Neiva, 25 de julio de 2019

Señores

CENTRO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD SURCOLOMBIANA

Ciudad

El (Los) suscrito(s):

Pedro Francisco Morales García, con C.C. No. 7.711.484

Jairo Alfonso Hermosa Trujillo , con C.C. No. 1.079.173.721,

Autor(es) de la tesis y/o trabajo de grado; Titulado **Análisis comparativo de las pruebas saber 11 en las comunas 6 y 10 de Neiva mediante Regresión Beta.** Presentado y aprobado en el año 2019 como requisito para optar al título de Especialista en Estadística;

Autorizo (amos) al CENTRO DE INFORMACIÓN Y DOCUMENTACIÓN de la Universidad Surcolombiana para que, con fines académicos, muestre al país y el exterior la producción intelectual de la Universidad Surcolombiana, a través de la visibilidad de su contenido de la siguiente manera:

- Los usuarios puedan consultar el contenido de este trabajo de grado en los sitios web que administra la Universidad, en bases de datos, repositorio digital, catálogos y en otros sitios web, redes y sistemas de información nacionales e internacionales “open access” y en las redes de información con las cuales tenga convenio la Institución.

- Permita la consulta, la reproducción y préstamo a los usuarios interesados en el contenido de este trabajo, para todos los usos que tengan finalidad académica, ya sea en formato Cd-Rom o digital desde internet, intranet, etc., y en general para cualquier formato conocido o por conocer, dentro de los términos establecidos en la Ley 23 de 1982, Ley 44 de 1993, Decisión Andina 351 de 1993, Decreto 460 de 1995 y demás normas generales sobre la materia.

- Continúo conservando los correspondientes derechos sin modificación o restricción alguna; puesto que, de acuerdo con la legislación colombiana aplicable, el presente es un acuerdo jurídico que en ningún caso conlleva la enajenación del derecho de autor y sus conexos.

De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, “Los derechos morales sobre el trabajo son propiedad de los autores” , los cuales son irrenunciables, imprescriptibles, inembargables e inalienables.

EL AUTOR/ESTUDIANTE:

Pedro Francisco Morales Garcia

Firma: Pedro Fco Morales

EL AUTOR/ESTUDIANTE:

Jairo Alfonso Hermosa Trujillo

Firma:

TÍTULO COMPLETO DEL TRABAJO: Análisis comparativo de las pruebas saber 11 en las comunas 6 y 10 de Neiva mediante Regresión Beta.

AUTOR O AUTORES:

Primero y Segundo Apellido	Primero y Segundo Nombre
Morales García	Pedro Francisco
Hermosa Trujillo	Jairo Alfonso

DIRECTOR Y CODIRECTOR TESIS:

Primero y Segundo Apellido	Primero y Segundo Nombre
Sánchez Hernández	Alfonso

ASESOR (ES):

Primero y Segundo Apellido	Primero y Segundo Nombre
Sánchez Hernández	Alfonso

PARA OPTAR AL TÍTULO DE: Especialista en Estadística

FACULTAD: Ciencias Exactas y Naturales

PROGRAMA O POSGRADO: Especialización en Estadística

CIUDAD: **Neiva** **AÑO DE PRESENTACIÓN:** 2019 **NÚMERO DE PÁGINAS: 42**

TIPO DE ILUSTRACIONES (Marcar con una **X**)**:**

Diagramas___ Fotografías___ Grabaciones en discos___ Ilustraciones en general___ Grabados___ Láminas___ Litografías___ Mapas___ Música impresa___ Planos___ Retratos___ Sin ilustraciones___ Tablas o Cuadros_x_

SOFTWARE requerido y/o especializado para la lectura del documento:

MATERIAL ANEXO:

PREMIO O DISTINCIÓN *(En caso de ser LAUREADAS o Meritoria)*:

PALABRAS CLAVES EN ESPAÑOL E INGLÉS:

Español	Inglés	Español	Inglés
1. Saber 11	Saber 11th	**6.** Estadística no paramétrica	Nonparametric Statistics
2. Comuna	Commune	**7.** Comparación múltiple	Multiple comparison
3. Estudio comparativo	Comparative study	**8.** Regresión beta	Beta regression
4. Análisis descriptivo	Descriptive analysis	**9.** Inferencia Bayesiana	Bayesian inference
5. Prueba de normalidad	Normality test	**10.** Distribución multivariada	Multivariate distribution

RESUMEN DEL CONTENIDO: (Máximo 250 palabras)

La investigación realizada, se realiza un estudio comparativo de los resultados de la prueba saber once, en las áreas de matemáticas, lectura crítica, inglés y sociales, entre instituciones educativas oficiales de las comunas seis y diez de la ciudad de Neiva, para la realización de este estudio comparativo, se desarrolló un análisis descriptivo, pruebas de normalidad, pruebas de comparación múltiple, estadística no paramétrica y por ende en uso de la regresión Beta, dado que los resultados utilizados se basan en índices.

Sé lograron determinar los mejores modelos de regresión Beta, usando áreas del conocimiento y factores asociados, estableciendo medidas de influencia estadística para medir la sensibilidad de los modelos estimados. Los factores asociados a cada uno de los índices, resultaron ser los apropiados y altamente significativos. Con estos modelos se pueden establecer condiciones y evaluar acciones que lleven a formular estrategias para el mejoramiento de los resultados en las pruebas saber once de las instituciones educativas oficiales de las comunas seis y diez de Neiva.

ABSTRACT: (Máximo 250 palabras)

The hereby research makes a comparative study of the results of the Prueba Saber Once in mathematics, critical reading, English and social studies between the official educational institutions of the communes six and ten of the city of Neiva in Huila, Colombia.

For the comparative performance, normality analysis, multiple comparison tests, non-parametric statistics and the use of Beta regression are analyzed, since the results are based on indexes.

This study made possible to determine the best Beta regression models, using areas of knowledge and associated factors, establishing statistical influence measures to measure the sensitivity of the estimated models. The associated factors to each one of the indexes, proved to be adequate and highly significant. By using these models, conditions can be established and actions can be evaluated to carry out formulas for the improvement of the results in the Prueba Saber Once into the official educational institutions of the communes six and ten of Neiva.

APROBACION DE LA TESIS

Nombre Presidente Jurado: Jaime Polanía Perdomo

Firma: _____

Nombre Jurado: Carlos Arturo Monje Álvarez

Firma: _____

UNIVERSIDAD
SURCOLOMBIANA

ANÁLISIS COMPARATIVO DE LAS PRUEBAS SABER 11 EN LAS COMUNAS 6 Y 10 DE NEIVA MEDIANTE REGRESION BETA

Pedro Francisco Morales
Jairo Alfonso Hermosa

Especialización en Estadística
Facultad de Ciencias
UNIVERSIDAD SURCOLOMBIANA
NEIVA
2019

ANÁLISIS COMPARATIVO DE LAS PRUEBAS SABER 11 EN LAS COMUNAS 6 Y 10 DE NEIVA MEDIANTE REGRESION BETA

Pedro Francisco Morales
Jairo Alfonso Hermosa

Trabajo de grado para optar al título de
Especialista en Estadística

Director:
Alfonso Sánchez Hernández
MSc. en Investigación Operativa y Estadística

UNIVERSIDAD SURCOLOMBIANA
FACULTAD DE CIENCIAS
NEIVA
2019

Índice General

Índice de tablas

Lista de Figuras

1 Introducción

Las pruebas Saber Pro (11), anteriormente conocidas como las pruebas del ICFES, son el preámbulo para que el futuro bachiller tenga la facilidad de ingresar a la educación superior. En cualquier ámbito se podría llegar a pensar que sólo se trata de un examen de conocimientos, no obstante el verdadero ejercicio con estas pruebas consiste en medir la capacidad de abstracción, habilidad, comprensión y conocimientos alcanzados por el estudiante, en las distintas áreas del conocimiento, basadas en la formación obtenida por el individuo en las dos primeras fases de estudio.

En todas las instituciones de educación básica y media, tanto públicas como privadas del país, éstas pruebas son de caracter obligatorio, e involucran no sólo al indivividuo, sino también a los profesores, directivos, a la familia y por ende a la sociedad.

El Ministerio de Educación Nacional, a través del Instituto Colombiano para el Fomento de la Educación Superior (ICFES), en el primer semestre del año 2019, lanzó la propuesta para que los grupos de investigación, adscritos a Instituciones públicas y privadas del país, inscritas en Colciencias, utilicen las bases de datos de la pruebas Saber, a través de proyectos de investigación que promuevan el estudio y desarrollo de la Ciencia y la Tecnología.

Es inevitable reconocer que aquellas regiones del país en las que los resultados de las pruebas Saber son mayores, el desarrollo industrial, tecnológico y académico supera con creces a las regiones menos favorecidas y en las que los resultados

no son del todo, los mejores.

Cabe resaltar que en cada región del país y en su entorno se presentan en mayor o en menor grado, situaciones que afectan de manera directa o indirecta un buen resultado en las pruebas Saber, pueden ser problemas de índole familiar, personal, social, económico e incluso afectivo.

Tomando como base las anteriores reflexiones, usando la base de datos suministrada por el ICFES, el presente proyecto pretende realizar un estudio comparativo de los resultados de las pruebas Saber 11, en las áreas de Matemáticas, Lectura Crítica, Inglés y Sociales, entre instituciones pertenecientes a las comunas 6 y 10 de Neiva.

Dentro de las técnicas estadísticas que se pretender utilizar en el presente proyecto están: análisis descriptivo, pruebas de normalidad, pruebas de comparación múltiple, estadística no paramétrica, y por ende el uso de la Regresión Beta, dado que los resultados utilizados se basan en índices. Vale la pena destacar que este tipo de herramienta estadística, se ha venido utilizando con éxito desde 2004, año en el cual Ferrari y Cribari-Neto (2004, [6]), proponen esta herramienta con el fin de estudiar: tasas de homicidios, hurtos, accidentalidad, reclamos bancarios y de la construcción en Brasil, obteniendo resultados asombrosos que fueron posteriormente comparados con análisis bayesianos, los cuales permitieron descubrir la alta fiabilidad de los resultados obtenidos a través de este tipo de regresión.

2 Objetivos

2.1 Objetivo General

Realizar un estudio comparativo de las pruebas saber 11, en las Comunas 6 y 10 de Neiva mediante *Regresión Beta*, años 2014 - 2018.

2.2 Objetivos Específicos

- Desarrollar un estudio estadístico descriptivo general que permita comparar promedios y variaciones en los índices para cinco áreas del conocimiento.
- Identificar factores de interés asociados con el rendimiento en las cinco áreas del conocimiento en las comunas 6 y 10 de Neiva.
- Implementar pruebas de normalidad para los índices de conocimiento y verificar relaciones y comparaciones entre los mismos.
- Utlizar herramientas estadísticas alternativas con el fin de establecer relaciones y comparaciones, en donde no se pueda verfificar normalidad.
- Determinar los mejores modelos de regresión beta, usando áreas del conocimiento y factores asociados, estableciendo medidas de influencia estadística para medir las sensibilidad de los modelos estimados.

3 Marco Teórico

3.1 Modelo de Regresión Beta Univariado

El presente marco teórico se basa fundamentalmente en los trabajos de Ferrari y Cribari-Neto (2004) quienes propusieron un modelo de regresión, en el que la variable respuesta es continua y está restringida al intervalo $(0, 1)$ la cual está relacionada con otras variables, a través de una estructura de regresión. El modelo propuesto está basado en suponer, que la variable respuesta tiene distribución beta, utilizando una parametrización de la ley beta, de acuerdo a la media y al parámetro de precisión.

La distribución beta es muy flexible en situaciones en donde la variable dependiente es continua y está restringida al intervalo $(0, 1)$ debido a que su función de densidad puede asumir diferentes formas, dependiendo del parámetro que la acompañe. Una función de densidad de distribución beta, con parámetros a, b está definida como:

si $y \sim Beta(a, b)$

$$f_Y(y) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \quad 0 < y < 1 \quad a, b > 0$$

como:

$$Beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

se tiene que:

$$\frac{1}{B(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

entonces:

$$f_Y(y;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \quad 0<y<1 \quad a,b>0$$

de donde:

$$\mu = E(y) = \frac{a}{a+b} \;\; y \;\; Var(y) = \frac{ab}{(a+b)^2(a+b-1)}$$

luego:

$$\mu = \frac{a}{a+b} \quad sea \;\; \phi = a+b \Longrightarrow \mu = \frac{a}{\phi} \Longrightarrow \mu\phi = a$$

es decir:

$$\phi = \mu\phi + b \Longrightarrow \phi - \mu\phi = b \Longrightarrow \phi(1-\mu) = b$$

Ahora:

$$f(y;a,b) = \frac{\Gamma(\mu\phi + \phi(1-\mu))}{\Gamma(\mu\phi)\Gamma\phi(1-\mu)} y^{\mu\phi-1}(1-y)^{\phi(1-\mu)-1}$$

$$f(y;a,b) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\phi(1-\mu)}\, y^{\mu\phi-1}(1-y)^{\phi(1-\mu)-1} \; 0<y<1 \; 0<\mu<1 \; \phi>0$$

Así la media y la varianza están dadas por:

$$E(y) = \mu \quad y \quad Var(y) = \tfrac{V(y)}{1+\phi}$$

Donde $V(\mu) = \mu(1-\mu)$ está dada en función de la varianza, μ es la media de la variable respuesta y ϕ puede ser interpretado como un parámetro de precisión, cuanto mayor sea el valor de ϕ, menor es la varianza de y. Entonces el modelo propuesto por Ferrari y Cribari-Neto (2004) queda de la siguiente manera:

$$y_i|\mu_i,\phi \sim Beta(\mu_i,\phi), \;\; i = 1,2,3,...,n$$

$$g(\mu_i) = \eta_i = \sum_{k=1}^{p} x_{ik}\beta_k,\ p < n$$

donde n es el número de observaciones, p es el número de coeficientes de regresión, g(.) es una función estrictamente monótona y dos veces diferenciables que mapea el intervalo $(0, 1)$ en $\Re$, $\beta^T = (\beta_1, ..., \beta_p)$ es un vector de coeficientes de regresión y $x_{i1}, ..., x_{ip}$ son observaciones de p covariables, $i = 1, 2, ..., n$

3.2 Inferencia Bayesiana en el modelo de Regresión Beta Univariado

Para estimar las parámetros del modelo propuesto por Ferrari y Cribari-Neto (2004), ellos adoptaron un enfoque frecuentista, para este caso se hace de una manera totalmente Bayesiana, en la explicación del modelo y se terminará asignando distribuciones a priori, $p(\beta, \phi)$ para $\beta\ y\ \phi$. Para realizar inferencia Bayesiana sobre estos parámetros, se supone que se tiene información sobre las variables aleatorias de $Y_1, ..., Y_n$, cuyos valores pertenecen al intervalo $(0, 1)$, el cual es equivalente al intervalo (c, d), transformando las variables en $(Y_1/(d-c), ..., Y_n/(d-c))$. Sea $\beta = (\beta_1, ..., \beta_p)^T$ y $g(\mu_i) = logit(\mu_i)$, con esta información, se tiene que la densidad de la distribución beta, sobre la parametrización, del modelo propuesto por Ferrari y Cribari-Neto (2004), para la i-ésima observación queda dado de la siguiente manera:

$$
\begin{aligned}
f(y_i|\boldsymbol{\beta},\phi) =& \frac{\Gamma(\phi)}{\Gamma(\mu_i(\boldsymbol{\beta})\phi)\Gamma((1-\mu_i(\boldsymbol{\beta}))\phi)} y_i^{\mu_i(\boldsymbol{\beta})_{\phi-1}}(1-y_i)^{[1-\mu_i(\boldsymbol{\beta})]_{\phi-1}} \\
& \propto \exp\big\{log\Gamma(\phi) - [log\Gamma(\mu_1(\boldsymbol{\beta})\phi + log\Gamma((1-\mu_i(\boldsymbol{\beta}))\phi)] + \\
& \mu_i(\boldsymbol{\beta})\phi logit(y_i) + \phi log(1-y_i)\big\}
\end{aligned}
$$

Por consiguiente se tiene que la función de verosimilitud $L(\beta,\phi)$, para n observaciones independientes es:

$$
\begin{aligned}
L(\boldsymbol{\beta},\phi) =& \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta},\phi) \\
=& \Gamma(\phi)^n \prod_{i=1}^{n} \frac{1}{\Gamma(\mu_1(\boldsymbol{\beta})\phi)\Gamma((1-\mu_i(\boldsymbol{\beta}))\phi)} y_i^{\mu_i(\boldsymbol{\beta})_{\phi-1}}(1-y_i)^{[1-\mu_i(\boldsymbol{\beta})]_{\phi-1}}
\end{aligned}
$$

Entonces, considerando una densidad a priori $p(\beta,\phi)$, se tiene que la densidad a posteriori de estos parámetros es tal que:

$$
\begin{aligned}
p(\beta,\phi|y) &\propto L(\beta,\phi)p(\beta,\phi) \\
&\propto [\Gamma(\phi)]^n \prod_{i=1}^{n} \frac{1}{\Gamma(\mu_1(\beta)\phi)\Gamma((1-\mu_i(\beta))\phi)} y_i^{\mu_i(\beta)_{\phi-1}}(1-y_i)^{[1-\mu_i(\beta)]_{\phi-1}} p(\beta,\phi).
\end{aligned}
$$

Hay diferentes formas para escoger la densidad a priori $p(\beta,\phi)$. Una de esas formas es considerar $p(\beta,\phi) = p(\beta)p(\phi)$, lo que equivale a ϕ y β independientes a priori. Por ejemplo se puede adoptar $\beta_k \sim N(m_k,\sigma_k^2)$ y $\phi \sim$

$Gama(a,b),\ pues\ \beta_k \in (-\infty,\infty)$,

$k = 1,...,p\ \ y\ \ \phi > 0$, donde resulta el modelo de regresión beta univariado, dado por:

$$y_i|\mu_i,\phi \sim Beta(\mu_i(\beta),\phi),\ \ i = 1,2,3,...,n$$

$$g(\mu_i) = \eta_i = \sum_{k=1}^{p} x_{ik}\beta_k,\ \ p < n$$

$$\beta_k \sim N(m_k,\sigma_k^2)$$

$$\phi \sim Gama(a,b),$$

con distribuciones a priori para ϕ y β_k, k= 1,...,p.

Los términos desarrollados $\Gamma(\mu_i(\beta)\phi)$ y $\Gamma(\phi)$ indican que la densidad a posteriori de β y ϕ no poseen forma cerrada. Para generar muestras de esa distribución, se utiliza el método de Monte Carlo vía cadenas de Markov. El uso de las distribuciones a priori mencionadas, llevan a la obtención de distribuciones condicionales completas a posteriori para ϕ y β también de forma desconocida. Por lo tanto un método eficaz para la simulación de generar distribuciones a posteriori es el algoritmo Metrópolis-Hastings.

Las distribuciones condicionales completas a posteriori del modelo descrito anteriormente quedan de la siguiente manera:

$$p(\phi|\boldsymbol{\beta},\, \mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\beta},\phi)\, \mathbf{p}(\phi)$$
$$\propto \exp\Bigg\{ n\, log\boldsymbol{\Gamma}(\phi) - \sum_{i=1}^{n}\big[log\boldsymbol{\Gamma}(\mu_i(\boldsymbol{\beta})\phi) + log((1-\mu_i(\boldsymbol{\beta}))\phi)\big] +$$
$$\phi \sum_{i=1}^{n}[\mu_i(\boldsymbol{\beta})\text{logit}(y_i) + log(1-y_i)] \Bigg\} \times \phi^{a-1}\exp(-b\phi)$$

y

$$p(\boldsymbol{\beta}_k|\beta_{-k}\,\phi,\, \mathbf{y}) \propto L(y|\boldsymbol{\beta},\phi)p(\beta_k)$$
$$\propto \exp\Bigg\{ n\, log\boldsymbol{\Gamma}(\phi) - \sum_{i=1}^{n}\big[log\boldsymbol{\Gamma}(\mu_i(\boldsymbol{\beta})\phi) + log((1-\mu_i(\boldsymbol{\beta}))\phi)\big] +$$
$$\phi \sum_{i=1}^{n}\big[\mu_i(\boldsymbol{\beta})\text{logit}(y_i) + log(1-y_i)\big] \Bigg\} \times \exp\Bigg\{ -\frac{1}{2}\sum_{k=1}^{p}\Big(\frac{\beta_k - m_k}{\sigma^k}\Big)^2,$$

para $k = 1, ..., p$ donde $\beta_{(-k)}$ es obtenido eliminando el k-ésimo término del vector β.

3.3 Distribuciones multivariadas con marginales beta

Hay situaciones en las que existen más de una variable dependiente y a su vez están relacionadas entre sí. El análisis multivariado se considera muy apropiado para establecer dicha asociación entre esas variables. Por lo tanto habrá una

mejor ganancia en la precisión de cantidades de interés, cuando los modelos son analizados conjuntamente y no de manera individual.

Hay varias distribuciones en la literatura denomidas beta multivariadas, pero la más apropiada sería una distribución conjunta cuya marginales posean distribución beta univariada.

Olkin y Liu (2003) definieron una función de densidad beta bivariada, utilizando una relación entre la distribución Beta y Gama de la siguiente manera:

sean:

$U \sim Gama(a,1)$ $V \sim Gama(b,1)$ y $W \sim Gama(c,1)$, variables aleatorias independientes.

Dónde:

$X = \frac{U}{U+W}$ y $Y = \frac{V}{V+W}$

Así: $X \sim Beta(a,c)$ y $Y \sim Beta(b,c)$

Ahora:

$$Si \;\; U \sim Gama(a,1) \Longrightarrow f_U(u) = \frac{1^a}{\Gamma(a)} u^{a-1} e^{-u} \, I_{(0,\infty)}(u)$$

$$Si \;\; V \sim Gama(b,1) \Longrightarrow f_V(v) = \frac{1^b}{\Gamma(b)} v^{b-1} e^{-v} \, I_{(0,\infty)}(v)$$

$$Si \;\; W \sim Gama(c,1) \Longrightarrow f_W(w) = \frac{1^c}{\Gamma(c)} w^{c-1} e^{-w} \, I_{(0,\infty)}(w)$$

Es decir:

$$f(u,v,w) = \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} u^{a-1} v^{b-1} w^{c-1} e^{-u} e^{-v} e^{-w} \;\; u,v,w > 0$$

Como:

$$X = \frac{U}{U+W} \;\; y \;\; Y = \frac{V}{V+W}$$

Entonces:

$x(u+w) = u \Longrightarrow xu+xw = u \Longrightarrow xw = u-xu \Longrightarrow xw = u(1-x) \Longrightarrow u = \dfrac{x}{1-x}w$

$y(v+w) = v \Longrightarrow yv+yw = v \Longrightarrow yw = v-yv \Longrightarrow yw = v(1-y) \Longrightarrow v = \dfrac{y}{1-y}w$

Reemplazando:

$$f(x,y,w) = \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)}\left(\frac{x}{1-x}w\right)^{a-1}\left(\frac{y}{1-y}w\right)^{b-1} w^{c-1}e^{-wq} \quad 0<x,y<1, w>0$$

Dónde:

$q = \frac{x}{1-x} + \frac{y}{1-y} + 1 = \frac{(1-xy)}{(1-x)(1-y)}$

Ahora integrando:

$$f(x,y) = \int_0^\infty f(x,y,w)dw$$
$$f(x,y) = \int_0^\infty \left(\frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)}\left(\frac{x}{1-x}w\right)^{a-1}\left(\frac{y}{1-y}w\right)^{b-1} w^{c-1}e^{-wq}\right)dw$$
$$f(x,y) = \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)}\left(\frac{x}{1-x}\right)^{a-1}\left(\frac{y}{1-y}\right)^{b-1}\int_0^\infty w^{a+b+c-3}e^{-wq}dw$$
$$f(x,y) = \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)}\frac{x^{a-1}}{(1-x)^{a-1}}\frac{y^{b-1}}{(1-y)^{b-1}}\int_0^\infty w^{a+b+c-3}e^{-wq}dw$$

Como:

$u = \frac{x}{1-x}w \quad y \quad v = \frac{y}{1-y}w$ calculando su jacobiano se tiene:

$$J(u,v) = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{w}{(1-x)^2} & 0 \\ 0 & \frac{w}{(1-y)^2} \end{vmatrix} = \frac{w^2}{(1-x)^2(1-y)^2}$$

Ahora

$$
\begin{aligned}
f(x,y) =& \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}}{(1-x)^{a-1}} \frac{y^{b-1}}{(1-y)^{b-1}} \frac{1}{(1-x)^2(1-y)^2} \int_0^\infty w^{a+b+c-3} w^2 e^{-wq} dw \\
=& \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}}{(1-x)^{a+1}} \frac{y^{b-1}}{(1-y)^{b+1}} \int_0^\infty w^{a+b+c-1} e^{-wq} dw \\
=& \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}}{(1-x)^{a+1}} \frac{y^{b-1}}{(1-y)^{b+1}} \Gamma(a+b+c) \left[\frac{(1-xy)}{(1-x)(1-y)}\right]^{-(a+b+c)} \\
=& \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}}{(1-x)^{a+1}} \frac{y^{b-1}}{(1-y)^{b+1}} \Gamma(a+b+c) \frac{(1-xy)^{-(a+b+c)}}{(1-x)^{-a-b-c}(1-y)^{-a-b-c}} \\
=& \frac{1}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}}{(1-x)^{-c-b+1}} \frac{y^{b-1}}{(1-y)^{-a-c+1}} \Gamma(a+b+c)(1-xy)^{-(a+b+c)} \\
=& \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)} \frac{x^{a-1}y^{b-1}(1-x)^{b+c-1}(1-y)^{a+c-1}}{(1-xy)^{a+b+c}} \quad 0<x,y<1
\end{aligned}
$$

Esta función de densidad puede ser generalizada para el caso de k variables aleatorias, definidas así:

$X_i = \frac{U_i}{U_i+W} \quad i = 1,2,3,...,K$ donde $U_1 \sim gama(a_1,1),...,U_K \sim gama(a_K,1), W \sim gama(a,1)$, encontrando su distribución conjunta $X_1, X_2, ..., X_K, W$.

La correlación de variables $X_i,\ i = 1,2,3,...,K$ restringe la correlación a ser siempre positiva, lo que no es adecuado para situaciones más generales. Aquí también se puede notar que para valores de a y b pequeños y c grande ocasionan correlaciones próximas a cero, en cuanto los valores de a y b grandes con c pequeño llevan correlaciones próximas a uno.

Una reparametrización de f(x,y) en términos de parámetros de media y precisión, más apropiado para una regresión beta está dada por:

$$f(x,y) = \frac{\Gamma(\phi_1+\phi_2-c)}{\Gamma(cu_1/(1-u_1))\Gamma(cu_2/(1-u_2))\Gamma(c)}$$
$$= \frac{x^{cu_1/(1-u_1)-1}y^{cu_2/(1-u_2)-1}(1-x)^{\phi_2-1}(1-y)^{\phi_1-1}}{(1-xy)^{\phi_1+\phi_2-c}},\ 0<x,y<1$$

En este caso la distribución beta bivariada depende todavía de un parámetro c.

3.4 Construcción de distribución conjunta vía cópulas

Otra posibilidad para obtener una distribución beta multivariada consiste en unir las marginales betas a través de una función cópula, la cual es una de las herramientas más útiles para trabajar con distribuciones multivariadas cuando las marginales son dadas o conocidas. El uso de funciones cópulas permiten la representación de los diferentes tipos de dependencia entre las variables.

- Definición. (Nelsen, 2006, p. 48) Una cópula está definida como una función de distribución conjunta

$$C(u_1,...,u_K) = P(U_1 \leq u_1,...,U_K \leq u_K),\ 0 \leq u_i \leq 1$$

con $U_i \sim U(0,1),\ i = 1,2,3,...,K$

- Teorema (Nelsen, 2006, p. 46) Sea H una función de distribución acumulada K-dimensional con marginales $F_1,...,F_K$ entonces existe una cópula

K-dimensional C tal que para todo $(y_1, ..., y_K) \in [-\infty, \infty]^K$,

$$H(y_1, ..., y_K) = C(F_1(y_1), ..., F_K(y_K))$$

- Corolario (Nelsen, 2006, p. 46) Sea H una función de distribución acumulada K-dimensional con marginales $F_1, ..., F_K$ y sean $F_1^{-1}, ..., F_K^{-1}$ inversas de $F_1, ..., F_K$ respectivamente. Entonces, para cualquier $u_i \in [0, 1]$, existe una cópula K-dimensional C, tal que

$$C(u_1, ..., u_K) = h(F_1^{-1}(u_1), ..., F_K^{-1}(u_K)).$$

Por el teorema anterior se puede encontrar una distribución conjunta de K variables aleatorias $Y_1, ..., Y_K$. Supóngase que tales variables aleatorias son continuas con funciones de distribuciones marginales $F_1, ..., F_K$ respectivamente y función de distribución conjunta $H(y_1, ..., y_K)$. Entonces la función de densidad conjunta está dada por:

$$\begin{aligned}
h(y_1, ..., y_K) =& \frac{\partial^K H(y_1, ..., y_K)}{\partial y_1, ..., \partial y_K} \\
h(y_1, ..., y_K) =& \frac{\partial^K C(F_1(y_1), ..., F_K(y_K))}{\partial F_1(y_1), ..., \partial F_K(y_K)} \frac{\partial F_1(y_1)}{\partial y_1} \times ... \times \frac{\partial F_k(y_K)}{\partial y_K} \\
h(y_1, ..., y_K) =& c(F_1(y_1), ..., F_K(y_K)) \prod_{i=1}^{K} f_i(y_i)
\end{aligned}$$

donde

$$c(F_1(y_1), ..., F_K(y_K)) = \frac{\partial^K C(F_1(y_1), ..., F_K(y_K))}{\partial F_1(y_1), ..., \partial F_K(y_K)} \ y \ f_i(y_i) = \frac{\partial F_i(y_i)}{\partial y_i}, \ i = 1, 2, ..., K$$

Para cualquier cópula $C(u_1, ..., u_K)$ los límites de Frechet-Hoftding inferior y superior son, respectivamente, dados por funciones $M(u_1, ..., u_K)$ y $W(u_1, ..., u_K)$, definidas como:

- $M(u_1, ..., u_K) = min(u_1, ..., u_K)$, que es una cópula para todo $K \geq 2$
- $M(u_1, ..., u_K) = max(u_1+, ..., +u_K - 1, 0)$, que es una cópula para $K = 2$

  Entonces

  $M(u_1, ..., u_K) \leq C(u_1, ..., u_K) \leq W(u_1, ..., u_K)$, $(u_1, ..., u_K) \in [0, 1]^K$

Ahora considerando la familia de funciones cópulas de Farlie-Gombel-Morgenstern, que en el caso bivariado tiene la forma

$$C_\theta(u, v) = uv + \theta uv(1 - u)(1 - v)$$

donde $\theta \in [-1, 1]$, $0 < u, v < 1$, supóngase que Y_1 y Y_2 son variables aleatorias con función de distribución F_1 y F_2. Entonces por el teorema anterior, una distribución conjunta de Y_1 y Y_2 está dada por:

$$\begin{aligned} H(y_1, y_2) =& C(F_1(y_1), F_2(y_2)) \\ =& F_1(y_1)F_2(y_2) + \theta F_1(y_1)F_2(y_2)[1 - F_1(y_1)][1 - F_2(y_2)]. \end{aligned}$$

Suponiendo que Y_1 y Y_2 son continuas, se tiene que su función de densidad conjunta h(.) es obtenida haciendo:

$$\begin{aligned} h(y_1, y_2) =& \frac{\partial^2 H(y_1, y_2)}{\partial y_1 \partial y_2} = \frac{\partial}{\partial y_2}\left(\frac{\partial H(y_1, y_2)}{\partial y_1}\right) \\ =& f_1(y_1)f_2(y_2)\left\{1 + \theta[1 - 2F_1(y_1)][1 - 2F_2(y_2)]\right\} \end{aligned}$$

donde $f_1(.)$ y $f_2(.)$ son las funciones de densidades marginales de Y_1 y Y_2.

Supóngase que $Y_1 \sim Beta(u_1, \phi_1)$ y $Y_2 \sim Beta(u_2, \phi_2)$ y que las densidades

de sus funciones tengan una reparametrización como lo hizo Ferrari y Cribari Neto, entonces para n observaciones independientes de $(y_{i1}, y_{i2}),\ i = 1, 2, ..., n$ se tiene que la función de verosimilitud es:

$$L(\mu_1, \mu_2, \phi_1, \phi_2, \theta) = \prod_{j=1}^{2}\prod_{i=1}^{n} \frac{\Gamma(\phi_i)}{\Gamma(\mu_j\phi_j)\Gamma((1-\mu_j)\phi_j)} y_{ij}^{\mu_j\phi_j - 1}(1-y_{ij})^{\phi_j(1-\mu_j)-1}$$

$$\times \prod_{i=1}^{n} \{1 + \theta[1 - 2F_1(y_1)][1 - 2F_2(y_2)]\}$$

donde $F_1\ y\ F_2$ son las funciones de distribuciones acumuladas de $Y_1\ y\ Y_2$.

3.5 Estadística No Paramétrica

En el presente trabajo se realizan algunas pruebas no paramétricas, cuando los datos no cumplen el supuesto de normalidad, tal es el caso de los índices de matemáticas (Y_1), ciencias sociales (Y_3) y lectura crítica (Y_5), como se verá más adelante.

3.5.1 Estadísticas para dos muestras

Sea $X_1, X_2, \ldots, X_{n_1}$ una muestra aleatoria proveniente de una población con función de distribución acumulada F y $Y_1, Y_2, \ldots, Y_{n_2}$ otra muestra aleatoria

proveniente de una función de distribución acumulada G. Sea también $n_1 + n_2$ el tamaño de la muestra total. El caso general requiere que las variables sean medidas al menos en escala ordinal. La hipoótesis de interés está dada por:

$$H_0 : \ F(t) = G(t) \quad \text{vs} \quad F(t) < G(t) \tag{1}$$

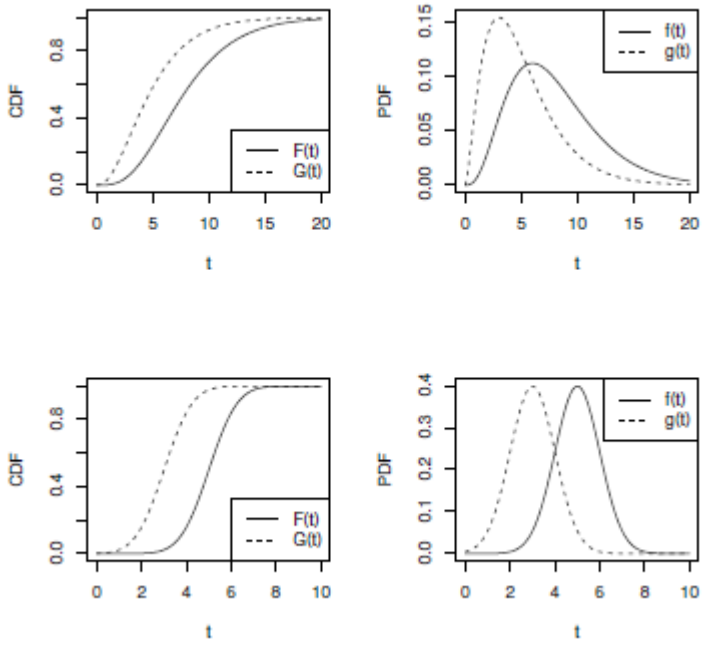


Figure 1: Funciones de Distribuciones y Probabilidad, Fuente: [11], pp. 52.

Sea $R(Y_j)$ el rango de Y_j en la muestra combinada. Luego la estadística de *Wilcoxon* es:

$$T = \sum_{j=1}^{n_2} R(Y_j) \tag{2}$$

Una segunda formulación de la prueba de *Wilcoxon* consiste en considerar todas

las diferencias $\{Y_j - X_i\}$ y sea T^+ el número de diferencias positivas.

$$T^+ = \#_{i,j}\{(Y_j - X_i) > 0\} \tag{3}$$

Se tiene la identidad

$$T^+ = T - \frac{n_2(n_2+1)}{2} \tag{4}$$

Y la estadística (3) es conocida como la prueba de *Mann-Whitney*.

3.5.2 Estadísticas para varias muestras

Cuando se trata de comparar más de dos muestras, se utiliza un análisis alternativo análogo al análisis de varianza en estadística paramétrica, denominado el análisis de *Kruskall-Wallis*. Esta prueba se basa en el modelo de localización:

$$Y_{ij} = \mu + e_{ij} \quad j = 1, 2, \ldots, n_i \quad i = 1, 2 \ldots, k \tag{5}$$

A causa de que la función de dispersión es invariante a localización, la dispersión en el modelo reducido es la dispersión de las obsrvaciones $D(0)$, el cual es llamado $D_\varphi(RED)$. La reducción en dispersión es entonces $RD_\varphi = D_\varphi(RED) - D_\varphi(FULL)$ y el ensayo estadístico está dado por:

$$F_\varphi = \frac{RD_\varphi/(k-1)}{\hat{\tau}_\varphi/2} \tag{6}$$

en donde $\hat{\tau}_\varphi$ es una estimación del parámetro de escala. Y cuando los puntajes de *Wilcoxon* se utilizan, se cambia el parámetro φ por W, quedando:

$$F_W = \frac{RD_W/(k-1)}{\hat{\tau}_W/2} \tag{7}$$

4 Análisis de Datos

4.1 Análisis Preliminares

Para el análisis de datos en el presente trabajo, se utilizó el Software R, mediante el uso de las librerías betareg, gamlss, xtable, agricolae, reshape, reshape2 y ggplot. Se utilizaron distintas pruebas comenzando con las pruebas de normalidad para cada uno de los índices.

Indice	**P. de Shapiro**	**Probabilidad**	**Significancia**	**Decisión**
Matemáticas	0.8707	0.004469	$***$	Se rechaza
C. Naturales	0.95226	0.2818		No se rechaza
C. Sociales	0.90362	0.02202	$**$	Se rechaza
L. Crítica	0.97241	0.7066		No se rechaza
Inglés	0.91835	0.047	$*$	Se rechaza

Table 1: Pruebas de Normalidad

En la tabla 1 se evidencia que en los índices correspondientes a *Ciencias Naturales* y *Lectura Crítica*, no se rechaza la hipótesis nula de normalidad

$$H_0 : \ I_i \sim N(\cdot\,,\cdot) \quad \text{con} \quad 1 = 1, 2, \ldots, 5$$

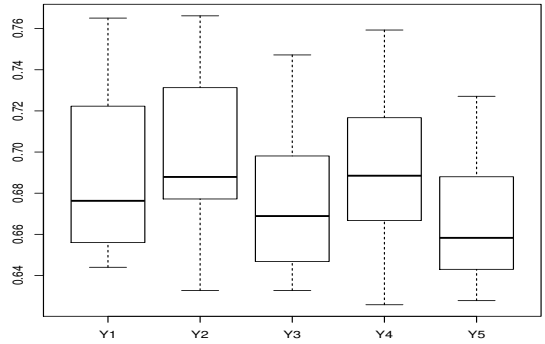
En los datos correspondientes a los índices de Matemáticas, Ciencias Sociales e Inglés, se verifica que no se cumple el supuesto de normalidad. Tomando la anterior evidencia como base, se procede a desarrollar unas pruebas de compara-

ción descriptivas, discriminadas por Comuna, específicamente para los índices en los que no se rechaza la prueba de normalidad.

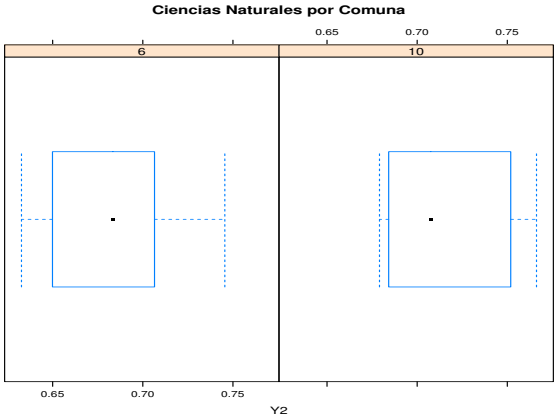
En la figura 2 se puede observar el diagrama de cajas y bigotes para las cinco áreas evaluadas, las cuales han sido designadas como Y_1, Y_2, Y_3, Y_4 y Y_5, para Matemáticas, Naturales, Sociales, Lectura Crítica e Inglés en su respectivo orden, se evidencia que el puntaje medio más alto es el de Ciencias Naturales y el más bajo corresponde a Inglés (ver figura 1 (a)). También el puntaje medio de Ciencias Naturales por Comuna es de aproximadamente 0.68 en la Comuna 6, mientras es de 0.70 en la Comuna 10 (ver figura 1 (b)). Por su parte el puntaje medio por comuna en el área de Lectura Crítica es de 0.68 en la Comuna 6 y de 0.72 en la Comuna 10 (ver figura 1 (c)).

Teniendo en cuenta la validación de las hipótesis resumidas en la tabla 1 se realiza a continuación un análisis de varianza con el fin de comparar en forma simultánea, las medias de las áreas Ciencias naturales y Lectura crítica, en cada uno de los niveles de los factores: año, colegio y clasificación. Es de recordar que el análisis de varianza convencional se realiza con base en modelos con matriz de diseño de rango incompleto, los cuales utilizan como fundamento el supuesto de normalidad de los errores, y por ende la normalidad de la variable respuesta ó dependiente.

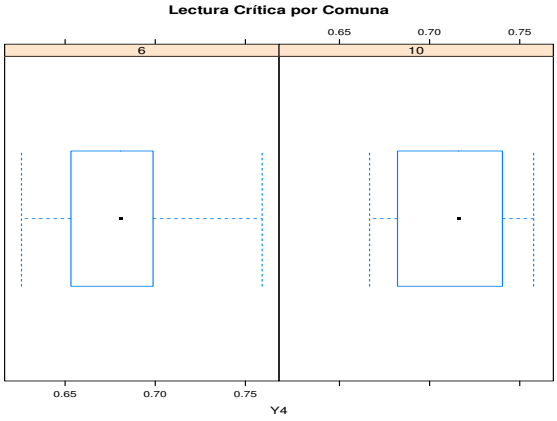
Para las variables que no cumplen el supuesto de normalidad, caso de Matemáticas, Sociales e Inglés, se realiza un análisis más apropiado, es decir un análisis *no paramétrico*.



(a) Areas



(b) Naturales



(c) Lectura Crítica

Figure 2: Areas por Comunas

4.2 Análisis de Varianza y Comparaciones de Medias

Se realiza un análisis de varianza para cada una de las dos variables normales: Ciencias Naturales (Y_2) y Lectura Crítica (Y_4), con respecto a cada uno de los factores: Comuna (X_2), Nombre del Establecimiento (X_4) y Clasificación (X_6). Posteriormente se realizan las pruebas de comparación de medias para cada uno de los factores, mediante el procedimiento de Tukey.

Factor	**G.L.**	**SS**	**SSM**	**Valor F**	**Pr**(>F)	$Significancia$
X2	1	0.01	0.01	36.38	0.0000	$***$
X4	4	0.02	0.01	32.69	0.0000	$***$
X5	2	0.00	0.00	11.14	0.0008	$***$
Residuales	17	0.00	0.00			

Table 2: Anova Ciencias Naturales

Se puede evidenciar en la tabla 2 que para la variable dependiente Ciencias Naturales, los tres factores de clasificación: Comuna, Nombre del Establecimiento y Clasificación del colegio son altamente significativas, por ende se pueden establecer las comparaciones de medias de cada uno de estos factores.

En la tabla 3 se pueden observar las diferencias más significantes para los factores: Comuna, en la que se evidencia una diferencia moderada, en los Colegios se observa que la mayor difrencia está entre las instituciones Humberto Tafur Charry y Agustín Codazzi, mientras que la menor diferencia se encuentra entre

Factor	**Diferencia**	**Linf.**	**Lsup.**	**Pr**(>F)	Signif.
Comuna$(10-6)$	0.03364	0.002267399	0.0650126	0.0366963	**
Colegio					
HTC-AC[1]	0.08178	0.041292414	0.12226759	0.0000586	***
OLB-AC[2]	0.07204	0.031552414	0.11252759	0.0002850	***
HTC-EL[3]	0.07840	0.037912414	0.11888759	0.0001009	***
OLB-EL[4]	0.06866	0.028172414	0.10914759	0.0004986	***
HTC-EOH[5]	0.05680	0.016312414	0.09728759	0.0036089	***
OLB-EOH[6]	0.04706	0.006572414	0.08754759	0.0178894	***
Clasificación					
B-A	-0.04795000	-0.07333318	-0.02256682	0.0002778	***
C-B	-0.08843182	-0.11228549	-0.06457815	0.0000000	***
C-B	-0.04048182	-0.06232109	-0.01864255	0.0003444	***

Table 3: Comparación de medias por Factor

[1] Humberto Tafur Charry - Agustín Codazzi [2] Oliverio Lara Borrero - Agustín Codazzi [3] Humberto Tafur Charry - El Limonar [4] Oliverio Lara Borrero - El Limonar [5] Humberto Tafur Charry - Enrique Olaya Herrera [6] Oliverio Lara Borrero - Enrique Olaya Herrera

las instituciones Oliverio Lara Borrero y Enrique Olaya Herrera. Para el factor Clasificación se evidencian diferencias altamente significativas.

Factor	**G.L.**	**SS**	**MSS**	**Valor F**	**Pr**(>F)	**Significancia**
X2	1	0.01	0.01	18.87	0.0004	$***$
X4	4	0.01	0.00	11.57	0.0001	$***$
X5	2	0.01	0.00	10.90	0.0009	$***$
Residuales	17	0.01	0.00			

Table 4: Anova Lectura Crítica

Algo similar a lo que ocurre con la variables Ciencias Naturales, ocurre también con la variable Lectura Crítica, los factores: Comuna, Nombre del Establecimiento y Clasificación de los mismos son altamente significativos, por tanto se procede a realizar la comparación multiple de medias, para así mismo determinar las principales diferencias significativas de Lectura Crítica discriminada por factor, utilizando el procedimiento de Tukey.

En la tabla 5 se evidencian diferencias altamente significativas en Comuna $(10-6)$, entre establecimientos educativos las mayores diferencias significativas se presentan entre Humberto Tafur Charry y Agustín Codazzi, Humberto Tafur Charry y el Limonar, finalmente entre Oliverio Lara Borrero y el Limonar. Por clasificación no se presentan diferencias, excepto una ligera diferencia entre las clasificaciones C y A.

Se puede también observar que para la variable Lectura Crítica, no hay sino 4 diferencias entre Establecimientos Educativos, mientras que para la variable Ciencias Naturales, se presentaron 6 diferencias. Lo mismo ocurrió con Clasificación, mientras que para Ciencias Naturales todas las diferencias resultaron

Factor	**Diferencia**	**Linf.**	**Lsup.**	**Pr**(>F)	Signif.
Comuna$(10-6)$	0.03155667	0.016229	0.04688433	0.0004415	$***$
Colegio					
HTC-AC[1]	0.055354667	0.021112235	0.08959710	0.0010734	$***$
OLB-AC[2]	0.044108667	0.009866235	0.07835110	0.0084847	$***$
HTC-EL[3]	0.052337333	0.018094901	0.08657977	0.0018636	$***$
OLB-EL[4]	-0.011246000	-0.045488432	0.07533377	0.0147452	$***$
Clasificación					
B-A	-0.008516042	-0.03317067	0.016138584	0.6560516	
C-A	-0.021687576	-0.04485659	0.001481436	0.0686128	$*$
C-B	-0.013171534	-0.03438397	0.008040899	0.2755213	

Table 5: Comparación de medias por Factor

[1] Humberto Tafur Charry - Agustín Codazzi [2] Oliverio Lara Borrero - Agustín Codazzi [3] Humberto Tafur Charry - El Limonar [4] Oliverio Lara Borrero - El Limonar

significativas, para Lectura Crítica, sólo hay una moderada diferencia entre las clasificaciones C y A. Con respecto a las comunas, parece que las diferencias son altamente significativas tanto para Ciencias Naturales como para Lectura Crítica. Es de aclarar que los valores que se están manejando son índices que varían entre 0 y 1, y que aunque las diferencias parezcan insignificantes, en sentido estadístico se pueden identificar con fines académicos.

4.3 Análisis no paramétricos, variables no normales.

Para las variables no normales: Matemáticas (Y_1), Sociales (Y_3) e Inglés (Y_5), se realiza la prueba alternativa de Kruskall-Wallis, para la comparación de medias por Comuna, Nombre del Establecimiento y Clasificación. En la tabla 6 se

Coeficientes	**Estimación**	**Error.Est**	**t.value.**	**Pr**(>F)	Signif.
Comuna(10)	0.0118	0.0048	2.4240	0.02679	$*$
Colegio					
EL[1]	-0.0218	0.0063	-3.4383	0.0031	$**$
EOH[2]	-0.0252	0.0066	-3.7883	0.0014	$***$
HTC[3]	0.0254	0.0095	2.6806	0.0158	$*$
OLB[4]	0.0335	0.0079	4.2074	0.0005	$***$
Clasificación					
B	-0.0309	0.0067	-4.5770	0.0002	$***$
C	-0.0434	0.0090	-4.8253	0.0001	$***$

Table 6: Comparación de medias por Factor

[1] El Limonar [2] Enrique Olaya Herrera [3] Humberto Tafur Charry [4] Oliverio Lara Borrero

puede evidenciar que hay una significancia estadística alta en la Comuna 10, los colegios Enrique Olaya Herrera y Oliverio Lara Borrero, así como en las clasificaciones B y C. Esto utilizando como variable respuesta el índice de matemáticas. En la tabla 7 se puede evidenciar que hay una significancia estadística alta en la Comuna 10, los colegios Humberto Tafur Charry y Oliverio Lara Borrero, así

Coeficientes	**Estimación**	**Error.Est**	**t.value.**	**Pr**(>F)	Signif.
Comuna(10)	1.7825e-02	4.7404e-03	3.7602	0.0015	$***$
Colegio					
EL[1]	4.6574e-05	6.1776e-03	0.0075	0.9940725	
EOH[2]	3.5666e-03	6.4938e-03	0.5492	0.5899920	
HTC[3]	5.2429e-02	9.2602e-03	5.6617	2.814e-05	$***$
OLB[4]	3.2286e-02	7.7736e-03	4.1532	0.0006656	$***$
Clasificación					
B	-1.8576e-02	6.5953e-03	-2.8165	0.0118841	$*$
C	-2.6253e-02	8.7684e-03	-2.9940	0.0081577	$**$

Table 7: Comparación de medias por Factor

[1] El Limonar [2] Enrique Olaya Herrera [3] Humberto Tafur Charry [4] Oliverio Lara Borrero

como en las clasificación C. Esto utilizando como variable respuesta el índice de sociales.

En la tabla 8 se puede evidenciar que hay una significancia estadística mediaa en la Comuna 10, en los colegios El Limonar y altamente significativo en el colegio Oliverio Lara Borrero, así como en las clasificacines B y C. Esto utilizando como variable respuesta el índice de inglés.

Coeficientes	**Estimación**	**Error.Est**	**t.value.**	**Pr**(>F)	Signif.
Comuna(10)	0.0098316	0.0054275	1.8115	0.0877756	·
Colegio					
EL[1]	0.0139687	0.0070730	1.9749	0.0647456	·
EOH[2]	-0.0013158	0.0074350	-0.1770	0.8616244	
HTC[3]	0.0149212	0.0106023	1.4074	0.1773448	
OLB[4]	0.0263528	0.0089003	2.9609	0.0087542	**
Clasificación					
B	-0.0342472	0.0075512	-4.5353	0.0002928	***
C	-0.0507630	0.0100393	-5.0565	9.739e-05	***

Table 8: Comparación de medias por Factor

[1] El Limonar [2] Enrique Olaya Herrera [3] Humberto Tafur Charry [4] Oliverio Lara Borrero

4.4 Análisis de Modelos de Regresión Beta

Teniendo en cuenta los índices presentados para cada una de las asignaturas evaluadas, basándonos en que estos índices son valores entre 0 y 1, se proponen tres modelos de Regresión Beta para cada uno de los índices no normales, es decir: matemáticas (Y_1), sociales (Y_3) e inglés (Y_5). Este análisis lo permite realizar la librería *gamlss* del paquete *R*. Aquí se tiene en cuenta que los índices de ciencias naturales (Y_2) y lectura crítica no entran, pues en ellos no se evidencia la falta de normalidad.

Factor	Estimate	Std. Error	t value	Pr(>|t|)	Signific.
Intercepto	0.91	0.03	26.76	0.00	* * *
Comuna 10	0.05	0.02	2.82	0.01	*
EL	-0.09	0.02	-4.04	0.00	* * *
EOH	-0.11	0.02	-4.49	0.00	* * *
HTC	0.13	0.03	3.64	0.00	**
OLB	0.16	0.03	5.58	0.00	* * *
B	-0.16	0.03	-6.25	0.00	* * *
C	-0.22	0.03	-6.60	0.00	* * *

Table 9: Regresión Beta por Factor, Matemáticas

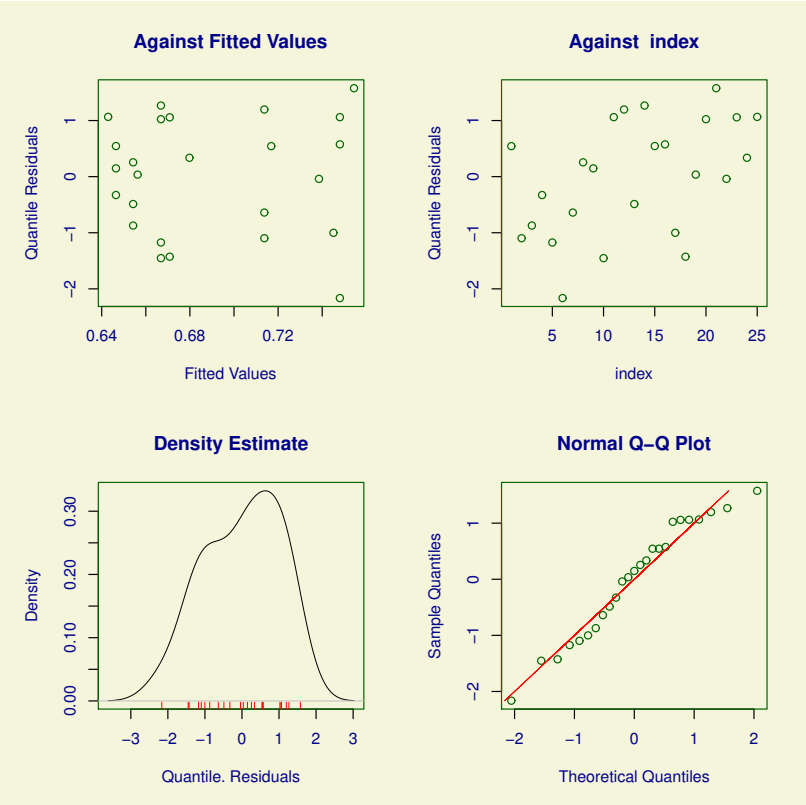


Figure 3: Influencia Estadística Modelo Matemáticas

Se evidencia en la tabla 9, la robustez del modelo de Regresión Beta, para la variable independiente índice de Matemáticas versus Comuna, Colegio y Clasificación. Todos los factores son altamente significativos, se obtuvo una variación estimada de $\hat{\sigma} = e^{-4.1302} = 0.016$.

En la Figura 3, se evidencia un comportamiento normal de los cuantiles residuales, se encuentran en el intervalo $[-2, 2]$ para media y varianza y el ajuste de la recta es aceptable.

Factor	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
Intercepto	0.70	0.03	24.99	0.00	* * *
Comuna 10	0.09	0.01	5.82	0.00	* * *
EL	0.00	0.02	0.00	1.00	
EOH	0.01	0.02	0.57	0.58	
HTC	0.25	0.03	8.65	0.00	* * *
OLB	0.15	0.02	6.24	0.00	* * *
B	-0.08	0.02	-4.03	0.00	* * *
C	-0.12	0.03	-4.25	0.00	* * *

Table 10: Regresión Beta por Factor, Sociales

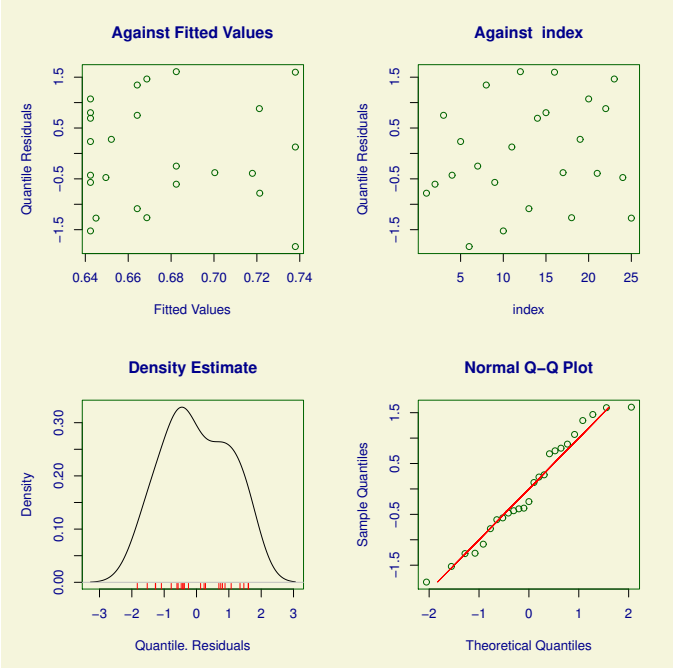


Figure 4: Influencia Estadística Modelo Sociales

Se evidencia en la tabla 10, una vez másla un buen modelo de Regresión Beta, para la variable independiente índice de Sociales versus Comuna, Colegio y Clasificación. Todos los factores son altamente significativos, excepto los colegios El Limonar y Enrique Olaya Herrera, se obtuvo una variación estimada de $\hat{\sigma} = e^{-4.3109} = 0.013$.

En la Figura 4, se evidencia un comportamiento normal de los cuantiles residuales, se encuentran en el intervalo $[-2, 2]$ para media y varianza y el ajuste de la recta es aceptable.

Factor	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
Intercepto	0.78	0.05	16.09	0.00	***
Comuna 10	0.03	0.03	1.27	0.22	
EL	0.07	0.03	2.17	0.05	*
EOH	0.00	0.03	0.12	0.91	
HTC	0.06	0.05	1.18	0.26	
OLB	0.13	0.04	3.09	0.01	**
B	-0.16	0.04	-4.38	0.00	***
C	-0.23	0.05	-4.81	0.00	***

Table 11: Regresión Beta por Factor, Ingles

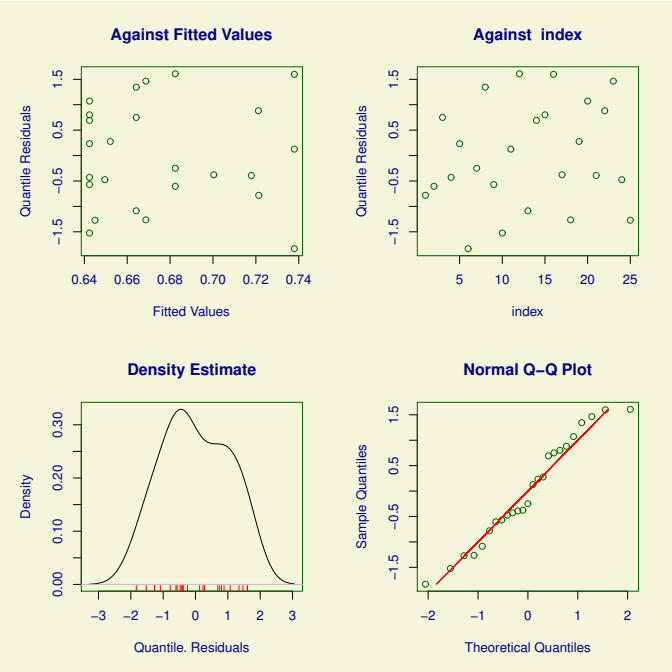


Figure 5: Influencia Estadística Modelo Ingles

Se evidencia en la tabla 11, una vez másla un buen modelo de Regresión Beta, para la variable independiente índice de Ingles versus Comuna, Colegio y Clasificación. Son significativos sólo el colegio Oliverio lara Borrero, las clasificaciones B y C, se obtuvo una variación estimada de $\hat{\sigma} = e^{-3.7432} = 0.023$.

En la Figura 5, se evidencia un comportamiento normal de los cuantiles residuales, se encuentran en el intervalo $[-2, 2]$ para media y varianza y el ajuste de la recta es aceptable.

5 Conclusiones

Del presente trabajo se han logrado extraer las siguientes conclusiones:

- Sorprende encontrar una metodología robusta que permite modelar información histórica, aunque el tamaño de los datos no sea grande.
- Se puede trabajar con Estadística clásica. siempre y cuando la distribución de los datos sea normal ó aproximadamente normal.
- Existen alternativas No Paramétricas modernas que permiten evaluar estadísticas en datos no normales, con una eficiencia buena.
- El análisis permitió establecer tres modelos de regresión Beta, cuando el supuesto de normalidad sobre las variables respuesta no se cumple.
- Los factores asociados a cada uno de los índices resultaron ser los apropiados y altamente significativos. Con estos modelos se pueden establecer condiciones y evaluar condiciones que lleven a formular estrategias por parte de los colegios.
- La comuna con mayor rendimiento académico en las pruebas saber 11 resultó ser la 10.
- La Regresión Beta del paquete *R*, ya fué sustituída por la librería *gamlss*.

Bibliografía

[1] Ferrari, S. L. P. and Cribari-Neto, F. (2004), *Beta regressión for modelling rates and proportions*. Journal of Applied Statistics, **31**, 799-815.

[2] Brasil, C. (2009), *Modelo de Regresión Beta para Analizar el Origen de los Problemas en Sistemas de Construcción* . Trabajo de Grado (Instituto de Matemáticas, Departamento de Estadística). Porto Alegre (Brasil): Universidad Federal de Rio Grande do Sul.

[3] Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering* . Second Edit, Jonh Wiley and Sons, London.

[4] Cribari-Neto F, Zeileis, A (2010), *Beta Regression in R*. Journal of Statistical Software. **34**(2), 1-24

[5] Espinheira, P. L., Ferrari, S. L. P. e CRIBARI-NETO, F. (2008a) *Influence Diagnostics in Beta Regression*. Computational Statistics and Data Analysis, **52**, 4417-4431.

[6] Ferrari, S. L. P. and Cribari-Neto, F. (2004), *Beta regressión for modelling rates and proportions*. Journal of Applied Statistics, **31**, 799-815.

[7] Ferreira do Souza, A. D. (2011), *Regresión Beta Multivariada con Aplicaciones en Pequeñas Áreas*. Trabajo de Grado (Instituto de Matemáticas). Rio de Janeiro (Brasil) : Universidad Federal de Rio de Janeiro.

[8] Miyashiro, E. S. (2008), *Modelos de Regressão Beta e Simplex para Análise de Proporções* . Trabajo de Grado (Instituto de Matemáticas e Estatística). São Paulo: Universidade de São Paulo.

[9] Nelsen, R. B. (2006), *An Introduction to Copulas*. Second Edit., Springer, New York.

[10] Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*. Wiley, Athens (Greece).

[11] Kloke, J. and McKean, J.W. (2015), *Non Parametric Statistical Methods Using R*. CRC Press, Taylor & Francis Group, A CHAPMAN & HALL BOOK, Boca Raton, FL.

[12] Liberal, T. (2010), *Regresión Beta Inflacionada: Inferencia y Aplicaciones*. Trabajo de Grado (Centro de Ciencias Exactas, Departamento de Estadística). Pernambuco (Brasil): Universidad Federal de Pernambuco.

[13] Olkin, I. and LIU, R. (2003), *A Bivariate Beta Distribution*. Statistics and Probability Letters, **62**, 407-412.

[14] Osina, R. and FERRARI, S. L. P. (2010), *Inated Beta Distributions*. Statistical Papers, **51**, 111-126.

[15] Paula, G. A. *Modelos de regressão com apoio computacional*. *http.www.//ime.usp.br/* $\sim$ *giapaula*.