.

Este método de postestratificacion, ofrece una buena precisión frente a otras técnicas de muestreo, se dan algunas condiciones para garantizar su efectividad:

a).- *La muestra debe ser suficientemente grande.* Si los estratos varían mucho en tamaño, la muestra debe ser más grande para garantizar que caen suficientes observaciones en cada estrato.

b).- Las ponderaciones W_h no están exentas

de errores, pues N_h suele ser una estimación, pero se supone que el nivel de error cometido en esta estimación es despreciable.

El estimador de la media y su varianza son:

1).- $\bar{y}_{post}$ es un estimador insesgado de $\bar{y}$

2).- $V(\bar{y}_{post}) \approx \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2$

El estimador de la varianza del estimador de la media es:

$$\hat{V}\left(\bar{y}_{post}\right) = \frac{1-f}{n}\sum_{h=1}^{m} W_h . s_h^2 + \frac{1}{n^2}.\sum_{h=1}^{m}(1-W_h).s_h^2$$

Donde es s_h^2 información muestral.

Para mostrar las aplicaciones de las técnicas mencionadas, acudimos a una información muestral de una encuesta de Uso de la tierra en el Departamento de Chuquisaca, año 2008, cuya muestra estaba conformada por 834 registros.

La variable que se analizo fue "Superficie cultivada" (Has.) y se procedió a la estimó de la media y el Total, utilizando un nivel de confianza del 95%.

PRECISIÓN POR EL MUESTREO ALEATORIO SIMPLE.-

Estimación de la superficie media cultivada.

N=73.388 Unidades de Producción

n = 84 Unidades de Producción.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2{,}0814 \text{ Has.}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 7{,}2944$$

$$\hat{V}(\bar{y}) = \frac{s^2}{n}.\frac{(N-n)}{N} = 0{,}008646$$

$$\sqrt{\hat{V}(\bar{y})} = \sqrt{\frac{s^2}{n}.\frac{(N-n)}{N}} = 0{,}09299 \text{ Has.}$$

$$\hat{\bar{Y}}_{Inferior} = \bar{y} - z\sqrt{\hat{V}(\bar{y})} = 1{,}8991$$

$$\hat{\bar{Y}}_{Superior} = \bar{y} + z\sqrt{\hat{V}(\bar{y})} = 2{,}2637$$

$$CV(\bar{y}) = \frac{\sqrt{\hat{V}(\bar{y})}}{\bar{y}} * 100 = 4{,}4677$$

Estimación del total de Superficie Cultivada.

$$\hat{Y} = N\,\bar{y} = 73388\,(2{,}0814) = 152.749{,}8 \text{ Has.}$$

$$\hat{Y}_{Inferior} = N\left(\bar{y} - z\sqrt{\hat{V}(\bar{y})}\right) = 139.371$$

$$\hat{Y}_{Superior} = N\left(\bar{y} + z\sqrt{\hat{V}(\bar{y})}\right) = 166.128$$

$$\hat{V}(\hat{Y}) = \hat{V}(N\bar{y}) = N^2\hat{V}(\bar{y}) = 46.565.614{,}21$$

$$\sqrt{\hat{V}(N\bar{y})} = \sqrt{46565614{,}21} = 6.823{,}90$$

$$CV(N\bar{y}) = \frac{\sqrt{\hat{V}(N\bar{y})}}{\hat{Y}} * 100 = 4{,}4677\ \%$$

PRECISIÓN POR EL ESTRATIFICADO.

En el ***muestreo estratificado***, para el procedimiento de estimación, se consultó sobre una estratificación en el área de estudio, dando el siguiente corte de productores, de 0 a 1 Has.(Pequeños), de 1 a 4 Has.(Medianos) y Mayor a 4 Has.(Grandes),bajo esta referencia se dividido en tres estratos.

Estimación de la media de población:

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h$$

La varianza de la estimación $\bar{y}_{st}$:

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

$\hat{Y}_{st} = N.\bar{y}_{st}$ es la estimación del Total de la población Y, su varianza es:

$$V(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h}$$

La estimación insesgada de la varianza de $\bar{y}_{st}$ es:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{\hat{s}_h^2}{n_h}$$

Media de la población: $\bar{y}_{st} \pm z\sqrt{\hat{V}(\bar{y}_{st})}$

Total de la población:

$$N.\bar{y}_{st} \pm N\,z\sqrt{\hat{V}(\bar{y}_{st})} = N\left(\bar{y}_{st} \pm z\sqrt{\hat{V}(\bar{y}_{st})}\right)$$

Con la información muestral proporcionada, se armó el siguiente Cuadro:

Cuadro 1.

ESTRATIFICACIÓN DE LA SUPERFICIE CULTIVADA (Has.)

	ESTRATO 1	ESTRATO 2	ESTRATO 3	TOTAL
nh	339	410	85	**834**
$\bar{y}_h$	0,64011	2,1374	7,56011	**2,0814**
s_h^2	0,09981	0,4951	30,7518	
S_h	0,3159	0,7036	5,5454	
CV	49,28	32,87	72,91	
Nh	29.831	40.038	3.519	**73.388**
W_h	0,40648	0,54557	0,04795	**1**
$W_h\,\bar{y}_h$	0,260196	1,16612	0,362512	**1,78883**
$\hat{V}(\bar{y}_{st})$				**0,0012155**

Fuente: Elaboración Propia

El estimador de la media estratificada:

$$\bar{y}_{st} = \sum_{i=1}^{L} W_h\bar{y}_h = 1{,}78883$$

El error de estimación:

$$\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_h(N_h - n_h)\frac{\hat{S}_h^2}{n_h} = 0{,}001216$$

$$Lim\ Inf\ (\bar{y}) = \bar{y}_{st} - z\sqrt{\hat{V}(\bar{y}_{st})} = 1{,}7205$$

$$Lim\ Sup\ (\bar{y}) = \bar{y}_{st} + z\sqrt{\hat{V}(\bar{y}_{st})} = 1{,}8572$$

$$\hat{CV}(\bar{y}_{st}) = \frac{\sqrt{\hat{V}(\bar{y}_{st})}}{\bar{y}_{st}}100 = \frac{\sqrt{0{,}001216}}{1{,}78883}100 = 1{,}9494\%$$

El Total estimado de Superficie cultivada:

$$\hat{Y}_{st} = N\,\bar{y}_{st} = 73.388\,(1{,}78883) = 131.278{,}6\ Has.$$

El error de estimación:

$$\hat{V}(\hat{Y}_{st}) = N^2\hat{V}(\bar{y}_{st}) = 73.388^2(0{,}001216) = 6.549.13$$

$$Lim\ Inf\ (\hat{Y}) = \hat{Y}_{st} - z\sqrt{\hat{V}(\hat{Y}_{st})} = 126.262{,}72$$

$$Lim\ Sup(\hat{Y}) = \hat{Y}_{st} + z\sqrt{\hat{V}(\hat{Y}_{st})} = 136.294{,}47$$

$$\hat{CV}(\hat{Y}_{st}) = \frac{\sqrt{\hat{V}(\hat{Y}_{st})}}{\hat{Y}_{st}}100 = \frac{\sqrt{6.549.131{,}03}}{131.278{,}6}100 = 1{,}94$$

PRECISIÓN POR POSTESTRATIFICACIÓN

Utilizamos el estimador de la media estratificada.

$$\bar{y}_{st} = \sum_{i=1}^{L} W_h\bar{y}_h = 1{,}78883$$

El error de esta estimación:

$$\hat{V}(\bar{y}_{post}) \approx \frac{1-f}{n}\sum_{i=1}^{L} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)s_h^2$$

$$\hat{V}(\bar{y}_{post}) = 0{,}0008971$$

$$\sqrt{\hat{V}(\bar{y}_{post})} = \sqrt{0{,}0008971} = 0{,}0290517$$

$$Lim\ Inf\ (\bar{y}) = \bar{y}_{st} - z\sqrt{\hat{V}(\bar{y}_{post})} = 1{,}731888$$

$$Lim\ Sup\ (\bar{y}) = \bar{y}_{st} + z\sqrt{\hat{V}(\bar{y}_{post})} = 1{,}845771$$

$$\hat{CV}(\bar{y}_{post}) = \frac{\sqrt{\hat{V}(\bar{y}_{post})}}{\bar{y}_{post}}100 = 1{,}6743\ \%$$

La superficie cultivada total estimada:

$$\hat{Y}_{post} = N\,\bar{y}_{post} = 73.388\,(1{,}78883) = 131.278{,}6$$

El error de estimación:

$$\hat{V}\left(\hat{Y}_{post}\right) = N^2\hat{V}\left(\bar{y}_{post}\right) = 4.831.599{,}9$$

$$\sqrt{\hat{V}\left(\hat{Y}_{post}\right)} = \sqrt{4.831.599{,}9} = 2.198{,}09$$

$$Lim\ Inf\,(\hat{Y}) = \hat{Y}_{post} - z\sqrt{\hat{V}\left(\hat{Y}_{post}\right)} = 126.970{,}34$$

$$Lim\ Sup(\hat{Y}) = \hat{Y}_{post} + z\sqrt{\hat{V}\left(\hat{Y}_{post}\right)} = 135.586{,}86$$

$$\hat{CV}\left(\hat{Y}_{post}\right) = \frac{\sqrt{\hat{V}\left(\hat{Y}_{post}\right)}}{\hat{Y}_{post}}100 = 1{,}6743\ \%$$

Se trata de un error muy aceptable. Los resultados de las tres técnicas podemos sintetizarlo en el siguiente cuadro.

Cuadro 2

ANÁLISIS DE LAS ESTIMACIONES

ESTIMACIÓN DE SUPERFICIE CULTIVADA (Has.), DESVIACIÓN ESTANDAR, COEFICIENTE DE VARIACIÓN E INTERVALO DE CONFIANZA EN EL DEPARTAMENTO DE CHUQUISACA, SEGÚN VARIABLE DE ESTUDIO, ENCUESTA DE USO DE LA TIERRA AÑO 2008.

TÉCNICA	ESTIMADOR	ESTIMACIÓN DE SUPERFICIE. CULTIVADA	DESVIACIÓN ESTÁNDAR	COEFICIENTE DE VARIACIÓN (%)	INTERVALO DE CONFIANZA 95%	
					LÍMITE INFERIOR	LIMITE SUPERIOR
M.A.S	Media	2,0814	0,0922	4,4677	1,8991	2,2637
Estratificado	Media	1,78883	0,03486	1,9494	1,7205	1,8572
Postestratificado	Media	1,78883	0,0290517	1,6743	1,731888	1,845771
M.A.S	Total	152.749	6.823,90	4,4677	139.371,15	166.128,4
Estratificado	Total	131.278,6	2.559,12	1,9494	126.262,72	136.294,47
Postestratificado	Total	131.278,6	2.198,09	1,6743	126.970,34	135.586,86

Fuente: Elaboración Propia

En el cuadro 2 se puede observar que la superficie cultivada, utilizando el muestreo aleatorio es 152.749 Has. conteniendo un error estándar de 6.823 Has., siendo su coeficiente de variación de 4,46 por ciento.

El Estratificado, da un valor estimado de 131.278 Has., con un error estándar de 2.559 Has. y su coeficiente de variación es 1,94 por ciento, y se ve que el Método de Post estratificación, estima la superficie cultivada en 131.278 Has, con un error estándar de 2.198 Has y un coeficiente de variación de 1,67 por ciento.

Se puede apreciar $\hat{V}_{POST} < \hat{V}_{ESTRA} < \hat{V}_{MAS}$ el estratificado en variabilidad es menor que el Muestreo aleatorio simple y la variabilidad es más menor en el Post estratificado, mostrándonos que hay mayor precisión, es decir su estimación está más cerca del parámetro poblacional.

Corresponde determinar una estimación con cierto nivel de error de muestreo que sea útil para la toma de decisiones, de acuerdo con el grado de fiabilidad que precisa, y se ve que analizando el Método de Postestratificacion, ofrece una mayor precisión.

BIBLIOGRAFÍA

1. Cochran, Willian G. (1996). Técnicas de muestreo. Ed. Continental, 10ed, México, 513p.
2. Pérez López, Cesar. (2000). Técnicas de muestreo estadístico. Alfaomega, México, 603p.
3. Kish, Leslie. (1979). Muestreo de encuestas. Trillas, México, 739p.
4. Azorín Poch, Francisco. (1972). Curso de muestreo y aplicaciones. Aguilar, Madrid, 375p.

MODELO DE RESPUESTA ALEATORIZADA DE WARNER PARA INCREMENTAR LA PROBABILIDAD DE OBTENER RESPUESTAS SINCERAS A PREGUNTAS SENSIBLES

M. Sc. Dindo Valdez Blanco

✉ *dindovaldez@hotmail.com*

RESUMEN

En esta investigación se estudia el modelo de respuesta aleatorizada propuesto por Warner (Warner, 1965) con el propósito de disminuir el sesgo de respuesta cuando se formulan preguntas sensibles y/o delicadas, la aplicación se realiza en la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés. Para la aplicación se considera preguntas sensibles o delicadas, como el comportamiento a realizar trampa en los exámenes y el consumo de drogas, aplicando el método de pregunta aleatorizada de Warner y el método de pregunta directa. Esencialmente el método de Warner involucra una técnica de aleatorización de tal manera que el entrevistado debe responder a las preguntas sensibles de acuerdo al resultado que arroje el método, dichos procedimientos pueden ser: juegos de monedas, maso de cartas ruletas giratorias, entre otros. La técnica de aleatorización permite calcular el estimador de la proporción de manera indirecta, dando al entrevistado el anonimato y permitiendo una respuesta sincera. Por lo tanto, se establece una relación probabilística entre una respuesta dada y la pregunta sensible. Finalmente se realiza una comparación entre el método de la pregunta directa que establece los estimadores usuales para la proporción de éxitos y el método de respuesta aleatorizada de Warner, llegando a determinar que el método de Warner es efectivo.

PALABRAS CLAVE

Pregunta sensible, modelo de respuesta aleatorizada, sesgo de respuesta, encuestas por muestreo.

ABSTRACT

In this research, the randomized response model proposed by Warner (Warner, 1965) is studied with the purpose of reducing the response bias when sensitive and / or delicate questions are asked, the application is carried out in the Faculty of Pure and Natural Sciences of the Universidad Mayor de San Andrés. For the application, sensitive or delicate questions are considered, such as behavior to cheat in exams and drug use, applying the Warner randomized question method and the direct question method. Essentially the Warner method involves a randomization technique in such a way that the interviewee must answer sensitive questions according to the result of the method, such procedures can be: coin games, deck of rotating roulette cards, among others. The randomization technique allows to calculate the estimator of the proportion indirectly, giving the interviewee anonymity and allowing a sincere answer. Therefore, a probabilistic relationship is established between a given answer and the sensitive question. Finally, a comparison is made between the direct question method that establishes the usual estimators for the proportion of successes and Warner's randomized response method, determining that Warner's method is effective.

KEYWORDS

Sensitive question, randomized response model, response bias, sample surveys.

1. INTRODUCCIÓN

En estudios sobre temáticas delicadas o muy personales, se presentan dos problemas: no dan respuesta y/o no contestan con veracidad.

Estos dos problemas generan sesgos de muestreo. Por lo cual radica la importancia de estudiar metodologías que mejoren la tasa de respuesta veraz ante este tipo de preguntas sin comprometer al entrevistado ante estas preguntas delicadas.

El objetivo principal es estudiar el modelo de respuesta aleatorizada propuesto por Warner y aplicar el mismo a los estudiantes de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés, que se encuentran matriculados en la gestión 2019 para analizar el comportamiento de los alumnos de la facultad frente a preguntas sensibles.

2. MATERIALES Y MÉTODOS

La metodología para la presente investigación comprende de dos partes: la implementación de una encuesta y el método de Warner para la estimación de la proporción de personas con una característica sensible. La encuesta permitirá aplicar el método de entrevista directa a los estudiantes de la universidad matriculados en la gestión 2019, para comparar los resultados con el método de Warner. A partir de la en cuesta se podrá analizar si el método de respuesta aleatorizada asegura realmente el anonimato de los sujetos y aumenta la probabilidad de obtener respuestas sinceras a preguntas sensibles reduciendo el error.

El Modelo de Respuesta Aleatorizada de Warner

Los individuos de una población pueden diferenciarse en cuanto si son portadores de un rasgo sensible X o no. Luego se busca la proporción π de los portadores de características sensibles en la población, donde

$$\pi = P(X = 1) \text{ y } 1 - \pi = P(X = 0),$$

también se puede describir como la probabilidad de llevar la característica sensible. A los encuestados se les presentan dos declaraciones siguiendo el siguiente esquema:

Declaración A: Soy el portador de la característica sensible X.

Declaración B: No soy el portador de la característica sensible X.

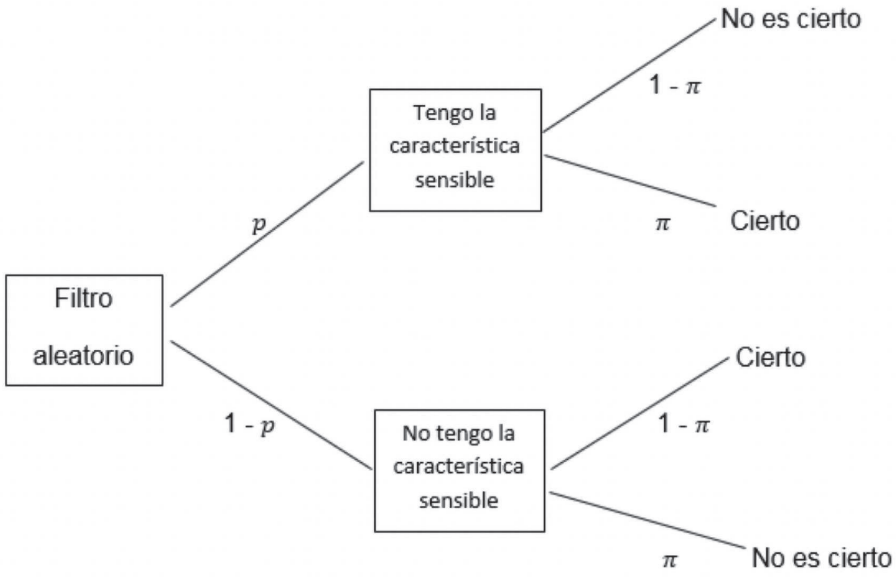
La selección modelo de Warner debe satisfacer dos condiciones:

1. Las probabilidades (p) de selección de las dos afirmaciones se conocen de antemano y no son iguales a 0.5.

2. El entrevistador no conoce el resultado del experimento aleatorio, solo el encuestado sabe cuál de las dos afirmaciones fue seleccionada. Luego solo indica si la declaración seleccionada se aplica a él o no. La figura 1 muestra esquemáticamente el principio de la encuesta a partir de diagrama de árbol.

Figura 1

Representación de la técnica de respuesta aleatoria según Warner (1965).



Fuente: Elaboración Propia

Cálculo de la probabilidad de tener la característica sensible λ:

$$\lambda = p\pi + (1-p)(1-\pi) \;\; ; \;\; p \neq 0.5$$

Donde

λ, probabilidad de una respuesta afirmativa

π, probabilidad de tener la característica sensible

p, probabilidad de responder la pregunta sensible

Por tanto, la estimación de la proporción de personas que tienen la característica sensible es

$$\hat{\pi} = \frac{\lambda + p - 1}{2p - 1}$$

Y su varianza es

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}$$

Estimador de la pregunta directa

En el caso de preguntas directas, las estimaciones de la proporción de casos que responden si a la pregunta sensible son:

$$\hat{\pi} = \frac{\sum x_i}{n}$$

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

Con $x_i \sim Bernoulli(\pi)$.

Comparación del Modelo de respuesta aleatorizada de Warner con el modelo de pregunta directa

La equivalencia entre el modelo de Warner y el modelo de entrevista directa, se proporciona una base para comparar el modelo de Warner con el Modelo de respuesta directa.

A partir de sus varianzas:

$$Var(\hat{\pi}_w) = Var(\hat{\pi}_D) + \frac{p(1-p)}{n(2p-1)^2}$$

Donde Var($\hat{\pi}_w$) es la varianza del Modelo de Warner. Utilizando el criterio de varianzas se obtiene lo siguiente:

$$Var(\hat{\pi}_w) - Var(\hat{\pi}_D) = \frac{p}{n(1-p)(2p-1)^2} h_{WT}(p|\pi)$$

Con:

$$h_{WT}(p|\pi) = (4\pi - 3)p^2 + (2 - 4\pi)p + \pi$$

Para encontrar los valores de p para diferentes valores de π primero se resuelve la ecuación de segundo grado de la función $h_{WT}(p|\pi)$ en términos de p:

$$p = \frac{2\pi - 1 \pm \sqrt{1-\pi}}{4\pi - 3} \quad ; \quad \pi \neq \frac{3}{4}$$

Eficiencia relativa del Modelo de Warner y el Modelo Directo

En este sentido se considera que la eficiencia relativa del modelo de Warner (p≠0.5) para el modelo directo; es decir,

$$ER_{W\to D}(\pi, p) = \frac{\text{Var}(\hat{\pi}_W)}{\text{Var}(\hat{\pi}_D)} = 1 + \frac{p(1-p)}{\pi(1-\pi)(2p-1)^2}$$

La eficiencia relativa es el cociente de varianzas y es independiente del tamaño de la muestra n, y solo depende de los parámetros π y p, para el modelo de Warner como para el modelo Directo como se muestra a continuación en la tabla 1, se observa que a medida que la probabilidad de responder a la pregunta sensible se aproxima a 0.5, la eficiencia relativa del método directo aumenta exponencialmente, la eficiencia del modelo directo es mayor que la del modelo de Warner, en particular cuando ($0.48 \leq p < 0.50$) que es el rango óptimo para que la privacidad de los encuestados este protegido, la eficiencia del modelo directo es de aproximadamente 2500 hasta 6942 veces mayor que del modelo de Warner.

Tabla 1

Eficiencia Relativa del método de Ward en relación al método directo para varias combinaciones de π y p

π	p									
	0,10	0,20	0,30	0,40	0,49	0,51	0,60	0,70	0,80	0,90
0,10	2,56	5,94	15,58	67,67	6942,67	6942,67	67,67	15,58	5,94	2,56
0,20	1,88	3,78	9,20	38,50	3905,69	3905,69	38,50	9,20	3,78	1,88
0,30	1,67	3,12	7,25	29,57	2976,00	2976,00	29,57	7,25	3,12	1,67
0,40	1,59	2,85	6,47	26,00	2604,13	2604,13	26,00	6,47	2,85	1,59
0,49	1,56	2,78	6,25	25,01	2501,00	2501,00	25,01	6,25	2,78	1,56
0,51	1,56	2,78	6,25	25,01	2501,00	2501,00	25,01	6,25	2,78	1,56
0,60	1,59	2,85	6,47	26,00	2604,13	2604,13	26,00	6,47	2,85	1,59
0,70	1,67	3,12	7,25	29,57	2976,00	2976,00	29,57	7,25	3,12	1,67
0,80	1,88	3,78	9,20	38,50	3905,69	3905,69	38,50	9,20	3,78	1,88
0,90	2,56	5,94	15,58	67,67	6942,67	6942,67	67,67	15,58	5,94	2,56

Fuente: Elaboración Propia

Grado de protección de la privacidad

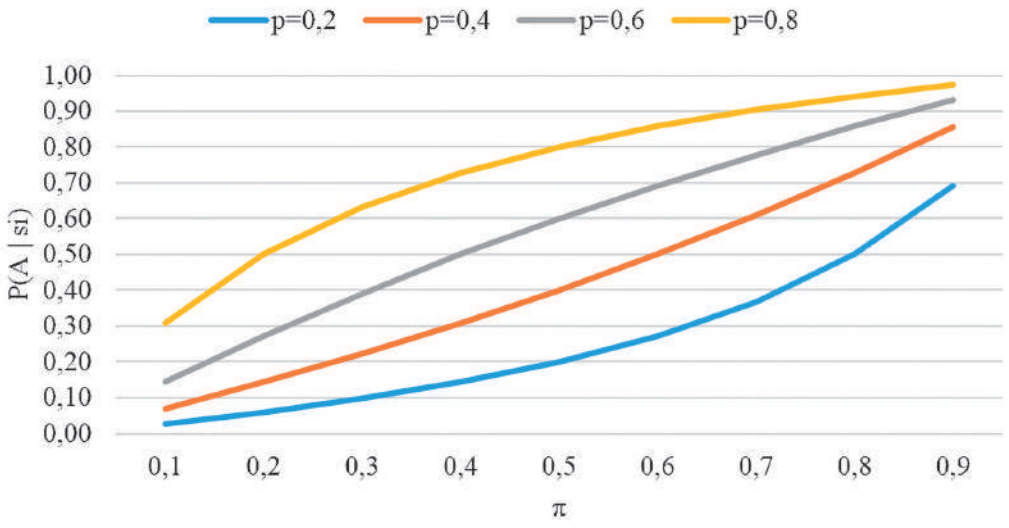
Se define la probabilidad condicional, que el encuestado realmente posea la característica sensible dado que responde afirmativamente.

$$\mathrm{GPP} = \mathrm{P(A|si)} = \frac{\pi p}{\pi p + (1-\pi)(1-p)}$$

Donde A indica que el entrevistado posee la característica sensible.

Figura 2

Grado de proteción de la privacidad para los que responden Si



Fuente: Elaboración Propia

Figura 3

Modelo de preguntas con respuesta aleatorizada de Warner

A continuación, contesta la opción **A** si tu cédula de identidad termina en 1 o 2, en caso contrario contesta la opción **B.**

opción **A**	opción **B**
Nunca hice trampa en un examen de la universidad	Alguna vez hice trampa en un examen de la universidad
SI ◯	NO ◯

A continuación, contesta la opción **A** si tu cédula de identidad termina en 8 o 9, en caso contrario contesta la opción **B.**

opción **A**	opción **B**
Nunca he consumido drogas	He consumido drogas en alguna ocasión
SI ◯	NO ◯

Fuente: Elaboración Propia

3. RESULTADOS

Tabla 2

Estimaciones de las preguntas sensibles por el método de Warner y el método Directo

Pregunta sensible	Método de Warner		Método Directo	
	π	IC	π	IC
¿Alguna vez hiciste trampa en un examen de la universidad?	0,5017	0,4290 - 0,5740	0,4160	0,3605 - 0,4715
¿Has consumido drogas en alguna ocasión?	0,1667	0,1002 - 0,2332	0,0594	0,0327 - 0,0860

Fuente: Elaboración Propia

Tabla 3

Estimaciones de las preguntas sensibles por el método de Warner por sexo

Pregunta sensible	Masculino		Femenino	
	π	IC	π	IC
¿Alguna vez hiciste trampa en un examen de la universidad?	0,4817	0,3808 - 0,5825	0,5233	0,4190 - 0,6276
¿Has consumido drogas en alguna ocasión?	0,2333	0,1377 - 0,3289	0,0950	0,0038 - 0,1861

Fuente: Elaboración Propia

4 . CONCLUSIONES

1. La aplicación del modelo de respuesta aleatorizada de Warner ha demostrado que es una técnica que permite obtener mejores resultados en encuestas con preguntas sensibles.

2. La estimación de la proporción de estudiantes que hacen "trampa en los exámenes" utilizando el estimador de Warner es 50,17% mientras que en el diseño de entrevista directa es de 41,6%.

3. La estimación de "Consumo de drogas en alguna ocasión" utilizando el estimador de Warner es de 16,67%, mientras que con el diseño de entrevista directa es de 5,94%.

4. La desventaja principal del método de Warner radica en lo que respecta a la capacitación de los encuestadores y el tiempo que requiere explicar al entrevistado el cuestionario.

A la fecha existen otros métodos de estimación frente a preguntas sensibles, algunos de estos modelos son: los modelos de respuesta no aleatorizada y el modelo triangular de respuesta no aleatorizada. Se recomienda estudiar dichos métodos para futuros trabajos de investigación.

BIBLIOGRAFÍA

1. Basulto, J. (1982). El diseño de respuesta aleatorizada de Warner. Un modelo de superpoblación., (96), 51 a 62.
2. Guo-Liang Tian, J.-W. Y. (2007). A new non-randomized model for analysis sensitive questions with binary outcomes. Statistics in medicine, 26(23), 4238-52. http://doi.org/10.1002/sim.2863.
3. Liang, T. G., & Lai, T. M. (2008). Two new models for survey sampling with sensitive characteristic: desing and analysis. Metrica, 251-263.
4. Nayak, T. K. (1994). On randomized response surveys for estimating a proportion. Communications in Statistics - Theory and Methods, 23(11), 3303-3321. http://doi.org/10.1080/03610929408831448.
5. Warner, S. L. (1965). Randomized response: a survey tecnique for eliminating evasive answer bias. Journal of Applied Psychology (Vol. 60).

Calle 27 de Cota Cota
Bloque F.C.P.N. - Primer Piso

La Paz - Bolivia