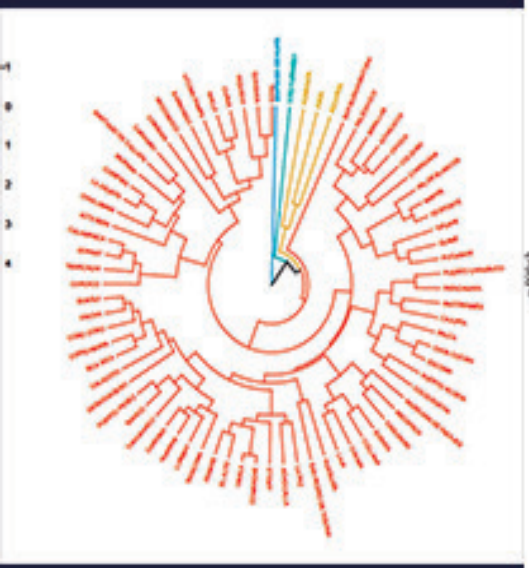
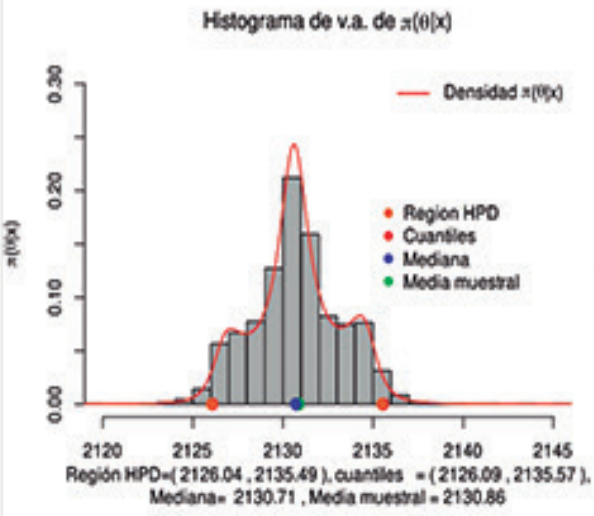
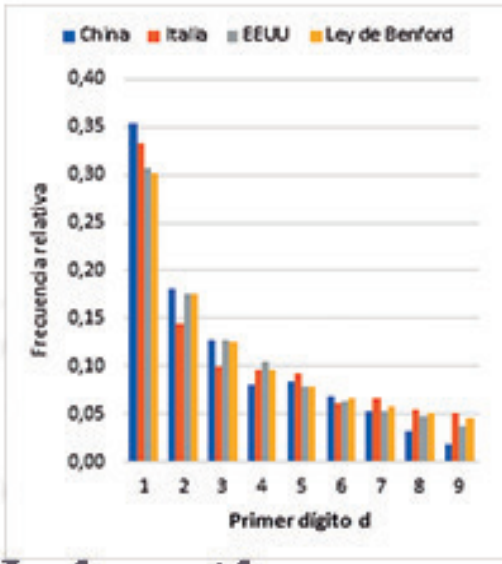
Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales
Carrera de Estadística

Varianza

Revista del Instituto de Estadística Teórica y Aplicada







$$E[X^2]-2X\mu+\mu^2]$$
$$=E[X^2]-2\mu E[X]+\mu^2$$
$$=E[X^2]-2\mu^2+\mu^2$$
$$=E[X^2]-\mu^2$$
$$=E[X^2]-E$$

Varianza

Revista de Investigación de la Carrera de Estadística

Publicación del Instituto de Estadística Teórica y Aplicada

Número 18
Octubre, 2021
La Paz - Bolivia

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

ISSN 2789-3529 VERSIÓN EN LÍNEA

DEPÓSITO LEGAL
4-1-285-2021 P.O**.**

REVISTA VARIANZA
Nº 18 - Octubre, 2021

DIRECTOR CARRERA DE ESTADÍSTICA
M. Sc. Juan Carlos Flores López

DIRECTOR a.i. I.E.T.A. - EDITOR
M. Sc. Fernando Oday Rivero Suguiura

DIAGRAMACIÓN Y DISEÑO
Lic. María Zulema Vargas Cerrudo

Los artículos presentados son entera responsabilidad de los autores

VISIBILIDAD REVISTAS BOLIVIANAS

La Paz - Bolivia
Edificio Bloque FCPN - Campus Cota Cota
Teléfonos: 2612824 -2612844
Email: ieta@umsa.bo
Página web: https://ieta.umsa.bo/ediciones-varianza

COMITÉ EDITORIAL NACIONAL

COMITÉ EDITORIAL INTERNACIONAL

Dedicada en memoria a los fallecidos de nuestra casa superior de estudios a consecuencia de la pandemia

PRESENTACIÓN

La revista Varianza pertenece a la carrera de Estadística y al Instituto de Estadística Teórica y Aplicada (IETA), de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés. Es una publicación destinada a la difusión de resultados de los trabajos de investigación logrados en la carrera y por los investigadores nacionales e internacionales.

A partir de la presente revista Varianza N° 18, será de publicación semestral y abierta para recibir contribuciones de autores nacionales e internacionales. De esta manera se pretende generar un espacio favorable para el ámbito académico, análisis, desarrollo difusión de las técnicas y metodología que tiene la estadística. Por lo tanto, sus receptores son todos los profesionales e investigadores de cualquier área que involucre la temática.

Esperando sea de utilidad para la comunidad universitaria y el conjunto de nuestra sociedad, damos a conocer los distintos temas de investigación desarrollados.

La dirección a mi cargo, agradece y reconoce la colaboración del Comité Editorial por sus aportes, en el análisis crítico de los artículos, así como sus observaciones. Así mismo, agradecer al personal administrativo de la carrera, del IETA y a todo el personal que trabaja en la edición de la revista, para asegurar la calidad y difusión de su contenido a nivel nacional e internacional.

M. Sc. Juan Carlos Flores López
DIRECTOR CARRERA DE ESTADÍSTICA

ÍNDICE

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 1-7

MODELOS MARGINALES APLICACIÓN A DATOS DE PANEL

MARGINAL MODELS APPLICATION TO PANEL DATA

Ramiro Coa Clemente[1]

Instituto de Estadística Teórica y Aplicada, Universidad Mayor de San Andrés, La Paz - Bolivia

✉ *clementecoa@gmail.com*

Artículo recibido: 2021-08-15
Artículo aceptado: 2021-09-07

RESUMEN

El objetivo de este artículo es aplicar el modelo marginal en el análisis de datos tipo panel sobre la situación nutricional de los recién nacidos. Luego de examinar sucintamente los aspectos centrales de los modelos marginales, se revisa brevemente el método de ecuaciones de estimación generalizada (EEG), un método apropiado para la estimación de este tipo de modelos. Con base en un modelo marginal logístico con patrón de correlación intercambiable, se concluye que fumar durante el embarazo y un servicio prenatal inadecuado incrementan significativamente la probabilidad de un nacimiento con bajo peso al nacer.

Palabras clave: *Modelo marginal, ecuaciones de estimación generalizada, datos tipo panel*

ABSTRACT

The aim of this article is to apply the marginal model in the analysis of panel data on the nutritional status of newborns. After a brief review of the central aspects of marginal models, the generalized estimating equations (GEE) method, an appropriate method for estimating this type of model, is briefly reviewed. Based on a logistic marginal model with exchangeable correlation pattern, it is concluded that smoking during pregnancy and inadequate prenatal care significantly increase the probability of a low birth weight.

Keywords: *Marginal model, generalized estimating equations, panel data*

INTRODUCCIÓN

Los modelos lineales generalizados constituyen una clase unificada de modelos para análisis de regresión con respuesta discreta o continua y observaciones independientes. No es posible, sin embargo, la aplicación directa de estos modelos a datos tipo panel debido a la correlación entre las observaciones obtenidas en las mismas unidades. Por tal razón, se consideran extensiones para datos panel. Hay muchas formas de extender los modelos lineales generalizados a fin de tomar en cuenta la correlación entre las observaciones. En este trabajo se revisa y aplica uno de ellos: los modelos marginales.

Los modelos marginales constituyen una metodología para analizar datos de panel cuando la variable respuesta es discreta o continua. Esta metodología no

[1] Ex-Director de Investigación en la Unidad de Análisis y Política Social de Bolivia (UDAPSO). Ex-Director Nacional de la Encuesta de Demografía y Salud, en 1989 y 1998. M.Sc. Estadística, Pontificia Católica de Chile. Candidato a Doctor en Demografía. Docente investigador de la carrera de Estadística, Universidad Mayor de San Andrés. ORCID: 0000-0002-2955-0204

requiere supuestos distribucionales para las observaciones. El método depende solamente del supuesto de cómo la respuesta media está relacionada con las covariables. La evasión de supuestos distribucionales conduce al método de estimación conocido como ecuaciones de estimación generalizada, EEG. El enfoque de EEG es una buena alternativa a la estimación de máxima verosimilitud.

MÉTODO

El modelo

Para el análisis de los datos tipo panel sobre la situación nutricional de los recién nacidos se usa el modelo marginal. Según Fitzmaurice et al. (2011), el modelo marginal para datos de panel es especificado en tres partes:

i. La esperanza o media condicional de cada respuesta depende de las covariables a través de una función de enlace g conocida

$$g(\mu_{it}) = \eta_{it} = X'_{it}\beta$$

ii. La varianza condicional de la respuesta en cada ocasión es función de la media

$$Var\,(Y_{it}/X_{it}) = \phi\, h\,(\mu_{it})$$

donde $h\,(\mu_{it})$ es una función-varianza conocida y ϕ es un parámetro de escala.

iii. Se asume que existe asociación entre pares de observaciones dentro de los *clusters*. A partir de un modelo para las correlaciones por pares, la correspondiente matriz de varianzas-covarianzas de trabajo se construye como:

$$V_i = A_i^{1/2}\, Corr\,(Y_i)\, A_i^{1/2}$$

donde A_i es una matriz diagonal con $Var(Y_{it}/X_{it}) = \phi h(\mu_{it})$ a través de su diagonal y $Corr(Y_i)$ representa la matriz de correlación.

Es importante resaltar algunas características particulares de estos modelos. Los modelos marginales son una forma muy natural de extender los modelos lineales generalizados para tratar respuestas tipo panel correlacionadas y permiten realizar inferencias sobre medias poblacionales. La palabra marginal se usa para enfatizar que el modelo para la respuesta media en cada ocasión no depende de efectos aleatorios. En estos modelos se asume que las respuestas para los distintos *clusters* son independientes entre sí, pero, las medidas repetidas para el mismo *cluster* no son independientes. No requieren supuestos distribucionales para la respuesta, esto porque no hay una especificación completa de la distribución multivariada conjunta para respuestas discretas. La anulación de supuestos distribucionales conduce al método de estimación conocido como ecuaciones de estimación generalizada, una buena alternativa a la estimación de máxima verosimilitud.

Para el caso particular de una respuesta binaria, la especificación completa del modelo marginal, de acuerdo a Fitzmaurice et al. (2011), es expresada como:

i. $ln\,(\mu_{it}/(1 - \mu_{it})) = \eta_{it} = X'_{it}\beta$

ii. $Var\,(Y_{it}/X_{it}) = \mu_{it}\,(1 - \mu_{it})$

iii. $ln\,[\,\mathrm{RC}\,(Y_{it}, Y_{it'})/(X_{it}, X_{it'})] = \alpha_{tt'}$

donde la razón de chances entre dos respuestas para los momentos t y t' es definido como:

$$RC(Y_t, Y_{t'}) = \frac{Pr(Y_t = 1,\ Y_{t'} = 1)Pr(Y_t = 0,\ Y_{t'} = 0)}{Pr(Y_t = 1,\ Y_{t'} = 0)Pr(Y_t = 0,\ Y_{t'} = 1)}$$

Las ecuaciones de estimación generalizada

Cuando las variables respuesta son discretas, no es posible una especificación conveniente de su distribución mutivariante conjunta. Por este hecho, para los modelos marginales se requiere un método alternativo de estimación, alternativo al de máxima verosimilitud. El enfoque de ecuaciones de estimación generalizado representa esta alternativa. Este método proporciona un enfoque muy general y unificado para analizar respuestas correlacionadas que pueden ser discretas o continuas. La idea esencial detrás del enfoque de EEG es generalizar y extender las habituales ecuaciones de verosimilitud para un modelo lineal generalizado incorporando la matriz de varianzas-covarianzas del vector de respuestas. Inicialmente esta idea fue introducida por Wedderburn (1974). En el artículo de Wedderburn se asume independencia entre las observaciones y que la forma de la función varianza es una función conocida de la media sin la exigencia formal de que ellos se originen a partir de una distribución específica. Este es un supuesto menos fuerte que el proveniente de una distribución específica. En consecuencia, uno es libre de elegir cualquier parametrización de las funciones media y varianza, y aplicarlos en la ecuación de estimación.

El término ecuación de estimación generalizada indica que una ecuación de estimación no es el resultado de una derivación basada en la verosimilitud, es obtenido por la generalización de otra ecuación de estimación. La modificación que se hace para obtener una ecuación de estimación generalizada es introducir componentes de varianza de segundo orden directamente en la ecuación de estimación.

Para datos tipo panel, la ecuación de estimación de máxima quasiverosimilitud (MQV) para un modelo lineal generalizado, según Hardin and Hilbe (2013), es expresada como:

$$\Psi(\beta) = \sum_{i=1}^{n} X_{ji}'\ D_i\ V_i^{-1}\ \left(\frac{Y_i - \mu_i}{a(\phi)}\right) = 0$$

donde:

$$D_i = diag\left(\frac{\partial \mu_{it}}{\partial \eta_{it}}\right) \quad t = 1, \ldots, T_i$$

$$V_i = diag\left[h^{1/2}(\mu_{it})\right]\ I_{T_i * T_i}\ diag\left[h^{1/2}(\mu_{it})\right]$$

Notar que V_i es matriz diagonal, por lo que esta ecuación de estimación trata a las observaciones dentro de cada *cluster* como independientes.

Liang and Zeger (1986) propusieron una EEG que es una modificación de la ecuación de estimación de MQV. La modificación consiste en remplazar la matriz identidad con una matriz de correlación más general

$$\mathrm{V_i} = \mathrm{diag}\left[\mathrm{h}^{1/2}(\mu_{it})\right]\ \mathrm{R}(\alpha)_{\mathrm{T_i * T_i}}\ \mathrm{diag}\left[\mathrm{h}^{1/2}(\mu_{it})\right.$$

Adicionalmente, también se tiene una ecuación de estimación para los parámetros auxiliares α, la cual se la expresa como (Hardin and Hilbe, 2013).

$$\Psi(\alpha) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(\frac{\partial \varepsilon_{\mathrm{i}}}{\partial \alpha}\right)' \mathrm{H_i^{-1}}(\mathrm{W_i} - \varepsilon_{\mathrm{i}}) = 0_{\mathrm{q*1}}$$

donde $q = \binom{T_i}{2}$, W_i y ε_i son vectores de dimensión $q*1$, H_i es una matriz diagonal $q*q$, definidos como:

$$W_i = \left[r_{i1}r_{i2}, r_{i1}r_{i3}, \ldots, r_{iT_i-1}r_{iT_i}\right]'$$

$$H_i = Diag\ [Var\ (W_{it})]$$

$$\varepsilon_i = E(W_i)$$

y r_{it} es el *it-ésimo* residuo de Pearson.

Combinando las ecuaciones de estimación para los parámetros de la regresión y para los parámetros auxiliares, la completa EEG para modelos marginales está dada por:

$$\Psi(\beta,\alpha)=\begin{bmatrix}\sum_{i=1}^{n} X_{ji}'\ D_i\ V_i^{-1}\ \left(\frac{Y_i-\mu_i}{a(\phi)}\right)\\ \sum_{i=1}^{n}\left(\frac{d\varepsilon_i}{d\alpha}\right)'\ H_i^{-1}(W_i-\varepsilon_i)\end{bmatrix}$$

donde:

$$V_i=\mathrm{D}[h^{1/2}(\mu_{it})]\ \mathrm{R}(\alpha)\ \mathrm{D}[h^{1/2}(\mu_{it})]$$

En cada paso primero se estima R(α) y luego se lo usa para estimar β. Se declara convergencia cuando el cambio en las estimaciones de los parámetros es menor a algún valor o vector de valores pre-definidos, o cuando el cambio en la suma de los cuadrados de las devianzas es inferior a un valor pre-determinado.

Estructuras para la matriz de correlación

Hardin and Hilbe (2013) plantean varias estructuras estándar para la estimación de la correlación dentro de los *clusters*, algunas de esas estructuras son:

Correlación intercambiable

Como una extensión simple a la estructura independiente se puede lanzar la hipótesis de que las observaciones dentro de un panel tienen una correlación común. En este caso, la matriz de correlación tiene la siguiente estructura

$$R_{tt'}=\begin{cases}1 & si\ \ t=t'\\ \alpha & si\ \ t\neq t'\end{cases}$$

Correlación autorregresiva

Si las observaciones repetidas dentro de los paneles tienen un orden natural puede ser más razonable asumir una dependencia del tiempo para la asociación. En este caso, la matriz de correlación es la estructura de correlación autoregresiva de orden *k*. En particular, para un AR(1) se tiene:

$$R_{tt'}=\begin{cases}1 & si\ \ t=t'\\ \alpha^{|t-t'|} & si\ \ t\neq t';\end{cases}$$

Correlación estacionaria

Como una alternativa a la hipótesis de autocorrelación, se puede postular que existen las correlaciones para algún número pequeño de unidades de tiempo. En esta hipótesis se especifica una diferencia de tiempo máxima para la cual las observaciones pueden estar correlacionadas de modo que la matriz de correlación esté acotada. En este caso, α es un vector de correlaciones de hasta *k* rezagos y la matriz de correlación puede ser descrita como:

$$R_{tt'}=\begin{cases}\alpha_{|t-t'|} & si\ \ |t-t'|\leq k\\ 0 & en\ otro\ caso\end{cases}$$

APLICACIÓN

El propósito de la aplicación es determinar si el hecho de fumar cigarrillo durante el embarazo y la calidad del control prenatal afectan significativamente la probabilidad de que el nacimiento resulte con bajo peso al nacer. Otras variables incluidas en el análisis son la edad, educación y estado civil de la madre, además del momento en el que se realiza la atención prenatal y el sexo del

nacido. Se cuenta con un total 648 madres y 3 nacimientos por cada madre.

Dado que Y_{it} es una respuesta binaria que toma valores de 1 (bajo peso al nacer) y 0 (no bajo peso), interesa relacionar los cambios en la $E(Y_{it} / X_{it})$ con los cambios en las covariables. El modelo marginal especificado es un modelo de regresión logística con un patrón de asociación intercambiable. Los resultados se presentan en el siguiente cuadro.

Cuadro No. 1
Regresión logística con patrón de correlación intercambiable

Bajo Peso al nacer	*Razón de chances*	*Error estándar robusto*	*P > z*	*Intervalo de confianza del 95%*	
Fumó	*2,77*	*0,95*	*0,003*	*1,41*	*5,43*
Sexo del nacido	*0,67*	*0,19*	*0,149*	*0,39*	*1,15*
Edad de la madre	*0,95*	*0,04*	*0,201*	*0,88*	*1,03*
Educación de la madre	*0,90*	*0,07*	*0,220*	*0,77*	*1,06*
Madre casada	*0,55*	*0,22*	*0,127*	*0,25*	*1,19*
Prenatal de calidad intermedia	*2,50*	*0,93*	*0,014*	*1,20*	*5,19*
Prenatal de calidad inadecuada	*5,49*	*2,66*	*0,000*	*2,13*	*14,17*
Sin control prenatal	*0,83*	*0,59*	*0,797*	*0,20*	*3,38*
1er C. prenatal en 2do trimestre	*0,68*	*0,27*	*0,331*	*0,31*	*1,49*
1er C. prenatal en 3er trimestre	*0,07*	*0,07*	*0,014*	*0,01*	*0,58*
Constante	*0,60*	*0,53*	*0,568*	*0,11*	*3,42*

Fuente: Elaboración propia

Claramente el efecto de fumar sobre el bajo peso al nacer es altamente significativo (valor - p = 0,003). Cuando la madre fuma, la chance de tener bajo peso al nacer es 2,8 veces más que cuando la madre no fuma. La educación de la madre no tiene un efecto significativo (valor - p = 0,220) sobre la probabilidad de nacer con bajo peso, sin embargo, el control prenatal inadecuado tiene un efecto altamente significativo sobre la probabilidad de nacer con bajo peso (valor - p = 0,000).

DISCUSIÓN

Muchos estudios se han realizado a fin de determinar el efecto de fumar durante el embarazo sobre el peso al nacer, entre los cuales se encuentran los trabajos de Alonso et al. (2005) y Caraballoso (1999). En ambos trabajos se concluye que fumar durante el periodo de gestación afecta negativamente el peso del recién nacido. En el trabajo de Alonso et al. (2005) se aplican dos modelos, por una parte, un modelo de regresión lineal para explicar el peso de los recién nacidos en función de la condición de fumar de la madre y de su pareja, además de la edad gestacional y, por otra parte, se aplica una regresión logística con las mismas variables explicativas. En el estudio de Caraballoso (1999) también se aplicó un modelo de regresión logística con una serie de variables explicativas, entre las que se consideró, además del hábito de fumar durante el embarazo, variables como las vinculadas al alcoholismo, el deseo del embarazo y los antecedentes de hijos con bajo peso al nacer.

Hay, sin embargo, dos diferencias metodológicas entre los dos trabajos citados y el presente estudio. Primero, en este estudio se usan datos de panel, los cuales son apropiados para un control más efectivo de los otros factores que podrían afectar el peso al nacer, como el consumo de alcohol durante el embarazo y el estado nutricional de la madre. Segundo, el uso de un modelo marginal, un modelo apropiado para el análisis de datos tipo panel.

Si bien existen diferencias metodológicas para abordar un mismo objetivo – explicar el efecto de consumo de tabaco sobre el peso al nacer - es claro que las conclusiones a las que se arriban en los distintos trabajos son similares. Este y los otros estudios muestran claramente el efecto significativo del consumo de tabaco durante el embarazo sobre el bajo peso al nacer. Sin duda que estos hallazgos son útiles para fines de políticas públicas vinculadas a la salud y nutrición de los infantes.

CONCLUSIÓN

El modelo marginal es bastante flexible en el sentido que no es necesario especificar una distribución de probabilidad conjunta para las respuestas. Permite hacer inferencias sobre las medias poblacionales, pero, requiere de un método apropiado de estimación de los parámetros, el denominado método de ecuaciones de estimación generalizada. Este método está basado en el concepto de ecuaciones de estimación y proporciona un enfoque muy general y unificado para analizar respuestas correlacionadas, una característica muy frecuente en los datos tipo panel.

Su aplicación permitió evidenciar estadísticamente que fumar durante el embarazo y un control prenatal inadecuado están asociados con un bajo peso al nacer, luego de controlar el efecto de otras variables como la edad y la educación de la madre. Será importante ver, en un posterior trabajo de investigación, si el uso de otros métodos de análisis de datos de panel conduce a las mismas conclusiones.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2002). Categorical data análisis (2 ed.). John Wiley & Sons Inc.

Alonso, A., Cano, J., Girón, A., Yep, G. y Sánchez, M. (2005). Peso al nacimiento y tabaquismo familiar. Asociación Española de Pediatria, vol. 63, No. 2, pp. 116-119.

AndreB, H., Golsch, K. and Schmidt, A. (2013). Applied panel data analysis for economic and social surveys. Springer.

Baltagi, B. (2013). Econometric analysis of panel data (5 ed.). John Wiley & Sons.

Banerjee, M. and Frees, E. (1997). Influence diagnostics for linear longitudinal models. Journal of the American Statistical Association vol. 92, pp. 999-1005.

Bickel, P. and Doksum, K. (1977). Mathematical Statistics. Holden-Day.

Bijleveld, C. and Van der Kamp, L.(1998). Longitudinal data analysis: Designs, models and methods. Sage Publications.

Caraballoso, M. (1999). Bajo peso al nacer y tabaquismo. Revista Cubana de Salud Pública vol. 25, No. 1.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). Analysis of longitudinal data (2 ed.). Oxford University Press.

Fitzmaurice, G., Laird, N. and Ware, J. (2011). Applied longitudinal análisis (2 ed.). John Wiley & Sons.

Frees, E. (2004). Longitudinal and Panel Data: Analysis and applications for the social sciences. Cambridge University Press.

Gosho, M., Hamada, C. and Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation: Approach for longitudinal balanced data. Communications in Statistics-Theory and Methods No. 40(21), pp. 3839-3856.

Gosho, M. (2014). Criteria to select a working correlation structure for the generalized estimating equations method in SAS. Journal of Statistical Software, vol. 57.

Hardin, J. and Hilbe, J. (2013). Generalized estimating equations (2 ed.). Chapman & Hall.

Hsiao, Cheng. (2003). Analysis of panel data (2 ed.). Cambridge University Press.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. Biometrika No. 1, vol. 73, pp. 13-22.

McCullagh, P. and Nelder, J. (1989). Generalized linear models (2 ed.). Chapman & Hall.

Skrondal, A. and Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. Chapman & Hall.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models and the gauss-newton method. Biometrika No 61, vol. 3, pp. 439.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. Statistics in Medicine, vol. 23, pp. 859-880.

Artículo Original
Revista Varianza -Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 9-19

ESTRATIFICACIÓN ASIMÉTRICA EN ENCUESTAS ELECTORALES

ASYMMETRIC STRATIFICATION IN ELECTORAL POLLS

Ronal Edwin Condori Huanca[1]

Programa Doctorado en Políticas Públicas, Universidad Mayor de San Andrés, La Paz -Bolivia

✉ *ronal.c.huanca@gmail.com*

Artículo recibido: 2021-07-30
Artículo aceptado: 2021-08-30

RESUMEN

La importancia de las encuestas electorales radica en su precisión sobre los resultados oficiales, en particular, el diseño de muestreo que define la selección y el proceso de estimación es un componente neurálgico en esta operación estadística. Estrategias como la estratificación, y el uso de dos o más etapas son habituales, sin embargo, las regiones bisagra o lo que aquí se denomina como estratificación electoral asimétrica es una alternativa para mejorar la precisión de las estimaciones de encuestas de opinión en Bolivia. Con el objetivo de mejorar la precisión de las estimaciones, se experimentó la inclusión de variables electorales para la estratificación y la incorporación de otras variables en la Post-estratificación, mediante técnicas multivariantes como el de componentes principales y análisis *cluster* no jerárquico, esto genera mejoras en la desviación media absoluta (DMA) de 6.9 a 4.8, y en otras medidas de precisión, comparadas con las estimaciones de resultados de la 2da Encuesta de la iniciativa TuVotocuenta para las Elecciones Generales de Bolivia en 2020.

Palabras clave: *Post-estratificación, encuesta, asimétrica, bisagra, elecciones, Bolivia, TuVotoCuenta.*

ABSTRACT

The importance of electoral polls lies in their precision on the official results, in particular, the sampling design that defines the selection and the estimation process is a neuralgic component in this statistical operation. Strategies such as stratification and the use of two or more stages are common; however, hinge regions or what it here call asymmetric electoral stratification is an alternative to improve the precision of opinion poll estimates in Bolivia. In order to improve the precision of the estimates, the inclusion of electoral variables for stratification and the incorporation of other variables in Post-stratification were experimented with, using multivariate techniques such as principal components and non-hierarchical cluster analysis, this generates improvements in the absolute mean deviation (DMA) from 6.9 to 4.8, and in other precision measures, compared with the estimates of the results of the 2nd Survey of the TuVotocuenta initiative for the Bolivian General Elections in 2020.

Palabras clave: *Post-stratification, survey, asymmetric, hinge, elections, Bolivia, TuVotoCuenta.*

INTRODUCCIÓN

Indistintamente del país, región o temporalidad, dentro de todo proceso electoral, las encuestas son un instrumento que permiten dar desde un panorama, hasta influir en la población, sobre los resultados que estos actos electorales generan (Galvez, 2011).

[1] Profesional/consultor en Estadística(s), candidato al Doctorado en Políticas Públicas de la UMSA, ha ejercido laboralmente en el área de estadística de entidades privadas para la elaboración de estadísticas, y en entidades públicas como el INE-Bolivia, Ministerios de: Salud, Desarrollo Productivo, Educación, Economía, etc. https://orcid.org/0000-0003-2557-7079

En particular las encuestas por muestreo, son usadas de forma masiva en estos procesos, dado que tienen la propiedad de inferir los resultados oficiales de manera económica y rápida, en contra posición con los resultados oficiales que pueden demorar bastante e incluso en ocasiones, con postergaciones de entrega de resultados oficiales, que generan susceptibilidad en la población.

Existen diversos tipos de encuestas electorales, sin embargo las más importantes según su temporalidad son: *i) encuestas de intención de voto, ii) encuestas en boca de urna, iii) conteos rápidos por muestreo* (ODCA, 2006).

En los últimos años la región latinoamericana, y en particular Bolivia ha pasado por varios procesos electorales, entre los cuales, algunos de ellos han sido cuestionados a nivel nacional e internacional (CEPR, 2019), (OEA, 2019), conduciendo así a incrementar la tensión política, social e institucional que acompaña todo acto electoral.

En este contexto, instituciones públicas y privadas, ya sean con fines académicos, de comunicación, de acompañamiento electoral u otro, han presentado estimaciones sobre los resultados electorales oficiales. Según la Ley de Régimen Electoral (Ley 026, 2010) y los *Reglamentos de Elaboración de Estudios de Opinión*, todas estas operaciones estadísticas deben especificar el diseño de muestreo que se aplica (OEP, 2020), para cualquiera de los tres tipos de encuestas que se aplique.

Los diseños de muestreo permiten construir metodológicamente un sustento teórico estadístico, para la cuantificación de dos valores esenciales: *i) el estimador,* y *ii) la estimación del error,* esto a través de la definición y cálculo de los niveles de estratificación, unidades de muestreo que se hayan aplicado, y los factores de expansión o ponderadores calculados.

Antecedentes

En Bolivia, según las fichas metodológicas disponibles para el Órgano Electoral Plurinacional (OEP), y elaboradas por las diferentes instituciones, las encuestas electorales, generalmente incorporan un conjunto de variables en el proceso de estratificación, estas abarcan aspectos geográficos y político-administrativos.

De forma paralela, se incorporan otras variables más para poder aplicar una post–estratificación, según aspectos sociodemográficos como edad, sexo, u otros (CiesMori, 2019). Debido a los costos y tiempos algunas instituciones incluso aplican un muestreo por cuotas dentro de las Unidades Primarias de Muestreo, denominadas como UPM's (Mercados y Muestras, 2019).

Las variables de conglomeración para las diferentes etapas usualmente manejan desde municipios, áreas censales, manzanos, comunidades o incluso áreas sobre los recintos de votación habilitados (ViaCiencia, 2019).

Problemática

Concretamente en Bolivia, para las Elecciones Generales de 2019, se realizaron un total de 14 encuestas de intención de voto[2] , un conteo rápido y no se realizaron encuestas en boca de urna. Las primeras en su mayoría aplican un muestreo en más de una etapa, y con estratificación (Cuadro No. 1).

[2] Informes y fichas metodológicas centralizadas en el OEP (https://www.oep.org.bo/elecciones-generales-2019/)

Cuadro No. 1
2019: Diseños encuestas electorales 2019

Instituciones	*Rondas*	*Diseño**
ViaCiencia	*3*	*ME/ES/AS-C*
Mercados&Muestras	*3*	*ME/ES/AS-C*
CiesMori	*3*	*ME/ES/AS-C*
IPSOS	*2*	*C*
Tal Cual	*1*	*ME/ES*
Captura Consulting	*1*	*ME/SS*
Fund. Misky Utaha	*1*	*ME/ES*
Total	***14***	

Fuente: Informes/fichas técnicas OEP, elaboración propia
(*) ME: Múltiples etapas, ES: Estratificado, AS: Aleatorio simple, SS: Selección sistemática, C: Por cuotas.

Las instituciones en su mayoría mantienen la metodología entre rondas, para asegurar su comparabilidad interna, debido a que, si existieran cambios en la metodología, esto repercuten en los resultados y pierde su comparabilidad temporal.

Cuadro No. 2
Resultados y estimaciones electorales 2019

Fuente	*Estimación*		*Diferencia*		*DMA**
	MAS	*CC*	*MAS*	*CC*	
Resultado Oficial £	***46,6***	***36,8***	**--**	**--**	
ViaCiencia	*44,3*	*32,4*	*2,3*	*4,4*	*3,4*
CiesMori	*44,9*	*33,3*	*1,7*	*3,5*	*2,6*
IPSOS	*49,0*	*27,0*	*2,4*	*9,8*	*6,1*
M.&Muestras	*44,6*	*35,1*	*2,0*	*1,7*	*1,8*
Fund.Misky Utaha	*38,1*	*37,3*	*8,5*	*0,5*	*4,5*

Fuente: Informes/fichas técnicas OEP, elaboración propia,
(*) Desviación media absoluta (DM),
(£) Resultados nacionales.

Paralelamente, y con fines de acompañamiento en 2019 la Organización de Estados Americanos (OEA), sostuvo que los resultados de un conteo rápido, bajo la aplicación de una muestra aleatoria, y a cargo de una comisión técnica interna, encontraron diferencias significativas respecto a los resultados oficiales (OEA, 2019), sin embargo no se aclaran los pormenores del muestreo aplicado.

Este hecho también es contrastado con las últimas rondas encuestas previas sobre intención de voto, presentaron diferencias considerables sobre los resultados oficiales, algunos con casi 10 puntos de diferencia para el partido CC. Usando la Desviación Media Absoluta (DMA), como medida de precisión (Beltran & Valdivia, 1999), permite cuantificar para el caso boliviano, errores de entre 1,8 a 6,1 puntos (Cuadro No. 2).

Todo esto plantea una problemática, sobre sí los diseños de muestreo que se aplicaron y se aplican actualmente, en las encuestas por muestreo con fines electorales, son los mejores.

Por otro lado, en ninguna de las fichas técnicas se determina la aplicación de estratificación sobre variables electorales, en algunas sólo se especifica que se aplicó estratificación, sin definir que estratos se usaron, y en otras no se menciona el uso de estratos de manera precisa.

Con esto la pregunta de investigación se define como: "*¿La incorporación de una estratificación asimétrica permite mejorar las estimaciones sobre los resultados oficiales?*", pues éste es el objetivo de la estratificación, definiendo estratos donde la variable de interés se comporta más homogéneamente, mejorando la precisión de las estimadores (Lohr, 1999), y sin descuidar la cobertura de la muestra.

Hipótesis

Incorporar un nivel de estratificación asimétrica posterior o post estratificación, mejora la precisión de las estimaciones sobre los resultados oficiales de las Elecciones Generales de Bolivia en 2020.

MATERIAL Y MÉTODOS

Alternativas de estratificación

La literatura sobre la estratificación para encuestas electorales, aborda diversas alternativas, sin embargo, se enfatiza solo en aquellas que utilicen no solamente variables o aspectos geográficos o demográficos, sino los que incluyan aspectos electorales. Por ejemplo, en España, se maneja estratificación según "competencia electoral", diferenciando entre provincias donde hay más o menos diversidad en representación electoral (CIS, 2016).

El caso colombiano, se demostró que las diferencias existentes del comportamiento electoral entre municipios es casi ancestral (Bautista-Sierra, 2005), ya que la composición y comportamiento de votos entre los partidos de corte liberal, conservador o izquierda se conservan a lo largo de varias décadas (Pacheco & Bautista, 1989).

Una variable de estratificación que se incluye de forma transversal en casi todas las encuestas, es: Nivel Socio Económico (NSE), o también llamado Clase Social (CS), que es de vital importancia en el diseño de muestreo. En Perú se relacionan directamente la opinión pública con el NSE (IPSOS-Perú, 2021), bajo varios criterios en base a información de las Oficinas Nacionales de Estadística de cada país (Tuesta, 1997).

Del mismo, modo sucede en algunas encuestas en Canadá, donde IPSOS-Canadá, utilizan el NSE y otras variables sociodemográficas y culturales como lengua materna, pero bajo un proceso de post-estratificación (Durand & Blais, 2019).

Talvez el ejemplo más famoso de la inclusión de variables electorales en el proceso de estratificación para fines de muestreo, está en Estados Unidos. En donde su sistema electoral desde 1860 posee una composición bipartidista entre republicanos y demócratas. Este sistema permite a cada distrito elegir por mayoría simple a representantes, y estos a su vez eligen al presidente posteriormente (DoS-EEUU, 2012).

Este comportamiento electoral se incluyó en los últimos sondeos de las elecciones de 2020, donde *Joe Biden* consiguió la victoria, colocando los antecedentes electorales de las regiones en el diseño de muestra (PEW R.C., 2021). Esta estratificación diferencia[3] dos tipos de regiones: *i) estados bisagra (o pendulares)*, donde el comportamiento electoral tiende a ser muy reñido, y con diferencias estrechas; *ii) estados seguros* donde el favoritismo republicano o demócrata está muy consolidado. Algunos de estos estados bisagra son: Arizona, Colorado, Florida, Georgia, y otros más (EOM, 2020).

Este enfoque de estratificación plantea una disyuntiva estadística, ya que puede interpretarse en poner más atención a las encuestas regionales en estas regiones bisagra, que en las mismas encuestas nacionales (Forsberg & Payton, 2015).

Para el proceso de identificación de los estados bisagra, se tiene varios criterios y combinaciones entre ellos, de entre éstos los que presentan un criterio cuantitativo y más conocido son: *i) margen de diferencia, ii)margen de variación, y iii) método de Bellwether* (Clayton, 2019).

[3] En ingles denominados: Safe States, y Swin States, estos últimos también se denominan Battleground States.

Estratificación asimétrica

El muestreo estratificado tiene como principales objetivos, mejorar la precisión o reducir la varianza de los estimadores y mantener una cobertura en cada estrato definido. Estos estratos deben de ser en lo posible homogéneos dentro de ellos y heterogéneos entre sí (Blaconá, Marí, & Méndez, 2009).

Un problema en el muestreo en general, se dá cuando existen unidades que se pueden considerar bastante grandes respecto al resto, y su omisión o presencia dentro de la muestra (Srinath & Hidiroglou, 1981), inciden en una subestimación o sobreestimación, además de afectar a la varianza estimada. En poblaciones donde sucede esto, suelen ser denominadas "*poblaciones asimétricas*" (Fuller, 1970).

Para afrontar este problema desde el enfoque del muestreo estadístico, se estableció la creación de un estrato de inclusión forzosa, donde todas sus unidades son seleccionadas en la muestra (Hidiroglou, 1986).

Hidiroglou(1986) desarrolló el estimador, su varianza, el tamaño de muestra mínimo, y el valor del punto óptimo, que permite definir el estrato de inclusion forzosa, definiendo una poblacion finita ordenada como $y_{(1)} \leq y_{(2)} \ldots \leq y_{(N)}$, de donde se selecciona una muestra de tamaño $n(t)$, con t unidades de inclusion forzosa, y $n(t)$-t unidades de selección aleatoria simple. Un estimador del total poblacional "*Y*", y su varianza estan dado por:

$$\hat{Y} = \frac{N-t}{n(t)-t}\sum_{i=1}^{n(t)-t} z_i + \sum_{i=N-t+1}^{N} y_{(i)} \qquad (1)$$

$$V(\hat{Y}) = \frac{(N-t)*(N-n(t))}{n(t)-t} S^2_{[N-t]} \qquad (2)$$

Donde: $S_{[N-t]}$ es la desviacion estandar dentro la población de los primeros N-t terminos, y la muestra aleatoria esta dada por valores Z, tales que: $y_{(1)} \leq z_{(i)} \leq y_{(N-t)}$. Para esto, es rápido verificar que esta descomposición genera 2 estratos, y la misma puede extenderse si se desea generar más estratos de selección. Bajo esta linea, el aspecto esencial es el definir los "*L-1*" limites de los *L* estratos.

Según la literatura se desarrollaron diversos métodos en la estratificación de poblaciones asimétricas, de los cuales se destacan: *i) Geometrico, ii) Lavallée-Hidiroglou, iii) Lavallée-Hidiroglou Geométrico iv) Kozak* (Blaconá, Marí, & Méndez, 2009). Todos tienen el objetivo de determinar los límites inferiores y superiores de los estratos: $E_h = (k_{h-1}, k_h >$ para $h = 1,2,\ldots,L$, y su relación con la población ordenada como: $k_0 = y_{(1)}$ y $k_L = y_{(N)}$.

Precisión de las estimaciones

Las encuestas electorales son de las pocas encuestas por muestreo, en donde se pueden cuantificar el sesgo de manera precisa. Ya que el resultado electoral oficial, brinda datos con exactitud para así poder obtener los parámetros desconocidos de la población.

En esta línea, existen diversas formas de evaluar la precisión de las estimaciones, las cuales se desarrollaron en "*The National Council of Public Polls*"[4]. (Warren, 1998). De las cuales, se hace énfasis en tres, estas son: *i) desviación media absoluta, ii) Error del margen, y iii) Chi-cuadrado*.

i. $DMA = \frac{1}{k}\sum_{i=1}^{k}\left|\hat{\theta}_i - \theta_i\right|$

ii. $EM = \left|\left(\hat{\theta}_1 - \hat{\theta}_2\right) - (\theta_1 - \theta_2)\right|$

iii. $\chi^2 = \sum_{i=1}^{k}\frac{(\hat{\theta}_i - \theta_i)^2}{\theta_i}$

Otros criterios se centran en evaluar su consistencia interna, ya sea entre

estimaciones de las diferentes rondas de la misma institución, o las estimaciones entre todas las rondas e instituciones, analizando así tendencias o medidas semejantes (Mateos & Penadés, 2013).

Información disponible

Se utilizan tres fuentes de información:

Primera, la encuesta de intención de voto, realizada por la iniciativa "*Tu-Voto-Cuenta*" en colaboración con la UMSA y la Fundación Jubileo, en su segunda ronda 2020.

Segunda, el padrón electoral de 2020, los resultados oficiales publicados para 2014, 2019 y el referéndum de 2016, disponible en el portal web del OEP.

Tercera, la encuesta de hogares de 2019, elaborada por el INE-Bolivia, que permitirá incorporar una composición poblacional estimada, al proceso de post estratificación sobre niveles educativos en Bolivia.

Diseño de la investigación

Se considera de tipo experimental, ya que el diseño de muestra y los resultados de la iniciativa *TuVotoCuenta*, se consideran como un grupo control o contra factual, siendo estos comparados con los ajustes aquí planteados.

Este diseño de muestreo "*de control*" es de tipo probabilístico, en múltiples etapas, pero particularmente con una estratificación por departamento, área, y tamaño del asiento electoral[5].

Implementación de la estratificación

El padrón electoral 2020, será el insumo base para definir el marco de muestreo y su estratificación está a nivel de municipios. Esto debido a limitantes en la identificación de las unidades de muestreo de la encuesta "*Tu-Voto-Cuenta*".

A pesar que, desde el último Censo de Población y Vivienda de 2012, se han creado regiones autonómicas, municipios y otras más, se optó por mantener la estructura de 339 municipios dentro del marco de muestreo (INE, 2015), esto para fines comparativos con los resultados electorales previos.

Desde el punto de vista numérico de clasificación, en el sistema electoral Boliviano, no existe un bipartidismo o multipartidismo como tal, sino más bien un esquema de "*un partido dominante*" (Villafuerte, 2012), el cual está liderado por el Movimiento al Socialismo (MAS), quien capitaliza la mayoría de votos en los últimos 14 años.

Esto orientó a elaborar un indicador del tipo "*margen de diferencia electoral*", para cada uno de los actos electorales a lo largo del tiempo (Clayton, 2019). La misma se define como $b_t = 50 - mas_t / val_t$, dando a entender que, si esta brecha estaría cercana a cero, entones habría una mayor competitividad, si fuera muy elevada, entonces el partido MAS tiene una mayor ventaja, y si fuera muy negativa, los partidos de oposición poseen una ventaja respecto al MAS.

Tras obtener estas brechas para 2014, 2016

[4] Estas surgieron en las controvertidas elecciones de 1948 en EEUU, donde firmas como Gallup, Roper, y otras más, fallaron en sus estimaciones.

[5] Ficha Técnica 2da encuesta: http;//www.oep.org.bo/wp-content/uploads/2020/10/Tu_Voto_Cuenta_UMSA_2_EG_2020.pdf.

y 2019, se sintetizó un indicador aplicando un Análisis de Componentes Principales, extrayendo solo una componente principal. Posteriormente se clasificaron a los municipios mediante un Análisis Cluster no jerárquico de *k-medias*, definiendo de antemano 3 estratos, esto último para obtener los municipios bisagra, y municipios seguros para el MAS o sus opositores.

Debido a su tamaño poblacional asimétrico respecto al resto de municipios, y heterogeneidad interna entre los distritos electorales que los componen. Los 9 municipios capitales y el municipio de El Alto, se consideran como un estrato de inclusión forzosa.

Clasificados los municipios dentro de cada uno de los "*estratos electorales*", se obtuvieron sus totales poblacionales en base al padrón electoral de 2020, que está disponible previamente al acto electoral.

Con este último insumo y con la distribución según nivel educativo estimado de la EH-2019 se obtuvo los "*estratos educativos*", ya que es una variable que debe ser controlada y es un proxy del NSE.

Finalmente, se procede a realizar una calibración de los ponderadores, mediante una post-estratificación por el método Raking (Deville, Sarndal, & Sautory, 1993). Esto dentro de cada departamento y usando los ponderadores que posee la base como "*factores base*", y los totales de los "*estratos electorales y estratos educativos*".

RESULTADOS

Los estratos electorales conformados, permitieron diferenciar entre municipios que pertenezcan a cada uno de las situaciones favorables o negativas al partido MAS, y los municipios Bisagra.

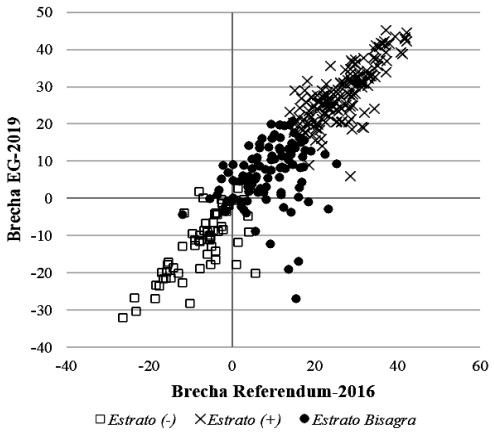
El comportamiento de las brechas electorales promedio y la componente principal, guarda consistencia entre sí (Cuadro No. 3). Gráficamente las brechas 2019 y 2016, se concentran al origen en el estrato bisagra, se acercan al primer cuadrante en el estrato seguro (+), y al tercer cuadrante en el estrato seguro (-) para el partido MAS (Figura No.1).

Cuadro No. 3
Promedios de brechas y tamaño por cada estrato

Variable	*Total*	*Estrato*		
		Seguro (-)	*Bisagra*	*Seguro (+)*
Brecha 2014	*26,7*	*-0,3*	*23,7*	*40,0*
Brecha 2016	*14,4*	*-7,0*	*9,2*	*26,7*
Brecha 2019	*13,3*	*-11,8*	*6,6*	*28,2*
Componente (Y1)	*0,0*	*-1,5*	*-0,3*	*0,8*
# Municipios	*329*	*67*	*103*	*159*

Fuente: Resultados electorales 2014-2019, elaboración propia.

Figura No. 1
Dispersión de municipios entre brechas electorales Referéndum 2016 y Elecciones generales 2019



Fuente: Resultados electorales 2014-2019, elaboración propia.

Luego de la construcción de los estratos y su incorporación en el proceso de post estratificación, las nuevas estimaciones

se evaluan mediante los tres criterios de precisión mencionados antes.

Se planteó tres escenarios de ajuste post estratificación de los factores originales, estos son: *i) estratificación electoral, ii) estratificación electoral/educativa,* y *iii) estratificación educativa*.

Cuadro No. 4
Precisión de las estimaciones 2020

Ponderador	*DMA*	*EM*	χ^2
Inicial/original	*6,9*	*16,9*	*4,4*
Estr, Electoral	*6,3*	*15,3*	*3,7*
Estr, Electoral & educativa	*4,8*	*10,7*	*2,4*
Estr, Educativa	*5,4*	*11,7*	*3,2*

Fuente: Segunda Encuesta *TuvotoCuenta* 2020, elaboración propia.

En estos tres escenarios, todos obtuvieron mejores niveles de precisión nacional respecto a las estimaciones originales. Particularmente el segundo escenario que combina la incorporación de los 2 niveles de estratos, tiene los mejores resultados, reduciendo de 16,9 a 10,7 en el Error del Margen (EM), de 6,9 a 4,9 en la Desviación Media Absoluta (DMA), y de 4,4 a 2,4 en la distancia Chi cuadrado (χ^2) (Cuadro No. 4).

DISCUSIÓN

Los resultados generados tras la incorporación de uno ó más niveles de estratificación permitieron mejorar la precisión de las estimaciones y la reducción de los sesgos posteriores, sin embargo, se debe revisar estas mejoras por cada una de las regiones o departamentos.

El mantener simplemente estratos geográficos, ignora el comportamiento electoral histórico de las regiones o distritos electorales. Además, el sólo centrarse en la hipotética brecha electoral urbana-rural, ignora la similitud de ciudades intermedias y las localidades rurales, pero este hecho debe ser manejado con más detalle dado el crecimiento poblacional de municipios como Montero, El Alto o Sacaba.

Aunque por un muestreo por cuotas, los niveles socioeconómicos deben de ser incluidos y controlados en las encuestas electorales, con alguna variable *proxy* como el nivel educativo, tenencia de activos u otras.

Los resultados obtenidos ahora, pueden ser cuestionados, ya que la encuesta utilizada, tiene una muestra bastante grande, y este tamaño brinda mucho margen para realizar ajustes a las estimaciones que se obtengan.

Si bien la estratificación asimétrica debe de aplicarse desde la misma planificación y selección de la muestra, en este artículo se la aplicó posteriormente, debido a la limitante de la información disponible.

La normativa actual del OEP, menciona que se debe señalar que tipo de muestreo se utiliza, sin embargo, no indica que se deba específicar las variables de estratificación, conglomeración o post-estratificación que se implementan, ya que las instituciones en su totalidad no reportan esto, y hay cierto grado de secretismo en las mismas, que lleva a critica por ocultar información pública.

La difusión de datos en Bolivia está a cargo principalmente del INE (DL No. 14100, 1976), pero vale analizar que el INE participe o no en el acompañamiento de éstas encuestas, para que exista consistencia entre todas las encuestas sin que todas arrojen cifras iguales (Penades, 2015).

CONCLUSIONES

Concretamente para la segunda encuesta de la iniciativa "*Tu-Voto-Cuenta*", todos los ajustes por post-estratificación con variables electorales obtuvieron mejores resultados en términos de precisión sobre las cifras oficiales en las Elecciones Generales de Bolivia en 2020, verificando la veracidad de la hipótesis planteada.

El incluir esta estratificación previa o posterior, en diversas encuestas de carácter político o electoral, permitirá mejorar potencialmente las estimaciones obtenidas, y así respaldar la credibilidad tanto en dichas herramientas estadísticas como en los resultados oficiales.

Cualquier resultado sesgado daña tanto a la imagen institucional de los que recaban información, los organismos electorales, como a la credibilidad de los candidatos y partidos políticos que concursan. Ya que todos ellos se respaldan en el grado de consistencia entre cifras estimadas y cifras oficiales.

De manera contrapuesta, cabe aclarar que, si todas las instituciones que levantan encuestas arrojaran cifras similares, también generan el efecto opuesto, porque es en la diversidad de opiniones y percepciones, donde la democracia se ve consolidada.

Un último aspecto, es el de mejorar la transparencia de información disponible para todo el público, con el debido cuidado de anonimizar a los informantes, y apegándose a normativas semejantes a las que maneja el Instituto Nacional de Estadística (INE), pero éstas, en un sentido más abocado a los estudios de opinión pública.

REFERENCIAS BIBLIOGRÁFICAS

Bautista-Sierra, L. (2005). Estrategia de muestreo para la estimacion de la tasa de favoritismo en la eleccion presidencial. Revista Colombiana de Estadıstica Vol. 28 No. 1, pp. 39-62.

Beltran, U., & Valdivia, M. (1999). Accuracy and Error in Electoral Forecasts: The case of Mexico. International Journal of Public Opinion Research Vol. 11 No. 2, pp. 115-134.

Blaconá, M. T., Marí, G., & Méndez, F. (2009). Estratificación de Poblaciones Asimétricas. Jornadas Investigaciones en la Facultad de Ciencias Económicas y Estadística - UNR, No. 14.

CEPR. (2019). Qué sucedió en el recuento de votos de las elecciones de Bolivia de 2019. Washington, DC: Center for Economic and Policy Research.

CiesMori. (2019). Primera Encuesta Pre Electoral con Miras a las Elecciones Generales 20 de Octubre de 2019. La Paz Bolivia.

CIS. (2016). Informe Metodologico Pre-Electoral y Post-Electoral: Elecciones Generales 2015. Madrid: Centro de Investigaciones Sociologicas.

Clayton, J. (2019). What Makes a State Swing? Research Association For Interdisciplinary Studies jun-2019, pp. 151-161.

Deville, J.-C., Sarndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures

in Survey Sampling. Journal of the American Statistical Association, Vol. 88, No. 423, pp. 1013-1020.

DL No. 14100. (8 de noviembre de 1976). Ley del Sistema Nacional de Información Estadística. Bolivia: Ley del Sistema Nacional de Información Estadística, DL No. 14100.

DoS-EEUU. (2012). EE.UU. Elecciones en síntesis. Washington, DC: Departamento de Estado de Estados Unidos.

Durand, C., & Blais, A. (2019). Quebec 2018: A Failure of the Polls? Canadian Journal of Political Science 2019, pp. 1-18.

EOM. (6 de Octubre de 2020). ¿Qué son los swing states, los estados "bisagra" de Estados Unidos? Obtenido de El Orden Mundial: https://elordenmundial.com/que-son-los-swing-states-los-estados-bisagra-de-estados-unidos/.

Forsberg, O., & Payton, M. (2015). Analysis of Battleground State Presidential Polling Performances, 2004–2012. Statistics and Public Policy, No. 2:1, pp. 1-10.

Fuller, W. (1970). Simple Estimators for the Mean of Skewed Populations. Iowa: Technical Report prepared for the U.S. Bureau of the Census.

Galvez, L. (2011). Las encuestas electorales y el debate sobre su influencia en las elecciones. Revista Mexicana de Opinión Pública, No. 11, pp. 25-43.

Hidiroglou, M. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. The American Statistician Vol. 40, No.1, pp. 27-31.

INE. (2015). CNPV-2012: Características de la Población. La Paz: Instituto Nacional de Estadística.

IPSOS-Perú. (2021). Informe de Resultados Estudio de opinión El Comercio - Ipsos Gestión y Elecciones generales Perú, febrero de 2021. Lima: IPSOS-Perú.

Ley026. (30 de junio de 2010). Ley del Régimen Electoral. Estado Plurinacional de Bolivia.

Lohr, S. (1999). Muestreo: Diseño y Analisis. Phoenix-Arizona: International Thomson Editores.

Mateos, A., & Penadés, A. (2013). Las encuestas electorales en la prensa escrita (2008-2011). Errores, sesgos y transparencia. Metodología de Encuestas Vol. 15, pp. 99-119.

Mercados y Muestras. (2019). Informe 1ra Encuesta Nacional 2019. La Paz: Mercados y Muestras SRL.

ODCA. (2006). Manual de campaña electoral: marketing y comunicación política. Buenos Aires: Konrad Adenauer Stiftung.

OEA. (2019). Análisis de Integridad Electoral Elecciones Generales en el Estado Plurinacional de Bolivia 2019. Washington, D.C.: Organizacion de Estados Americanos.

OEP. (21 de diciembre de 2020). Reglamento de Elaboración y difusión de Estudios de Opinión en materia Electoral en Procesos Electorales. Organo Electoral Plurinacional.

Pacheco, P., & Bautista, L. (1989). Análisis de la Evolución del Comportamiento Electoral departamental en los últimos años: Aplicación de los Métodos Factoriales al Estudio de Series

temporales cortas. Revista Colombiana de Estadística No. 19-20, pp. 94-112.

Penades, A. (2015). Especial encuestas: errores, cocina y predicción. Salamanca: Fundacion Alternativas: Zoom Politico.

PEW R.C. (30 de Junio de 2021). Behind Biden's 2020 Victory: Methodology. Obtenido de PEW RESEARCH CENTER: https://www.pewresearch.org/politics/2021/06/30/validated-voters-methodology/.

Srinath, K., & Hidiroglou, M. (1981). Some Estimators of a Population Total From Simple Random Samples Containing Large Units. Journal of the American Statistical Association, Vol. 76, No. 375, pp. 690-695.

Tuesta, F. (1997). No Sabe/No Opina (Encuestas Politicas y Medios). Lima: Fundación Konrad Adenaue.

ViaCiencia. (2019). Ficha Tecnica Encuesta Intencion de Voto octubre 2019. Santa Cruz de la Sierra: VIACIENCIA SRL.

Villafuerte, V. (2012). Crisis y Colapso de los Sistemas de Partidos en los Países Andinos, desde 1990 hasta 2009. Quito: FLACSO.

Warren, M. (1998). Was 1996 a Worse Year for Polls Than 1948? The Public Opinion Quarterly, Vol. 62, No. 2, pp. 230-249.

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 21-33

APLICACIÓN DE MACHINE LEARNING SIN SUPERVISIÓN

MACHINE LEARNING APPLICATION WITHOUT SUPERVISION

Juan Carlos Flores López[1]
Instituto de Estadística Teórica y Aplicada-UMSA, La Paz -Bolivia
✉ *caarloslopez1@gmail.com*

Artículo recibido: 2021-08-16
Artículo aceptado: 2021-09-12

RESUMEN

El presente artículo, tiene como objetivo principal el desarrollo de la aplicación de *machine learning* no supervisado. La aplicación de esta metodología se realiza considerando la Encuesta Sociodemográfica del Departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc. De esta se considera los 75 municipios del Departamento de La Paz y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideran: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

Los resultados muestran que los municipios Santiago de Huata, y Tito Yupanqui, muestran similares características respecto a los indicadores: tasas de participación, relación de masculinidad y tasa de alfabetismo. Otro *cluster* definido por la customización son los municipios Humanata, Alcapata y Ayata y son parecidos en distribución de hogares según disponibilidad de dormitorios por persona y tasa de alfabetismo. Se concluye que la customización es la mejor forma de clasificación por la forma en que se presenta en forma mucho más clara que otras formas de clasificar consideradas en el estudio.

Palabras clave: *Aprendizaje no supervisado, clasificación, dendograma*

ABSTRACT

The main objective of this article is the development of the unsupervised machine learning application. The application of this methodology is carried out considering the Sociodemographic Survey of the Department of La Paz, carried out in 2015. The database considered has data on migration, health, education, employment, income, agriculture, housing, etc. Of this, the 75 municipalities of the Department of La Paz and the educational, employment, demographic and housing indicators are considered and within this are considered: the literacy rate, participation rate, distribution of households according to availability of bedrooms and masculinity ratio.

The results show that the municipalities of Santiago de Huata and Tito Yupanqui show similar characteristics regarding the indicators: participation rates, masculinity ratio and literacy rate. Another cluster defined by the customization are the municipalities Humanata, Alcapata and Ayata and they are similar in distribution of households according to availability of bedrooms per person and literacy rate. It is concluded that customization is the best form of classification due to the way it is presented in a much clearer way than other forms of classification considered in the study.

Keywords: *Unsupervised learning, classification, dendrogram*

[1] M. Sc. en Estadística, M.Sc. en Educación Superior, Dr.c. en Educación Superior e Investigación Transdisciplinar, investigador del Instituto de Estadística Teórica y Aplicada de la carrera de Estadística, tutor de varias tesis de pregrado. Docente CEPIES (Centro Psicopedagógico y de Investigación en Educación Superior). https://orcid.org/0000-0002-5522-1949

INTRODUCCIÓN

El aprendizaje estadístico es un conjunto de herramientas para comprender los datos. Estas herramientas pueden clasificarse como supervisadas o no supervisadas. El aprendizaje estadístico supervisado implica construir un modelo estadístico para predecir, o estimar, una salida basada en una o más entradas. Este tipo de problemas ocurren en diferentes campos por ejemplo en la economía, educación, negocios, medicina, astrofísica, política pública, etc. Con el aprendizaje estadístico no supervisado, hay entradas, pero sin salida de supervisión; sin embargo, podemos aprender relaciones y estructura de tales datos.

OBJETIVO

Como objetivo principal del presente artículo mostrar los resultados de la aplicación del aprendizaje estadístico *machine learning* (aprendizaje automático) considerando el aprendizaje sin supervisión, de la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad de los 75 municipios de la ciudad de La Paz.

METODOLOGÍA

La aplicación de esta metodología se realiza considerando la encuesta sociodemográfica del departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc.

Machine learning se interpreta como aprendizaje automático y está estrechamente relacionado muy a menudo, con estadística computacional; una disciplina que también se especializa en hacer predicciones. El aprendizaje automático se emplea en una variedad de disciplinas científicas.

El aprendizaje sin supervisión, es aquel que solo proporciona datos de salida, sin ninguna entrada. El objetivo es descubrir una "estructura interesante" en los datos; esto a veces se llama descubrimiento del conocimiento. A diferencia del aprendizaje supervisado, no se indica cuál es el resultado deseado para cada entrada. En cambio, se formaliza la tarea como una de estimación de densidad, es decir, queremos construir modelos de la forma $P(X_i \mid \theta)$. Hay dos diferencias con respecto al caso supervisado. Primero, se escribe $P(X_i \mid \theta)$ en vez de $P(y_i \mid X_i, \theta)$; es decir, el aprendizaje supervisado es una estimación de densidad condicional, mientras que el aprendizaje no supervisado es una estimación de densidad incondicional. Segundo, X_i es un vector de características, por lo que se necesita crear modelos de probabilidad multivariados. Por el contrario, en el aprendizaje supervisado y_i generalmente es solo una variable que se está tratando de predecir, esto significa que, para la mayoría de los problemas de aprendizaje supervisado, se puede utilizar modelos de probabilidad univariados (con parámetros dependientes de la entrada), lo que simplifica significativamente el problema (Murphy, 2012).

Agrupamiento (*clustering*)

El agrupamiento es una de las técnicas más utilizadas para el análisis exploratorio de datos. En todas las disciplinas, desde las ciencias sociales hasta la biología y la informática, las personas intentan tener una primera intuición sobre sus datos identificando grupos significativos entre los puntos de datos. Por ejemplo, los biólogos

computacionales agrupan genes sobre la base de similitudes en su expresión en diferentes experimentos; los minoristas agrupan a los clientes, en función de sus perfiles de clientes, para fines de marketing dirigido; y los astrónomos agrupan estrellas en función de su proximidad espacial. (Shalev, 2014).

El primer punto que se debe aclarar es, naturalmente, ¿qué es la agrupación? Intuitivamente, la agrupación es la tarea de agrupar un conjunto de objetos de manera que los objetos similares terminen en el mismo grupo y los objetos diferentes se separen en grupos diferentes. Claramente, esta descripción es bastante imprecisa y posiblemente ambigua. Sorprendentemente, no está nada claro cómo llegar a una definición más rigurosa.

Hay varias fuentes para esta dificultad. Un problema básico es que los dos objetivos mencionados en la declaración anterior pueden en muchos casos contradecirse. Matemáticamente hablando, la similitud (o proximidad) no es una relación transitiva, mientras que el *cluster* compartido es una relación de equivalencia y, en particular, es una relación transitiva. Más concretamente, puede darse el caso de que haya una larga secuencia de objetos, $x_1, \ldots, x_m$ de manera que cada x_i es muy similar a sus dos vecinos, x_{i-1} y x_{i+1}, pero x_1 y x_m son muy diferentes. Si deseamos asegurarnos de que cada vez que dos elementos sean similares compartan el mismo grupo, entonces debemos colocar todos los elementos de la secuencia en el mismo grupo. Sin embargo, en ese caso, terminamos con elementos diferentes (x_1 y x_m) que comparten un *cluster*, lo que viola el segundo requisito (Shalev, 2014).

Un modelo de agrupación

Las tareas de agrupación pueden variar en términos del tipo de entrada que tienen y el tipo de resultado que se espera que calculen. Para concretar, nos centraremos en la siguiente configuración común:

Entrada

Un conjunto de elementos χ, y una función de distancia sobre él. Es decir, una función d: $\chi \times \chi \rightarrow R_+$ que es simétrica, satisface $d(x,x)=0$ para todo $x \in \chi$ y, a menudo, también satisface la desigualdad del triángulo. Alternativamente, la función podría ser una función de similitud s: $\chi \times \chi \rightarrow [0,1]$ que es simétrica y satisface $s(x, x) = 1$ para todo $x \in \chi$. Además, algunos algoritmos de agrupación también requieren un parámetro de entrada k (que determina el número de agrupaciones requeridas).

Salida

Una partición del dominio establece χ en subconjuntos. Es decir, $C = (C_1, \ldots, C_k)$ donde $\bigcup_{i=1}^{k} C_i = x$ y para todo i ≠ j, $C_1 \cap C_k = \emptyset$. En algunas situaciones, la agrupación es "blanda", es decir, la partición de χ en los diferentes grupos es probabilística donde la salida es una función que asigna a cada punto de dominio, $x \in \chi$, un vector $(p_1(x), \ldots, p_k(x))$, donde $p_1(x) = P[x \in C_i]$ es la probabilidad de que x pertenezca al grupo C_i. Otra salida posible es un dendograma de agrupación (del griego dendron = árbol, gramma = dibujo), que es un árbol jerárquico de subconjuntos de dominio, que tiene los conjuntos únicos en sus hojas y el dominio completo como raíz (Shalev, 2014).

RESULTADOS

Para la aplicación de *machine learning* no supervisado, se toma en cuenta la información de la Encuesta Sociodemográfica

del Departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc.

De esta base de datos se considera los 75 municipios y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

La tasa de alfabetismo se determina bajo el siguiente criterio:

$$Tasa\ de\ alfabetismo = \frac{Población\ del\ grupo\ edad\ i\ que\ sabe\ leer\ y\ escribir}{Población\ total\ del\ grupo\ edad\ i}\%$$

Se define como la magnitud relativa de la población que sabe leer y escribir. La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio), sexo y grupos de edad. Y cuyo código en la base de datos es ED040.

La tasa de participación fue determinada bajo el siguiente criterio:

$$Tasa\ global\ de\ participación = \frac{Población\ económicamente\ activa}{Población\ en\ edad\ de\ trabajar}\%$$

Se define como el porcentaje de la población en edad de trabajar que forma parte de la población económicamente activa.

La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio), sexo. Y cuyo código en la base de datos es EI0102.

Índice de disponibilidad de dormitorios, es la distribución de hogares según disponibilidad de dormitorios por persona, dado por:

$$Tasa\ de\ disponibilidad\ de\ dormitorios = \frac{Nº\ de\ miembros\ del\ hogar}{Nº\ de\ dormitorios\ existentes}\%$$

La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio). Cuyo código en la base de datos es VV0101.

Índice de masculinidad fue determinado bajo el siguiente criterio:

$$Índice\ de\ masculinidad = \frac{Población\ masculina}{Población\ femenina} \times 100$$

Se define la proporción de hombres frente a las mujeres.

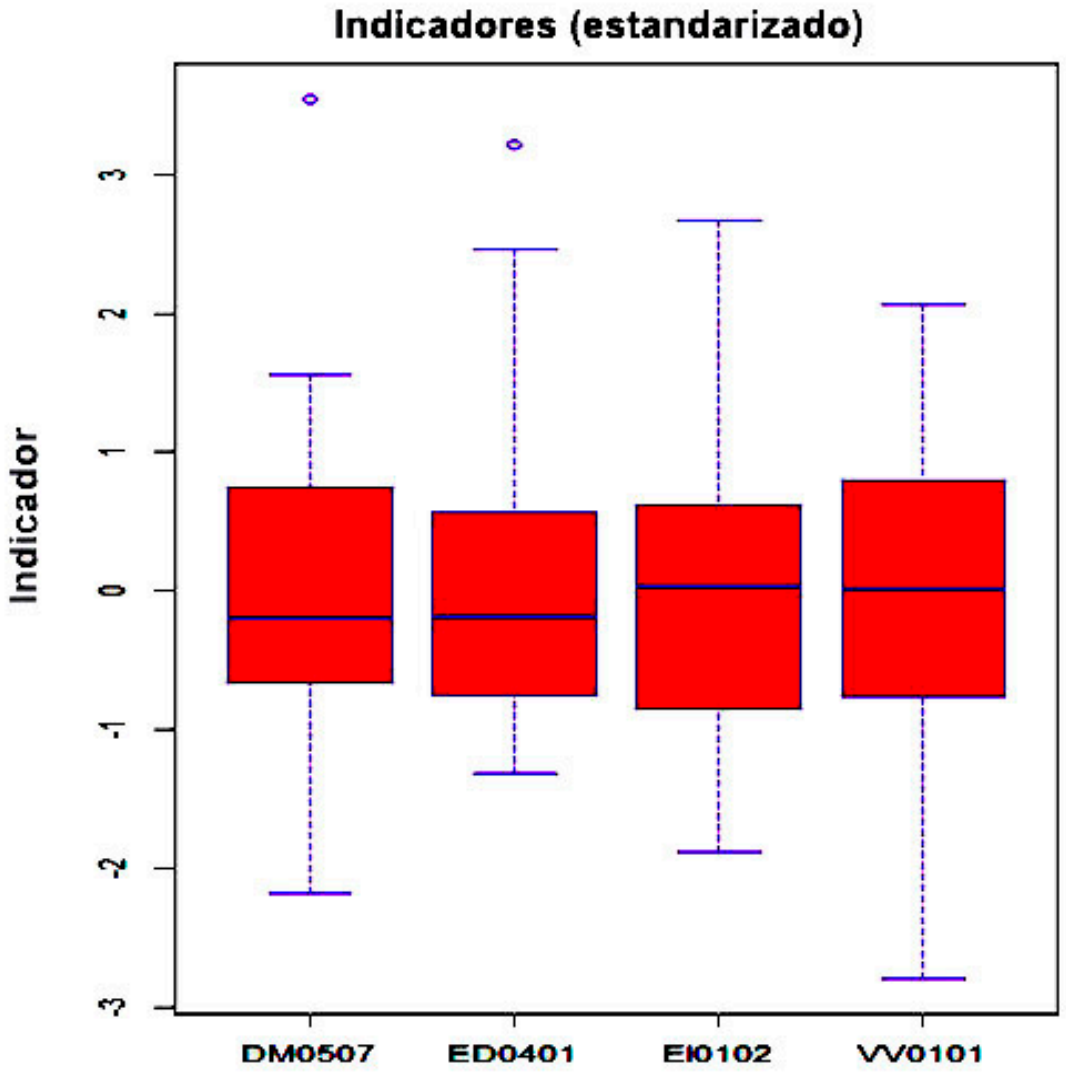
La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio). Cuyo código en la base de datos es DM0507.

A continuación, se presenta los resultados obtenidos de la investigación que fueron los siguientes:

En la Figura No. 1 se observa los indicadores en forma estandarizada, esta estandarización se realizó con el fin de realizar la aplicación de *machine learning* no supervisada.

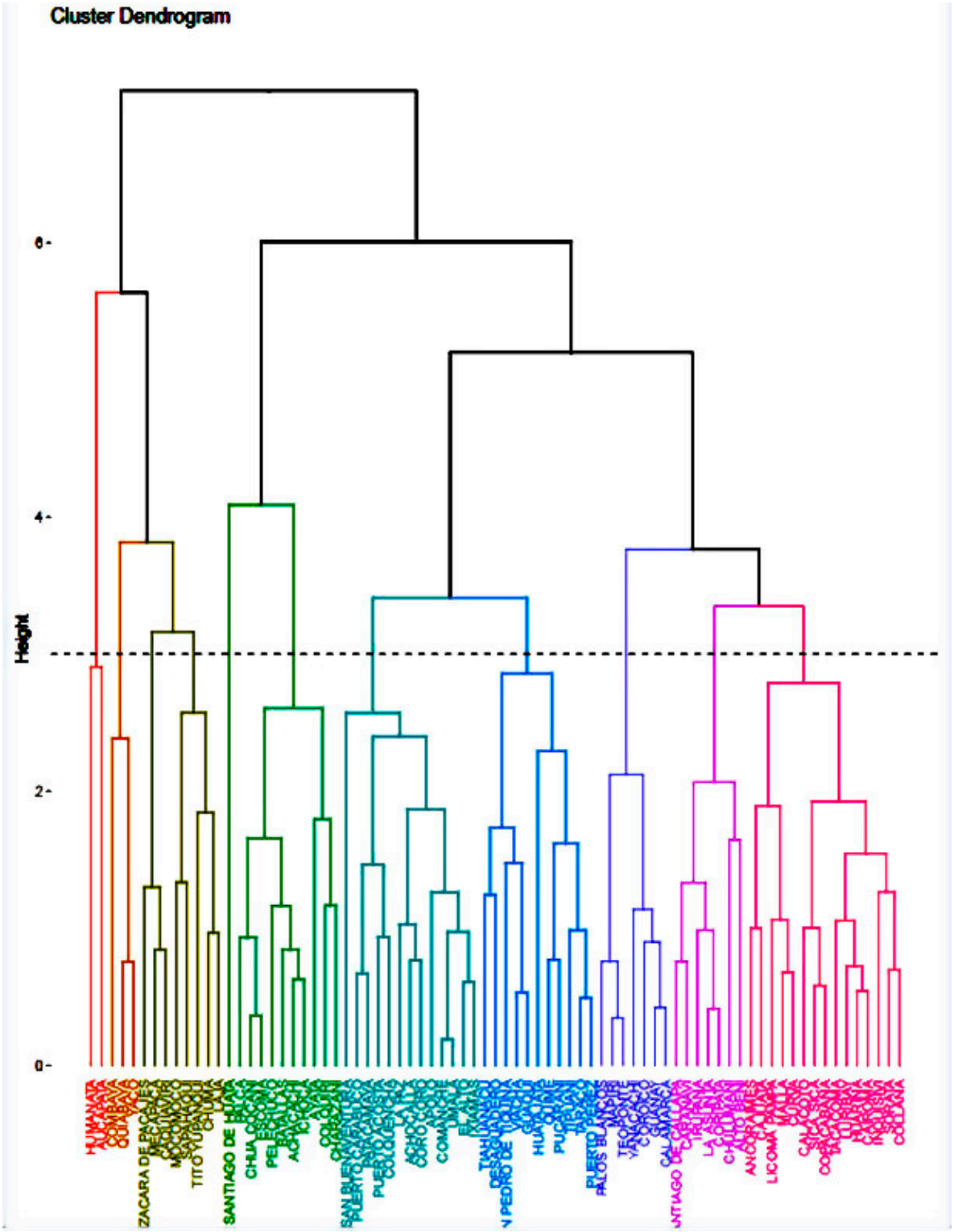
La aplicación de agrupación jerárquica que es una alternativa a los métodos de agrupación de *clusters* de particiones que no requiere que se pre-especifique el número de *clusters*, y en el presente estudio, se muestra en la Figura No. 2.

Figura No. 1
Indicadores de estudio, estandarizado



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Figura No. 2
Agrupaciones K-*medoids clustering*

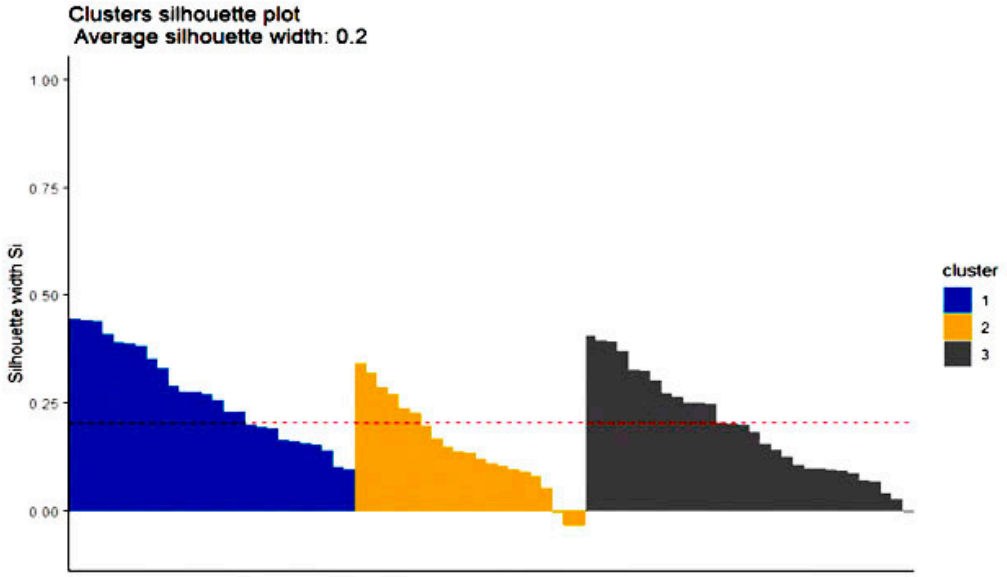


Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Para la validación interna de los *clusters*, se considera la homogeneidad (también llamada *compactness* o cohesión) sea lo mayor posible, a la vez es necesario la separación entre *clusters*. Cuantificar estar dos características es una forma de evaluar como de bueno es el resultado obtenido.

En esta investigación se utilizó el índice de *silhouette widht*. Cuyo resultado se muestra en la Figura No. 3.

Figura No. 3
Cluster silhoutte



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

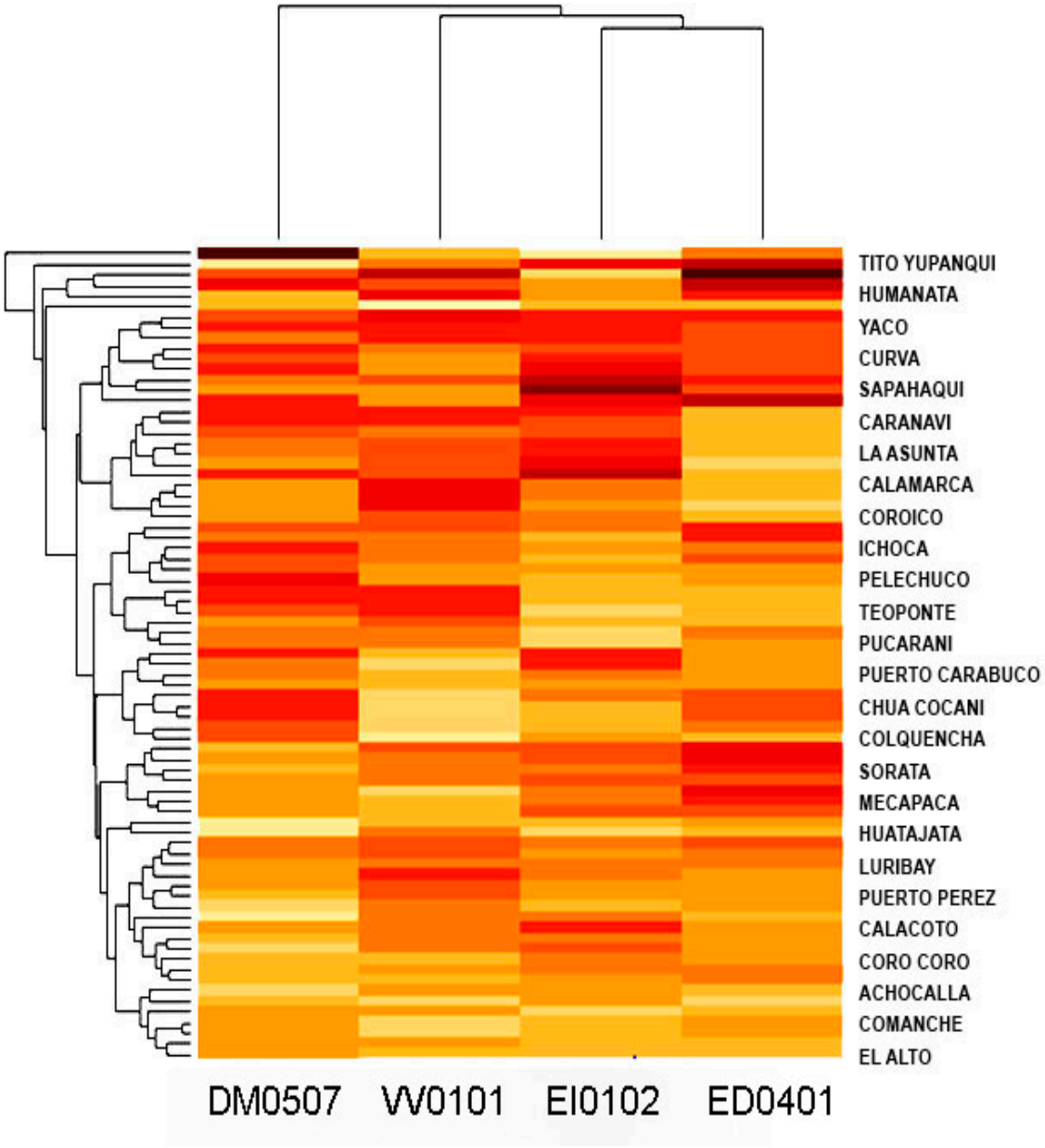
El *cluster* 2 muestra cierta dificultad por tener algunos valores negativos. Lo que implica que esas observaciones podrían tener no clasificadas correctamente.

A continuación, se presenta un *heatmaps* (mapa de calor). Los *heatmaps* son el resultado obtenido al representar una matriz de valores en la que, en lugar de números, se muestra un gradiente de color proporcional al valor de cada variable en cada posición.

La combinación de un dendograma con un *heatmap* permite ordenar por semejanza las filas y o columnas de la matriz, a la vez que se muestra con un código de colores el valor de las variables. Se consigue así representar más información que con un simple dendograma y se facilita la identificación visual de posibles patrones característicos de cada *cluster*.

En este estudio, se muestra un ejemplo de mapa de calor, de los datos de la encuesta sociodemográfica donde se considera los 75 municipios y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad. Como se muestra en la Figura No. 4.

Figura No. 4
heatmap



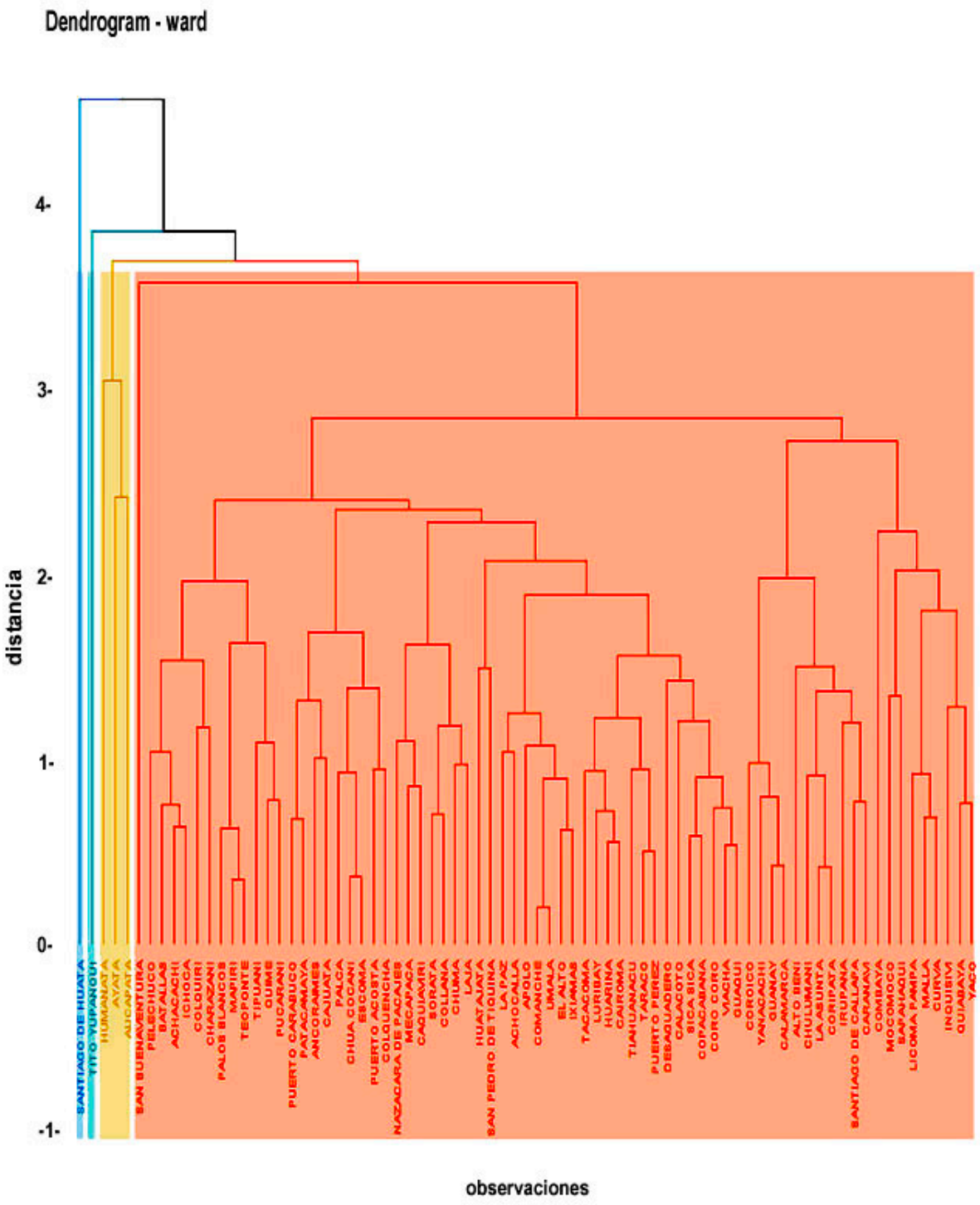
Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Es una forma de presentación a través de heatmap (*stats*) estos datos pueden ser comparables.

Prosiguiendo con la presentación de resultados, consideramos la customización de dendogramas que significa que toda observación pertenece al menos a uno de los k *cluster*. Que refiere a modificar algo de acuerdo a las preferencias personales.

Puede decirse, por lo tanto, que customizar un objeto es lo mismo que personalizarlo. Como se muestra en la Figura No. 5.

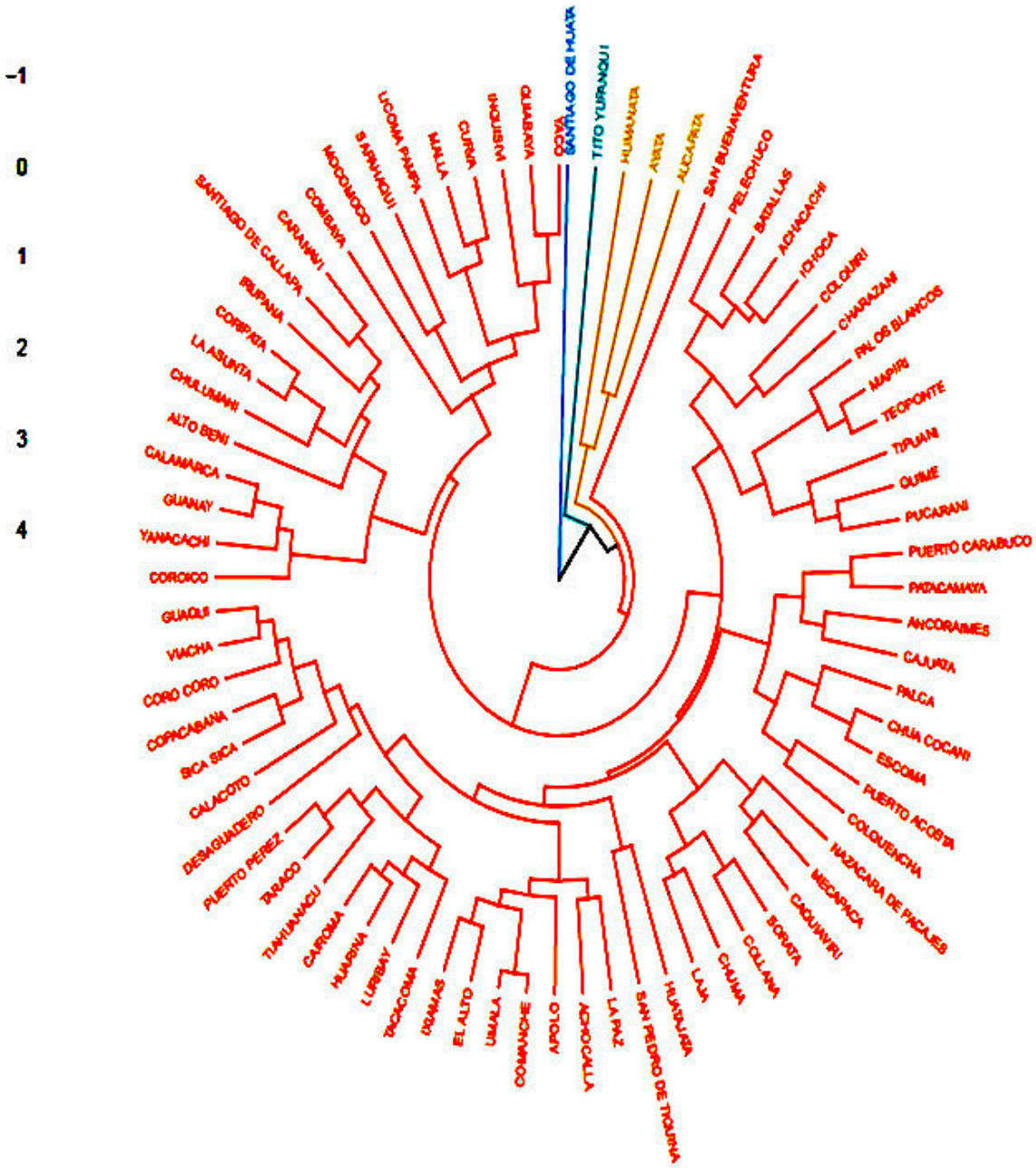
Figura No. 5
Dendograma customizado



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Existen varias formas de presentar, por ejemplo, dendograma circular como se muestra en la Figura No. 6.

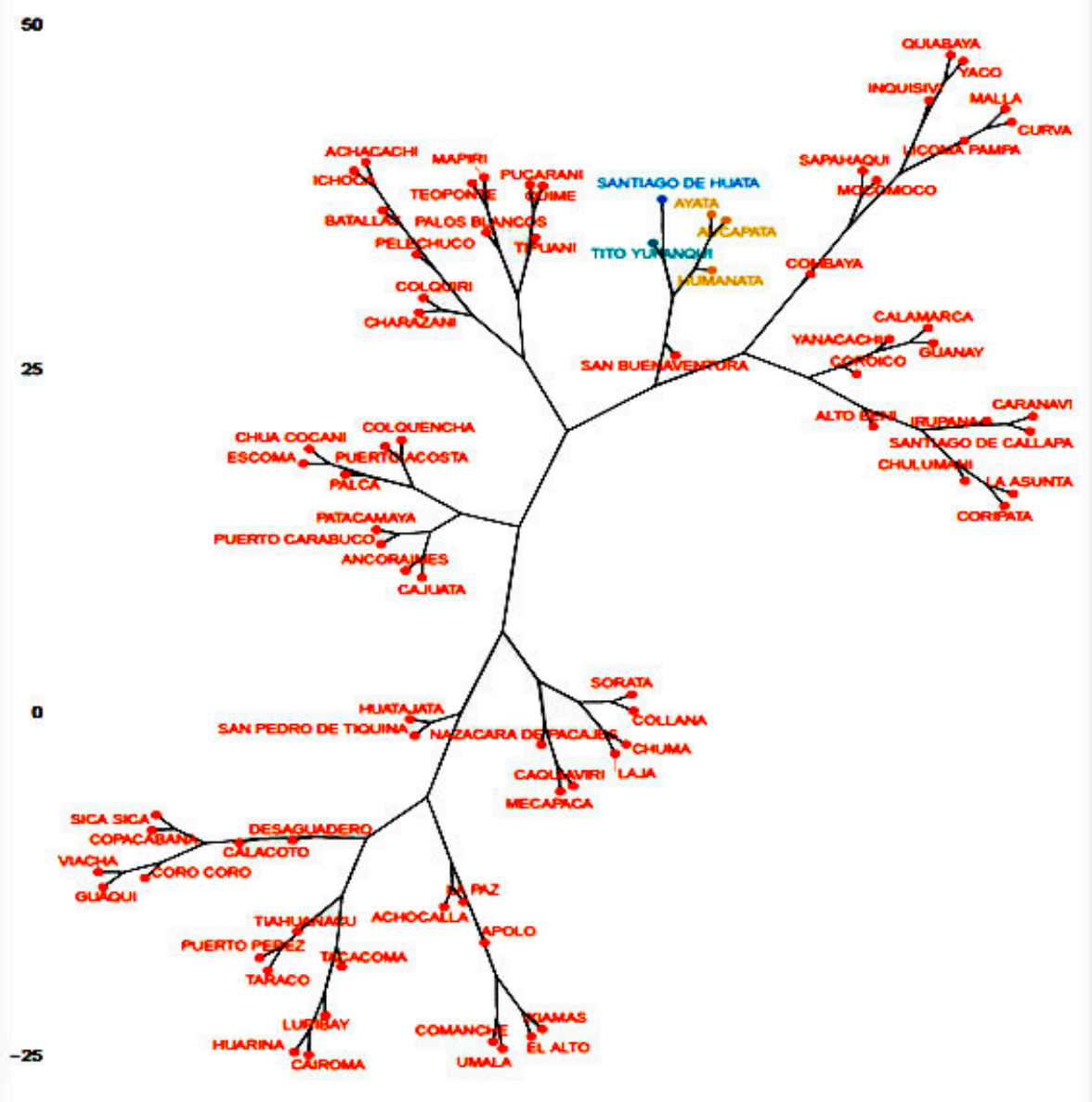
Figura No. 6
Dendograma circular (customizado)



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

También se presenta los resultados de la investigación a través de dendograma en forma de árbol filogenético, como se muestra en la Figura No. 7.

Figura No. 7
Dendograma en forma de árbol filogenético



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Las Figuras No. 5, 6 y 7 son formas distintas de presentar los resultados estudiados de los datos de la Encuesta Sociodemográfica del Departamento de La Paz, realizada en el año 2015, de los indicadores tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

CONCLUSIONES Y DISCUSIÓN

En el presente estudio se llega las siguientes conclusiones:

- La metodología *machine learning* es muy importante porque proporciona información que aporta en la descripción, el análisis y su posterior toma de decisiones.

- El presente estudio tomó en cuenta los datos de la encuesta sociodemográfica realizado el 2015 considerando los 75 municipios del departamento de La Paz y los indicadores educativos, empleo, demográficos y vivienda, dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

- Se debe notar que, la metodología *machine learning* es tipo supervisado y no supervisado, en el presente estudio se implementó estudio de *machine learning* no supervisado.

- Inicialmente tomó en cuenta al análisis *cluster* clásico el cual proporcionó la clasificación de agrupamiento de municipios similares respecto a la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad. Que no necesariamente fue muy claro por tener muchos municipios. La clasificación fue bastante confusa y dificultosa para su análisis.

- Se concluye que la customización es la mejor forma de clasificación por la forma en que se presenta en forma mucho más clara que las otras formas de clasificar.

- Existe la alternativa de clasificación como el dendograma circular, el dendograma en forma de árbol filogenético. Estas alternativas de clasificación son muy buenas porque permiten realizar las interpretaciones más claras y más contundentes.

- La clasificación de los municipios respecto al alfabetismo, tasa de participación, relación de masculinidad y la distribución de hogares según disponibilidad de dormitorios, dio como resultado, lo siguiente:

Los municipios Santiago de Huata, y Tito Yupanqui, muestran similares características respecto a los indicadores: tasas de participación, relación de masculinidad y tasa de alfabetismo.

Otro *cluster* definido por la customización son los municipios Humanata, Alcapata y Ayata y son parecidos en distribución de hogares según disponibilidad de dormitorios por persona, tasa de alfabetismo.

El resto de los *clusters* contiene a todos los restantes municipios cuya clasificación indica que tienen una similitud respecto a la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

Recomendaciones

- Se sugiere seguir estudiando *machine learning* no supervisado con otras alternativas para enriquecer este tipo de estudios y metodologías innovadoras.

- Seguir estudiando *machine learning* supervisado las cuales permitirán enriquecer este tipo de estudios y metodologías innovadoras.

- Como esta metodología está involucrada con *big data*, minería de datos, se sugiere seguir profundizando con las herramientas sugeridas.

- Se recomienda el manejo de *Python* para la ampliación de *machine learning*.

REFERENCIAS BIBLIOGRÁFICAS

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2° ed.) Springer-Verlag, New York. Cambridge, MA, 1997. Mit Press.

Cristianini, N. and Shawe- Taylor, J. (2000). An Introduction to Support Vector Machines

Harig, A.L. y Fausch, K.D.(2002). Minimum habitat requirements for establishing translocated cutthroat trout populations. Ecol. Appl.12 (2): pp. 535-551.

Murphy, K.P (2012). Introduction to Support Vector Machines

Nagelkerke, N.J. (1991). A note on a general definition of the coefficient of determination. Biometrika 78: pp. 691-692.

Osuna E., Freund, R. and Girosi, F. 2007. "An Improved Training Algorithm for Support.

Platt, J.C. (1997). "Fast Training of Support Vector Machines Using Sequential Minimum.

Río, M. del; Bravo, F.; Pando, V.; Sanz, G. & Sierra, R.(2004). Influence of individual tree and stand attributes in stem straightness in Pinus pinaster Ait. Stands. Ann. Sci. For.61(2): pp. 141-148.

Shalev, S. (2014). Understanding Machine Learning: From Theory to Algorithms.

Smola, A.J. and Schölkopf, B. (2004). "A tutorial on Support Vector Regression," Neuro COLT2.

Vaptnik, V. Golowich, S. and Smola, A. (1996). "Support vector method for function".

Vaptnik, V. N.(1995). The nature of Statical Learning Theory, New York: Wiley.

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 35-49

GENERACIÓN Z. AFECTACIONES A LA SALUD ASOCIADO AL USO DE LA TECNOLOGÍA

GENERATION Z. HEALTH EFFECTS ASSOCIATED WITH THE USE OF TECHNOLOGY

Elisa Mendoza G.[1]
Universidad de Panamá, Ciudad de Panamá, Panamá
✉ *elisa.mendoza@up.ac.pa*

Edilberto De León.[2]
Universidad de Panamá, Ciudad de Panamá, Panamá
✉ *edilberto.deleonm@up.ac.pa*

Pablo Moreira[3]
Universidad de Panamá, Ciudad de Panamá, Panamá
✉ *pablomoregonza1234@gmail.com*

Melanie Ortíz[4]
Universidad de Panamá, Ciudad de Panamá, Panamá
✉ *melaniedivina073@gmail.com*

Artículo recibido: 2021-07-20
Artículo aceptado: 2021-09-01

RESUMEN

El uso de las tecnologías es un tema trascendental en los últimos años, cada vez más evidente en todas las generaciones. Hasta el momento se afirmaba que las nuevas generaciones, particularmente la generación Z, era el grupo poblacional con mayor consumo de la tecnología, redes sociales y todos los equipos y accesorios que satisfacen la necesidad de una comunicación continua y permanente. Debido a esto, se plantea la interrogante si existe algún efecto a la salud y el comportamiento por el uso de la tecnología en la generación Z en comparación con las generaciones X e Y.

Por lo tanto, se hace una exploración para determinar las diferencias en cuanto al tiempo promedio de uso de la tecnología y sus posibles efectos en la salud entre los grupos generacionales: Z (los nacidos desde 1996) y, la X (1961-1980) y Y (1981-1995). La muestra no probabilística responde a un instrumento electrónico, en el que participaron 273 universitarios. El análisis realizado correspondió a los métodos estadísticos no paramétricos para determinar la asociación entre las variables de interés.

Entre los grupos generacionales, el tiempo promedio de uso de las tecnologías, no mostró diferencias estadísticamente significativas; respecto a las afectaciones a la salud, la ansiedad se presentó con mayor frecuencia en la generación Z (40%) comparado con las generaciones X e Y (34%); un comportamiento muy similar al trastorno del sueño 41% y 36%, respectivamente. También se observó que alrededor del 50% de la generación Z presenta problemas oculares.

Palabras clave: *Trastornos a la salud, Ansiedad, Millennials, Generaciones Z, Redes sociales, Problemas oculares.*

[1] Docente, investigadora y directora del Departamento de Estadística, Facultad de Ciencias, Naturales, Exactas y Tecnología. ORCID: 0000-0003-0089-6436
[2] Docente, investigador y Coordinador de Maestría en Cálculos y Técnicas Actuariales, Facultad de Ciencias, Naturales, Exactas y Tecnología. ORCID: 0000-0002-2696-9587
[3] Licenciado en Registros Médicos y Estadística de Salud. Escuela de Estadística. ORCID: 0000-0002-2826-7554
[4] Licenciada en Registros Médicos y Estadística de Salud. Escuela de Estadística. ORCID: 0000-0001-5015-0427

ABSTRACT

The use of technologies is a transcendental issue in recent years, increasingly evident in all generations. Until now, it was affirmed that the new generations, particularly generation Z, were the population group with the highest consumption of technology, social networks and all the equipment and accessories that satisfy the need for continuous and permanent communication. Because of this, the question arises whether there are any health and behavioral effects of the use of technology in Generation Z compared to Generations X and Y.

Therefore, an exploration is made to determine the differences in terms of the average time of use of the technology and its possible effects on health between the generational groups: Z (those born since 1996) and, X (1961-1980) and Y (1981-1995). The non-probabilistic sample responds to an electronic instrument, in which 273 university students participated. The analysis carried out corresponded to non-parametric statistical methods to determine the association between the variables of interest.

Among the generational groups, the average time of use of the technologies did not show statistically significant differences; regarding health effects, anxiety was presented more frequently in generation Z (40%) compared to generations X and Y (34%); behavior very similar to sleep disorder 41% and 36%, respectively. It was also observed that around 50% of generation Z have eye problems.

Keywords: *Health disorders, Anxiety, Millennials, Generations Z, Social networks, Eye problems.*

INTRODUCCIÓN

En los últimos años se ha observado un importante desarrollo de la ciencia, la tecnología y la comunicación a nivel global, que ha envuelto a la sociedad debido al uso continuo, casi permanente, de los equipos, artículos y demás accesorios tecnológicos. Esto como una forma de facilitar las tareas y el acceso a la información y a la comunicación en tiempo real, afectando en alguna medida la salud y el comportamiento de las personas (López-Barbosa, 2019).

Por lo general, es de interés caracterizar a las personas de acuerdo con sus características, actitudes y aptitudes frente a algo, por ejemplo, hacia su desarrollo personal, en el campo laboral o en competencias de mercado, por mencionar algunos. Sin embargo, en esta investigación es de interés estudiar las principales enfermedades que se presentan en las personas específicamente de acuerdo con su generación, esto debido a que cada una está relacionada con eventos trascendentales en el mundo entre ellos el desarrollo de la tecnología y bajo la hipótesis de que el uso de esta incide en su salud como lo son los problemas oculares, audición, ansiedad, estrés, obesidad, trastornos del sueño, entre otros.

Con el avance de las TIC, la sociedad en términos generales se adaptó a nuevos estilos de vida, en lo cotidiano, educativo y laboral. Estos cambios en la sociedad revierten en dos tendencias: En lo positivo, se observa un importante impacto en cuanto a facilitar tareas, la comunicación, la innovación y emprendimientos; en lo negativo, ha conllevado al uso adictivo de los dispositivos de comunicación como el celular, además de los videojuegos y sus accesorios, además de una necesidad de respuesta inmediata (o de estar conectados).

Por otra parte, se asocia de forma negativa el uso de la tecnología y la distracción con el incremento de los accidentes de tránsito. Según estudios realizados en la Universidad Politécnica de Tulancingo, los accidentes de tránsito a nivel mundial en los cuales están

involucrados los dispositivos móviles son más de 48 millones, de los cuales muchos enfrentarán lesiones de por vida (Cárdenas-Franco, 2018).

En personas que usan su teléfono móvil cinco horas o más al día, el riesgo de padecer obesidad puede ser hasta de un 43% superior que en aquellas que lo usan poco. Así lo explica una investigación de científicos de la Universidad Simón Bolívar de Barranquilla, Colombia, presentada en la última Conferencia Latinoamericana de Cardiología (Vásquez, 2019). De modo, que las afectaciones asociadas a la tecnología son diversas, así las generaciones no solo evidencian características de conducta o competencias tecnológicas propias de la época, sino que también cambios en la morbilidad y posibles causas de mortalidad.

De acuerdo con Soni (2016), el estudio de estas generaciones no es nada nuevo, pero resulta de mucho interés ya que existen diferencias en las características de estas que inciden en distintas formas en la sociedad, por ejemplo, en lo laboral los intereses de las nuevas generaciones (Gen Z) son muy distintos de las generaciones X y Y. Por lo que los procesos y sistemas de reclutamiento de personal deben ser eficaces para atraer y retener a su personal, independientemente de su edad (Soni, 2016, p.56). La generación X corresponden a los nacidos entre 1961 y 1980, quienes actualmente tendrán entre 40 y 60 años, aún en edad productiva laboral, que nace después de la guerra y practica valores de paz, amor, tolerancia y se promueven los derechos humanos. La generación Y, son los nacidos entre 1981 y 1995 y que representan a la generación que se adaptó a la tecnología y a las nuevas formas de comunicación, actualmente con edades entre 25 y 40 años. La generación Z, corresponden a todos los nacidos desde 1996 a la fecha, una generación joven nativa digital, quienes se encuentran en un proceso de incursión al mercado laboral, culminación de estudios superiores y quienes han vivido en mayor medida la revolución vertiginosa de la tecnología, siendo esta parte inclusive de sus hogares (Trincado, 2020).

Diversos estudios sobre estas cohortes generacionales y su relación con el uso de la tecnología muestran diferencias en diversos aspectos particularmente en su conducta, competencias y en ocasiones se mencionan algunos problemas de salud. Por ejemplo, en Madrigal Moreno (2018) se refieren a los *Millennials*, para referirse a la generación Y como una generación con un estilo informal al vestir en el entorno laboral, jóvenes competitivos y con expectativas muy altas respecto a los salarios. Esta generación no dura más de tres años en el mismo trabajo, ya que no busca permanencia, sino desafíos en el ámbito privado y profesional o un pago más elevado que el anterior, por eso los gerentes contratan a trabajadores maduros, son más confiables y comprometidos que los Millennials (Madrigal Moreno, Ávila Carreón, & Madrigal Moreno, 2018).

Para ampliar aún más sobre estas clasificaciones de las generaciones, se debe referir a Chávez quien señala que:

> Cada generación comparte contextos históricos, sociales y culturales; tienen su época de protagonismo y decadencia, en la medida que alcanzan una posición mayoritaria en el mercado del trabajo y luego, a medida que envejecen, van disminuyendo su protagonismo dando paso a las generaciones futuras (Chávez, 2018, p. 12).

El aporte de Chávez define características poblacionales de las cohortes generacionales marcadas por las experiencias, hitos o tendencias de la época, que hasta cierto punto, se puede observar en mayor medida en la participación en el plano laboral; podría

decirse que estas características inciden en la dinámica de los mercados y el relevo generacional.

Por otro lado, Chirinos (2009) cita a Gilburg (2007) indicando que Generación es "un grupo de edad que comparte a lo largo de su historia un conjunto de experiencias formativas que los distinguen de sus predecesores" (Chirinos, 2009, p.137), lo que permite en alguna medida diferenciar los distintos grupos poblacionales que vivieron situaciones muy particulares marcados por hechos mundiales o tendencias de la época, en algunos casos, hechos bélicos, recesión económica, el desarrollo tecnológico, por mencionar algunos.

Ante este panorama, donde es claro que estas generaciones están caracterizadas por situaciones particulares y mayormente asociadas con el desarrollo económico, científico y tecnológico, entre otros, y resulta de interés poder responder respecto a las condiciones de salud, lo siguiente: ¿Cuáles son las principales afectaciones a la salud que caracteriza a la población llamada 'generación Z' en comparación con las generaciones X y Y?

Por lo tanto, se tiene como principal objetivo analizar las principales enfermedades que afectan a los grupos poblacionales generación Z en comparación con las generaciones X y Y, y su relación con el uso de las tecnologías, cabe señalar que en Chirinos (2009) se hace referencia a la generación X, como una generación "independientes desde niños", mientras que la Generación Y o *Millenials*, los define como aquellos "que han crecido con las vidas totalmente planificadas", además de tener "fácil acceso a la información a través de la tecnología". También, en Dutra (2017) se define a los *Millennials* como personas *multitasting*, es decir, con la capacidad de realizar varias tareas a la vez, además de manejar muy bien las redes sociales.

Toda vez, que estas tres cohortes generacionales están asociadas con el sector productivo, económico y social; en donde, mientras una generación nueva se va incorporando al sistema, otra se encuentra en procesos de acogerse a su pensión o retiro laboral (Trincado, 2020).

Afectaciones a la salud

En Hodelin (2021) se hace un análisis muy interesante de las afectaciones de la salud debido al uso prolongado de ordenadores, en el cual se comprende y entiende que se está viviendo una época en donde la tecnología es parte de la vida, de la sociedad y de lo laboral de las poblaciones; además, no se desconocen los problemas de salud que se presentan en la población a "causa del uso inadecuado de la computadora". Entre los trastornos de salud señalados están: riesgo de cáncer, diabetes y problemas cardíacos, obesidad, estrés ocular, desórdenes de tendón, desórdenes de la vista, debilitamiento muscular, entre otros. Si se observa con detenimiento estos trastornos se puede relacionar además con la falta de actividades físicas y herramientas que se contrapongan a los efectos negativos de lo que mal bien se puede llamar un "mal necesario" puesto que la tecnología es en mucho de los casos la herramienta de trabajo, y las competencias laborales giran en torno a esas destrezas de uso, de procesamiento de la información, de generar resultados en forma inmediata, conllevando a estas generaciones a una competencia innata al continuo apego a la información y a los equipos tecnológicos.

Una importante distracción en la cual se incurre casi imperceptible es precisamente la afectación a la salud, contradiciendo el concepto de estado salud que ha definido la

Organización Mundial de la Salud (OMS) como: "un estado completo bienestar físico, mental y social, y no la simple ausencia de enfermedad" (Rodríguez, 2020).

Otros problemas que son hasta cierto punto un "escándalo" se refieren a los accidentes de tránsito cuyo posible factor causal, directa e indirectamente, se asocia al uso del celular. Aunque, no es motivo de este estudio demostrarlo, si se puede señalar que, en estas generaciones, particularmente Y y Z, existe esa necesidad de la comunicación continua y no solo eso, sino de respuesta inmediata, en tiempo real (día y noche) ocasionando trastornos de sueño, de concentración (distracciones) y otros (Hodelín, 2021; Illescas, 2021; León-Pluas, et al., 2019). De forma tal que estudios realizados muestran que los conductores se distraen enviando mensajes de textos mientras conducen. De acuerdo con Illescas (2021), los jóvenes entre 17 y 29 años leen y responden mensajes mientras manejan. En Panamá, el 1% de las infracciones reportadas por la Autoridad de Tránsito y Transporte Terrestre (2021) durante el período enero – mayo de 2021, promediaban, 642 infracciones por "hablar por teléfono al conducir", situaciones que pueden ocasionar accidentes de tránsito, más por la distracción del conductor que por otros factores ambientales o geográficos e infraestructuras vial.

En Panamá, las principales causas de muerte al 2018 correspondieron a los tumores con una tasa por cien mil habitantes de 75.5, seguidas por las enfermedades isquémicas del corazón (43.2), enfermedades cerebrovasculares (43.1), accidentes, lesiones autoinfligidas, agresiones y otras violencias (36.4) y en quinto lugar, la diabetes mellitus (32.8), (INEC, 2019), en todos los casos, con mayor tasa en hombres que en mujeres. Aquí, se debe hacer un énfasis, en los problemas de salud relacionados con la salud mental, puesto que las estadísticas oficiales señalan que las principales afectaciones en los grupos de 15 a 24 años, la principal causa, la representan los accidentes, lesiones autoinfligidas, agresiones y otras violencias, con 311 defunciones (50.1%), en el grupo de 25 a 44, esta misma causa fue la principal causa de muerte, representando un 32.4%; cuando se sigue a los otros grupos de edad, de otras generaciones, se evidencia el cambio de los tipos de afectaciones como lo son los tumores malignos, entre otros (INEC, 2016).

Alteraciones visuales

Algunos de los trastornos asociados al uso frecuente de los aparatos y equipos electrónicos, computacionales, e incluso video juegos, se refieren a los visuales, e incluso a los postural (Vargas, 2004), en donde los problemas visuales se presentan en mayor medida en la población desde muy temprana edad. Jiménez (2021) determinó que una de las afectaciones está relacionada con la agudeza visual interocular observada en el 31.2% de los jóvenes escolares de entre los 4 a 12 años de edad. De forma tal, que estos estudios concluyen y abogan por políticas públicas en las que se promuevan hábitos e higiene dirigidas a la protección visual, tomando en cuenta que la tecnología no es el enemigo del desarrollo económico y social, sino de la salud de las poblaciones (Bustamante-López, 2021).

Alteraciones musculoesqueléticas

El uso cotidiano de las redes sociales entre las más comunes: *Facebook, Twitter, Instagram y WhatsApp*, que hasta cierto punto han permitido que las personas utilicen sus teléfonos celulares de manera obsesiva. Esto produciendo sin darse cuenta en importantes afectaciones a su sistema

musculo esquelético, siendo la enfermedad del túnel carpiano la mayor prevalente en la población de 25 a 40 años. Sin embargo, se han encontrado casos de jóvenes de 18 años y hasta menos, demostrando que son los jóvenes los máximos consumidores de estos equipos en el mercado.

La enfermedad del túnel carpiano puede provocar hormigueo, entumecimiento y daños musculares en la muñeca y los dedos en donde el movimiento repetitivo, permanente, ocasiona enfermedades de tipo progresivo que pueden afectar la movilidad y desempeño a la hora de utilizar las manos como lo es el Síndrome del túnel del carpo (STC), el cual es definido como: "Esta es una afección en la cual hay presión excesiva sobre el nervio mediano, el nervio de la muñeca que permite la sensibilidad y el movimiento a partes de la mano" (Villa-Martínez, 2014), siendo un mecanismo de prevención y de recuperación, en primer lugar reconocer el problema y en segundo la realización de ejercicios de muñeca, dedos, manos, brazos, articulaciones en general. En Rodríguez (2020) se señala como una afectación que se ha incrementado de manera importante es el síndrome del manguito rotador, de hasta un 118%, además llega a la conclusión, que estos problemas de salud no se asocian a factores socioeconómicos, sino más bien al uso de computadora, afectando no solo las manos, sino a los miembros superiores y cuello.

Trastornos a la salud mental

Para referirnos a los trastornos asociados a la salud mental, se introduce el concepto de adicción, para referirse al consumo o uso excesivo de algo y que en ocasiones más que serles de utilidad y de un bien, puede ser contraproducente al producir ansiedad, trastornos del sueño y en el peor de los casos el *cyberbullying* (Trincado, 2018) De manera que las generaciones X, Y y Z están expuestas a este tipo de afectaciones de forma involuntaria, acarreando problemas de adicción a las nuevas tecnologías caracterizada al consumo excesivo de: teléfonos celulares, videojuegos, computadoras y redes sociales y muchas veces ocasionada por presión social, de grupos de amigos o incluso familiares (Secades-Villas, 2012; Cruzado-Díaz, 2006).

Por ejemplo, la OMS señala que una de cada cuatro personas, sufre trastornos relacionados con las nuevas adicciones. Algunas manifestaciones de la adicción al uso de la tecnología, en particular del celular, es por ejemplo, cuando nunca quieren salir y relacionarse con otras personas o en otras actividades. Sin importar en donde estén estas personas prefieren estar pegados a sus teléfonos celulares, a las redes sociales o a cualquier aparato electrónico, para no convivir con los demás (Cruzado-Díaz, 2006; Martín, 2021), dominando su voluntad de hacer otras actividades, pues el *like* se ha convertido en una necesidad dominando de las generaciones *millennials*.

En el estudio realizado por Martín (2021) se pudo determinar que las generaciones millennials, particularmente, la Z, prefiere usar *Instagram* y *WhatsApp* como sus redes sociales favoritas, además de permanecer más de 3 horas al día conectados. Los diferentes estudios que tratan de la adicción a la tecnología enfatizan ese impulso involuntario a las redes sociales como una puerta de aceptación al mundo exterior, ocasionando en ellos efectos negativos como: "ansiedad, depresión, estrés o fuertes sentimientos de soledad" (Jiménez et al., 2020).

MATERIALES Y MÉTODOS

En este estudio participaron profesores y estudiantes de la Universidad de Panamá de

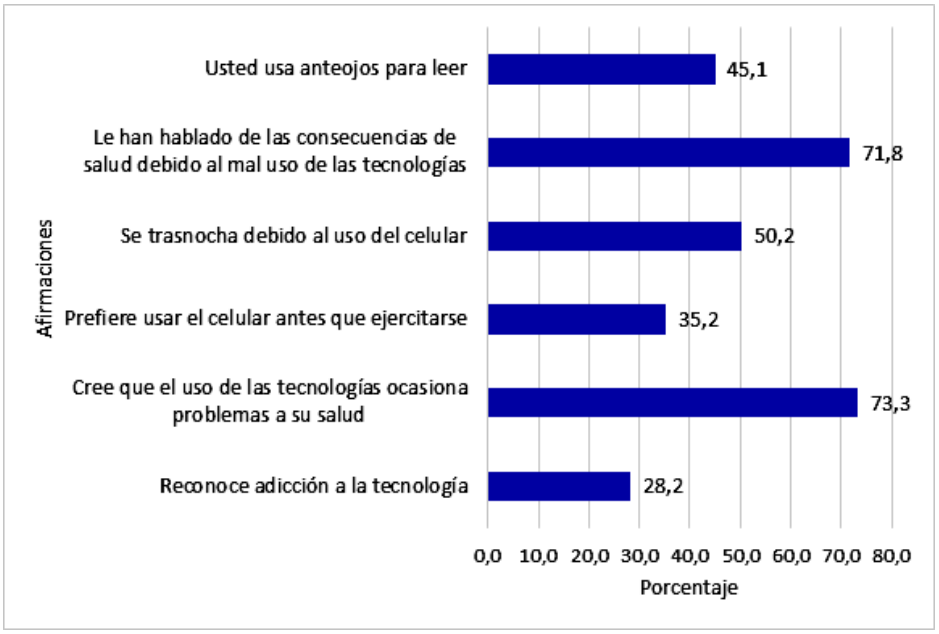
las facultades de Humanidades y Ciencias Naturales, Exactas y Tecnología, menores de 60 años, representando una muestra no probabilística de 273 universitarios, de los cuales 185 correspondieron al grupo de la generación Z y 88 al grupo de la generación X y Y. Se empleó un instrumento electrónico, el que fue enviado mediante correo a los decanos de ambas facultades para su distribución. El cuestionario constó de una gama de preguntas estructurada en 3 secciones y catorce preguntas claves las cuales recogían información de: identificación de la generación, actividades que realiza, uso de la tecnología, tiempo de uso, afectaciones a la salud, horas de estudio, horas de descanso, redes sociales, Nivel de dificultad del uso de la tecnología. Para el análisis de los datos se utilizó *Excel* versión 2016 y *Jamovi*. Se emplearon pruebas estadísticas no paramétricas con un nivel de significancia del 5%, entre ellas, la prueba de Chi-cuadrado para asociar las variables de salud y las generaciones Z vs las generaciones X y Y, y la prueba de Kruskal-Wallis, con la finalidad de evidenciar diferencias en cuanto al tiempo de uso de las tecnologías en los tres grupos.

RESULTADOS

Los principales resultados del estudio muestran que para las cohortes generacionales X, Y y Z, la edad promedio se determinó en 55, 39 y 20 años, respectivamente. En estos grupos la proporción de mujeres fue mayor, 54, 62 y 66%.

El Internet, es el medio que más utilizan para las actividades académicas, según el 73% de los participantes en este estudio, el restante mencionó otros recursos como: periódicos, revistas, libros y bibliotecas. Un 65% informó que el uso de la tecnología es un importante distractor para hacer otras actividades; aunado a esto, también, consideran que es necesario reducir este tiempo que se emplean

Gráfico No. 1
Afirmaciones respecto al uso de la tecnología por los universitarios que representan las generaciones X, Y y Z. Septiembre-octubre de 2020.



Fuente: Encuesta aplicada a estudiantes y profesores de la Facultad de Humanidades y de Ciencias Naturales, Exactas y Tecnología de la Universidad de Panamá. Sept.-oct. de 2020.

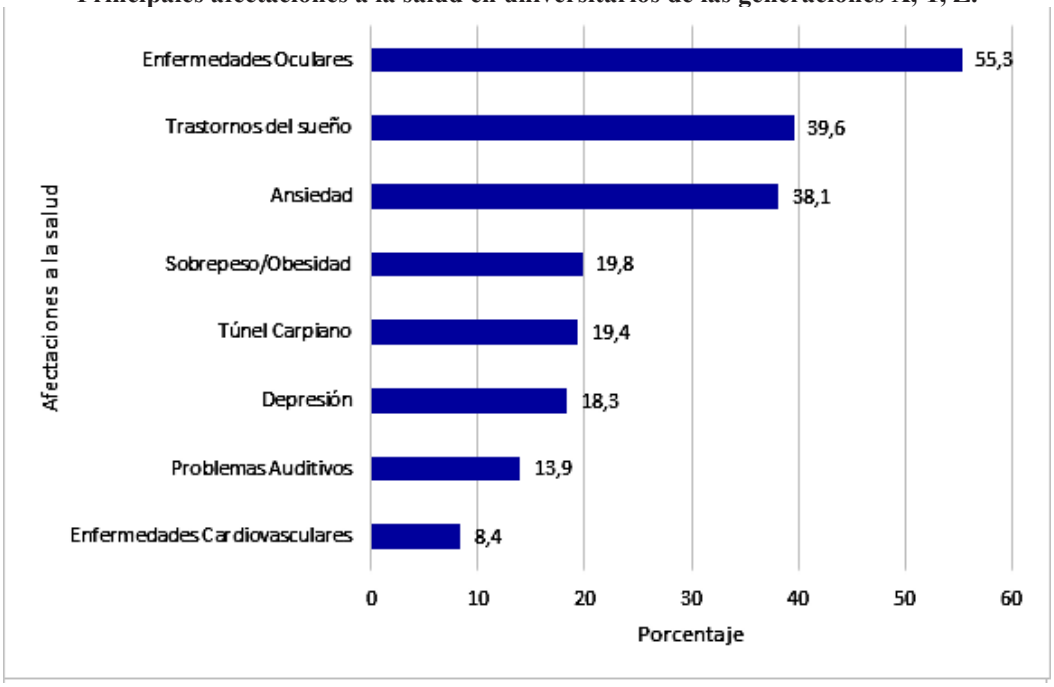
en el uso de las redes y el celular (79%), una causa de este fenómeno es una afectación a las relaciones y comunicación familiar (79%).

En el estudio también se afirmaron algunos factores asociados al uso de la tecnología como es si tienen conocimiento de las consecuencias de salud debido al mal uso de las tecnologías, en el que un 71.8% afirmó tener esta información y además, el 73.3% reconoció que el uso de las tecnologías ocasiona problemas a la salud; sin embargo, también, el 50.2% asevera que se trasnocha debido al uso del celular, pero solo el 28.2% reconoció adicción a la tecnología y un 35.2%, prefiere usar el celular antes que realizar ejercicios físicos.

Otra información relevante al uso de la tecnología, en particular a las redes sociales consultadas, se identificaron por los usuarios universitarios como la más utilizadas, el *WhatsApp* (93%) y el *Instagram* (75%), en tercer y cuarto lugar, el *Facebook* (31%) y el *Twitter* (15%), respectivamente.

Con relación a los trastornos o afectaciones a la salud de los universitarios correspondientes a estas generaciones X, Y, y Z, se identificaron como la más prevalente las relacionadas con las enfermedades oculares (55.3%), seguido

Gráfico No. 2
Principales afectaciones a la salud en universitarios de las generaciones X, Y, Z.



Fuente: Encuesta aplicada a estudiantes y profesores de la Facultad de Humanidades y de Ciencias Naturales, Exactas y Tecnología de la Universidad de Panamá. Sept.-oct. de 2020.

por problemas de trastornos de sueño (39.6%) y ansiedad (38.1%), dejando evidencias de hacia donde orientar las políticas sanitarias para la atención de estas afectaciones en poblaciones económicamente activas como lo representan estas tres generaciones.

Una prueba estadística que relaciona algunas de estas afectaciones con los grupos generacionales, dejan evidenciado que las enfermedades cardiovasculares y los problemas oculares están asociadas estadísticamente. Es relevante, destacar que,

Tabla No. 1
Análisis de asociación mediante el estadístico Chi-cuadrado, de algunas afectaciones a la salud y la cohorte generacional

Enfermedades que padece	*Cohorte generacional*		*Chi Cuadrado*	*p-valor*
	Generación Z	*Generaciones X y Y*		
Total	**185 (100%)**	**88 (100%)**		
Ansiedad				
Sí	74 (40.0)	30 (34.1)	0.883	0.3473
No	111 (60.0)	58 (65.9)		
Trastorno del sueño				
Sí	76 (41.1)	32 (36.4)	0.555	0.4562
No	109 (58.9)	56 (63.6)		
Depresión				
Sí	34 (18.4)	16 (18.2)	0.002	0.9686
No	151 (81.6)	72 (81.8)		
Obesidad				
Sí	31 (16.8)	23 (26.1)	3.306	0.069
No	154 (83.2)	65 (73.9)		
Enfermedades Cardiovasculares				
Sí	7 (3.8)	16 (18.2)	16.023	0.0001(*)
No	178 (96.2)	72 (81.8)		
Problemas Oculares				
Sí	93 (50.3)	58 (65.9)	5.901	0.0153 (*)
No	92 (49.7)	30 (34.1)		

(*) significancia estadística al nivel del 5%.
Fuente: Encuesta aplicada a estudiantes y profesores de la Facultad de Humanidades y de Ciencias Naturales, Exactas y Tecnología de la Universidad de Panamá. Sept.-oct. de 2020.

en la cohorte de la Generación Z, el porcentaje de universitarios con problemas oculares es de 50.3%, uno de los problemas de salud más alarmante en este grupo poblacional, con una edad promedio de 20 años.

Con respecto al tiempo de uso de la tecnología, particularmente, celulares, computadoras y accesorios, se comparó el tiempo promedio de uso de las tres generaciones (Tabla No.2 y Gráfico No. 3), en donde se puede observar que los tiempos promedios de las

Tabla No. 2
Comparación del tiempo promedio de uso de la tecnología y la cohorte generacional.

Estadísticas	*Generación X (1961 – 1980)*	*Generación Y (1981 – 1995)*	*Generación Z (nacidos después del (1996)*
N	36	52	185
Media	6.36	6.56	7.35
Mediana	6	6.5	8
Desviación estándar	3.25	3.53	3.44

Fuente: Encuesta aplicada a estudiantes y profesores de la Facultad de Humanidades y de Ciencias Naturales, Exactas y Tecnología de la Universidad de Panamá. Sept.-oct. de 2020.

Gráfico No. 3
Principales afectaciones a la salud en universitarios de las generaciones X, Y, y Z.

Fuente: Encuesta aplicada a estudiantes y profesores de la Facultad de Humanidades y de Ciencias Naturales, Exactas y Tecnología de la Universidad de Panamá. Sept.-oct. de 2020.

generaciones X. e Y son muy parecidas, 6.36 y 6.56 horas; y la generación Z presentó un tiempo promedio de 7.35 ± 3.44 horas un poco más elevada que las otras generaciones, sin embargo, estadísticamente no son diferentes ($p > 0.05$) de acuerdo con la prueba de Kruskal-Wallis.

DISCUSIÓN

Las definiciones de las cohortes generacionales están marcadas por hitos relevantes acontecidos en el mundo y compartidos por las poblaciones nacidas en dichas épocas y en el que Strauss & Howe (2008) citado por (Chávez y Barrera, 2018) señalan que dichos acontecimientos marcaron aquellas características comunes: personales, de vida, de comportamiento sociales, culturales, debido al momento que vivieron y que en alguna medida influyeron en su vida adulta, familiar, social, laboral, además de sus hábitos de autocuidados y de salud, tema sobre el cual se enfoca este estudio.

Por ejemplo, Bruce (2016) señala que la era de los *millennials* (para referirse a los nacidos después de 1978) corresponden a la época de grandes cambios entre ellos la globalización, la vertiginosa revolución tecnológica, el acelerado flujo de información y ese creciente y continuo ritmo de comunicación en la humanidad. Si bien se observa en los jóvenes actuales (e inclusive adultos que se han adaptado a estos cambios) en el que existe casi una dependencia total a la comunicación casi permanente y la necesidad de respuestas instantáneas a la información en tiempo real (Villa-Martínez, 2014), determinado en este estudio por poco más de la mitad (50.2%) de los universitarios, quienes además reconocen el efecto negativo a su salud por el uso desmedido de la tecnología (73.3%).

En la actualidad, en la mayoría de los países, particularmente en Panamá se está atravesando por un proceso de transición demográfica (Muñoz-Núñez, 2020; Mendoza et al., 2015); Soria-Romero y Montoya-Arce (2017) destacan que este fenómeno se evidencia por un cambio en la estructura de la pirámide poblacional en la cual se afecta el envejecimiento poblacional de forma tal que se observa una menor tasa de fecundidad y mayor esperanza de vida (reducción de la mortalidad), observándose mayor longevidad en su población (Mendoza

y Cosme, 2017; INEC, 2016). Ante este panorama, y las características de las generaciones actuales, cabe evidenciar las principales afectaciones de la población, tal como se identificó en el estudio, el 45.1% utiliza anteojos para leer, considerando que la edad promedio de los participantes en este estudio era de 38 años; y que estos hábitos del uso de la tecnología no solo afectan la salud sino sus actividades físicas, puesto que el 35.2% indicó que prefieren usar el celular a ejercitarse. Al respecto la OMS (2018) informa que entre 2015 y 2050, el porcentaje de los habitantes del planeta mayores de 60 años casi se duplicará, pasando del 12% al 22%, esto implica que habrá un mayor número de personas de 60 años o más que niños menores de 5 años y, por lo tanto, exige un reto importante en los sistemas sanitarios para la atención de estas poblaciones. Se señalan como principales afectaciones: pérdida de la audición, cataratas, dolores de espalda y cuello, depresión, entre otros. Se mencionan una serie de factores asociados a dichas afectaciones, sin embargo, no se menciona manera directa como factor causal el uso excesivo de la tecnología. Más bien, se refieren a estas como una herramienta que bien aprovechada puede ser un aliado en el campo de la salud (OMS, 2015), además, se enfatiza en la importancia de la realización de actividades físicas periódicas, dieta equilibrada, hábitos saludables.

Las estadísticas de salud analizadas en este estudio correspondiente a enfermedades no transmisibles mostraron afectaciones relevantes como las enfermedades oculares (55.3%), esto representa un problema de salud y de calidad de vida (Pérez, et al., 2021; Hodelin, 2016), pues como se ha dicho, existen pocos estudios que enfocan las afectaciones a la salud y su causal potencial, el uso desmedido de las computadoras. En otros casos, se alinean con problemas de la ergonometría justificando el hecho de pasar horas sentado frente a un computador y su efecto en las actividades laborales (Martínez-Moreno, 2021; Hodelín, 2016) en donde el uso y posturas incorrectas deterioran la salud, tal como se determinó en este estudio, en el cual el 19.4% de los encuestados indicaron tener afectaciones en el túnel carpiano para referirse a problemas de salud (dolores en las manos y/o muñecas). Otros, señalaron problemas auditivos (13.9%), además de problemas relacionados con el estado de salud mental y emocional como lo son: trastornos del sueño (39.6%), ansiedad (38.1%), Depresión (18.3%) y un problema de salud muy relacionada con enfermedades de tipo cardiovascular que es la obesidad.

Aquí es necesario hacer un paréntesis y puntualizar en que las generaciones Y y aún más la generación Z particularmente, presentan mayor promedio de horas en el uso de la tecnología, por ejemplo, mientras que en la generación X se determinó un promedio de 6.4 ± 3.3; en la generación Y fue de 6.6 ± 3.5 y en la generación Z de 7.4 ± 3.4 horas. Esto conlleva a una falta de actividades físicas, es decir, mucho tiempo sentados o en el peor de los casos acostados (Hodelín, 2016). En este mismo artículo se indica una realidad, “es imposible encontrar un trabajo que no involucre en mayor o menor medida el uso de un ordenador”, lo que lleva a un mayor número de horas frente a una pantalla, sentados, exponiendo los ojos a dicha luz. Hodelín (2016) va más allá de las enfermedades no transmisibles y enlista enfermedades graves como el riesgo de cáncer, diabetes, problemas cardíacos y la obesidad que es un factor relevante asociado a los problemas cardiovasculares, que usualmente, se presentan en personas adultas y adultas mayores y que podría convertirse en un problema grave en poblaciones más jóvenes (*millennials*).

CONCLUSIONES

La tecnología en sí mismo no es un problema a discusión, puesto que el desarrollo de las tecnologías, de la comunicación y de la información en definitiva imprime un factor innovador al ser humano, por ejemplo, agiliza procesos, permite la comunicación en tiempo real sin importar la distancia; sin embargo desconocer, las afectaciones a la salud debido a ésta es un serio problema complejo que incide en los sistemas sanitarios; puesto que solo se mira el problema como un resultado de la edad, o en todo caso de una transición demográfica-epidemiológica sin atender de forma preventiva y con toda la relevancia que amerita los hitos que han marcado las generaciones X, Y y Z, entre ellos globalización, desarrollo tecnológico, innovación en la comunicación y el acceso a la información.

Múltiples estudios se han realizado para caracterizar a estas generaciones en su forma de actuar, de manejarse ante sus empleadores, necesidades, competencias y potencialidades; más se han ignorado los problemas de salud asociados a ellos, por lo que, los problemas visuales, auditivos, motrices, salud mental (depresión, ansiedad, otros), obesidad entre otros con el paso de los años han ganado terreno y a mediano plazo podrían convertirse en un problema de salud irremediable.

No se encontró diferencias en el tiempo promedio de uso de la tecnología entre las tres generaciones estudiadas, esto es que, aunque se hacen llamados de atención de manera enfática en las nuevas generaciones por la persistencia al uso de los aparatos tecnológicos, lo cierto es que las generaciones X e Y también se han adaptado al uso frecuente de esta superando las seis horas diarias en promedio, tal vez por cuestiones de competitividad laboral o la necesidad de estar comunicados con las nuevas generaciones (sus hijos), tema que queda pendiente de discusión.

Las enfermedades cardiovasculares, son un problema de salud significativamente importante en las generaciones X e Y, lo mismo que los problemas oculares; sin embargo, se debe resaltar que la mitad del grupo generacional Z ya se autoidentificó con esta afectación tomando en cuenta que es una población joven, lo que podría convertirse en un importante problema de salud pública a corto plazo.

REFERENCIAS BIBLIOGRÁFICAS

ATTT (2021). Reporte estadístico de accidentes de tránsito, Panamá, 2021. Recuperado de: http://www.transito.gob.pa/sites/default/files/estadistica_accidentes_junio-fusionado.pdf.

Bruce, T. (2016). No todo el mundo merece un trofeo. Cómo liderar millennials de manera efectiva. Grupo editorial Patria. México.

Bustamante et al. (2021). Higiene y protección visual en el uso de las Tecnologías de la Información y Comunicaciones. Revista Cubana De TecnologíA De La Salud, 12(2), 191-198. Recuperado de http://revtecnologia.sld.cu/index.php/tec/article/view/2101.

Cárdenas-Franco, J. A. (05 de Julio de 2018). Milenio 2020. Obtenido de Uso excesivo de dispositivos móviles: https://www.milenio.com/opinion/varios-autores/universidadpolitecnica-de-tulancingo/

uso-excesivo-de-dispositivos-moviles

Chávez D., B. y Barrera V., G. (2018). Reporte especial: Emprendimiento en las 4 generaciones: Baby Boomers, X, Millennials, Z. Región de Los Ríos 2017-2018. INACAP. Recuperado de: http://www.inacap.cl/web/2019/flippage/reportes-gem/generacion/los-rios-generacion/files/LOS-RIOS-ID-GENERACION-07052019-PAGINAS.pdf.

Chirinos, N. (Julio-diciembre 2009). Características generacionales y los valores. Su impacto en lo laboral. Observatorio Laboral Revista Venezolana. Vol. 2, N°4: pp. 133-153.

Martín-Critikián, D. y Medina Núñez, M. (2021). Redes sociales y la adicción al like de la generación z. Revista de Comunicación y Salud, 11, 55-76. Recuperado de: https://doi.org/10.35669/rcys.2021.11.e281.

Cruzado-Díaz, L., Matos-Retamozo, L., y Kendall-Folmer, R. (2006). Adicción a internet: Perfil clínico y epidemiológico de pacientes hospitalizados en un instituto nacional de salud mental. Revista Medica Herediana, 17(4), pp. 196-205. Recuperado en 20 de julio de 2021, de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1018-130X2006000400003&lng=es&tlng=es.

Dutra, Ma.F. (2017). Generación Z: entre las nuevas formas de organización del trabajo y la convivencia generacional. (Trabajo de grado: Universidad de la República). Uruguay. Recuperado de: https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/10918/1/Dutra%2c%20Ma.%20Florencia.pdf.

Hodelín, Y., de los Reyes, Z., Hurtado, G., Batista, M. Riesgos sobre tiempo prolongado frente a un ordenador. Rev Inf Cient [Internet]. 2016 [citado 20 Jul 2021];, 95(1):[aprox. 15 p.]. Disponible en: http://www.revinfcientifica.sld.cu/index.php/ric/article/view/149.

Illescas, E. (2021). Análisis del uso del celular al conducir un vehículo en la ciudad de Cuenca. (]Trabajo de pregrado). Universidad Politécnica Salesiana. Sede Cuenca. Recuperado de:https://dspace.ups.edu.ec/bitstream/123456789/19912/1/UPS-CT008983.pdf.

INEC. (2016). Situación de la población. El proceso de Transición Demográfica en Panamá. Sección 221, año 2016. Panamá.

INEC. (2016). Comentarios. Recuperado de: https://www.inec.gob.pa/archivos/P7731Comentarios.pdf.

INEC. (2019). Estadísticas vitales. Volumen III. Defunciones. Año 2018. Panamá.

Jiménez, V., De Jesús Ruiz, Ma., Huerta, M. y Alcantar, Ma.L. (2020). Dependencia al uso del celular en estudiantes universitarios de la ciudad de Morelia. Revista Eureka.

Jiménez, M. (2020). Estudio poblacional de las alteraciones visuales infantiles en el área escolar de Sevilla. (Trabajo Fin de Grado Inédito). Universidad de Sevilla, Sevilla.

León-Pluas et al., (2019). Análisis de causas de accidentes de tránsito en el Ecuador utilizando Minerías de Datos. Revista Ibérica de Sistemas e Tecnologías de Información. pp. 540-547.

Madrigal-Moreno, F., Ávila-Carreón, F., y Madrigal-Moreno, S. (2018). Retos y oportunidades del comportamiento organizacional de los millennials como fuerza de trabajo. Universidad Michoacana de San Nicolás de Hidalgo, pp. 86-95.

Martínez, P., Aguirre, M. y González, W. (2015). Estudio ergonómico como parte de la responsabilidad social en trabajadores del centro regional de informática de la Universidad Veracruzana, México. Inquietud Empresarial. Vol. XV(2), pp. 87-114.

Mendoza, E. y Cosme, M. (2017). Protocolo de Análisis, Políticas Públicas para Panamá y Lineamientos de Comunicación. Informe Ministerio de Seguridad. Panamá.

Mendoza, E., Rodríguez, R., Quintero, E., Góndola, E. & Cruz, C. (2019). Situación de Salud de la Región de San Miguelito. 2006-2015. Centros: Revista Científica Universitaria. 8(1), pp. 101-114. Recuperado a partir de https://revistas.up.ac.pa/index.php/centros/article/view/483.

Muñoz-Núñez, N. (2020). Percepción de salud biopsicosocial y laboral: caso Universidad de Panamá, Veraguas. Centros: Revista Científica Universitaria. 9(2), pp. 114-157.

Vargas, M. (2004) Incidencia de uso de los videojuegos en alteraciones visuales ergonómicas, en niños de 9 a 14 años. Cienc Tecnol Salud Vis Ocul. (3): pp. 37-51.

OMS (5 de febrero de 2018). Envejecimiento y salud. Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/envejecimiento-y-salud.

OMS (2015). Informe Mundial sobre el Envejecimiento y la Salud.

Pérez Tejeda, A., Acuña Pardo, A., y Rúa Martínez, R. (2008). Repercusión visual del uso de las computadoras sobre la salud. Revista Cubana de Salud Pública, 34(4) Recuperado en 19 de julio de 2021, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-34662008000400012&lng=es&tlng=es.

Rodríguez Espinosa, K. X. (2020). Trastornos musculoesqueléticos en personal administrativo. Ergonomía, Investigación Y Desarrollo, 2(2), pp. 151-162. Recuperado a partir de http://revistasacademicas.udec.cl/index.php/Ergonomia_Investigacion/article/view/2413.

Secades-Villas, R. (2012). Psicologia de las Adicciones. Ovideo: Universidad de Ovideo Facultad de Psicologia.

Soni, H. y Ashish, A. (2016). Undestanding Generation Gap at Work Place. IOSR. Journal of Business and Management. Volume 18, Issue 8. Ver. 1 PP.56-58 https://www.researchgate.net/profile/Harvinder-Soni/publication/306272889_Understanding_Generation_Gap_at_Work_Place/links/5f8ea138458515b7cf8dda03/Understanding-Generation-Gap-at-Work-Place.pdf.

Soria-Romero, Z. y Montoya-Arce, B. (2017). Envejecimiento y factores asociados a la calidad de vida de los adultos mayores en el Estado de México. Papeles de Población, 2(93). pp. 59-93.

Trincado-Rivas, Javiera C., (2020). Cyberbullying. Impacto de las redes sociales en la generación Z. Universidad de

Chile. Recuperado de: http://repositorio.uchile.cl/bitstream/handle/2250/175859/Cyberbullyng-impacto-de-las-redes-sociales.pdf?sequence=1.

Villa-Martínez, S. (2014). Los smartphones y su incidencia en el síndrome del túnel carpiano. (Trabajo de pregrado) Universidad de San Buenaventura de Cartagena. Recuperado de: http://bibliotecadigital.usbcali.edu.co/bitstream/10819/2347/1/Los%20smartphone%20y%20su%20incidencia_Sol%20Villa_USBCTG_2014.pdf.

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 51-62

MODELO LOGÍSTICO MULTINOMIAL CONDICIONES SOCIOECONÓMICAS DE LAS PERSONAS QUE HABITAN EN LA CIUDAD DE EL ALTO

MULTINOMIAL LOGISTIC MODEL SOCIOECONOMIC CONDITIONS OF PEOPLE THAT INHABIT IN EL ALTO CITY

M. Sc. Fernando Rivero Suguiura[1]
Instituto de Estadística Teórica y Aplicada - UMSA, La Paz - Bolivia
✉ *friverosuguiura2004@gmail.com*

Artículo recibido: 2021-08-05
Artículo aceptado: 2021-09-13

RESUMEN

El objetivo es investigar las condiciones de vida de las personas que habitan en la ciudad de El Alto de La Paz, a partir de la referencia de algunas variables demográficas y socioeconómicas como son: edad, nivel de educación, seguridad alimentaria, ingreso, gasto mensual y ocupación, con la aplicación del modelo logístico multinomial que clasifica a la población en nivel socioeconómico alto, medio y bajo. Además, el modelo permite la medición probabilística de pertenecer a dichas categorías con mayor o menor influencia.

Palabras clave: *Condiciones socioeconómicas, modelo logístico multinomial, ciudad de El Alto.*

ABSTRACT

The objective is to investigate the living conditions of people living in El Alto city, located in the department of La Paz, based on the reference of some demographic and socioeconomic variables such as: age, education level, food security, income, monthly expenditure and occupation, implicating the multinomial logistic model that classifies the population into high, medium and low socioeconomic level. In addition, the model allows the probabilistic measurement of belonging to these categories with greater or lesser influence.

Keywords: *Socioeconomics conditions, multinomial logistic model, El Alto city*

INTRODUCCIÓN

La ciudad de El Alto se encuentra ubicada en la meseta del altiplano norte del departamento de La Paz en la cuarta sección de la provincia Murillo, al pie de la cordillera oriental a 12 km del centro de la ciudad de La Paz a una altura de 4.500 m.s.n.m. y tiene una extensión de 387,56 km², según el Plan de Desarrollo de la Ciudad de El Alto. Ciudad creada el 6 de marzo de 1985 como urbe segunda más poblada de Bolivia con una población aproximada de 943.600 habitantes, donde el 51,4% son mujeres y el 48,6% son hombres, de estos, el 54,2% es menor de 30 años, según los datos del último Censo de Población y

[1] Docente de la carrera de Estadística, Facultad de Ciencias Puras y Naturales de la UMSA. Consultor en muestreo, censos y análisis estadístico en entidades nacionales e internacionales. Magister en Ciencias de la Estadística. https://orcid.org/0000-0001-9095-7778

Vivienda (INE, 2012).

Respecto al sector educativo, en 2018, la población matriculada en los niveles inicial, primaria y secundaria en la educación pública y privada llegó a 300.927 personas. El porcentaje de matriculados que cumplieron con el requisito mínimo para un curso inmediatamente superior (tasa de promoción) alcanzo al 97%, los que abandonaron la escuela o colegio (tasa de abandono) fueron 1,6% y los que no cumplieron con la nota mínima de aprobación (tasa de reprobados) 1,3%. (El Alto en Cifras 2020 INE).

El Alto se caracteriza por su dinámico movimiento comercial y productivo basado en el crecimiento del sector de la micro, pequeña y mediana empresa y artesanía productiva, calificada por ello, como una de la segunda ciudad industrial de Bolivia (Chuquimia, 2008), tal es el caso de la feria 16 de Julio que se instaura domingo tras domingo ofreciendo el comercio de productos desde un alfiler hasta una maquinaria industrial sofisticada.

En la ciudad de El Alto se encuentran las principales vías de salida de las mercancías de exportación, siendo estas el Aeropuerto Internacional y la Zona Franca Industrial y Comercial. El movimiento de mercancías por estas aduanas llegó a un total de 1.047,7 millones de dólares en 2019, siendo la principal vía de salida el Aeropuerto de El Alto (INE, 2020).

El Producto Interno Bruto (PIB) per cápita de esta ciudad alcanza aproximadamente al 25% del PIB del departamento de La Paz que es aproximadamente de 16.558 millones de bolivianos y al 6% del PIB de Bolivia (Estrategia de Desarrollo Económico Local e Informe Estadístico del municipio de El Alto, 2020). Este porcentaje de aporte del PIB, especialmente proviene del sector servicios, comercio y transporte, como el de manufactura en: minería, refinerías de petróleo, fábricas de azúcar y aceite, entre otros.

En el tema de pobreza, la ciudad de El Alto es una de las ciudades metrópoli más pobres de Bolivia, según el índice de Necesidades Básicas Insatisfechas (NBI) del Censo de Población y Vivienda del año 1992 al 2001, según el INE, reduce de un 73,8% a un 66% y posteriormente en el año 2005 alcanza a un 47,5%. Respecto a la desigualdad, se relaciona principalmente con la falta de servicios básicos como: alcantarillado, agua, energía eléctrica y vivienda propia.

En el tema de desigualdad habitacional, el Índice de Calidad de Vivienda reporta que el 52% de la población alteña cuenta con una vivienda catalogada como de nivel alto, el 47% de nivel medio y 1% de nivel bajo, sin embargo, este incremento en calidad no deja en expectativa la desigualdad de los habitantes de esta ciudad en este tema (INE, 2005).

A todo lo anteriormente señalado, la investigación se concluye con un estudio de las condiciones de vida de los pobladores de la ciudad de El Alto en base a información recopilada de la Encuesta de Hogares del INE Bolivia año 2018, mediante la aplicación del modelo logístico multinomial. Este último permite un análisis exhaustivo del tema de acuerdo al uso de algunas características sociales y económicas.

METODOLOGÍA

Modelo logístico multinomial

La regresión logística multinomial, es

utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías y es una extensión multivariante del modelo logístico binario. Las variables independientes pueden ser tanto continuas como categóricas o no métricas. En esta aplicación sobre las condiciones socioeconómicas de las personas habitantes de la ciudad de El Alto, se construye la variable dependiente condición socioeconómica, mediante el método multivariante *cluster* análisis no jerárquico con tres categorías (alta, media y baja). Esta variable o factor, considera las subvariables tales como: a) materiales de construcción de la vivienda (techo, piso, pared); b) hacinamiento (número de personas por dormitorio); c) agua y saneamiento básico (disponibilidad de agua, alcantarillado, baño); d) insumos energéticos (combustible, energía eléctrica entre otros); e) salud y educación (acceso, atención de la salud por personal calificado, años de escolaridad y alfabetismo), fuente Encuesta de Hogares (INE, 2018).

Formulación del modelo

Sea la variable dependiente Y categórica nominal politómica con probabilidades p_1, p_2, … , p_k para las k categorías que compone la variable Y. Si se requiere el análisis del efecto de variables independientes X_1, X_2, …, X_p se define el modelo siguiente:

$$p_c = \frac{1}{1 + e^{-X\beta}} \qquad (01)$$

haciendo algunos cambnios en (01) se tiene

$$p_c = \frac{1}{1 + \frac{1}{e^{X\beta}}} = \frac{1}{\frac{1 + e^{X\beta}}{e^{X\beta}}}$$

$$= \frac{e^{X\beta}}{1 + e^{X\beta}} \qquad (02)$$

donde p_c es la probabilidad de que la categoria $Y = c$; $c = 1, 2, \ldots, k$ se dé, con $\sum_{c=1}^{k} p_c = 1$. Además, en (02) se tiene que:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{pmatrix}$$

Donde X es la matriz de dimensión n x (p+1) de observaciones y variables, como β que es el vector de dimensión p+1 de parámetros a estimar en el modelo. Si $q_c = 1 - p_c$ es la probabilidad complementaria de p_c, tal que $p_c + q_c = 1$, entonces:

$$q_c = 1 - p_c$$

$$= 1 - \frac{e^{X\beta}}{1 + e^{X\beta}} \qquad (03)$$

resulta de reemplazar (02) en (03). Luego

$$q_c = \frac{1 + e^{X\beta} - e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{1 + e^{X\beta}}$$

La razón entre p_c/q_c se define como la razón

$$r = \frac{p_c}{q_c} = \frac{\frac{e^{X\beta}}{1 + e^{X\beta}}}{\frac{1}{1 + e^{X\beta}}} = e^{X\beta}$$

considerándose un modelo no lineal.

Método de estimación de máxima verosimilitud

La función de verosimilitud, está definida como:

$$L = \prod_{i=1}^{n} \left(p_{1i}^{Y_{1i}} p_{2i}^{Y_{2i}} \dots p_{ki}^{Y_{ki}}\right) = \prod_{i=1}^{n} \left[\left(\frac{p_{1i}}{p_{ki}}\right)^{Y_{1i}} \left(\frac{p_{2i}}{p_{ki}}\right)^{Y_{2i}} \dots \left(\frac{p_{k-1,i}}{p_{ki}}\right)^{Y_{k-1,i}} p_{ki}\right] \quad (04)$$

Aplicando a la función de razón de verosimilitud generalizada se tiene

$$\Lambda = -2ln(L) = -2\sum_{i=1}^{n} \left[Y_{1i} ln\left(\frac{p_{1i}}{p_{ki}}\right) + Y_{2i} ln\left(\frac{p_{2i}}{p_{ki}}\right) + \dots + Y_{k-1,i} ln\left(\frac{p_{k-1,i}}{p_{ki}}\right) + ln(p_{ki})\right] \quad (05)$$

La maximización en (04) es equivalente a minimizar (05). Por la complejidad de la función Λ, esta se resuelve por métodos numéricos de forma iterativa para los estimadores $\hat{\beta}$ del vector β.

Significación del modelo

Para probar si el modelo es significativo o no en su estructura global, se aplica el test estadístico de contraste de hipótesis, teniendo en cuenta que la diferencia entre el valor inicial y el valor final de la función de razón de verosimilitud generalizada Λ tiene distribución Chi-cuadrado con grados de libertad igual al número de regresores multiplicado por el número de categorías menos uno.

La hipótesis nula (H_0) de que no existe efecto de las variables regresoras, contra la hipótesis alterna (H_1), se define como:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$
$$H_1: \beta_1 \neq \beta_2 \neq \dots \neq \beta_p \neq 0 \quad (06)$$

El p' valor del test para H_0 vendrá dado por la siguiente probabilidad

$$p' = P\left(\chi^2_{(p-1)(k-1)} > \Lambda_0 - \Lambda_1\right)$$

se rechaza H_0 si p'<0,05.

RESULTADOS

Aplicación del modelo logístico multinomial

La aplicación del modelo logístico multinomial definido anteriormente, considera las siguientes variables.

La variable 1 (Clasific_Soc) corresponde ser la dependiente del modelo con las categorías: 1 (Alto), 2 (Medio) y 3 (Bajo). Las siguientes variables de la 2 a la 5 son de condición categórica: género, edad, nivel de educación y seguridad alimentaria. Las variables 6 a la

Tabla No. 1
Variables para el modelo

Nro.	Variable	Condición	Descripción	Categoría
1	Clasific_Soc	Dependiente	Clasificación Socioeconómica	1. Alto
				2. Medio
				3. Bajo
2	Gen	Categórica	Genero o Sexo	0. Mujer
				1. Hombre
3	Edad	Categórica	Edad en años	1. de 0 a 19
				2. de 20 a 39
				3. de 40 y más
4	Niv_educ	Categórica	Nivel Educativo	1. Ninguno
				2. Básico
				3. Superior
5	Seg_alim	Categórica	Seguridad Alimentaria	1. Inseguridad Severa y Moderada
				2. Inseguridad Leve
				3. Seguridad
6	I_h	Covariable	Ingreso del Hogar Mensual en Bs	
7	G_h	Covariable	Gasto del Hogar Mensual en Bs	
8	P_ocup	Covariable	Porcentaje de Ocupados por Hogar	

Fuente: Encuesta de Hogares 2018 (INE), elaboración propia.

8, son cuantitativas o llamadas también covariables tales como: ingreso y gasto del hogar mensual en Bs y el porcentaje de ocupados, estas últimas no categorizadas.

Especificación del modelo

Inicialmente se realiza la categorización de las variables independientes o explicativas para ser consideradas en el modelo logístico. Para ello, en algunos casos, se crean variables ficticias o *dummy* (dicotómicas), 1 si pertenece a la categoría y 0 cuando no pertenece, es decir:

$$X_j = \begin{cases} 1 & si\ pertenece\ a\ la\ categoría \\ 0 & si\ no\ pertenece\ a\ la\ categoría \end{cases}$$

X_3: Porcentaje de ocupados del hogar

X_4: Género mujer

X_5: Género hombre

X_6: Edad entre 0 y 19 años

X_7: Edad entre 20 y 39 años

X_8: Edad de 40 y más años

X_9: Nivel educación ninguno

X_{10}: Nivel educación básico

X_{11}: Nivel educación superior

X_{12}: Seguridad alimentaria (inseguridad severa y moderada)

X_{13}: Seguridad alimentaria (inseguridad leve)

X_{14}: Seguridad alimentaria (seguridad)

Las variables cuantitativas ingreso y gasto del hogar siguen un proceso de estandarización para que sean comparables ambas, mediante el siguiente método de transformación de variable Z_j

$$Z_j = \frac{X_j - \bar{X}_j}{S_j} \qquad j = 1, 2$$

Donde X_j tiene la media y desviación estándar muestral ($\bar{X}_j$, S_j). Por lo que se cuenta con las siguientes dos variables:

Z_1: Ingreso del hogar mensual en Bs, estándar

Z_2: Gasto del hogar mensual en Bs, estándar

Luego el modelo que genera las probabilidades de categorización, está dado por:

$$p = \frac{e^{\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5+\beta_6 X_6+\beta_7 X_7+\beta_8 X_8+\beta_9 X_9+\beta_{10} X_{10}+\beta_{11} X_{11}+\beta_{12} X_{12}+\beta_{13} X_{13}+\beta_{14} X_{14}}}{1+e^{\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5+\beta_6 X_6+\beta_7 X_7+\beta_8 X_8+\beta_9 X_9+\beta_{10} X_{10}+\beta_{11} X_{11}+\beta_{12} X_{12}+\beta_{13} X_{13}+\beta_{14} X_{14}}} \quad (07)$$

con la variable dependiente formulada como

$Y=X\beta$

$$=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5+\beta_6 X_6+\beta_7 X_7+\beta_8 X_8+\beta_9 X_9+\beta_{10} X_{10}+\beta_{11} X_{11}+\beta_{12} X_{12}+\beta_{13} X_{13}+\beta_{14} X_{14}$$

luego (07) se puede escribir como

$$p_c = \frac{e^{Y}}{1+e^{Y}}$$

X la matriz de dimensión n=3.965 (muestra de hogares ciudad de El Alto) por p+1=15 de variables, y β el vector de dimensión p+1=15 de parámetros a estimar en el modelo. Sin embargo, la regresión logística multinomial presenta k-1 modelos desagregados de acuerdo a la cantidad de categorías menos uno que tiene la variable dependiente Y, en este caso dos modelos diferentes.

Estimación y significación del modelo logístico

Aplicando la función de razón de verosimilitud generalizada (Λ), expresión (05), y realizando la maximización por métodos numéricos de forma iterativa, se consigue estimar el modelo logístico con coeficientes $\hat{\beta}$.

Se han probado diferentes combinaciones de variables categóricas y covariables, además de la inclusión y exclusión de éstas, para determinar el modelo más significativo a los datos analizados. Se han analizado 12 modelos posibles con las 14 variables y categorías descritas anteriormente, donde la variable ***Genero*** con categorías ***mujer*** y ***hombre*** parecían influyentes en el resultado, sin embargo, se pudo observar que dicha variable no es relevante en el modelo, ignorándola y quedando con el modelo final estimado en su versión 12, siguiente.

Los estimadores de β para las variables y categorías consideradas en los modelos, se presentan en la segunda columna de la Tabla No. 2. Se puede observar, que el coeficiente independiente $\hat{\beta}_0$ es negativo cercano a -3 para el primer caso, sin embargo para el segundo caso es -0.24 y poco significativo individualmente. Los coeficientes estimados de las variables cuantitativas Ingreso y Gasto estándar (Z_1 , Z_2), para ambas opciones de la variable dependiente Y de los modelos Bajo y Medio, son negativos y significativos según el estadístico de contraste de Wald (valor alto) y probabilidad Sig, presente en la 6ta columna que se mide por debajo de 0,05. La variable porcentaje de ocupados por hogar (X_3) en ambas categorías de Y son relativamente cercanas a cero, pero significativas en ambos modelos, de acuerdo a los estadísticos Wald y Sig.

La columna siete de la tabla No. 2 presenta los siguientes resultados más relevantes:

Tabla No. 2

Estimadores y características de los coeficientes β

Estimaciones de parámetro

VARIABLE DEPENDIENTE CLASIFICACIÓN SOCIOECONÓMICA[a]		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B) Límite inferior	Límite superior
Bajo	Intersección	-2,854	0,227	157,367	1	0,000			
	Ingreso estandarizado	-1,100	0,096	131,260	1	0,000	0,333	0,276	0,402
	Gasto estandarizado	-1,326	0,108	151,441	1	0,000	0,266	0,215	0,328
	PORCENTAJE DE OCUPADOS POR HOGAR	0,008	0,002	10,455	1	0,001	1,008	1,003	1,012
	[Edad recodificada dos=1]	0,171	0,137	1,568	1	0,210	1,187	0,908	1,551
	[Edad recodificada dos=2]	0,652	0,138	22,392	1	0,000	1,920	1,465	2,516
	[Edad recodificada dos=3]	0[b]			0				
	[NIVEL DE EDUCACIÓN GENERAL=1,00]	2,038	0,232	77,250	1	0,000	7,673	4,871	12,088
	[NIVEL DE EDUCACIÓN GENERAL=2,00]	1,035	0,170	37,167	1	0,000	2,816	2,019	3,927
	[NIVEL DE EDUCACIÓN GENERAL=3,00]	0[b]			0				
	[CODIGO SEGURIDAD ALIMENTARIA=1]	1,829	0,199	84,286	1	0,000	6,229	4,215	9,205
	[CODIGO SEGURIDAD ALIMENTARIA=2]	0,780	0,113	47,254	1	0,000	2,181	1,746	2,724
	[CODIGO SEGURIDAD ALIMENTARIA=3]	0[b]			0				
Medio	Intersección	-0,236	0,148	2,539	1	0,111			
	Ingreso estandarizado	-0,573	0,058	96,729	1	0,000	0,564	0,503	0,632
	Gasto estandarizado	-0,516	0,060	73,214	1	0,000	0,597	0,530	0,672
	PORCENTAJE DE OCUPADOS POR HOGAR	0,006	0,002	11,304	1	0,001	1,006	1,003	1,010
	[Edad recodificada dos=1]	0,123	0,105	1,373	1	0,241	1,131	0,921	1,388
	[Edad recodificada dos=2]	0,188	0,105	3,200	1	0,074	1,207	0,982	1,482
	[Edad recodificada dos=3]	0[b]			0				
	[NIVEL DE EDUCACIÓN GENERAL=1,00]	0,780	0,175	19,960	1	0,000	2,182	1,550	3,073
	[NIVEL DE EDUCACIÓN GENERAL=2,00]	0,232	0,111	4,342	1	0,037	1,261	1,014	1,000
	[NIVEL DE EDUCACIÓN GENERAL=3,00]	0[b]			0				
	[CODIGO SEGURIDAD ALIMENTARIA=1]	0,809	0,180	20,124	1	0,000	2,246	1,577	
	[CODIGO SEGURIDAD ALIMENTARIA=2]	0,302	0,091	10,941	1	0,001	1,352	1,131	
	[CODIGO SEGURIDAD ALIMENTARIA=3]	0[b]			0				

[a] La categoría de referencia es: Alto.
[b] El parámetro esta establecido en cero por que es redundante.

Fuente: Modelo logístico multinomial SPSS.

- $\exp(\beta) = \exp(2{,}038) = 7{,}673$

Para el primer modelo, significa que pertenecer a una condición socioeconómica baja frente a una condición socioeconómica alta, de personas que no tienen ninguna educación, están en 7,7 veces peor de los

que tienen educación básica. El valor real se encuentra en el intervalo de confianza al 95% entre (4,9; 12,1). Asimismo, se tiene que

- exp(β) = exp (1,829) = 6,229

para el primer modelo, significa que pertenecer a una condición socioeconómica baja frente a una condición socioeconómica alta, de personas que tienen inseguridad alimentaria severa y moderada, están en 6,2 veces peor de los que tienen inseguridad alimentaria leve. Otro caso, se tiene

- exp (β) = exp (0,78) = 2,182

para el segundo modelo, que pertenecer a una condición socioeconómica media frente a una condición socioeconómica alta, de personas que no tienen ninguna educación, están en 2,2 veces peor de los que tienen educación básica. El valor real se encuentra en el intervalo de confianza al 95% entre (1,6; 3,1). Finalmente

- exp (β) = exp (0,809) = 2,246

para el segundo modelo, significa que pertenecer a una condición socioeconómica media frente a una condición socioeconómica alta, de personas que tienen inseguridad alimentaria severa y moderada, están en 2,2 veces peor de los que tienen inseguridad alimentaria leve.

Así se pueden interpretar los demás coeficientes exp (β) con la compañía de sus intervalos de confianza.

Test de hipótesis de β individual

La prueba individual de los coeficientes β, se realiza en base al estadístico de Wald como se presenta en la Tabla No. 2, definido como

$$H_0: \beta_{jc} = 0 \text{ vs } H_1: \beta_{jc} \neq 0 \quad j = 1, 2, \ldots, 9 \quad c = 1, 2$$

Se supone distribución Normal en β

$$Z = \frac{\hat{\beta}_{jc} - \beta_{jc}}{\hat{\sigma}_{\hat{\beta}_{jc}}} \approx N(0,1)$$

El estadístico de contraste de hipótesis, denominado de Wald, es

$$W = Z^2 \approx \chi^2_{1-\frac{\alpha}{2}}$$

Se rechaza H_0 si $W > \chi_1^2$. Es decir, si el estadístico de Wald es superior al valor chi-cuadrado con 1 grado de libertad.

Los coeficientes estimados de β a nivel de variables con categoría ***Baja*** de la variable dependiente Socioeconómica (Y), son significativos excepto el coeficiente $\hat{\beta}_{6,1}$ ***Edad recodificada = 1*** no es significativo y

Tabla No. 3
Test conjunto de los coeficientes estimados de β

Información de ajuste de los modelos

	Criterios de ajuste de modelo			Pruebas de la razón de verosimilitud		
Modelo	AIC	BIC	Logarítmo de la verosimilitud - 2	Chi-cuadrado	gl	Sig.
Sólo intersección	8202,238	8214,809	8198, 238			
Final	7207,584	7333,289	7167,584	1030,654	18	,000

Fuente: Modelo logístico multinomial SPSS.

las categorías de las variables en el caso **0**^{b}, para el modelo desagregado denominado categoría ***Baja***.

Para la cualidad de *Y* ***Media***, se observa que las categorías de los coeficientes $\hat{\beta}_{0,2}$ de Intersección y $\hat{\beta}_{6,2}$ ***Edad recodificada = 1*** no son significativos con Sig por encima de 0,05 y los identificados con **0**^{b} de la Tabla No. 2.

La Tabla No. 3, proporciona información sobre el ajuste de los modelos en forma global mediante la hipótesis de que los coeficientes de β sean significativos de manera conjunta, es decir:

$$H_0: \beta_1 = \beta_2 = ... = \beta_p = 0$$

Los modelos estimados, son:

$$H_1: \beta_1 \neq \beta_2 \neq ... \neq \beta_p \neq 0$$

El nivel de significación p´ valor del test para H_0 vendrá dado por la siguiente probabilidad

$$p' = P\left(\chi^2_{(p-1)(k-1)} > \Lambda_0 - \Lambda_1\right)$$

se rechaza H_0 si p'<0,05.

La Tabla No. 3 en la columna ***Pruebas de la razón de verosimilitud*** demuestra que el valor del estadístico chi cuadrado es alto con p^' (Sig) tendiente a cero, por lo cual los modelos de las cualidades Baja y Media son altamente significativas de manera conjunta. El criterio de Akaike (AIC) justifica la medición relativa de calidad de ajuste de los modelos empleados.

- **Cualidad Baja modelo 1**

  Según lo señalado, la razón entre la probabilidad $\hat{p}_1$ y $\hat{p}_3$ se define como

$$r_1 = \frac{\hat{p}_1}{\hat{p}_3} = e^{X\hat{\beta}_1}$$

$$r_1 = e^{\hat{\beta}_{0,1}+\hat{\beta}_{1,1}Z_1+\hat{\beta}_{2,1}Z_2+\hat{\beta}_{3,1}X_3+\hat{\beta}_{6,1}X_6+\hat{\beta}_{7,1}X_7+\hat{\beta}_{9,1}X_9+\hat{\beta}_{10,1}X_{10}+\hat{\beta}_{12,1}X_{12}+\hat{\beta}_{13,1}X_{13}}$$
$$= e^{-2,854-1,1Z_1-1,326Z_2+0,008X_3+0,171X_6+0,652X_7+2,038X_9+1,035X_{10}+1,829X_{12}+0,78X_{13}}$$

- **Cualidad Media modelo 2**

  De igual manera la razón entre la probabilidad $\hat{p}_2$ y $\hat{p}_3$ es

$$r_2 = \frac{\hat{p}_2}{\hat{p}_3} = e^{X\hat{\beta}_2}$$

$$r_2 = e^{\hat{\beta}_{0,2}+\hat{\beta}_{1,2}Z_1+\hat{\beta}_{2,2}Z_2+\hat{\beta}_{3,2}X_3+\hat{\beta}_{6,2}X_6+\hat{\beta}_{7,2}X_7+\hat{\beta}_{9,2}X_9+\hat{\beta}_{10,2}X_{10}+\hat{\beta}_{12,2}X_{12}+\hat{\beta}_{13,2}X_{13}}$$
$$= e^{-0,236-0,573Z_1-0,516Z_2+0,006X_3+0,123X_6+1,88X_7+0,78X_9+0,232X_{10}+0,809X_{12}+0,302X_{13}}$$

Donde las probabilidades de ocurrencia de clasificación cualidad baja, media y alta se determinan mediante las siguientes relaciones:

$$\hat{p}_1 = \frac{r_1}{1 + r_1 + r_2}$$

$$\hat{p}_2 = \frac{r_2}{1 + r_1 + r_2}$$

y la probabilidad de $\hat{p}_3$ se determina por el complemento de las anteriores, es decir

$$\hat{p}_3 = 1 - \hat{p}_1 - \hat{p}_2$$

tal que $\hat{p}_1 + \hat{p}_2 + \hat{p}_3 = 1$.

Clasificación observada y esperada

La Tabla No. 4 muestra la clasificación de individuos que fueron observados en las categorías condición socioeconómica alta, media y baja y los valores esperados por el modelo logístico multinomial, es decir:

Tabla No. 4
Clasificación observada y esperada

Clasificación				
	Pronosticado			
Observado	Bajo	Medio	Alto	Porcentaje correcto
Bajo	191	654	24	22,0%
Medio	147	1574	233	80,6%
Alto	14	774	354	31,0%
Porcentaje global	8,9%	75,7%	15,4%	53,4%

Fuente: Modelo logístico multinomial SPSS.

De acuerdo a la Tabla N° 4, el 53,4% de las personas clasifican tanto en observación y pronostico por el modelo en las categorías baja, media y alta de condición socioeconómica. El 80,6% de las personas están en la categoría media observada y el 75,7% en la misma categoría de acuerdo al modelo. Hay un 46,6% de la población que no son coincidentes en la observación y proyección del modelo, sin embargo, no dejan de aproximarse a los resultados en la mayoría de los casos en un 99%.

RESULTADOS Y DISCUSIÓN

Al no contar con una variable categórica observada dependiente para el modelo logístico, denominada condición socioeconómica en los niveles alto, medio y bajo, se ha procedido a su determinación a partir del análisis *cluster* no jerárquico en base a las variables socioeconómicas: materiales de construcción de la vivienda, hacinamiento, agua y saneamiento básico, y otras, que son parte de las que componen el índice de necesidades básicas insatisfechas para la medición de pobreza estructural. Se puede notar luego, que al modelo logístico se le incorpora variables independientes que tienen relación con las condiciones de vida y pobreza, como: la ocupación, nivel de educación, seguridad alimentaria y otras que son de característica coyuntural como el ingreso y gasto. Al respecto, existe una predicción del modelo en un 53% en las categorías socioeconómicas de alta, media y baja para la población de la ciudad de El Alto con respecto a lo que presenta la variable observada. Sin embargo, esto no basta pues también el modelo proporciona la medición de la probabilidad de pertenecer a dichas categorías.

RECOMENDACIÓN

A lo referido anteriormente, es fundamental ampliar el análisis no solo a la categorización que pertenece la población en condición alta, media y baja; sino agrupar, además, cada condición, por intervalos de medición de probabilidad dadas por el modelo logístico e investigar a profundidad, cuál o cuáles de las variables influyen más a la determinación de la condición socioeconómica de la persona.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (1990). Categorical Data Analysis. John Wiley and Sons. New York.

Arguello, O. CELADE. Desarrollo Económico, Políticas Sociales y Población. CELADE.

Arguello, O. (1980). Pobreza y Desarrollo. Características Socio-Demográficas de las Familias Pobres en Venezuela. Santiago de Chile. Centro Latinoamericano de Demografía.

Arias, O, S., Bendini, M. (2006). Evaluación de la pobreza en Bolivia: Sentando las bases para un crecimiento a favor de los pobres. Programa Operacional y Analítico de la Región de América Latina y el Caribe, Banco Mundial.

Ayaviri, N, D., Alarcón, L, S. (2014). Clasificación socioeconómica de los municipios de Bolivia. Universidad Católica Boliviana "San Pablo", Unidad Académica Regional Cochabamba.

Banco de Desarrollo de America Latina (CAF). (2017). Agua y saneamiento en el Estado Plurinacional de Bolivia. Buenos Aires.

CELADE. (2002). Vulnerabilidad sociodemográfica: viejos y nuevos riesgos para comunidades, hogares y personas. Brasilia. CEPAL.

CELADE. (2005). Dinámica demográfica y desarrollo en América Latina y el Caribe. Santiago de Chile. CEPAL.

Cox, D. R. & Snell, E. J. (1989). The Analysis of Binary Data. Chapman and Hall. London.

De la Fuente, F, S. (2011), Análisis de conglomerados. Universidad Autónoma de Madrid (UAM).

UDAPE (2019). Dossier de Estadísticas Económicas y Sociales. Vol. 29. La Paz.

INE Bolivia (2020). El Alto en cifras.

Estado Plurinacional de Bolivia, Ministerio de Educación. (2010). Ley de la Educación "Avelino Siñani-Elizardo Pérez".

Gobierno Autónomo Municipal de El Alto. Secretaría Municipal de Desarrollo Económico del municipio de El Alto (2018). Fortalecimiento a las iniciativas económicas: Localización Distrital. El Alto.

Hair, J, F., Anderson, R, E., Tatham, R, L., Black, W,C. (1999). Análisis Multivariante. Madrid, España. Editorial Prentice Hall.

Hosmer, D. W. & Lemeshow, S. (1989). Applied Logistic Regression. Wiley Interscience. New York.

Informe Estadístico del Municipio de El Alto (2020). Ministerio de Desarrollo Productivo y Económico Plural.

Johnson, R. A. Applied Multivariate Statistical Analysis. Prentice Hall. New Jersey.

López, J.L. (2011). La Ecuación Diferencial Logística.

Medina, M. E. (2003). Modelos de elección discreta.

Menard, S. (2000). Coefficients of

Determination for Multiple Logistic Regression Analysis. The American Statistician.

Osorio, D. (2009). Planteamiento del Modelo Logístico Multinomial a travéz de la Función Canónica de Enlace de la Familia Exponencial.

Peña, Daniel (2002). Análisis de Datos Multivariantes. Mc Graw Hill.

Quispe, G.M. (2009). La formación de la ciudad de El Alto y sus consecuencias. Tesis Doctoral, Universidad Autónoma de Madrid.

UDAPE: Unidad de Análisis de Políticas Sociales y Económicas, INE: Instituto Nacional de Estadística. (2018). Migración interna en Bolivia. Organización Internacional para las Migraciones (OIM).

Uriel, E., Aldás, J. (2005). Análisis multivariante aplicado. Madrid, España. Editorial Thomson.

Villaroel, P., Hernani-Limarino, W. (2013). La evolución de la pobreza en Bolivia: un enfoque multidimensional. Revista Latinoamericana de Desarrollo Económico (LAJED).

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 63-70

UN VISTAZO A LA INFERENCIA BAYESIANA

A QUICK LOOK AT BAYESIAN INFERENCE

Lizbeth Román Padilla[1]

Facultad de Ciencias Actuariales
Universidad Anáhuac, Ciudad de México, México
✉ *lizroman@hotmail.com*

Artículo recibido: 2021-07-30
Artículo aceptado: 2021-09-07

RESUMEN

El enfoque Bayesiano de la estadística debe considerarse como una alternativa adicional al enfoque clásico, siendo ambos enfoques complementarios más no excluyentes. La estadística Bayesiana ofrece una gran variedad de métodos estadísticos similares en número a los proporcionados por el enfoque clásico.

La estadística Bayesiana debe su nombre al uso repetido del Teorema de Bayes: la distribución final o posterior es el resultado de aplicar el Teorema de Bayes a la información que proporcionan los datos (función de *verosimilitud*) y la información previa del parámetro de interés (*distribución inicial*). La distribución *posterior* es idónea para hacer *cualquier* tipo de inferencias sobre el parámetro de interés1, ya sea estimación puntual o por intervalo, pues incluye *toda* la información disponible acerca de θ una vez observados los datos junto con la información *inicial*.

El objetivo de este artículo es la ejemplificación de obtención el estimador puntual Bayesiano y la región creíble de la media (θ) de datos con distribución *Cauchy* (θ,*1*). Para este propósito se usarán los datos de precipitaciones anuales del estado mexicano de Tabasco. Adicionalmente, se utilizan técnicas de simulación de variables aleatorias e integración numérica.

Los resultados obtenidos mediante inferencia Bayesiana permitirán tener una aproximación a la verdadera media de precipitación (θ) desde que el estimador clásico se vuelve inestable conforme incrementa el tamaño de muestra. Con este simple ejercicio se pretende dar a conocer algunas ventajas de aplicar los métodos Bayesianos.

Palabras clave: *Algoritmo aceptación-rechazo; Estimador Bayesiano; Iniciales no Informativas; Inferencia Bayesiana; Regiones HPD; Simulación.*

ABSTRACT

The Bayesian approach to statistics should be considered as an additional alternative to the classical approach, both approaches being complementary but not exclusive. Bayesian statistics offers a great variety of statistical methods similar in number to those provided by the classical approach.

The origin of the term ‘Bayesian Statistics’ is due to the repeated use of the Bayes Theorem: the final or posterior distribution is the result of applying the Bayes Theorem to the information provided by data (likelihood function) and initial information about the parameter of interest (distribution initial).

Posterior distribution is ideal for making any kind of inferences about the parameter of interest, whether it can be a point estimate or by interval, since it includes all the information available about θ after data has been observed together with initial information.

[1] Lizbeth Román Padilla es doctora en Estadística (Estadística Bayesiana Objetiva) por la Universidad de Valencia, España. Maestra en Méts. Matemáticos en Finanzas (Universidad Anáhuac) y Actuaria (Facultad de Ciencias, UNAM). Ha hecho dos posdoctorados (Francia y México) y desde 2013 es docente en los niveles de licenciatura y posgrado. https://orcid.org/0000-0001-9673-4209

The objective of this article is to illustrate how to get a Bayesian point estimator and credible region for the mean (θ) of Cauchy data Cau(θ, 1). For this purpose, annual rainfall data of Tabasco (Mexican state) will be used. Additionally, random variable simulation techniques and numerical integration are employed.

The results obtained through Bayesian inference provides us with an approximation to the true mean of precipitation (θ) since the classical estimator becomes unstable as the sample size increases. This simple exercise is intended to show some advantages of applying Bayesian methods.

Keywords: *Acceptance-Rejection algorithm; Bayesian estimation; Non-Informative Priors; Bayesian Inference; HPD regions; Simulation*

INTRODUCCIÓN

Filosofía bayesiana

Al igual que en los métodos estadísticos clásicos de inferencias, existe otro acercamiento a la inferencia conocido como **inferencia bayesiana**. La cual tiene tres características fundamentales heredadas del enfoque bayesiano de la estadística:

- La inferencia bayesiana, y en general la estadística bayesiana, se fundamenta en la interpretación subjetiva de la probabilidad, es decir la probabilidad describe un *grado de creencia*, y al contrario del enfoque clásico, no se basa en el límite de las frecuencias relativas o en la descripción de problemas físicos (interpretación clásica y frecuentista de la probablidad, respectivamente).
- Bajo el supuesto de que los datos siguen un modelo estadístico indexado por algún parámetro θ (desconocido). En el enfoque bayesiano, los *parámetros son tratados como variables aleatorias*, es decir, se les atribuye un grado de *incertidumbre* y por tanto se les puede asociar una distribución de probabilidad.
- Y, finalmente, al obtener una distribución de probabilidad del parámetro de interés, se pueden hacer inferencias acerca de su verdadero valor o sobre cualquiera de sus propiedades.

Método bayesiano

Suponga que $\boldsymbol{x}' = (x_1, x_2, ..., x_n)$ es un vector de n observaciones cuya distribución de probabilidad $p(\boldsymbol{x}|\theta)$ que depende de k parámetros $\boldsymbol{\theta}' = (\theta_1, \theta_2, ..., \theta_k)$. Suponga, además que $\boldsymbol{\theta}'$ cuenta por sí misma con una función de distribución $\pi(\boldsymbol{\theta}')$. Entonces,

$$p(\mathbf{x}|\theta)\pi(\theta) = p(\boldsymbol{x}|\theta') = \pi(\theta|\boldsymbol{x})\,\mathrm{p}(\boldsymbol{x}). \quad (1)$$

Entonces, dados los datos observados $\boldsymbol{x}$, la distribución condicional de $\boldsymbol{\theta}$ es

$$\boldsymbol{\pi}(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\boldsymbol{\pi}(\boldsymbol{\theta})}{p(\mathbf{x})} \quad (2)$$

y $p(\mathbf{x})$ puede ser

$$p(\mathbf{x}) = \begin{cases} \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} & \boldsymbol{\theta}\ continua \\ \sum p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) & \boldsymbol{\theta}\ discreta. \end{cases}$$

donde la suma o la integral toma todos los posibles valores de θ.

La ec. (2) puede escribirse de manera equivalente como

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (3)$$

donde el símbolo "$\propto$" se refiere a que el lado derecho de la ecuación es aproximado, salvo la constante que normaliza a $\boldsymbol{\pi}(\boldsymbol{\theta} \mid \boldsymbol{x})$ para que el área bajo la curva (o la suma) sea uno. A la ec. (2) se le conoce como el

Teorema de Bayes y sus elementos son:

- La función se le llama **distribución inicial** o *prior*, y nos dice *todo lo que se sabe acerca* de θ antes de haber observado los datos.
- La función $\boldsymbol{\pi}(\boldsymbol{\theta} \mid \boldsymbol{x})$ es la distribución **posterior** o **final**, y es la que proporciona información acerca de θ dado el *conocimiento* de los datos.
- Dados los datos, la función $p(\mathbf{x} \mid \theta)$ puede verse como función solamente de θ, y se le llama función de **verosimilitud** de $\boldsymbol{\theta}$ dado $\boldsymbol{x}$, la cual se puede escribir como $l(\boldsymbol{\theta} \mid \boldsymbol{x})$.

Es decir, el Teorema de Bayes asegura que la distribución de probabilidad posterior de $\boldsymbol{\theta}$ dado $\boldsymbol{x}$ es *proporcional* al producto de la verosimilitud y la inicial:

$$\text{posterior} \propto \text{verosimilitud} \times \text{inicial}.$$

Esta expresión permite formular matemáticamente cómo la información inicial puede combinarse con la información que proveen los datos. La función de verosimilitud es la que *a través de ella los datos* $\boldsymbol{x}$ *modifican el conocimiento previo de* $\boldsymbol{\theta}$, es decir, es la información de $\boldsymbol{\theta}$ que proporciona los datos. Por tanto, juega un papel importante dentro del Teorema de Bayes.

El Teorema de Bayes describe el proceso de aprendizaje a través de la experiencia. (Box y Tiao, 1992, Secc. 1.2.2).

Inferencia bayesiana

La inferencia bayesiana hace uso de la *Teoría de la Decisión*, pues al elegir un estimador dentro de un conjunto de posibles estimadores del parámetro $\boldsymbol{\theta}$ (o de alguna función de este), a esa elección se le asocia una *ganancia* o *pérdida* que a su vez dependerá del *estado de la naturaleza*, es decir, el verdadero valor del parámetro $\boldsymbol{\theta}$. La distribución de probabilidad posterior refleja el conocimiento del tomador de desiciones, pues es el resultado de combinar la información inicial junto con la información que proveen los datos acerca del parámetro. Por tanto, se espera que el tomador de desiciones elija la acción que maximice (minimice) su beneficio.

La inferencia bayesiana proporciona una forma satisfactoria de introducir explícitamente y mantener los supuestos acerca del conocimiento previo o de la ignorancia. (Box y Tiao, 1992).

Estimadores bayesianos

Si uno desea hacer inferencia bayesiana puntual, comenzará por definir la *función de pérdida* y deberá encontrar el *estimador bayesiano* que la minimice (una mejor aproximación a la Teoría de Decisión bayesiana puede encontrarla en Mood et al., 1974 y Box and Tiao, 1992).

El estimador bayesiano más común es la esperanza de la distribución posterior $\boldsymbol{\pi(\theta|x)}$ pues es el que minimiza la pérdida esperada cuadrática. Sin embargo, el estimador bayesiano dependerá de la función de pérdida utilizada.

Conjuntos (creíbles) HPD

Otra forma común de inferencia es presentar intervalos de confianza para $\boldsymbol{\theta}$. El análogo bayesiano a los intervalos de confianza clásicos son los llamados *conjuntos creíbles* $100(1 - \alpha)$ % para $\boldsymbol{\theta}$. Un *conjunto creíble* es un subconjunto $C \in \Theta$ tal que $P(C \mid x) \geq 1 - \alpha$ (Def. 4 en Berger, 2010). Note que se está trabajando directamente con la distribución posterior $\boldsymbol{\pi(\theta|x)}$, por tanto, tiene sentido interpretarlo como *la probabilidad*

(posterior) de que θ se encuentre en C. Recuerde que los intervalos de confianza se interpretan en función de la cobertura de probabilidad, véase las secciones 1.6 y 4.1 de (Berger, 2010).

De los posibles conjuntos creíbles para θ se elige el conjunto con volumen más pequeño tal que contenga a los valores *de θ más probables*. Los intervalos **HPD** (siglas en inglés, *highest posterior density*) son *conjuntos creíbles* 100(1 - α) % de θ, $C \epsilon \Theta$, de la forma $C=\{\theta \epsilon \Theta : \pi(\theta \mid x) \geq k(\alpha)\}$ siendo $k(\alpha)$ la constante más grande que cumpla con que $P(C \mid \boldsymbol{x}) \geq 1 - \alpha$. Los conjuntos creíbles generalmente son fáciles de calcular y algunas veces on la única alternativa a sus contrapartes clásicas (Berger, 2010).

Inferencia sobre θ de *Cauchy* (θ,*1*)

Distribución Cauchy

$$p(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, \quad -\infty < x < \infty.$$

Consideraremos el problema de hacer inferencias acerca del parámetro de localización θ de la distribución Cauchy.

La distribución Cauchy pertenece a la familia de distribuciones de localización y escala, además de que carece de media y varianza asímismo no cuenta con función generadora de momentos. Las estimaciones muestrales de la media y la varianza crecen conforme se incrementa el tamaño de muestra y se vuelven inestables (Wikipedia contributors 2021, July 6) y el método de máxima verosimilitud implica encontrar las raíces de polinomios de grado mayor, donde las raíces pueden ser máximos locales pero no necesariamente globales. Otra característica de la distribución Cauchy es que pertenece a la familia de distribuciones estables. Una distribución de probabilidad se dice estable si una combinación lineal de dos variables aleatorias independientes de esta distribución tiene la misma distribución, salvo algún parámetro de localización o de escala, véase la Def. 16.20 en (Klenke, 2014).

La familia de distribuciones estables son adecuadas para modelar datos con colas pesadas y sesgadas (Ball *et al*., 2021) en hidrología. Específicamente, la distribución Cauchy se utiliza para modelar eventos extremos, tales como el máximo anual de caida de lluvia en un dia (Wikipedia contributors. 2021, July 6).

Inferencia bayesiana: Cauchy

Suponga que se tiene una muestra aleatoria $X_1, X_2, ..., X_n$ provenientes de una distribución Cauchy con parámetro θ desconocido y varianza conocida e igual a uno, $X \sim Ca(\theta, 1)$. El objetivo es hacer inferencias acerca del parámetro de localización θ. Se utilizará una distribución inicial **no informativa** $\pi(\theta) \propto 1$ definida en el espacio restringido de θ > 0 desde que θ es un parámetro de localización. Un resumen completo sobre la selección de las distribuciones iniciales puede verse en (Kass, R., y Wasserman, L.,1996). La densidad posterior de θ dado $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ estará dada por

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{\prod_{i=1}^{n} \frac{1}{\pi[1+(x_i-\theta)^2]}}{\int_0^\infty \prod_{i=1}^{n} \frac{1}{\pi[1+(x_i-\theta)^2]} d\theta} \\ &= \frac{\prod_{i=1}^{n} [1+(x_i-\theta)^2]^{-1}}{\int_0^\infty \prod_{i=1}^{n} [1+(x_i-\theta)^2]^{-1} d\theta}, \theta > 0. \quad (4) \end{aligned}$$

Ejemplo. Datos de precipitación

La Tabla No. 1 muestra las precipitaciones anuales registradas durante 36 años en el estado mexicano de Tabasco, el cual se caracteriza por ser uno de los estados

con más precipitaciones en un año. Suponga que las precipitaciones siguen una distribución Cauchy con parámetro de localización desconocido (θ) y varianza conocida e igual a uno ($\sigma^2 = 1$).

Tabla No. 1
Precipitaciones anuales (mm), años 1985-2020, del estado de Tabasco, México.

Año	*(mm)*	*Año*	*(mm)*	*Año*	*(mm)*
1985	*2493.2*	*1997*	*2134.7*	*2009*	*1730.5*
1986	*1733.3*	*1998*	*1956.2*	*2010*	*2561.9*
1987	*2126.6*	*1999*	*2476*	*2011*	*2496.5*
1988	*2468.4*	*2000*	*2511.2*	*2012*	*2069.7*
1989	*2190.2*	*2001*	*2451.4*	*2013*	*2811.9*
1990	*2324.4*	*2002*	*2297.2*	*2014*	*2394.4*
1991	*2180.7*	*2003*	*2076.1*	*2015*	*2426.3*
1992	*1775.1*	*2004*	*1964.1*	*2016*	*1747.7*
1993	*2130.6*	*2005*	*2036*	*2017*	*2013.9*
1994	*2170.9*	*2006*	*2676.2*	*2018*	*1965.5*
1995	*3366.9*	*2007*	*2554.1*	*2019*	*1916*
1996	*1996.2*	*2008*	*2617.3*	*2020*	*3017.9*

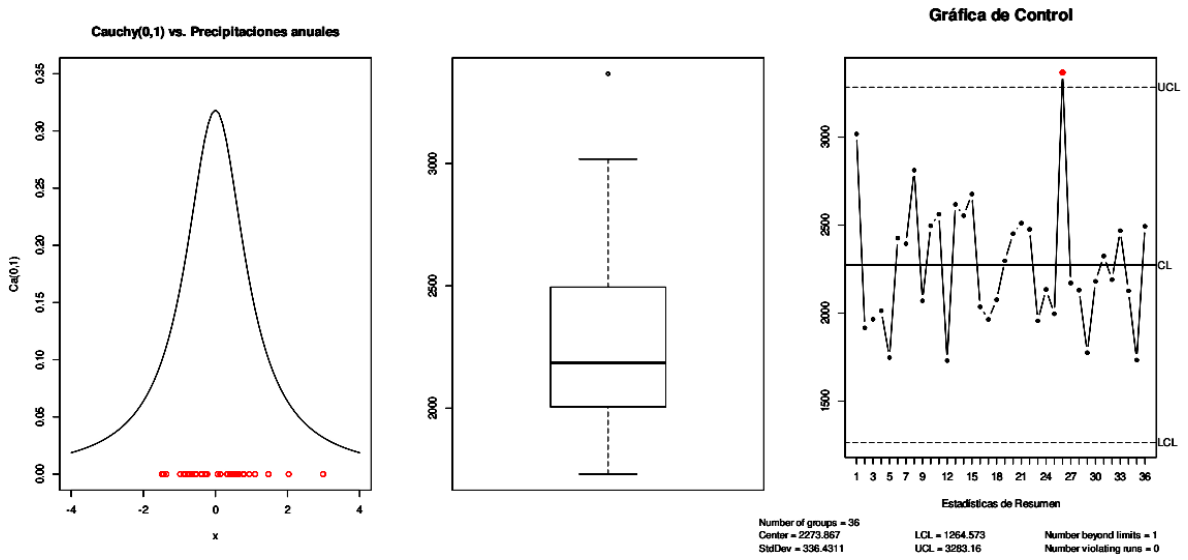
Fuente: https://smn.conagua.gob.mx/es/climatologia/temperaturas-y-lluvias/resumenes-mensuales-de-temperaturas-y-lluvias

En análisis descriptivo sobre los datos de precipitación puede verse en la Figura No.1. La gráfica izquierda muestra las precipitaciones normalizadas (*eje X*) *vs*. la densidad Cauchy(0,1). La gráfica de *caja y bigote* (centro) muestra una distribución ligeramente sesgada a la derecha e identifica una observación aberrante. Finalmente, la gráfica de control (derecha) muestra claramenta al *outlier* que rebasa tres desviaciones estándares mientras que el resto de las observaciones no rebasan dicha franja.

Aproximación numérica de $\pi(\theta \mid x)$

Mediante aproximación numérica (Narasimhan, B. (n.d.)) se obtiene una estimación de la constante de integración (denominador de la ec. (4)) y por tanto una aproximación a la distribución posterior $\pi(\theta|\boldsymbol{x})$, véase la Figura No. 2 (línea roja). Mediante el algoritmo de *aceptación-rechazo* (véase secc. 4.4 de Ross, 1999) se generaron $m = 10{,}000$ variables aleatorias de una distribución de cobertura y se *aceptaron* $n = 7782$ variables provenientes de la distribución $\pi(\theta \mid x)$, $\{\theta_1, \theta_2, ..., \theta_{7782}\}$, histograma de la

Figura No. 1
Análisis descriptivo de las precipitaciones anuales. Densidad Cauchy(0,1) vs. precipitaciones normalizadas (der.), gráfica de caja y bigote de las precipitaciones anuales (centro) y gráfica de control con resumen de estadísticas (der.)



Fuente: Elaboración propia

Figura No. 2. La región *HPD* al 95% se obtiene a partir de los valores simulados $\theta_i (i = 1,... 7782)$, al igual que se obtienen los cuantiles $q_{0.025}$ y $q_{0.975}$, (véase la función *hdi* dentro del paquete *HDInterval* del lenguaje de programación R, M., M., & J., K.).

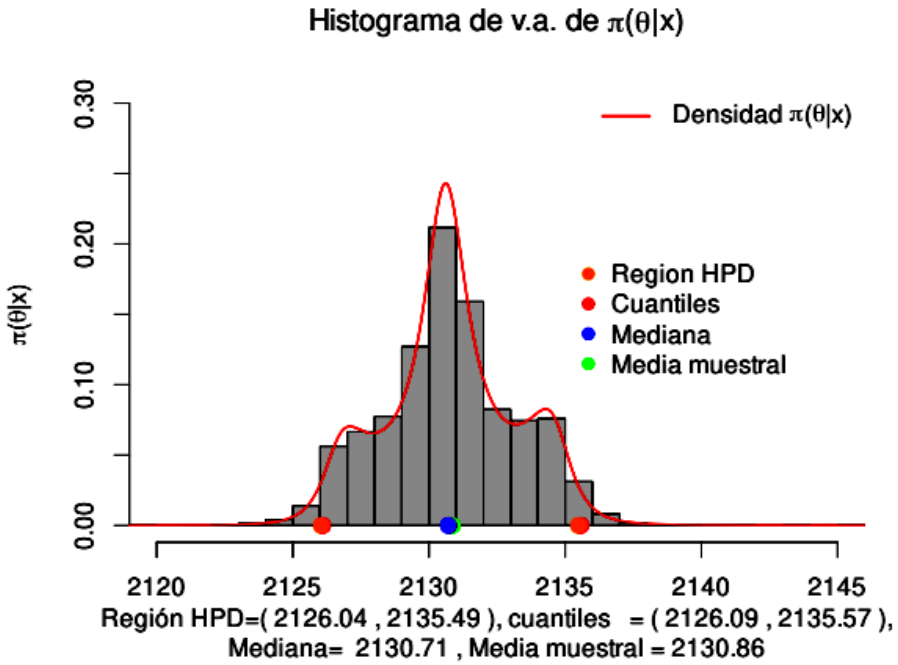
El estimador bayesiano puntual de θ puede ser la mediana o la media muestral posteriores, $\theta_{mediana}$ = 2130.71 y θ_{media} = 2130.86, respectivamente, ya que ambos son muy similares. Por otro lado, el intervalo *HPD* al 95% es (2126.09 , 2135.45) que difiere muy poco de los cuantiles $(q_{0.025}, q_{0.975})$ = (2126.09 , 2135.45). Sin embargo, el intervalo *HPD contendrá con una probabilidad del 95% al verdadero valor θ una vez que se han observado los datos*.

CONCLUSIONES

La inferencia bayesiana es una herramienta útil cuando se desea saber el verdadero valor del parámetro θ asociado a alguna distribución de probabilidad que siguen algunos datos. La idea principal es que la distribución posterior $\pi(\theta \mid \boldsymbol{x})$ contiene toda la información disponible sobre θ, información que proviene tanto de los datos como de la información inicial del parámetro. Por tanto, cualquier inferencia sobre θ puede obtenerse de $\pi(\theta \mid \boldsymbol{x})$. Incluso cuando se requiera estimar puntualmente o por intervalos.

El uso más simple de un proceso inferencial de la distribución posterior es reportar un estimador puntual para θ, la elección del estimador puntual se asocia a una función de pérdida, pues el proceso de selección

Figura No. 2
Histograma θi ~ π(θ|x) (*i = 1, ..., 7782*) obtenidas mediante simulación, la líneas roja continua corresponde a la distribución posterior π(θ|x) (ec. 4). En el eje x pueden verse la similitud entre las regiones HPD 95% y los cuantiles (q0.025, q0.975), al igual que los valores de la mediana y la media muestral.



Fuente: Elaboración propia

incurre en un pérdida (o ganancia) por haber obrado de tal manera. La media y la mediana de $\pi(\theta \mid \boldsymbol{x})$ son los *estimadores bayesianos* de las funciones de pérdida cuadrática y pérdida absoluta, respectivamente. En caso de requerir estimar un intervalo, los *intervalos* (*creíbles*) *HPD* son los equivalentes a los intervalos clásicos de confianza, salvo que difieren en su interpetación, pues las regiones creíbles contienen al valor de θ con una probabilidad posterior 1 - α, muy distinto a la interpretación frecuentista de su contraparte clásica.

Los resultados obtenidos mediante inferencia bayesiana permitien tener una aproximación a la verdadera media de precipitación (θ) desde que los parámetros de la distribución Cauchy no se corresponden a la media y varianza, *Wikipedia contributors*. (2021, July 6). Por tanto, no se pueden estimar mediante la media y varianza muestrales.

Finalmente, con este simple ejercicio se pretende dar a conocer algunas ventajas de aplicar los métodos bayesianos.

DISCUSIÓN

La inferencia bayesiana es una respuesta a las inferencia clásica cuando ésta no proporciona de una respuesta adecuada. Debido a que los métodos bayesianos funcionan de la misma forma; se determina una distribución inicial que proporcione la información del parámetro antes de observar los datos, el Teorema de Bayes permite construir la distribución (de probabilidad) final a partir de la información proporcionada por los datos y la distribución inicial. En este artículo, encontramos una solución al problema de inferir el verdadero valor del parámetro θ de una distribución $\mathcal{C}a(\theta, 1)$ cuando se trabaja desde el enfoque clásico, y como se mostró, la inferencia bayesiana proporcionó estimaciones puntuales y por intervalo.

En general, los métodos bayesianos son una alternativa válida a algunas deficiencias de los métodos clásicos (Gómez-Villegas, 2006). Sin embargo, ambos enfoques no son excluyentes, deben ser complementarios.

REFERENCIAS BIBLIOGRÁFICAS

Ball, C., Rimal, B., y Chhetri, S. (2021). A new generalized cauchy distribution with an application to annual one day maximum rainfall data. Statistics, Optimization and Information Com- puting, 9, pp.123–136. https://doi.org/10.19139/soic-2310-5070-1000.

Berger, J. (2010). Statistical decision theory and bayesian analysis (2nd). Springer-Verlag: New York, EEUU.

Box, G., y Tiao, G. (1992). Bayesian inference in statistical analysis. Wiley-Interscience.

Kass, R., y Wasserman, L. (1996). The selection of prior distributions by formal rules. Journal of the American Statistical Association, 91(435), pp. 1343–1340. https://doi.org/10.2307/2291752.

Klenke, A. (2014). Probability theory. A comprehensive course. (2nd). Springer-Verlag: London, UK.

Gómez-Villegas, M.A. (2006). ¿Por qué la inferencia estadística bayesiana? Boletín de la Sociedad de Estadística e Investigación Operativa, 22, 1, pp. 6-8.

M., M., y J., K. (n.d.). Highest (posterior) density intervals. Retrieved July 20,

2021, from https:// cran.r-project.org/ web/packages/HDInterval/ HDInterval. pdf.

Mood, A., Graybill, F., y Boes, D. (1974). Introduction to the theory of statistics (3rd). McGraw- Hill.

Narasimhan, B. (n.d.). Adaptive multivariate integration over hypercubes. Retrieved July 19, 2021, from https://bnaras.github.io/cubature/.

Ross, S. (1999). Simulación (2a.). Prentice Hall: México.

Wasserman, L. (2004). All of statistics. A concise course in statistical inference. Springer Science+Business Media: New York, EEUU.

Wikipedia contributors. (2021, July 6). Cauchy distribution. In Wikipedia, The Free Encyclopedia. Retrieved 00:23, August 31, 2021, from https://en.wikipedia.org/w/index.php?title=Cauchy_distribution&oldid=1032217044.

Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 71-75

LA LEY DE BENFORD Y LOS DATOS DEL COVID-19 EN BOLIVIA

BENFORD'S LAW AND COVID-19 DATA IN BOLIVIA

Dindo Valdez Blanco[1]
Instituto de Estadística Teórica y Aplicada, Universidad Mayor de San Andrés, La Paz -Bolivia

✉ *dvaldez@fcpn.edu.bo*

Artículo recibido: 2021-07-30
Artículo aceptado: 2021-09-11

RESUMEN

En la actualidad, con los datos de la pandemia COVID-19, existe la duda en creer que los estados o gobiernos estén informando datos confiables y precisos. En Bolivia, en particular, ocurre lo mismo, en vista que el sistema de salud es precario, se duda de la información reportada hasta el día de hoy, tanto en nuevos casos diarios, casos diarios recuperados y fallecidos, así como en los datos acumulados. Por lo tanto, el objetivo del trabajo de investigación radica en docimar si estos conjuntos de datos informados por el sistema de salud del país siguen la ley de Benford. La metodología del trabajo se basa en el procedimiento de un estudio de investigación de bondad de ajuste pues abarca el uso de dos pruebas de bondad de ajuste denominadas el test Chi cuadrado de Bondad de Ajuste y el test de bondad de ajuste de Kuiper. Los datos recopilados provienen de los reportes diarios del Ministerio de Salud del Estado Plurinacional de Bolivia entre el 1 de abril del 2020 y el 14 de julio del 2021. Para determinar si el primer dígito significativo del número diario de casos confirmados con COVID-19 en Bolivia se adecúa a la ley de probabilidad de Benford se realizan las pruebas de bondad de ajuste Chi cuadrado de Pearson y la prueba de Kuiper, en ambos casos se rechaza la hipótesis que los datos se ajustan a la ley de Benford, la diferencia significativa más grande es con el dígito 1, este hecho sugiere que existe una subestimación en los reportes diarios de casos confirmados.

Palabras clave: *Bondad de ajuste, Prueba Chi Cuadrado, Prueba de Kuiper, Análisis de datos.*

ABSTRACT

At present, with the data from the COVID-19 pandemic, there is doubt in believing that states or governments are reporting reliable and accurate data. In Bolivia, in particular, the same happens, given that the health system is precarious, the information reported to date is doubted, both in new daily cases, daily cases recovered and deaths, as well as in accumulated data. Therefore, the objective of the research work is to document whether these data sets reported by the country's health system follow Benford's law. The work methodology is based on the procedure of a goodness-of-fit research study, since it involves the use of two goodness-of-fit tests called the Chi-square Goodness-of-Fit test and the Kuiper goodness-of-fit test. The data collected comes from the daily reports of the Ministry of Health of the Plurinational State of Bolivia between April 1, 2020 and July 14, 2021. To determine if the first significant digit of the daily number of confirmed COVID-19 cases in Bolivia conforms to Benford's law of probability, Pearson's Chi-square goodness-of-fit tests and Kuiper's test are performed, in both cases the hypothesis that the data conform to Benford's law is rejected, the difference being more significant large is with the digit 1, this fact suggests that there is an underestimation in the daily reports of confirmed cases.

Keywords: *Goodness of Fit, Chi Square Test, Kuiper Test, Data analysis.*

[1] Maestría en Ciencias Estadísticas, Licenciado en Estadística. Profesor de Estadística - Universidad Mayor de San Andrés.
ORCID: 0000-0003-0704-0980

INTRODUCCIÓN

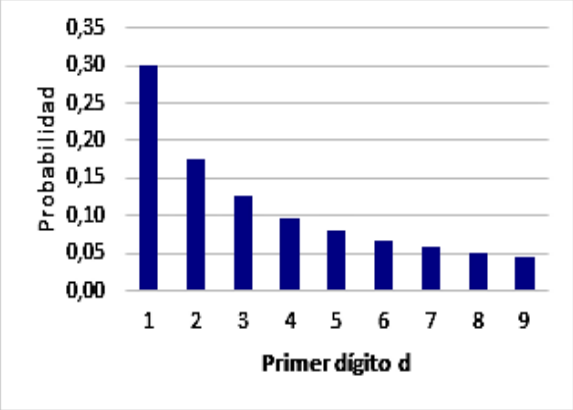
La ley de Benford establece que en muchos conjuntos de datos numéricos que ocurren naturalmente, el primer dígito de los datos no tiene la misma probabilidad de ser 1,..., 9, como cabría esperar, sino que se aproxima bastante por la ley logarítmica:

$$P(d) = \log\left(1 + \frac{1}{d}\right) \quad ; \quad d = 1,2,\ldots,9$$

Fuente: Newcomb (1881)

Donde d es el primer dígito del dato numérico y $P(d)$ es la probabilidad que el dato cuantitativo tenga como primer dígito significativo d. La Figura No. 1 muestra los valores de estas probabilidades.

Figura No. 1
Distribucion de probabilidad del primer dígito significativo según la ley de Benford



Fuente: Elaboración propia en base a la distribución de probabilidad logarítmica (Newcomb,1881)

La ley de Benford fue descubierta por primera vez por Simon Newcomb en su trabajo de 1881 en el *American Journal of Mathematics*. Benford (1938) redescubrió la ley en *Proceedings of the American Philosophical Society* y se le atribuyó el mérito. Descubrió que esta ley logarítmica era bastante precisa en muchas circunstancias; por ejemplo, las cantidades declaradas de impuestos, las longitudes de los ríos, los precios de las acciones, las constantes universales en física química, el número de habitantes de las grandes ciudades y muchas otras tablas de datos numéricos.

No todos los conjuntos de datos siguen la ley de Benford. Por ejemplo, los números de teléfono de una ciudad determinada no siguen dicha distribución probabilística porque el código de área es el mismo número.

La ley de Benford también se puede utilizar para detectar la manipulación de los estados financieros. Incluso se puede utilizar para detectar fraudes (impuestos, juegos de azar).

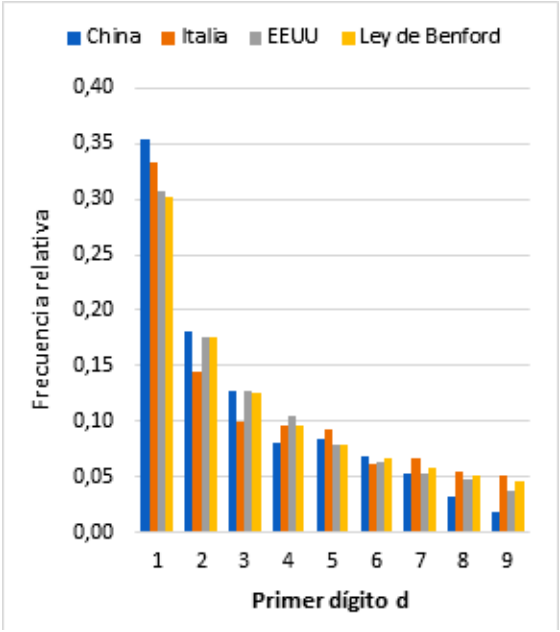
METODOLOGÍA

La metodología del trabajo se basa en el procedimiento de un estudio de investigación de bondad de ajuste pues baraca el uso de dos pruebas de bondad de ajuste denominadas el test Chi-Cuadrado de Bondad de Ajuste y el test de bondad de ajuste de Kuiper. Los datos para el estudio corresponden a los reportes diarios de casos confirmados de COVID-19 en Bolivia realizados por el ministerio de salud y que se encuentran disponibles en el repositorio de datos COVID-19 del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121) de la universidad Jhons Hopkins Situada en Baltimore, Estados Unidos.

La ley de Benford y los datos del COVID-19

Los conjuntos de datos del COVID-19 reportados en Estados Unidos, China e Italia tienden a ajustarse muy bien a la ley de Benford. La Figura No. 2 muestra el gráfico de la distribución del número de casos diarios confirmados en estos países (Koch y Okamura, 2020).

Figura No. 2
Distribución del primer dígito del número diario de casos confirmados con COVID-19 en EEUU, China, Italia.



Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121).probabilidad logarítmica (Newcomb,1881)

Prueba de bondad de ajuste para la ley de Benford

La prueba de bondad de ajuste más común es la prueba Chi-cuadrado de bondad de ajuste:

$$\chi^2 = \sum_{d=1}^{9} \frac{(f_d - e_d)^2}{e_d}$$

Dónde f_d denota la frecuencia observada de los dígitos y e_d es la frecuencia esperada de cada dígito según la ley de Benford.

Prueba de bondad de ajuste de Kuiper

Al aplicar la prueba Chi-cuadrado generalmente no se admite la distribución de Benford, esto ocurre porque la prueba de Chi-cuadrado es una prueba asintótica y tiende a rechazar la significación estadística incluso para pequeñas diferencias. Es por esta razón que es preferible aplicar la prueba de Kuiper definida como:

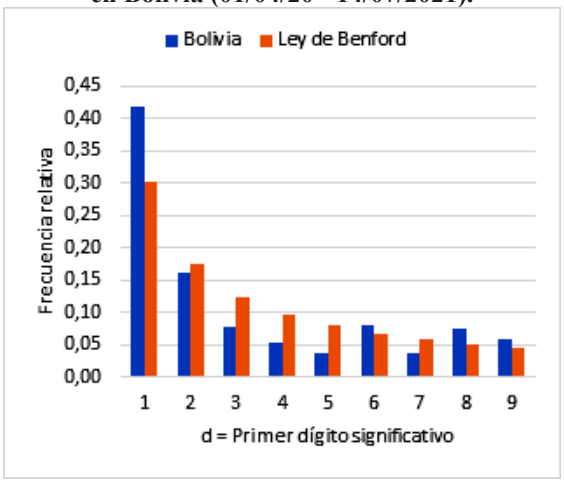
$$T = (D_n^+ + D_n^-)\left[\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right]$$

Donde $D_n^+ = \sup(F_d - E_d)$; $D_n^- = \sup(E_d - F_d)$ con F_d y E_d representan las frecuencias acumuladas para el primer dígito de los datos observados y esperados según la ley de Benford.

Evaluación de la ley de Benford y los datos del COVId-19 reportados en Bolivia

Los datos analizados para Bolivia provienen de los reportes diarios del Ministerio de Salud entre el 1 de abril del 2020 al 14 de julio del 2021.

Figura No. 3
Distribución del primer dígito del número diario de casos confirmados con COVID-19 en Bolivia (01/04/20 - 14/07/2021).



Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE, 2121).probabilidad logarítmica (Newcomb,1881)

Los datos analizados corresponden al número diario de casos confirmados. El cuadro indica la distribución de frecuencias relativas del primer dígito de los datos observados comparados con la ley de Benford.

Se observa un número elevado para la frecuencia relativa del primer dígito 1 en relación a lo que

se esperaría con la ley de Benford.

El Cuadro No.1 muestra los resultados al aplicar las pruebas de bondad de ajuste Chi Cuadrado y Kuiper.

Cuadro No. 1.
Prueba de bondad de ajuste de los casos diarios confirmados de COVID-19 en Bolivia respecto a la ley de Benford

País	*N*	*Test Chi cuadrado*	*Test de Kuiper*
Bolivia	*470*	*61.177**	*3.495**

** La prueba es significativa al 5%*

Fuente: Elaboración propia en base a los datos disponibles en el repositorio del Centro de Ciencias e Ingeniería de Sistemas (CSSE,2021).

En ambos casos se rechaza que los datos del número diario de casos confirmados con COVID-19 en Bolivia correspondan con la ley de Benford, esto es inquietante, en vista que en países como Estados Unidos si se corresponde con dicha ley de probabilidad.

DISCUSIÓN

En el trabajo realizado por Koch y Okamura (2020) se concluye que los datos diarios de casos confirmados por COVID-19 durante la gestión 2020 reportados por el CSSE de la Universidad Jhons Hopkins respecto a los países de Italia, Estados Unidos y China se adecúan a la Ley de Benford al realizar las pruebas de bondad de ajuste Chi cuadrado y Kuiper, sin embargo, en el análisis realizado con los datos reportados en Bolivia por el mismo centro de monitoreo de la Universidad Jhons Hopkins, se concluye que los datos no corresponden a la distribución de probabilidad de Benford. Este hecho puede tener diversos motivos que es necesario indagar en futuras investigaciones.

CONCLUSIÓN

Los datos analizados en Bolivia con respecto a la ley de Benford a primera vista no se corresponden. Esto no debería ser una sorpresa, en vista de la precariedad del sistema de salud del Estado Boliviano. Una cosa a tener en cuenta es que las frecuencias de los dígitos no disminuyen estrictamente a medida que el dígito aumenta para todas estas categorías, por ejemplo, un 6 es más común que un 5 o 4.

Al aplicar las pruebas de bondad de ajuste Chi-cuadrado y Kuiper la diferencia más significativa radica en el dígito 1, esto sugiere que los reportes del Ministerio de Salud pueden tener algún indicio de fraude, sin embargo, a medida que se reportan más y más datos de COVID-19, es posible que los datos se ajusten de mejor manera a la ley de Benford para ver si están reportando datos precisos y completos. Aunque es posible que por el sistema de información que se tiene y los recursos para realizar pruebas a tantas personas, los datos en Bolivia no sean los verdaderos y se está subestimando el reporte diario de casos nuevos confirmados de COVID-19.

REFERENCIAS BIBLIOGRÁFICAS

Balsari, S.; Buckee, C.; Khanna, T. Which (2020) «COVID-19 Data Can You Trust?» Harvard Business Review, 2020.

Benford, Frank (1938). «The Law of Anomalous Numbers». American Philosophical Society 78 (4): pp. 551-572.

Cho, T.W.; Gaines, B.J. (2007) «Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance». Am. Stat. 61, pp. 218–223.

Coronavirus Resource Center (2021). «COVID-19 Global Cases by the Center

for Systems Science and Engineering (CSSE) at Johns Hopkins University». https://coronavirus.jhu.edu/map.html.

Durtschi, C.; Hillison, W.; Pacini, C. (2004) «The Effective Use of Benford's law to assist in Detecting Fraud in Accounting Data». JFAR, 5, pp. 17–34.

Goodman, Q. (2016) «The promises and pitfalls of Benford's law». Significance, 13, pp. 38–41.

Koch, C.; Okamura, K. (2020) «Benford's Law and COVID-19». https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3586413.

Kuiper, N. H. (1960). «Tests concerning random points on a circle». Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A. 63: pp. 38–47.

Lee, K.B.; Han, S.; Jeong, Y. (2020) «COVID-19 flattening the curve, and Benford's law». Phys. A 2020.

Newcomb, Simon (1881). «Note on the Frequency of Use of the Different Digits in Natural Numbers». American Journal of Mathematics 4 (1): pp. 39-40.

Roukema, B.F. (2014) «A first-digit anomaly in the 2009 Iranian presidential election». J. Appl. Stat. 41, pp. 164–199.

World Health Organization. (2020) «Coronavirus Disease (COVID-19) Outbreak». WHO: Geneva, Switzerland, https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

REVISTA VARIANZA

ISSN 2789-3510 versión impresa
ISSN 2789-3529 versión en línea

MISIÓN

La misión de la revista Varianza es publicar artículos de investigación e interacción social de originalidad y alta calidad que cubran diferentes áreas de la estadística teórica, aplicada y de estudios interdisciplinarios.

VISIÓN

La visión de la revista Varianza es convertirse en una publicación oficial de la ciencia en investigación científica en estadística con aplicación en diferentes campos del conocimiento humano con referencia nacional e internacional.

POLÍTICA EDITORIAL

La revista Varianza es una publicación nacional e internacional, periódica de carácter electrónico, fue presentada por primera vez el año 2001 por el Instituto de Estadística Teórica y Aplicada, IETA de la Carrera de Estadística de la UMSA.

El objetivo, es contribuir al conocimiento sobre la enseñanza-aprendizaje de la estadística y establecer un foro permanente de discusión de ideas, conceptos, procedimientos y modelos concernientes al área estadística y sus aplicaciones.

El alcance de la revista es especializado en contribuciones de artículos del ámbito estadístico, y aplicaciones estadísticas.

A partir del 2021 esta revista se edita dos veces al año durante los meses de abril y octubre. Los trabajos publicados son producto de contribuciones originales y de alto rigor académico y científico, cuyo aporte son investigaciones de profundidad estadística teóricos y aplicados. Sin embargo, publica también temas selectos provenientes de disciplinas afines en estadística aplicada.

Se publican artículos en español donde el resumen y palabras clave deben ser escritos en español e inglés.

MODELO DE FINANCIAMIENTO

La revista tiene una asignación presupuestaria para su publicación impresa con recursos provenientes de la Universidad Mayor de San Andrés, bajo el presupuesto de la carrera de Estadística y el IETA (Instituto de Estadística Teórica y Aplicada). La publicación en la revista es gratuita bajo la modalidad Open Access.

PUBLICACIÓN

Los tipos de artículos que admite la revista son:

- Originales,
- Originales cortos,
- Reportes de caso,
- Artículos de revisión,
- Artículos de reflexión.

La aceptación de un artículo depende de la calidad, originalidad, rigor científico y ética de los escritos, mismos que son verificados mediante revisión de procesos ciegos de evaluación por pares. Las publicaciones son gratuitas y en ningún caso se cobrarán honorarios a los autores para la publicación de sus artículos.

La revista Varianza cuenta con registro ISSN 2789-3510 versión impresa; 2789-3529, versión en línea con depósito legal No. 4-1-285-2021 P.O. y sus artículos son presentados en formato APA. Asimismo, se encuentra indexada en Revistas Bolivianas (REVBOL).

PROCESO DE ARBITRAJE

El proceso inicial de revisión, pasa por el Comité Editorial, que revisa los requisitos básicos, así como los aspectos éticos para ser publicados en la revista Varianza. Esa revisión se realiza en un plazo de un mes. Al concluir el proceso, se notifica al autor o autores, vía correo electrónico.

La evaluación de los manuscritos se realiza con la ayuda de un formulario donde los evaluadores registran las observaciones y recomendaciones. Los evaluadores pueden aceptar o rechazar el artículo de manera definitiva o condicional dada la aclaración y/o rectificación de parte del autor. La comunicación con el autor es responsabilidad exclusiva del editor. Algunos aspectos que se toman en cuenta en la evaluación, son:

1. La originalidad e innovación en conceptos y técnicas estadísticas.
2. Si constituye una contribución en el área.
3. Su pertinencia y rigor científico.
4. Si los referentes teóricos y empíricos son apropiados.
5. Si hace un análisis bien fundamentado y está coherentemente argumentado.
6. Su aporte a la ciencia estadística y aplicaciones.
7. Se juzgará la calidad de la presentación, verificando si el resumen sintetiza el artículo en forma clara y adecuada.
8. Ajustes al formato en las normas APA.

La revisión de los artículos sigue un proceso bajo la modalidad de “doble ciego” con la participación de dos evaluadores con la especialidad requerida para realizar una apropiada evaluación. Esta revisión busca la originalidad e innovación de las publicaciones.

El Comité Editorial de la revista Varianza está conformado por un Consejo Editorial Nacional e Internacional cuya misión es la revisión de los artículos en la modalidad doble ciego.

Política editorial de la revista Varianza

ÉTICA DE PUBLICACIÓN

La revista Varianza tiene compromiso con la ética de la investigación, promueve los siguientes aspectos:

- Evitar conflictos de intereses,
- Evaluar objetivamente los manuscritos,
- Respetar los criterios de evaluación de los evaluadores,
- Conservar la confidencialidad de los autores y evaluadores, durante todo el proceso de revisión.

CONFLICTOS DE INTERESES

Para evitar conflictos de intereses en los procesos de evaluación de los manuscritos enviados, el Comité Editorial no selecciona como evaluadores a colegas que pertenezcan a la misma institución o a la misma red de investigación que los autores. Además, se solicita a los autores que declaren cualquier tipo de interés relacionado con algún miembro o miembros del Comité Editorial.

PRINCIPIOS ÉTICOS DE LA INVESTIGACIÓN

Los manuscritos recibidos son evaluados mediante el sistema de doble ciego, para evitar posible pérdida de objetividad por parte de los árbitros. Por otro lado, el Comité Editorial, se reserva el derecho de pasar por un detector de plagio, los manuscritos recibidos, a fin de identificar faltas éticas.

Cuando se detecta alguna violación a los principios éticos de la investigación, no se publica el artículo y el Comité Editorial informa las razones al autor o autores.

PROPIEDAD INTELECTUAL

Los autores al momento de tener la aceptación de la publicación, deben firmar la autorización de la publicación de la revista Varianza, en la que se estipula que son legítimos propietarios del artículo a publicar, que es una contribución original y que no existe problemas de derechos de autor con terceros y/u otros conflictos de naturaleza ética. Todo el contenido de la revista, excepto aquello que expresamente sea identificado, está bajo la licencia *Creative Commons*.

LICENCIAMIENTO

La revista Varianza se encuentra bajo licenciamiento *Creative Commons* atribución CC BY https://creativecommons.org/licenses/by/4.0/.

Atribución CC BY

La licencia permite que otros distribuyan, mezclen, adapten y construyan sobre su trabajo, incluso comercialmente, siempre que le reconozcan la creación original. Esta es la licencia más complaciente que se ofrece. Recomendado para la máxima difusión y uso de materiales con licencia.

INSTRUCCIONES PARA LOS AUTORES

Título

El título debe ser conciso e informativo de la investigación en mayúscula con no más de 12 palabras que contengan ya los descriptores, además debe ser redactado en sentido afirmativo, en idioma español e inglés.

Autoría

Debe ir a continuación del título, señalar el nombre y apellidos del (los) autor(es) acompañado de la(s) nota(s) a pie de página, según sea la posición en la autoría. A pie de página y con la numeración correspondiente se debe informar: profesión, filiación institucional y un breve curriculum vitae.

Resumen

El resumen debe contener el objetivo, metodología, material, métodos, resultados de la investigación haciendo énfasis en los logros alcanzados.

Palabras claves

Sirven para identificar el artículo en bases de datos internacionales de manera que un potencial usuario pueda llegar en forma efectiva al artículo. Van debajo del resumen, mínimo cuatro (4) y máximo siete (7) palabras clave que no deben hacer parte del título del artículo. Deben estar ordenadas alfabéticamente y separadas por comas que ayuden a identificar los aspectos importantes del artículo.

Abstract

Se escribe el resumen del artículo en inglés.

Key words

Se escriben las palabras clave en inglés.

Introducción

La introducción debe contener:

- ***Problema:*** Debe describir claramente lo que se resolverá con la investigación. Debe enunciar claramente el qué y el porqué de la investigación. Se debe desarrollar en uno o dos párrafos iniciales.

- ***Revisión de la literatura:*** Expone el marco referencial que da sustento al trabajo de investigación. A través de las citas se provee reconocimiento de estudios anteriores que se relacionan específicamente con el trabajo.

- ***Objetivo o hipótesis:*** Debe describir el objetivo en forma clara, debe indicar en forma inequívoca qué es lo que el investigador intenta observar y medir, redactados en forma

afirmativa y sujetos a una sola interpretación. La hipótesis debe expresar de manera clara, precisa y concisa una relación o diferencia entre dos o más variables, incluyendo, si corresponde, las variables del estudio y su efecto.

Materiales y métodos

Debe describir el universo de estudio, instrumentos y procedimientos con la precisión necesaria para permitir a los lectores una comprensión clara del artículo y la posibilidad de reproducir lo entendido. Debe especificarse el número de observaciones de los objetos a estudiar, los métodos y técnicas estadísticas para la generación y análisis de resultados.

Resultados

Presentarlos en secuencia lógica y que cuenten con un análisis estadístico o interpretativo en relación con el objetivo del estudio.

Discusión

La discusión debe ser sobre la base de los objetivos y los resultados para posteriormente comparar con el marco referencial. Discutir los aspectos nuevos y limitaciones que tiene su estudio, enunciando proyecciones o nuevas hipótesis si corresponde.

Conclusiones

Estas deben responder a los objetivos del estudio, limitándose a los datos encontrados sin citar referencias.

Referencias Bibliográficas

Contiene la referencia de libros y artículos consultados para el artículo científico. Las fuentes bibliográficas deben ser citadas a lo largo del texto, en formato APA.

Presentación

Los artículos deben ser presentados en doble columna, letra *Times New Roman* tamaño 12, espacio simple, los márgenes externos son: margen izquierdo 2,5 cm., derecho 2,0 cm., superior e inferior 2,0 cm.

Política editorial de la revista Varianza

Dirección: Calle 27 de Cota Cota
Bloque F.C.P.N. - Primer Piso
Email: ieta@umsa.bo
Página web: https://ieta.umsa.bo/ediciones-varianza
2021

La Paz - Bolivia