UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2014, Vol. 7, No. 2, pp. 1–23

Paquete ZOIP de R para analizar datos proporcionales inflados con ceros y/o unos

Package ZOIP of R to analyze proportional data inflated with zeros and / or ones

Juan Camilo Díaz Zapata[a]
jucdiazza@unal.edu.co

Freddy Hernández Barajas[b]
fhernanb@unal.edu.co

Olga Cecilia Usuga Manco[c]
olga.usuga@udea.edu.co

Resumen

Los datos obtenidos a partir de variables medidas bajo porcentajes, tasas y proporciones, son llamados datos proporcionales y estos se encuentran ubicados por lo general en el intervalo (0,1). Diferentes distribuciones se han desarrollado para caracterizar este tipo de variables, sin embargo existe la posibilidad que este tipo de variables puedan dar resultados de cero o uno. Haciendo que autores como Ospina & Ferrari (2012) y Rigby & Stasinopoulos (2005) implementaran una distribución beta inflada con ceros y/o unos, diferenciadas solo por su parametrización. Otros autores como Galvis & Lachos (2014) han trabajo con otras distribuciones para datos proporcionales inflados, como la distribución simplex, sin embargo, no se ha encontrado una distribución que reúna las principales distribuciones para datos proporcionales inflados con ceros y/o unos, en este trabajo se presenta el paquete **ZOIP** del sistema de computación R para la implementación de la distribución ZOIP (Zeros Ones Inflated Proportional) que reúne la distribución simplex y beta bajo diferentes parametrizaciones y estima los parámetros de dicha distribución. La estimación de los parámetros de la distribución ZOIP se hace vía máxima verosimilitud. Se realizan estudios de simulación que muestran la convergencia satisfactoria de los parámetros y se presenta el ajuste de una distribución ZOIP a datos reales.

***Palabras clave*:** Datos proporcionales, Distribución ZOIP, Máxima verosimilitud, R.

[a]Estudiante de maestría en estadística, Universidad Nacional de Colombia, Sede Medellín.
[b]Profesor asistente, Universidad Nacional de Colombia, Sede Medellín.
[c]Profesora asociada, Universidad de Antioquia, Medellín.

Abstract

The data obtained from variables measured by percentages, rates and proportions are called proportional data which its values are usually between the (0,1) interval. Although, different distributions have been developed to characterize these variables, it is possible that some variables take the extreme values zero and one. Making authors such as Ospina & Ferrari (2012) and Rigby & Stasinopoulos (2005) implement an inflated beta distribution with zeros and/or ones, differentiated only by their parameterization. Other authors such as Galvis & Lachos (2014) have worked with inflated proportional data, researchers have worked with other distributions such as simplex distributions. Nevertheless, it has not been found any distribution which combines the principals characteristic of several distributions for solving these issue. Thus, this paper presents the package Zeros Ones Inflated Proportional, **ZOIP**, developed on the computation system R, which combines both beta and simplex distribution and estimates its parameters given the parameters for the original distributions. Regarding the parameters estimations process, it is performed by maximum likelihood method. Finally, simulations studies have been performed which show the accuracy of the parameters convergence and the adjustment of a ZOIP distribution on real data.

***Keywords*:** Proportional data, ZOIP Distribution, Maximumn likelihood, R.

1. Introducción

En modelación estadística es posible encontrarnos con variables respuesta como proporciones, porcentajes o tasas que se encuentran en el intevalo (0, 1). La distribución más utilizada en la literatura para caracterizar este tipo de variables es la distribución beta con soporte en el intervalo (0,1), la cual ha sido reparametrizada por autores como Ferrari & Cribari-Neto (2004) y Rigby & Stasinopoulos (2005); otras distribuciones no tan comunes en la literatura pero que caracterizan este tipo de variables son la distribución simplex (Jørgensen 1997), beta-rectangular (Hahn 2008) y la distribución LogitSep (Hossain 2015). Por otra parte, es común que los porcentajes o proporciones puedan dar valores iguales a cero o uno, representando la ausencia o presencia total de la característica de interés, respectivamente, Las distribuciones descritas anteriormente no pueden ser admisibles para este tipos de variables, es por esto que se han desarrollado distribuciones infladas con ceros y/o unos, para tratar estos casos, como lo hizo Ospina & Ferrari (2010) quienes presentan una distribución beta inflada en la que hacen una combinación entre una distribución discreta para la parte de los valores que pueden tomar cero o uno y una parte continua para los valores continuos entre cero y uno. Rigby & Stasinopoulos (2005) incluyen dentro de sus modelos gamlss la distribución beta inflada con ceros y/o unos según su parametrización.

Esto ha dado pie para que diferentes autores hayan empezado a desarrollar diferentes modelos de regresión para tratar este tipo variables, Ospina & Ferrari (2012)

propusieran una clase general de modelos de regresión beta inflados con cero o uno, además Kosmidis et al. (2015) han estudiado dichos modelos inflados recientemente, pero con una distribución distinta a la presentada por Ospina & Ferrari (2012). Galvis & Lachos (2014) presentan modelos de regresión para diferentes distribuciones para datos proporcionales inflados con ceros y/o unos mediante metodologías de estimación bayesianas.

Muchos autores han implementado distribuciones para datos proporcionales en el software estadístico R, Zeileis et al. (2010) implementan el paquete **betareg** donde se encuentran los modelos de regresión beta propuestos por Ferrari & Cribari-Neto (2004), Qiu et al. (2008) implementan el paquete **simplexreg** para realizar análisis de distribución y regresión sobre una distribución simplex, para datos proporcionales no inflados, otros autores como Rigby & Stasinopoulos (2008) incluyen en el paquete **gamlss** la distribución beta inflada con ceros y/o unos y la posibilidad de realizar modelos de regresión sobre ellos.

Aunque muchos autores han implementado las distribuciones para datos proporcionales inflados con ceros y/o unos, ninguno ha presentado una propuesta como la de reunir en una sola distribución las diferentes distribuciones para datos proporcionales y sus diferentes parametrizaciones, además de implementarla en un solo paquete, como se presenta en el paquete **ZOIP** en R Core Team (2017) disponible en el repositorio web `GitHub`.

El artículo se encuentra organizado de la siguiente manera: primero se presentan las distribuciones más representativas para datos proporcionales, en la sección 3 se presenta la distribución para datos proporcionales inflados con ceros y/o unos ZOIP (Zeros Ones Inflated Proportional), seguido por el desarrollo analítico de la estimación de los parámetros de la distribución ZOIP vía máxima verosimilitud, en la sección 5 se aplica el ajuste de una distribución ZOIP en un estudio de simulación y para datos reales, por último se presenta el paquete **ZOIP**.

2. Distribución para datos proporcionales

Para los casos de modelación donde la variable de interés es una proporción, un porcentaje o una tasa. Este tipo de variables no pueden ser analizadas con la distribución normal, debido a que el soporte de la normal es la recta real $\mathbb{R}$, además en este tipo de variables es común la asimetría e incluso la bimodalidad, por esta razón en la literatura estadística se han propuesto distribuciones para este tipo de comportamientos, como la distribución beta, que cuenta con diferentes parametrizaciones (Ferrari & Cribari-Neto (2004) y Rigby & Stasinopoulos (2005)) y la distribución simplex propuesta por Jørgensen & Barndorff-Nielsen (1991), de igual manera otras distribuciones más particulares como la beta-rectangular (Hahn 2008) y LogitSep (Hossain 2015) se acoplan a este comportamiento, a continuación

se mostraran las funciones de densidad de probabilidad, la media, la varianza y dependencias de algunas de las distribuciones mencionadas anteriormente.

2.1. Distribución beta original

Si una variable aleatoria y definida entre cero y uno, tiene distribución beta con parámetros p y q se acostumbra a denotarla por $y \sim \text{Be}(p, q)$ y la función de densidad de probabilidad de la distribución es dada por:

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{1}$$

donde los parámetros $p > 0$, $q > 0$ y $\Gamma(\cdot)$ es la función gamma. El valor esperado y la varianza de y están dadas por:

$$E(y) = \frac{p}{p+q} \tag{2}$$

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{3}$$

2.2. Distribución beta parametrización Ferrari y Cribari-Neto (2004)

Ferrari & Cribari-Neto (2004) propusieron otra parametrización para la distribución beta en función de los parámetos μ y ϕ donde μ corresponde a la media de la distribución y ϕ es interpretado como un parámetro de precisión. Si $0 < y < 1$ y $y \sim \text{Be}(\mu, \phi)$ la función de densidad de probabilidad de la distribución está dada por:

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{4}$$

donde $0 < \mu < 1$ y $\phi > 0$. El valor esperado y la varianza de y están dados por:

$$E(y) = \mu \tag{5}$$

$$Var(y) = \frac{\mu(1-\mu)}{1+\phi} \tag{6}$$

Además note que la parametrizacion de la distribución beta original es equivalente a la de Ferrari & Cribari-Neto (2004) cuando:

$$p = \mu\phi \tag{7}$$

$$q = (1-\mu)\phi \tag{8}$$

2.3. Distribución beta parametrización Rigby y Stasinopoulos (2005)

Rigby & Stasinopoulos (2005) propusieron una nueva parametrización para la distribución beta con parámetros μ y σ donde μ es la media de la distribución y σ es interpretado como un parámetro de dispersión, se dice que $y \sim \text{Be}(\mu, \sigma)$ con $0 < y < 1$, si la función de densidad de probabilidad de la distribución está dada por:

$$f(y;\mu,\sigma) = B(\mu,\sigma)y^{\mu((1-\sigma^2)/\sigma^2)-1}(1-y)^{(1-\mu)((1-\sigma^2)/\sigma^2)-1} \tag{9}$$

donde $B(\mu,\sigma) = \frac{\Gamma((1-\sigma^2)/\sigma^2)}{\Gamma(\mu((1-\sigma^2)/\sigma^2))\Gamma((1-\mu)((1-\sigma^2)/\sigma^2))}$,

donde $0 < \mu < 1$ y $0 < \sigma < 1$. La media y la varianza de y están dadas por:

$$E(y) = \mu \tag{10}$$
$$Var(y) = \sigma^2\mu(1-\mu) \tag{11}$$

Además note que la parametrizacion de la distribucion beta original es equivalente a la de Rigby & Stasinopoulos (2005) cuando:

$$p = \frac{\mu(1-\sigma^2)}{\sigma^2}\phi \tag{12}$$
$$q = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \tag{13}$$

2.4. Distribución simplex

La distribución simplex fue introducida por Jørgensen & Barndorff-Nielsen (1991) y es un caso particular de los modelos de dispersión propuestos por Jørgensen (1997), dicha distribución depende de los parámetros μ que es la media de la distribución y σ^2 que es un parámetro de dispersión. Si $0 < y < 1$ y $y \sim S^-(\mu,\sigma^2)$ la función de densidad de probabilidad esta dada por:

$$f(y;\mu,\sigma^2) = \left\{2\pi\sigma^2[y(1-y)]^3\right\}^{-1/2}\exp\left\{-\frac{y(1-y)\mu^2(1-\mu)^2}{2\sigma^2(y-\mu)^2}\right\} \tag{14}$$

donde $0 < \mu < 1$ y $\sigma > 0$. Además el valor esperado y la varianza están dadas por:

$$E(y) = \mu \tag{15}$$

$$Var(y) = \mu(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2(1-\mu)^2}\right\} \tag{16}$$

donde $\Gamma(a,b)$ está dado por la función$-\Gamma$ incompleta definido como $\Gamma(a,b) = \int_b^\infty t^{a-1}b^t dt$. ver más en Zhang et al. (2016).

3. Distribución ZOIP (Zeros Ones Inflated Proporcional)

En las distribuciones vistas en la sección 2, se evidenciaron ciertas distribuciones que se ajustan al comportamiento de datos proporcionales, porcentajes o tasas que están en el intervalo (0,1), sin embargo es común que estos datos tomen valores en cero y/o uno que representarían la ausencia o presencia total de cierta característica, por lo que no sería posible ajustar los datos a las distribuciones vistas anteriormente y es por eso que en este trabajo se propone la distribución ZOIP, como un conjunto de distribuciones para datos proporcionales inflados con ceros y/o unos.

La distribución para datos proporcionales inflados con ceros y/o unos se compone de la mezcla de tres distribuciones, dos de ellas discretas, que son distribuciones degeneradas en cero y uno, y una tercera distribución continua que además es una función de densidad de probabilidad para datos proporcionales, como las presentadas anteriormente, si la variable aleatoria y tiene distribución ZOIP con parámetros μ, σ, p_0 y p_1, se denotará como $y \sim \text{ZOIP}(\mu, \sigma, p_0, p_1)$, la función de densidad de probabilidad está dado por:

$$g(y;\mu,\sigma,p_0,p_1) = \begin{cases} p_0 & \text{si } y = 0, \\ p_1 & \text{si } y = 1, \\ (1-p_0-p_1)f(y;\mu,\sigma) & \text{si } y \in (0,1) \end{cases} \tag{17}$$

donde $p_0 \geq 0$ representa la probabilidad que $y = 0$ y $p_1 \geq 0$ representa la probabilidad de que $y = 1$, además $0 \leq p_0 + p_1 \leq 1$ y $f(y;\mu,\sigma)$ representa alguna de las funciones de densidad de probabilidad para datos proporcionales, descritas en la sección anterior. La media y varianza de y, están dadas por

$$E(y) = p_1 + (1-p_0-p_1)E^*(y) \tag{18}$$

$$\begin{aligned} Var(y) = {} & p_1(1-p1)+ \\ & + (1-p_0-p_1)\left[Var^*(y) + (p_0+p_1)[E^*(y)]^2 - 2E^*(y)p_1\right] \end{aligned} \tag{19}$$

donde $E^*(y)$ es el valor esperado de una distribución para datos proporcionales como las descritas en (2), (5), (10) y (15). Además la $Var^*(y)$ es la varianza de una distribución para datos proporcionales como se muestra en (3), (6), (11) y (16).

Si para la distribución $ZOIP(\mu, \sigma, p_0, p_1)$ se elige la distribución beta con parametrización Ferrari & Cribari-Neto (2004) entonces el parámetro σ tomará el rol del parámetro ϕ de la distribución, si la parametrización es beta original μ y σ tomarán el rol de p y q respectivamente. En las demás distribuciones y parametrizaciones μ y σ tomaran los valores y dominios correspondientes a su distribución.

La distribución ZOIP se encuentra inflada con ceros y unos, es decir bilateralmente, pero existe la posibilidad que hayan casos de estudio en que se encuentren datos inflados con unos únicamente, por lo que $p_0 = 0$ y por lo tanto se estará llamando no una distribución ZOIP, sino una distribución OIP (Ones Inflated Proporcional) y si los datos se encuentran inflados con ceros únicamente, es decir $p_1 = 0$ se tendrá una distribución ZIP (Zeros Inflated Proporcional), Si los datos no se encuentran inflados, entonces $p_0 = p_1 = 0$ y la distribución ZOIP será una distribución para datos proporcionales clásica.

En la Figura 1 se muestran las densidades para varias de las distribuciones ZOIP-beta en sus diferentes parametrizaciones y ZOIP-simplex, es de aclarar que en las Figuras 1a. 1b. 1c. los valores de los parámetros son diferentes pero dan como resultado la misma distribución gracias a las ecuaciones descritas en (7), (8) para el caso Ferrari & Cribari-Neto (2004) y (12), (13) para el caso Rigby & Stasinopoulos (2005). Además se puede observar en la Figura 1d. como la distribución ZOIP-simplex hereda el comportamiento bimodal de la distribución clásica simplex, con valores mínimo y máximo en cero y uno respectivamente.

4. Inferencia estadística

Para estimar los parámetros de la distribución ZOIP se usa el método de máxima verosimilitud. La función de verosimilitud para $\boldsymbol{\theta} = (\mu, \sigma, p_0, p_1)^\top$, basado en una muestra de $\boldsymbol{y}_i$ observaciones independientes, es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\mathbf{y}_i; \mu, \sigma, p_0, p_1) \tag{20}$$

Para encontrar los estimadores de máxima verosimilitud (MLE) de la distribución ZOIP, se consideraran 2 casos:

1. **ZOIP-beta original**
   Considera la parametrización de la distribución beta original y la ecuación

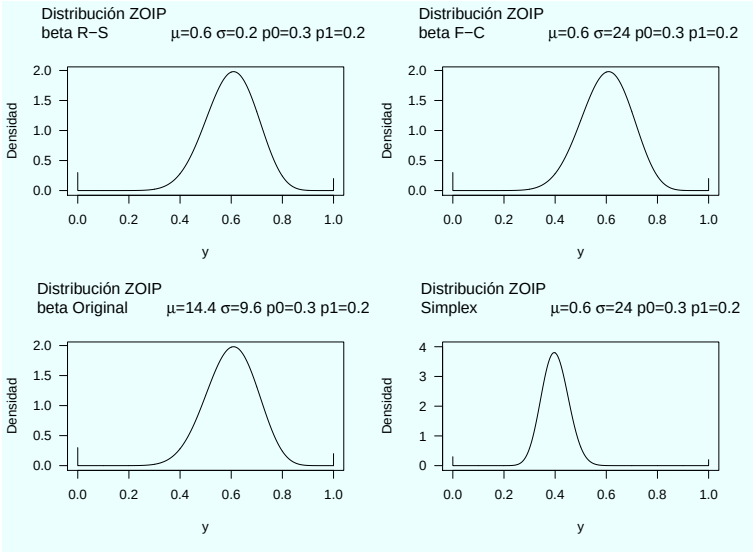

Figura 1: Densidades para la distribución ZOIP para algunos valores de los parámetros, donde R-S se refiere a Rigby & Stasinopoulos (2005) y F-C es Ferrari & Cribari-Neto (2004).

definida en (20) se tiene que:

$$\boldsymbol{\theta} = (p, q, p_0, p_1)^\top$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{\theta}|y_i) = L_1(p_0) \cdot L_2(p_1) \cdot L_3(p, q)$$

Note que la función de verosimilitud es factorizada en tres términos, dos de ellos del componente discreto y uno compuesto por p y q del componente continuo, por tanto los parámetros son separables (Pace & Salvan 1997), así la máxima verosimilitud puede ser tratada por separado.

$$L_1(p_0) = \prod_{i=1}^{n} p_0^{S_0(y_i)}(1-p_0)^{1-S_0(y_i)} = p_0^{\sum_{i=1}^{n} S_0(y_i)}(1-p_0)^{n-\sum_{i=1}^{n} S_0(y_i)}$$

donde:

$$S_j(y_i) = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{si } y_i \neq j \end{cases} \quad ; \quad j = 1, 2 \tag{21}$$

Ahora sacando logaritmo natural a la función de verosimilitud.

$$\ell_1(p_0) = \sum_{i=1}^{n} S_0(y_i) log(p_0) + (n - \sum_{i=1}^{n} S_0(y_i)) log(1-p_0)$$

$$\frac{\delta \ell_1(p_0)}{\delta p_0} = \frac{\sum_{i=1}^{n} S_0(y_i)}{p_0} - \frac{n - \sum_{i=1}^{n} S_0(y_i)}{1 - p_0} = \sum_{i=1}^{n} S_0(y_i) - p_0 n = 0$$

$$\hat{p_0} = \frac{1}{n} \sum_{i=1}^{n} S_0(y_i)$$

$$\therefore \hat{p_1} = \frac{1}{n} \sum_{i=1}^{n} S_1(y_i)$$

Ahora se halla MLE para los parámetros del componente continuo de la función.

$$\ell_3(p,q) = \sum_{i=1:y_i \in (0,1)}^{n} log(f(p,q|y_i)) = nlog(\Gamma(p+q)) - nlog(\Gamma(p)) - nlog(\Gamma(q))$$

$$+(p-1) \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) + (q-1) \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i)$$

entonces

$$\frac{\delta \ell_3(p,q)}{\delta p} = \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) + \frac{n \cdot \delta log(\Gamma(p+q))}{\delta p} - \frac{n \cdot \delta log(\Gamma(p))}{\delta p} - \frac{n \cdot \delta log(\Gamma(q))}{\delta p} = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta q} = \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i) + \frac{n \cdot \delta log(\Gamma(p+q))}{\delta q} - \frac{n \cdot \delta log(\Gamma(p))}{\delta q} - \frac{n \cdot \delta log(\Gamma(q))}{\delta q} = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta p} = \sum_{i=1:y_i \in (0,1)}^{n} log(y_i) - n(-\psi(p+q) + \psi(p)) = 0$$

$$\frac{\delta \ell_3(p,q)}{\delta q} = \sum_{i=1:y_i \in (0,1)}^{n} log(1-y_i) - n(-\psi(p+q) + \psi(q)) = 0$$

donde $\psi(\cdot) = \Gamma^{'}(\cdot)/\Gamma(\cdot)$
Este sistema de ecuaciones no tiene una solución de forma cerrada, por lo que para encontrar los MLE de p y q es necesario utilizar algoritmos iterativos, por ejemplo el método de Newton Raphson, mínimos cuadrados ponderados y en el paquete **ZOIP** se utiliza optimizadores a la función de verosimilitud mediante la función `nlminb` de R, sin embargo se puede garantizar que los puntos críticos encontrados serán máximos de la función de verosimilitud, ya que si hallamos la segunda derivada de la función se tiene que:

$$\frac{\delta^2 \ell_3(p,q)}{\delta p^2} = -n(\psi^{'}(p) - \psi^{'}(p+q)) < 0$$

$$\frac{\delta^2 \ell_3(p,q)}{\delta q^2} = -n(\psi^{'}(q) - \psi^{'}(p+q)) < 0$$

debido que la varianza de la transformación logarítmica de la variable es:

$$var(log(y)) = E[log^2(y)] - (E[log(y)])^2 = \psi^{'}(p) - \psi^{'}(p+q) > 0$$

$$var(log(1-y)) = E[log^2(1-y)] - (E[log(1-y)])^2 = \psi^{'}(q) - \psi^{'}(p+q) > 0$$

ver más en Owen (2008).

Para encontrar las estimaciones de los parámetros de beta en parametrizaciones de Ferrari & Cribari-Neto (2004) y Rigby & Stasinopoulos (2005), basta con encontrar los estimadores MLE anteriores de la parametrización original y utilizar las ecuaciones definidas en (7), (8) para el caso de Ferrari & Cribari-Neto (2004) y (12), (13) para el caso de Rigby & Stasinopoulos (2005).

2. **ZOIP-simplex**

Para este caso lo único que varía con respecto al anterior es la estimación en el componente continuo.

$$L_3(\mu,\sigma) = \prod_{i=1:y_i\in(0,1)}^{n} \left[2\pi\sigma^2[y_i(1-y_i)]^3\right]^{-1/2} exp\left(-\frac{1}{2\sigma^2}d(y_i;\mu)\right)$$

donde $d(y_i;\mu) = \frac{y_i(1-y_i)\mu^2(1-\mu)^2}{(y_i-\mu)^2}$

$$\ell_3(\mu,\sigma) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{3}{2}\sum_{i=1:y_i\in(0,1)}^{n} log(y_i(1-y_i)) - \sum_{i=1:y_i\in(0,1)}^{n} \frac{1}{2\sigma^2}d(y_i;\mu)$$

$$\frac{\delta \ell_3(\mu,\sigma)}{\delta\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu) = \sigma(-n\sigma^2 + \sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu)) = 0$$

no es admisible que $\sigma = 0$ entonces:

$$-n\sigma^2 + \sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu) = 0$$

$$\therefore \hat{\sigma^2} = \frac{1}{n}\sum_{i=1:y_i\in(0,1)}^{n} d(y_i;\mu)$$

El estimador MLE de σ^2 depende del valor estimado en μ, entonces:

$$\frac{\delta \ell_3(\mu,\sigma)}{\delta\sigma} = -\frac{1}{2\sigma^2} \sum_{i=1:y_i\in(0,1)}^{n} \frac{\delta d(y_i;\mu)}{\delta\mu} = 0$$

$$\frac{\delta d(y_i;\mu)}{\delta\mu} = \sum_{i=1:y_i\in(0,1)}^{n} \frac{y_i(1-y_i)\mu^2(1-\mu)^2}{2(y_i-\mu)^3} + \frac{2y_i(1-y_i)\mu(1-\mu)^2 - 2y_i(1-y_i)\mu^2(1-\mu)}{(y_i-\mu)^2} = 0$$

No tiene una solución cerrada analíticamente, entonces se deben utilizar algoritmos iterativos tal como Newton Raphson o mínimos cuadrados ponderados, en el paquete **ZOIP** se utiliza optimizadores para la función de verosimilitud mediante la función `nlminb` de R, para encontrar puntos críticos donde $\delta d(y_i;\mu)/\delta\mu = 0$.

5. Paquete ZOIP

En esta sección se presenta el paquete **ZOIP** de R alojado en `GitHub` y creado por los autores para analizar datos proporcionales inflados con ceros y/o unos y ajustar una distribución ZOIP.

5.1. Instalación

Para acceder a la última versión del paquete **ZOIP** se encuentra ubicada en `GitHub`, el cual es un alojamiento de repositorios Git, para obtener dicha versión es necesario ejecutar el siguiente código que instala el paquete **devtools**, que es necesario para descargar el paquete **ZOIP**.

```
if (!require('devtools')) install.packages('devtools')
devtools::install_github('jucdiaz/ZOIP', force=TRUE)
require(ZOIP)  # Carga el paquete
```

5.2. Funciones sobre distribución ZOIP

En el paquete **ZOIP** existen cuatro funciones llamadas `dZOIP`, `pZOIP`, `qZOIP` y `rZOIP` el cual corresponden a las funciones de densidad de probabilidad, la función de distribución acumulada, la función cuantil y la función generadora de números

aleatorios de la distribución ZOIP, respectivamente; en el siguiente código se observa como se halla la densidad de probabilidad en el punto 0.5 de una distribución ZOIP-beta con parametrización Rigby & Stasinopoulos (2005) descrita como ZOIP($\mu = 0.2$, $\sigma = 0.5$, $p_0 = 0.2$, $p_1 = 0.2$)

```
dZOIP(x=0.5, mu=0.2, sigma=0.5, p0=0.2, p1=0.2, family='R-S')
##[1] 0.3243543
```

Además se halla la probabilidad acumulada hasta el punto 0.5 de una distribución OIP-beta con parametrización Ferrari & Cribari-Neto (2004) dada por ZOIP($\mu = 0.2$, $\sigma = 3$, $p_0 = 0$, $p_1 = 0.2$)

```
pZOIP(q=0.5, mu=0.2, sigma=3, p0=0, p1=0.2, family='F-C')
##[1] 0.7181223
```

Se calcula el percentil en el punto 0.7 de una distribución ZIP-beta original dada por ZOIP($\mu = 0.6$, $\sigma = 2.4$, $p_0 = 0.2$, $p_1 = 0$)

```
qZOIP(p=0.7, mu=0.6, sigma=2.4, p0=0.2, p1=0, family='Original')
##[1] 0.2061418
```

Por último se generaron 8 valores aleatorios de una distribución ZOIP-simplex descrita como ZOIP($\mu = 0.6$, $\sigma = 3$, $p_0 = 0.2$, $p_1 = 0.2$). La función `set.seed` sirve para garantizar la repetición de los valores aleatorios generados en el ejemplo.

```
set.seed(12345)
rZOIP(n=8, mu=0.2, sigma=3, p0=0.2, p1=0.2, family='Simplex')
##[1] 0.3185479 1.0000000 0.3765073 1.0000000 0.1626598
##[6] 0.0000000 0.1138673 0.1840670
```

5.3. Función RM.ZOIP

La función `RM.ZOIP` estima los parámetros de un modelo ZOIP con y sin covariables vía máxima verosimilitud utilizando el optimizador deseado (`nlminb`,`optim`). La estructura de la función `RM.ZOIP` es la siguiente:

```
RM.ZOIP(
  formula.mu,
  formula.sigma = ~ 1,
  formula.p0 = ~ 1,
  formula.p1 = ~ 1,
  data,
  link = c('identity', 'identity', 'identity', 'identity'),
  family = 'R-S',
optimizer='nlminb'
)
```

Los argumentos de la función `RM.ZOIP` son:

- `formula.mu`: Formula que define la función de regresión para el parámetro μ, Un valor posible es `y` $\sim$ `x1 + x2`, es necesario definir la variable respuesta (y).
- `formula.sigma`: Formula que define la función de regresión para el parámetro σ, Un valor posible es $\sim$ `x1`. Por defecto $\sim$ `1`.
- `formula.p0`: Formula que define la función de regresión para el parámetro p_0, Un valor posible es $\sim$`x1`. Por defecto $\sim$`1`.
- `formula.p1`: Formula que define la función de regresión para el parámetro p_1, Un valor posible es $\sim$`x1`. Por defecto $\sim$`1`.
- `data`: es el conjunto de datos en formato `data.frame` donde debe contener las nombres de las columnas tal cual como están en las fórmulas.
- `family`: Elección de la parametrización de la distribución beta o distribución deseada en la parte continua de la distribución ZOIP, si toma el valor de `'R-S'` se utilizara la distribución beta con parametrización Rigby & Stasinopoulos (2005), si toma el valor de `'F-C'` se utilizara la distribución beta parametrización Ferrari & Cribari-Neto (2004), el valor de `'Original'` se utilizara la distribución beta con parametrización original, `'Simplex'` Utilizara la distribución simplex.
- `link`: Es un vector con las funciones enlace adecuadas para cada parámetro a estimar de acuerdo a las opciones escogidas en los parámetros de familia y formula. Si el modelo de regresión no posee covariables se debe utilizar como función enlace la opción `identity`, independientemente del valor escogido en familia, opciones posibles son `logit`, `log`. Por defecto `link=c('identity','identity','identity','identity')`.
- `optimizer`: Elección del optimizador, utilizado para la convergencia de la máxima verosimilitud.

En el siguiente ejemplo se mostrara el ajuste de una distribución ZOIP (modelo ZOIP sin covariables), para ello mostraremos la salida de la función `RM.ZOIP` de 1000 observaciones simuladas para la distribución ZOIP-beta parametrización Rigby & Stasinopoulos (2005).

```
yi <- as.data.frame(rZOIP(n=1000, mu=0.6, sigma=0.2,
                          p0=0.03, p1=0.05, family='R-S'))
mod <- RM.ZOIP(formula.mu=yi ~ 1, formula.sigma= ~ 1,
               formula.p0= ~ 1, formula.p1= ~ 1, data=yi,
               family='R-S')
summary(mod)
```

```
------------------------------------------------------------------
Fixed effects for identity(mu)
------------------------------------------------------------------
             Estimate Std. Error z value  Pr(>|z|)
(intercept) 0.6066914  0.0031636  191.78 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
------------------------------------------------------------------
Fixed effects for identity(sigma)
------------------------------------------------------------------
            Estimate Std. Error z value  Pr(>|z|)
(intercept) 0.196643   0.004322  45.498 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
------------------------------------------------------------------
Fixed effects for identity(p0)
------------------------------------------------------------------
             Estimate Std. Error z value Pr(>|z|)
(intercept) 0.0339992  0.0057308  5.9327 2.98e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
------------------------------------------------------------------
Fixed effects for identity(p1)
------------------------------------------------------------------
             Estimate Std. Error z value  Pr(>|z|)
(intercept) 0.0450005  0.0065556  6.8644 6.675e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
------------------------------------------------------------------
------------------------------------------------------------------
```

En el resultado anterior se obtienen los valores de $\hat{\mu} = 0.6066914$, $\hat{\sigma} = 0.196643$, $\hat{p_0} = 0.0339992$ y $\hat{p_1} = 0.0450005$, que son a los parámetros con los que se simulo y_i. Además cabe resaltar que en la función `RM.ZOIP` para ajustar distribuciones de probabilidad no es necesario colocar funciones de enlace ni espacio de busqueda de los parámetros, ya que estos son introducidas automáticamente de acuerdo a el

valor tomado en `family`, el parámetro que define las funciones enlace es introducido por el usuario cuando se quieren realizar modelos de regresión ZOIP y no ajustes de distribuciones ZOIP.

6. Aplicación

En esta sección se muestran varios resultados sobre el ajuste de una distribución ZOIP, primero se realizó un estudio de simulación para observar la convergencia de la estimación de los parámetros de la distribución, y en segunda instancia se ajustó una distribución ZOIP a datos reales sobre la utilización de una tarjeta de crédito de una entidad financiera.

6.1. Datos simulados

En este estudio de simulación se analizan diferentes aspectos de la capacidad de estimación que tiene el método de máxima verosimilitud sobre los parámetros de la distribución ZOIP. Se generaron muestras de una distribución ZOIP bajo las diferentes distribuciones y parámetrizaciones con tamaños de muestra n de: 5, 10, 15 y así sucesivamente hasta 500, y se realizaron 1000 réplicas para cada tamaño de muestra, posteriormente se calculó la mediana de cada una de las estimaciones de los parámetros, y así poder analizar la capacidad de convergencia de las metodologías implementadas en la distribución ZOIP y en el paquete **ZOIP**.

En el primer escenario del estudio de simulación se generaron los datos de una distribución ZOIP-beta($\mu = 0.6, \sigma = 0.2, p_0 = 0.03, p_1 = 0.05$) para el caso de la parametrización de Rigby & Stasinopoulos (2005), ZOIP-beta($\mu = 0.6, \sigma = 24, p_0 = 0.03, p_1 = 0.05$) para el caso de la parametrización de Ferrari & Cribari-Neto (2004), ZOIP-beta($\mu = 14.4, \sigma = 9.6, p_0 = 0.03, p_1 = 0.05$) en la parametrización original, cabe aclarar que las tres parametrizaciones anteriores generan exactamente la misma distribución, esto gracias a las ecuaciones definidas en (7), (8), (12) y (13), de igual manera se generó la misma cantidad de datos simulados para la distribución ZOIP-simplex($\mu = 0.4, \sigma = 0.2, p_0 = 0.03, p_1 = 0.05$).

En la Figura 2 se presentan las medianas de la estimación de los parámetros para cada tamaño de muestra, de esta figura se observa que independientemente de la distribución y parametrización escogida en la distribución ZOIP, todos las estimaciones convergen al valor verdadero del parámetro a medida que aumenta el tamaño de muestra n, de la Figura 2 se nota que las estimaciones de σ cuando son parámetros con significado de dispersión como es en la distribución beta con parametrización Rigby & Stasinopoulos (2005) y en la distribución simplex, tienden a dar valores subestimados, por otra parte, en las distribuciones que σ tiene significado de forma y precisión tienden a dar valores sobrestimados. Se observa que las estimaciones de los parámetros de inflación a pesar de que son pequeñas dan resultados muy satisfactorios y casi sin variación en su forma de estimación

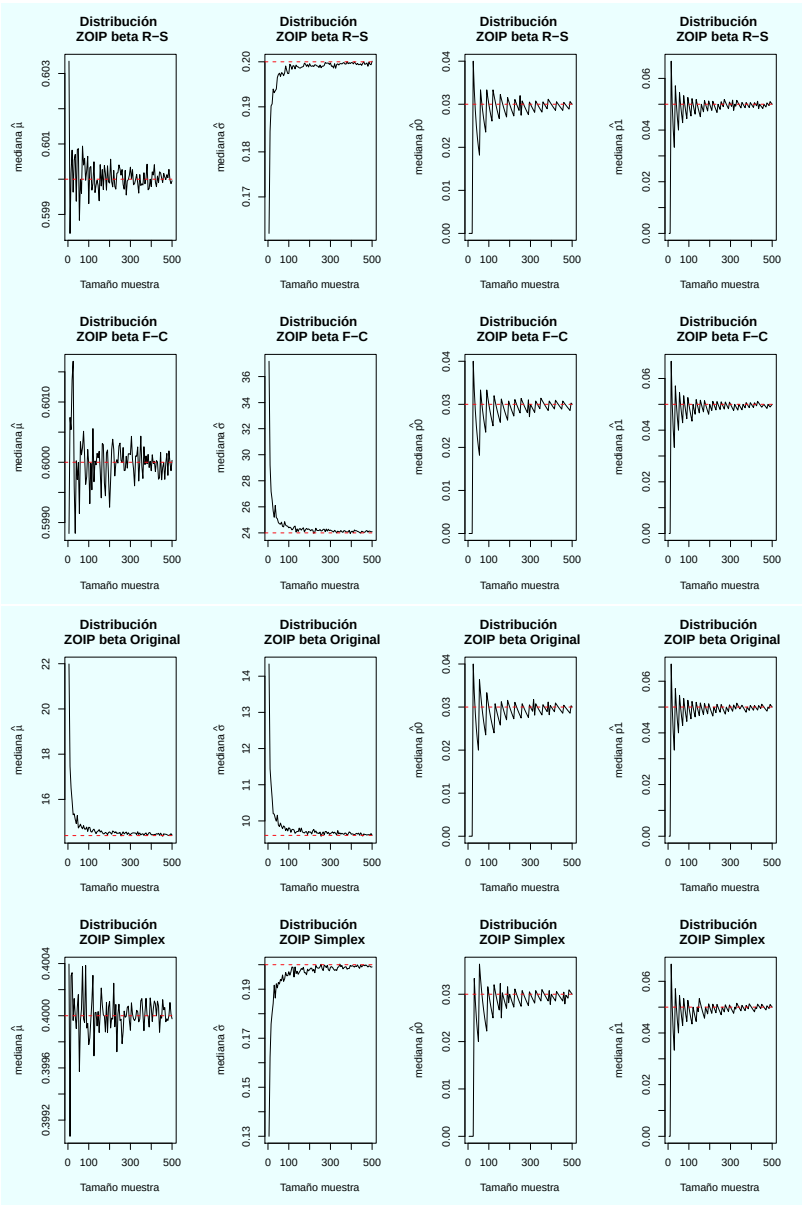

Figura 2: Mediana de los parámetros estimados en el escenario 1 para distintas parametrizaciones y valores de n, las líneas rojas representan el verdadero valor del parámetro.

de distribución a distribución.

Como medida global del proceso de estimación se eligió el MAPE (Error porcentual absoluto medio. $(\sum_{i=1}^{n} |y_i - \hat{y_i}/y_i|)/n)$ debido a los cambios de escala entre los diferentes parámetros de las diferentes distribuciones y parametrizaciones. Esta media se realizó como un promedio de los MAPES generados por cada uno de los parámetros de la distribución ZOIP en cada tamaño de muestra. En la Figura 4a se presenta el MAPE para las diferentes distribuciones y parametrizaciones estimadas, se observa como a medida que el tamaño de muestra aumenta, el MAPE va decreciendo rápidamente, aunque después de un tamaño de muestra de 200, el MAPE decrece de una manera más lenta, además los errores de estimación son muy parecidos entre los 4 casos de simulación, la estimación sobre los parámetros de la distribución ZOIP-simplex tiene un error un poco más grande, pero no es significativo sobre los demás casos.

En el segundo escenario de simulación se generó el mismo ejercicio de simulación anterior sobre las mismas distribuciones y parametrizaciones, solo que los valores de p_0 y p_1 cambian por 0.3 y 0.2, respectivamente. Dando así que el 50 % de los datos se vean contaminados por ceros y unos, esto para ver si de alguna forma afecta el aumento de la presencia de ceros y unos sobre las estimaciones de los parámetros de la parte continua de la distribución ZOIP.

En la Figura 3 se presentan las estimaciones de los parámetros de la simulación con inflaciones al 50 % para diferentes tamaños de muestras, en general se observa que no se ven cambios muy significativos sobre la Figura 2 en los parámetros de μ y σ, sin embargo, en la estimación de p_0 se tienden a dar valores subestimados con relación al estudio de simulación anterior y con el parámetro p_1 aunque las estimaciones son muy acertadas sobre el valor real desde tamaños de muestra pequeños, en algunas ocasiones se producen pequeñas perturbaciones no muy alejados del valor real.

Parámetro	Caso	MAPE escenario 1 %	MAPE escenario 2 %
μ	Caso R-S	0.61	0.86
	Caso F-C	0.50	0.85
	Caso original	0.53	0.70
	Caso simplex	0.47	0.63
σ	Caso R-S	2.53	3.40
	Caso F-C	5.10	6.90
	Caso original	5.30	6.98
	Caso simplex	5.30	7.37
p_0	Caso R-S	20.5	5.36
	Caso F-C	19.7	5.42
	Caso original	19.8	5.43
	Caso simplex	20.8	5.51
p_1	Caso R-S	15.2	7.28
	Caso F-C	16	7
	Caso original	15.7	7
	Caso simplex	16.2	7.12
	Promedio	10.57	5.26

Tabla 1: MAPE de las estimaciones para cada parámetro en diferentes parametrizaciones en los dos estudios de simulación.

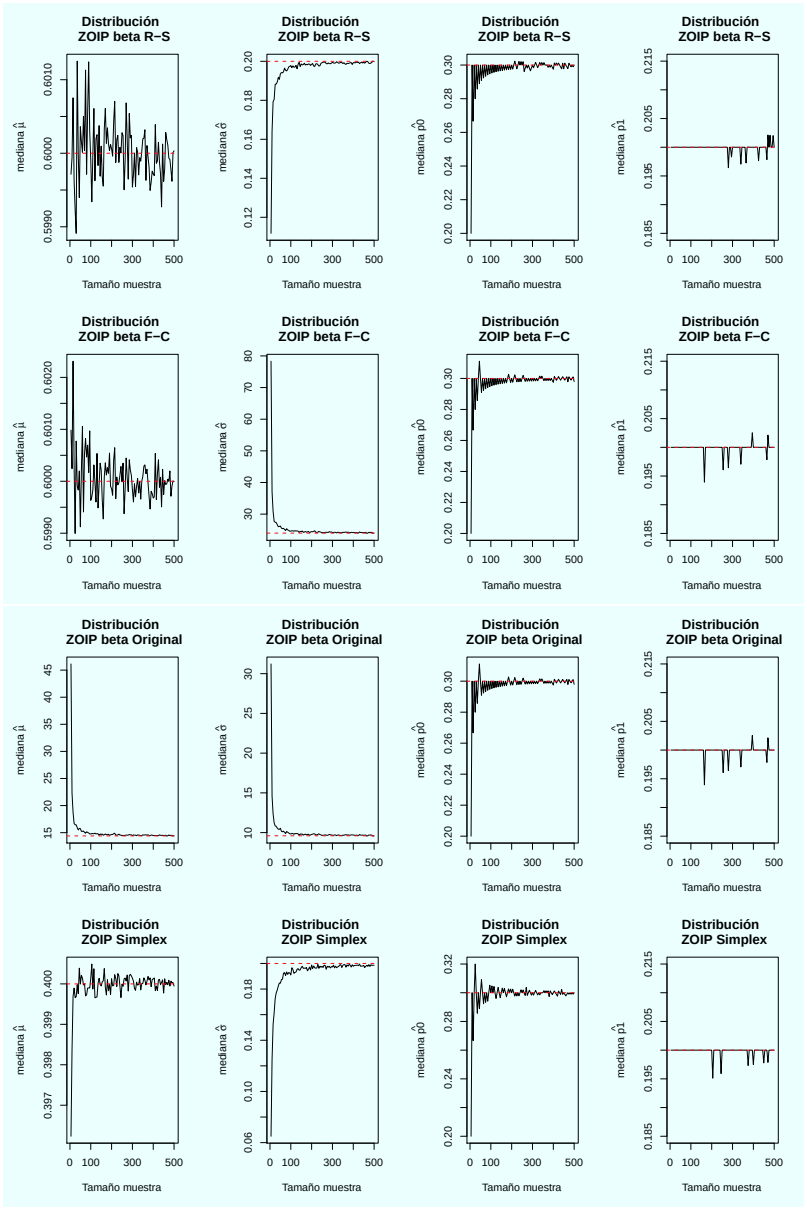

Figura 3: Simulación de distribución ZOIP para distintas parámetrizaciones con parámetros de inflación grandes, distribuciones y valores de n.

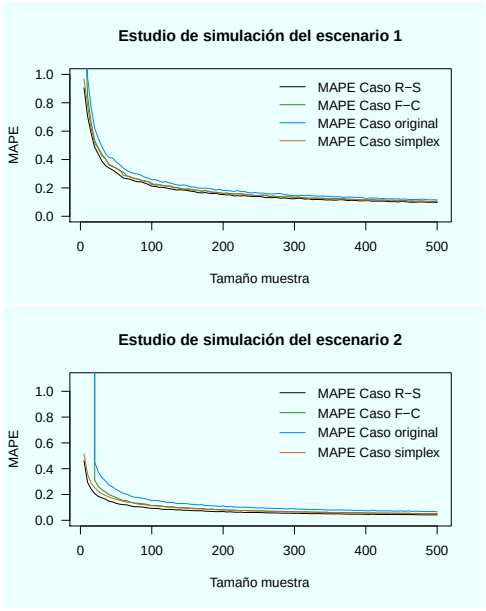

Figura 4: Mape (Error porcentual absoluto medio) para los dos escenarios de simulación y para distintas parametrizaciones y valores de n.

En la Figura 4b se presenta el MAPE para el estudio de simulación del escenario 2, se puede ver como se obtienen MAPES muy parecidos a los del estudio de simulación del escenario 1, pero cabe resaltar como se comete menos error sobre la estimación de los parámetros de la distribución beta con parametrización Rigby & Stasinopoulos (2005). En la tabla 1 se presenta el MAPE para cada parámetro de cada parametrización para ambos estudios de simulación, es claro ver como en general el estudio de simulación del escenario 2 produce un MAPE menor que el del escenario 1, esto es causado por que en el escenario 1 de simulación los errores de pronóstico son más grandes en los parámetros de inflación que en el escenario 2. Por todo lo visto anteriormente se puede concluir que el crecimiento de los parámetros de inflación no afecta de manera significativa la estimación de los parámetros de la parte continua de la distribución ZOIP, pero si en una mejor estimación de los parámetros de inflación.

6.2. Datos reales

En esta sección se presenta el ajuste de una distribución ZOIP a datos reales sobre la utilización de una tarjeta de crédito en un banco, para una entidad financiera grande como un banco es de vital importancia conocer el comportamiento del porcentaje de utilización de sus tarjetas de crédito (tdc), se define a y como el porcentaje de uso de una tdc, en la Figura 5 se presenta el histograma del por-

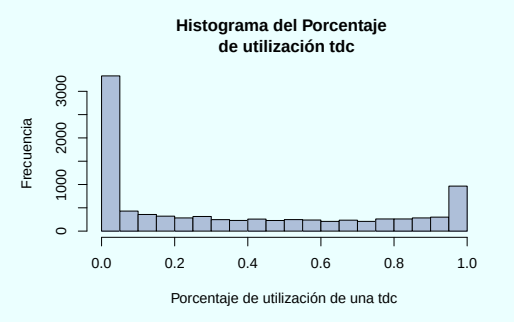

Figura 5: Histograma del porcentaje de utilización de las tdc en un banco.

centaje de utilización de las tdc y es claro notar que y se encuentra entre cero y uno, pero adicional es muy común ver que las tdc no sean utilizadas ($y = 0$) y también que las tdc sean utilizadas en la totalidad de su cupo asignado ($y = 1$), por lo que se trata a y como una variable aleatoria perteneciente a datos proporcionales inflados con ceros y unos. Se tiene un total de 9206 tdc, que representan el porcentaje de utilización de las tdc para un trimestre del año 2014 del banco. Se quiere estudiar el ajuste de una distribución ZOIP, para ello se utiliza el paquete en R llamado **ZOIP** mediante su función `RM.ZOIP`.

Familia	Parámetro	Estimación	Error estándar	Valor P	Log-Verosimilitud
R-S	μ	0.4040	0.0037	$< 2.2e^{-16}$	5854.067
	σ	0.6601	0.0027	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
F-C	μ	0.4040	0.0037	$< 2.2e^{-16}$	5854.067
	σ	0.4040	0.0037	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
original	μ	0.5233	0.0080	$< 2.2e^{-16}$	5854.067
	σ	0.7719	0.0130	$< 2.2e^{-16}$	
	p_0	0.2219	0.0043	$< 2.2e^{-16}$	
	p_1	0.0695	0.0027	$< 2.2e^{-16}$	
simplex	μ	0.5741	0.0010	$< 2.2e^{-16}$	54425.63
	σ	4885.4370	18.2430	$< 2.2e^{-16}$	
	p_0	0.1497	0.0032	$< 2.2e^{-16}$	
	p_1	0.0090	0.0004	$< 2.2e^{-16}$	

Tabla 2: Ajuste de diferentes distribuciones ZOIP en el porcentaje de utilización de una tdc.

En la Tabla 2 se muestran resultados de los cuatro parámetros estimados vía máxima verosimilitud para la distribución ZOIP, en ellas vemos cómo cambian los valores de los parámetros según la parametrización escogida, los valores de log-verosimilitud no indican que el mejor modelo ajustado es un ZOIP-beta, ya que es bastante menor el valor de log-verosimilitud de una distribución ZOIP-simplex,

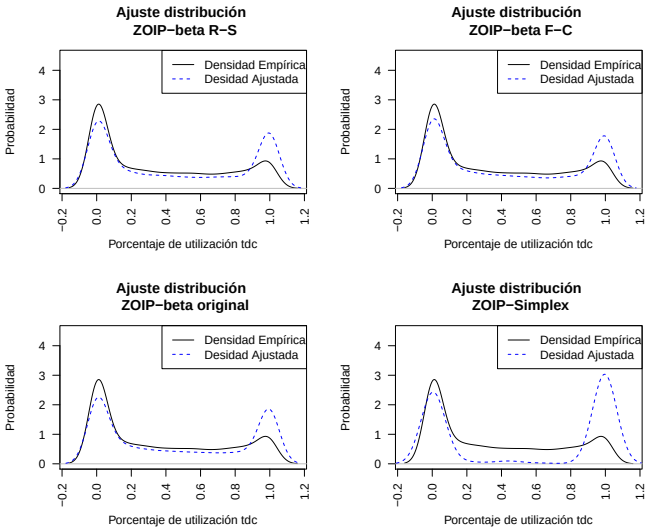

Figura 6: Ajuste de diferentes distribuciones y parametrizaciones ZOIP al porcentaje de utilización de una tdc.

además que en las estimaciones de los parámetros de la distribución ZOIP-simplex no se tuvo una convergencia, por lo tanto los valores son muy distintos para el parámetro de dispersión a los vistos en la distribución ZOIP-beta, inclusive muy elevados. Además, el valor de μ es mayor que las de la parametrización en Rigby & Stasinopoulos (2005) y Ferrari & Cribari-Neto (2004), un 17 % más.

En la Figura 6 se presenta gráficamente el ajuste de la distribución ZOIP para diferentes parametrizaciones al porcentaje de utilización de las tdc, la línea azul que representa la distribución ZOIP ajustada, es de notar que dicha línea azul es exactamente igual en las tres ocasiones que se ajusta la distribución ZOIP-beta y se ve como sigue el comportamiento original del porcentaje de utilización de las tdc. Es bueno resaltar también como en la Figura 1d de la Figura 6 no se nota un buen ajuste para los valores entre cero y uno. Por todo anterior es recomendable decir que el porcentaje de utilización de las tdc de este banco se comportan como una distribución ZOIP-beta con los parámetros descritos en la tabla 2, según la parametrización deseada y no como una distribución ZOIP-simplex.

7. Conclusiones

La distribución ZOIP y el paquete **ZOIP** de R permiten ajustar distribuciones para datos provenientes de porcentajes, tasas o proporciones que se encuentren infla-

dos con ceros y/o unos, dicha distribución está compuesta por cuatro parámetros, que son estimados vía máxima verosimilitud y en el cual de acuerdo a los estudios de simulación realizados estos convergen a los valores reales con un tamaño de muestra relativamente pequeño, además se observa como la estimación de los parámetros de la parte continua no se ven afectados por el aumento de la presencia de ceros y unos en los datos, pero si la estimación de los parámetros de la parte discreta. Por otra parte, se observa como el Ajuste de la distribución ZOIP-beta explica el comportamiento de la distribución del porcentaje de utilización de una tarjeta de crédito en un banco.

La distribución ZOIP y el paquete **ZOIP** de R permiten de una manera muy versátil utilizar y ajustar diferentes parametrizaciones y distribuciones para datos proporcionales. Además permite Utilizar y ajustar distribuciones para datos proporcionales que se encuentran inflados solo con ceros o solo con unos, de una manera práctica.

Referencias

Ferrari, S. & Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of applied statistics* **31**(7), 799–815.

Galvis, D. & Lachos, V. (2014), Bayesian analysis of regression models for proportional data in the presence of zeros and ones, PhD thesis, Estatal University of Campinas.

Hahn, E. (2008), 'Mixture densities for project management activity times: A robust approach to pert', *European Journal of operational research* **188**(2), 450–459.

Hossain, A. (2015), 'Modelling a proportion response variable using generalized additive models for location scale and shape', *International Workshop on Statistical Modelling* **2**, 137–140.

Jørgensen, B. (1997), 'The theory of dispersion models', *Computational statistics and Data analysis* **76**.

Jørgensen, B. & Barndorff-Nielsen (1991), 'Some parametric models on the simplex', *Journal of multivariate analysis* **39**(1), 106–116.

Kosmidis, I., Zou, A. & Jagan, K. (2015), 'Beyond beta regression: modelling percentages and fractions in the presence of boundary observations', *International Workshop on Statistical Modelling* **1**, 267–272.

Ospina, R. & Ferrari, S. (2010), 'Inflated beta distributions online: doi:10.1007/s00362-008-0125-4', *Statitical papers* .

Ospina, R. & Ferrari, S. (2012), 'A general class of zero-or-one inflated beta regression models', *Computational Statistics and Data Analysis* **56**(6), 1609–1623.

Owen, C. (2008), Parameter estimation for the beta distribution, Master's thesis, Brigham young university.

Pace, L. & Salvan, A. (1997), 'Principles of statistical inference from a neo-fisherian perspective. in: Advanced series on statistical science and applied probability', *World Scientific* **4**, 128.

Qiu, Z., Song, P. & Tan, M. (2008), 'Simplex mixed-effects models for longitudinal proportional data', *Scandinavian Journal of Statistics* **35**(4), 577–596.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/

Rigby, B. & Stasinopoulos, M. (2005), 'Generalized additive models for location, scale and shape', *Applied Statistical* **54**(3), 507–554.

Rigby, B. & Stasinopoulos, M. (2008), 'Instructions on how to use the gamlss package in r', *Computational statistics and Data analysis* **2**, 194–195.

Zeileis, A., Cribari-Neto, F. & Grün, B. (2010), 'Beta regression in r', *Journal of statistical software* **34**(2), 1–24.

Zhang, P., Qiu, Z. & Shi, C. (2016), 'simplexreg: An r package for regression analysis of proportional data using the simplex distribution', *Journal of Statistical Sftware* **71**(11).