APLICACIONES ESTADÍSTICAS

Socialización de Experiencias

Memorias 2016

Duitama, Boyacá
UPTC-Facultad Seccional Duitama
27 y 28 de mayo

Versión: Aplicaciones Estadísticas - 1. Socialización de Experiencias

ESPECIALIZACIÓN EN ESTADÍSTICA

Uptc® Universidad Pedagógica y Tecnológica de Colombia
ACREDITACIÓN INSTITUCIONAL DE ALTA CALIDAD MULTICAMPUS
RESOLUCIÓN 3910 DE 2015 MEN / 6 AÑOS
Posgrados Duitama
Grupo de Investigación en Estadística

Aplicaciones Estadísticas. Socialización de Experiencias
"ISSN: 2619-2888 (En línea)"
www.uptc.edu.co/enlaces/mem_aese

Diseño y Diagramación
Omar Velandia Castro - omarvelandia@hotmail.com
Eliana Leonor Valderrama Orozco – eliana.valderrama@uptc.edu.co
Luis Arbey Gómez Gómez – luis.gomez@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia
Facultad Duitama

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2016

ESPECIALIZACIÓN EN ESTADÍSTICA

Contacto

Universidad Pedagógica y Tecnológica de Colombia
Facultad Seccional Duitama
Escuela de Posgrados Sede Duitama
Teléfono: (57+8) 7624431
Conmutador (57 + 8) 7605306 Ext: 2838 - 2830
Carrera 18 Calle 22 Edificio Administrativo Piso 1
Duitama - Boyacá - Colombia

www.uptc.edu.co
posgrados.duitama@uptc.edu.co

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2016

ESPECIALIZACIÓN EN ESTADÍSTICA

PRESENTACIÓN

En calidad de coordinadora académica de la Especialización en Estadística, quisiera celebrar con ustedes este ejercicio de divulgación de los trabajos de aplicación desarrollados por los graduados de la primera cohorte. Se ha dispuesto de este espacio para intercambiar experiencias de la aplicación de técnicas estadísticas en ámbitos como la economía, educación, agronomía, administración, ingeniería, entre otros.

La información contenida en estas memorias es el fruto de un año de intenso trabajo por parte de nuestros estudiantes, agradecemos mucho por la confianza que depositaron en nuestra Institución, y confío en que con el paso del tiempo serán recompensados por decisión de cursar la Especialización. Expresamos nuestro reconocimiento a los profesores que dirigieron los trabajos de aplicación, gracias por el profesionalismo, dedicación y buena voluntad.

Es nuestro deseo que esta publicación sea fuente de consulta para profesionales que requieren el uso de técnicas estadísticas de dependencia e interdependencia para la solución de problemas en su área de trabajo.

Carmen Helena Cepeda Araque
Coordinadora Académica
Especialización en Estadística

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias
2016

ESPECIALIZACIÓN EN ESTADÍSTICA

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias
2016

TABLA DE CONTENIDO

Diseño y validación de un instrumento para la caracterización de una comuna de Duitama Boyacá Colombia 1
JULIÁN ANDRES SOLANO SIERRA - DAIRO SIGIFREDO GIL GIL
julian.solano@uptc.edu.co - dsigifre@gmail.com
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Análisis resultados de la prueba saber 11 de Duitama colegios calendario A-2014 8
LUCERO RODRÍGUEZ LÓPEZ - SANDRA PATRICIA CÁRDENAS OJEDA
lucero.rodriguez@uptc.edu.co - sandra.cardenas@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Planteamiento de un diseño experimental de mezclas para la fabricación de cemento con puzolana 22
DAVID JULIÁN SOTELO LÓPEZ - EDUARDO DÁVILA SANABRIA
david.sotelo@uptc.edu.co - eduardo.davila@talex.com.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Aplicación de un modelo de regresión logística con respuesta politómica ordinal en el análisis del desempeño académico en matemáticas 31
JHON JAIRO GONZÁLEZ GONZÁLEZ - CARMEN HELENA CEPEDA ARAQUE
jhosand01@yahoo.es - carmen.cepeda@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Metodología para evaluar la calidad de la información del componente de insumos del SIPSA 43
EMILCEN ROJAS PINZÓN - REINALDO ALARCÓN GUARÍN
emilcen.rojas@uptc.edu.co - reinaldo.alarcon@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Metodología para evaluar estadísticamente el efecto de un biopesticida elaborado con semillas de Melia Azedarach 53
ERIC GIOVANNY OSORIO OLEA - LUIS GUILLERMO DÍAZ MONROY
eric.osorio@uptc.edu.co – lgdmonroy@gmail.com
Universidad Pedagógica y Tecnológica de Colombia-Duitama

ESPECIALIZACIÓN EN ESTADÍSTICA

Diseño y análisis de un experimento para tutorado en arveja bajo presencia de sobredispersión respecto al modelo Poisson 60
MARÍA ELIANA DÍAZ SOSA - EDUARDO DÁVILA S.
mariaeliana.diaz@uptc.edu.co - eduardo.davila@talex.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Diseño y validación de un cuestionario para medir las características del visitante del anillo turístico del Lago Tota
CAMILO ERNESTO CAICEDO ESLAVA - DAIRO SIGIFREDO GIL GIL 70
kmilokicedo117@gmail.com – dgil_65@yahoo.es
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Aplicación de un modelo de regresión logística con respuesta politómica nominal en el análisis de preferencias alimentarias de aves
JORGE ALBERTO CHAPARRO PESCA - CARMEN HELENA CEPEDA ARAQUE 78
chapis_teto@yahoo.es - carmen.cepeda@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Resultados saber pro 2015 y su relación con variables sociodemográficas
YURI CAROLINA NIÑO CASTILLO - SANDRA PATRICIA CÁRDENAS OJEDA 91
yuricarolina.nino@uptc.edu.co - sandra.cardenas@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Series de precipitación pluviométrica en el municipio de Tota
WILMER ANTONIO MARTINEZ SUANCHA - ÁLVARO CALVACHE 106
wilmer.martinez@uptc.edu.co - acalvachea@gmail.com
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Aplicación de un modelo de sobrevida para tiempos de falla de generadores eléctricos
MARYLUZ CASTRO MORENO - CARMEN HELENA CEPEDA ARAQUE 118
maryluz.castro@uptc.edu.co - carmen.cepeda@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama

Análisis multivariado para el diagnóstico de habilidades matemáticas
DAYSY MAITE SÁNCHEZ BAREÑO - EDGAR FELIPE RUIZ ROBERTO 131
daysymaite.sanchez@uptc.edu.co - feruro42@gmail.com
Universidad Pedagógica y Tecnológica de Colombia-Duitama

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias
2016

ESPECIALIZACIÓN EN ESTADÍSTICA

Diseño y validación de un cuestionario para medir el grado de conciencia ambiental
INGRITH MARCELA QUINTERO FUENTES - REINALDO ALARCÓN GUARÍN
ingrith.quintero@uptc.edu.co - reinaldo.alarcon@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama
140

Descripción y cruce de variables relacionadas con la accidentalidad en la ciudad de Tunja
ELIANA IBETH MOYANO ALBA - SANDRA PATRICIA CÁRDENAS OJEDA
elim1220@hotmail.com - sandra.cardenas@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama
154

Caracterización de los automotores que ingresan al centro de diagnóstico automotriz Sogamoso LTDA CEDAS
ALBERTO ZEA HIGUERA - DAIRO SIGIFREDO GIL GIL
alzeahig10@gmail.com - dsigifre@gmail.com
Universidad Pedagógica y Tecnológica de Colombia-Duitama
171

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias
2016

ESPECIALIZACIÓN EN ESTADÍSTICA

DISEÑO Y VALIDACIÓN DE UN INSTRUMENTO PARA LA CARACTERIZACIÓN DE UNA COMUNA DE DUITAMA BOYACÁ COLOMBIA

Especialización en Estadística

Julián Andrés Solano Sierra[1,a], Dairo Sigifredo Gil Gil[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El objetivo de la investigación consistió en el diseño y validación de un instrumento para la caracterización de una comuna de Duitama Boyacá Colombia, se diseñó un cuestionario estructurado compuesto de 51 preguntas, el cual fue puesto a discusión por un grupo de expertos conformado por tres docentes universitarios (lingüista, sociólogo, estadista), además de esto se realizó una prueba piloto para validar el instrumento.

El instrumento aborda las características demográficas, socio-económicas y culturales, este instrumento es elaborado bajo la técnica cuestionario estructurado el cual parte de la hipótesis ?no se le pregunta a quien no sabe?, pero esa persona al formar parte de la muestra no se puede descartar por lo cual se le recaba información general (Demográfica, educacional, género, etc.), el instrumento obtuvo una calificación óptima.

***Palabras clave*:** Caracterización, validez, fiabilidad, comuna..

Abstract

The objective of the research was to design and validation of an instrument for the characterization of Duitama Boyacá Colombia commune, we designed a structured questionnaire composed of 51 questions, a group of experts (three university teachers: linguist, sociologist, statesman) discussed this instrument, in addition, a pilot test was conducted to validate the instrument.

The instrument addresses the demographic, socio-economic and cultural characteristics, this instrument is produced under the questionnaire technique structured which part of the hypothesis "not wondering who does not know", but that person to be part of the sample cannot be dismissed so it is collected general information (demographic, educational, gender, etc.), the instrument was optimal.

***Key words*:** characterization, reliability, validity, commune.

1. Introducción

En los últimos años no se encuentra una caracterización actualizada de la ciudad de Duitama. En el año 2009 (Dane n.d.), el sector salud realizó un estudio en el que se resaltan las características generales del municipio de Duitama (geográficas, antecedentes históricos y culturales, perfil político, administrativo y actividad económica). Además de estos aspectos, aborda otros en los que se da una mirada al estado de las

[a]Estudiante de Especialización en Estadística. E-mail: Julian.solano@uptc.edu.co

[b]Profesor asistente. E-mail: dgil$_{65}$@*yahoo.es*

viviendas y servicios con los que cuenta, resaltando la importancia de tener actualizada la información de las características de la ciudad. Se encuentra, también, un análisis estadístico por Hernández et al (2010), en el que se realiza una proyección poblacional a 2015, en el que se consideran (género, edad, pertenencia étnica, nivel educativo, situación educativa y según lugar de nacimiento, tipo de vivienda, servicios con que cuenta la vivienda, promedio de personas por hogar y actividad comercial).

En estos dos estudios se refleja una cercanía a la caracterización de la población de la ciudad de Duitama, pero es necesario actualizarlos para las autoridades municipales tengan elementos para la toma de decisiones para la solución de los problemas de las comunas. Se encuentran otras investigaciones respecto a caracterización como realizada a la población de Bajamar Isla departamento del Valle del Cauca, caracterización realizada a la Comuna 5 de la ciudad de Cali valle del Cauca, caracterización de la ciudad de Tunja Boyacá y caracterización realizada en la ciudad de Pereira Risaralda. En todas se observa que hacen énfasis en características demográficas, educativas, económicas y sociales (Echeverry, Suarez y Otros n.d.), (García Ramírez et al. n.d.), (Rincón Galvis, Agudelo Vega, Villate Corredor y Torres López n.d.).

2. Referente Conceptual

2.1. Análisis de instrumento estudio de fiabilidad

Según Pérez Juste, García Llamas, Gil Pascual y Galán González (2009) Pérez, R y otros (2009) la fiabilidad de las medidas se identifica con la precisión, de tal forma que decimos que un instrumento es fiable cuando mide algo con precisión independientemente de lo que se esté midiendo, ésta se puede interpretar como la constancia en las puntuaciones de los sujetos o bien la concordancia entre varias mediciones sucesivas de una misma realidad.

2.2. Procedimientos para determinar la fiabilidad

Para el cálculo de la fiabilidad existen diferentes procedimientos, fiabilidad como estabilidad, fiabilidad como equivalencia y fiabilidad como consistencia interna.

2.2.1. La fiabilidad como estabilidad del instrumento

Conocida como procedimiento de la repetición o del test-retest, se realiza la misma prueba dos veces a un grupo de sujetos esperando que pase un tiempo prudente, el cual, los investigadores sugieren sea entre 20 y 25 días de diferencia la aplicación de la prueba, de tal forma se observa la correlación que existe entre las puntuaciones obtenidas por los sujetos, la fórmula está dada por:

$$r_{xy} = \frac{n.\sum X.Y - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Donde,

$X =$ *Puntuación primera aplicación*,
$Y =$ *Puntuación segunda aplicación*.

Los valores por encima de 0,95 en una prueba afirman que tiene una muy buena fiabilidad.

2.2.2. La fiabilidad como equivalencia

También recibe la denominación de formas paralelas, consiste en aplicar dos pruebas diferentes que midan los mismos rasgos o características y se observa la correlación entre ambas pruebas, si las pruebas se aplican con una diferencia superior a 20 días podemos también calcular la estabilidad del instrumento, la fórmula

para este procedimiento está dada por:

$$r_{xy} = \frac{n.\sum X.Y - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

Donde,

$$X = \text{Pimera prueba},$$
$$Y = \text{Prueba Equivalente}.$$

La gran mayoría de los autores consideran buena fiabilidad como equivalencia a los valores superiores a 0,9.

2.2.3. La fiabilidad como consistencia interna

Conocido también como procedimiento de las mitades ya que se divide el test en dos mitades equivalentes particularmente se divide en puntuaciones obtenidas en los ítems pares e impares, de tal forma que estableciendo una relación entre ambas partes nos dará el coeficiente de fiabilidad (Pérez Juste, García Llamas, Gil Pascual y Galán González 2009), (Lohr y Velasco 2000).

2.2.4. Procedimiento de Sperman-Brown

En este procedimiento se observa la correlación entre las mitades, generalmente mediante el coeficiente de Pearson por ser correlación a mitades la fórmula para calcular la fiabilidad de la prueba es:

$$R_{xx} = \frac{2.r_{xx}}{1 + r_{xx}}.$$

El término

$$r_{xx}$$

es la correlación interna la cual se calcula por medio del coeficiente de correlación de Pearson entre las mitades donde se llamará

$$X_1$$

a las puntuaciones impares y

$$X_{21}$$

a las puntuaciones pares obteniendo:

$$r_{xx} = \frac{n.\sum X_1.X_2 - \sum X_1 \sum X_2}{\sqrt{\left[n\sum {X_1}^2 - (\sum X_1)^2\right]\left[n\sum {X_2}^2 - (\sum X_2)^2\right]}}.$$

Se dice que se tiene buena fiabilidad cuando nuestro $R_{xx} > 0,90$.

2.2.5. Procedimiento de Rulon

En este procedimiento observamos la varianza de las diferencias, comprobando la relación entre la varianza total del instrumento y la varianza existente entre ambas mitades, así:

$$r_{xx} = 1 - \frac{{s_d}^2}{{s_t}^2}$$

Donde debemos calcular la varianza de las diferencias ${s_d}^2$ y la varianza total ${s_t}^2$, así:

$${s_t}^2 = \frac{\sum T^2 - \frac{(\sum T)^2}{n}}{n-1};$$

donde T está dado por la suma de las puntuaciones obtenidas. Y

$$s_d{}^2 = \frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1},$$

donde d es la diferencia entre la puntuación impar y la par.

Se dice que se tiene buena fiabilidad cuando

$$r_{xx} > 0,90$$

2.2.6. Procedimiento de Kurder-Richerson

Se calcula teniendo en cuenta las inter-correlaciones de cada uno de los ítems, ya que se divide el instrumento en partes en tantos ítems posee obteniendo además del coeficiente de consistencia interna el coeficiente de homogeneidad:

$$r_{xx} = \left(\frac{n_e}{n_e - 1}\right) \cdot \left(\frac{s_t{}^2 - \sum p.q}{s_t{}^2}\right)$$

Donde

$$n_e$$

se refiere al número de elementos del que consta el instrumento, p es la proporción de sujetos que aciertan y $q = 1 - p$.

A pesar de que con éste procedimiento se obtiene un valor más bajo que con los otros procedimientos se considera buena fiabilidad cuando nuestro

$$r_{xx} > 0,90$$

2.2.7. Procedimiento Alfa-Cronbach

Por ser cuestionario el instrumento para esta investigación se utilizará el método de fiabilidad como consistencia interna mediante el procedimiento del alfa de Cronbach el cual está definido así:

$$\propto = \frac{n}{n-1}\ \left(1 - \frac{\sum s_i^2}{s_t^2}\right)$$

Donde:

n = Número de elementos o ítems de la prueba,

s_i^2 = Varianza de cada uno de los ítems,

s_t^2 = Varianza de las puntuaciones totales de la prueba.

3. Diseño y validación del instrumento

Con el fin de caracterizar una comuna cualquiera de las ocho en que está dividida la ciudad de Duitama, se diseñó un cuestionario estructurado compuesto de 51 preguntas.

3.1. Diseño del instrumento

Se utiliza la técnica moderna de Ícuestionario estructuradoÍ que parte de la hipótesis de que no se le pregunta al que no sabe, pero esa persona al formar parte de la muestra no se puede descartar por lo cual se le recaba información general (Demográfica, educacional, género, etc.) dirigida a realizar los diferentes cruces de variables que resulten necesarios.

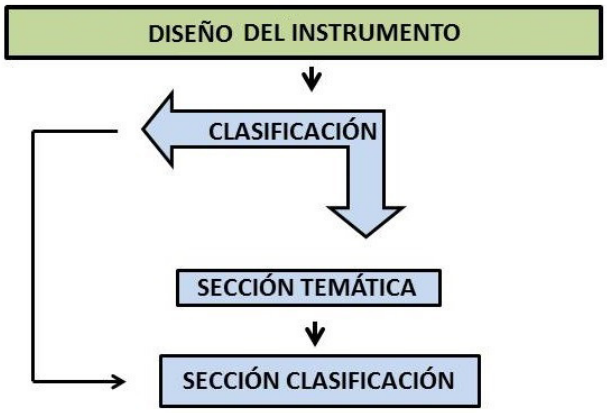


Figura 1: Diseño del Instrumento

El cuestionario consta de tres secciones:

1. La primera permite clasificar al respondiente si es o no residente de la ciudad de Duitama y por consiguiente, de la comuna en la que fue localizado. De no serlo, sólo se le solicitará información general que permita clasificar a la población en diferentes subpoblaciones.
2. La segunda (SECCIÓN TEMÁTICA), está dirigida a recolectar la información que caracteriza a los pobladores de la comuna en tres aspectos fundamentales:
   a. Aspectos demográficos de la comuna (composición étnica, tipo de vivienda, género, edad, etc.).
   b. Aspectos socio-económicos de la comuna (comerciales, servicios, industriales, empresariales, calidad de vida, riqueza, servicios, laboral y educacional, etc.).
   c. Aspectos culturales de la comuna (características religiosas, medio ambiental, recreativo, etc.)
3. La tercera (SECCIÓN DE CLASIFICACIÓN), enfocada a recolectar información demográfica, educacional, económica, etc., que permita realizar los cross tab para el estudio de las características asociadas y/o correlacionadas de la población.

3.2. Validación del instrumento

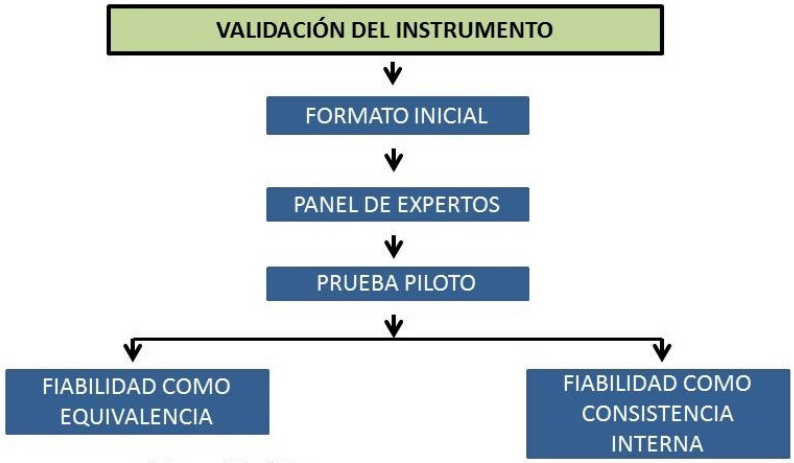


Figura 2: Validación del Instrumento

El instrumento está dirigido a medir simultáneamente una serie de variables que por su naturaleza (análisis multivariante) identifiquen las relaciones complejas que son difíciles de medir individualmente en los sujetos, el investigador debe preocuparse entonces por no incluir variables indiscriminadamente ni variables que estén altamente correlacionadas entre sí, para lograr parsimonia en el modelo y evitar enmascarar los efectos por la presencia de multicolinealidad. Debe estar dispuesto a atender a los errores a conocer los datos a establecer la diferencia existente entre significancia práctica y significancia estadística y a tener siempre presente el tamaño muestral y las implicaciones que conlleven las modificaciones de éste conocidas como efecto tamaño.

El investigador además de identificar las relaciones complejas entre las variables, debe siempre asegurar que existan observaciones suficientes por parámetro estimado para evitar el sobre ajuste de la muestra. Debe preocuparse también por evaluar la validez predictiva, validez concurrente y aparente así como la consistencia interna del conjunto de ítems del cuestionario.

Tratando de proporcionar un instrumento con alto grado de validez y de fiabilidad se diseñó el formato inicial del cuestionario, se sometió a autocrítica y a corregir la sintaxis, la eufonía y la concordancia en género y número, así como la ortografía.

Se solicitó los servicios de un panel de expertos conformado por tres docentes universitarios (lingüista, sociólogo, estadístico) quienes dieron sus conceptos sobre los constructos, fiabilidad, constancia y concordancia del instrumento. Debe advertirse que los criterios del panel fueron emitidos individualmente y no se concertó una reunión colectiva o conversatorio en el cual se debatieran por los expertos las diferentes falencias que pudiera tener el instrumento.

VARIABLES CUANTITATIVAS	Correlación de Pearson	alfa de Cronbach
¿Cuantas personas habitan en el hogar ?	0,78	0,704
Número de personas que estudian en el hogar	0,73	
integrantes del hogar con algún grado de escolaridad	0,84	
¿Electrodomésticos con que cuenta su vivienda son?	0,67	0,734
Número de mascotas en el hogar	0,63	
¿Cuántos garajes hay en la vivienda?	0,75	

Figura 3: Validación

En términos generales el instrumento fue bien calificado por el panel en todos los aspectos con observaciones únicamente de forma y ningún señalamiento que hiciera dudar sobre los constructos; por lo que puede afirmarse que el instrumento, como está, mide aquellas características para las cuales está diseñado.

Desde el punto de vista de la fiabilidad como consistencia interna, se utilizó el alfa de Cronbach como medida de la coherencia o consistencia en las respuestas del conjunto de sujetos a los diferentes elementos que integran el instrumento completo.

4. Conclusiones

Como para la validación del instrumento, además de las recomendaciones del panel de expertos, se realizó una prueba piloto n=50 con muestreo intencional, era de esperar que los diferentes indicadores especialmente la fiabilidad como equivalencia y la consistencia interna no dieran los valores propuestos por la teoría ($>$0,92 y $>$0,65) respectivamente, al no poder garantizar que la muestra cumpla las condiciones de ser:

Típica, representativa y suficientemente grande. Es de esperar que cuando se aplique el instrumento a la población objeto, utilizando muestreo aleatorio simple (MAS) con todas las técnicas, estos valores den igual

o superen a los estándares recomendados por la teoría ya que a pesar de las falencias de la encuesta piloto los valores se acercan relativamente a los estándares.

El cuestionario presenta un número relativamente alto (51) de preguntas, pero en la aplicación de la prueba piloto el tiempo máximo requerido fue de 18 minutos, tiempo en el que no se observaron síntomas de fatiga en el entrevistado.

El cuestionario transcurrió armónicamente, sin interrupciones ni reclamaciones por parte del entrevistado.

Puede afirmarse que el cuestionario recolecta información sobre todos los aspectos relevantes que permitan hacer una caracterización adecuada de una comuna para una ciudad de las características de Duitama Boyacá.

Referencias Bibliográficas

Alcaldía Mayor de Tunja (n.d.), 'Análisis de la situación de salud con el modelo conceptual de determinantes sociales de la salud'.
*http://www.boyaca.gov.co/SecSalud/images/Documentos/ASIS_2013/ASIS %20G %C3 %81MEZA %202013.pdf

Dane (n.d.), 'Censo poblacional ciudad de duitama'.
*http://www.dane.gov.co/files/censo2005/PERFIL_PDF_CG2005/15238T7T000.PDF

Echeverry, A., Suarez, C. y Otros (n.d.), 'Una mirada descriptiva a las comunas de Cali'.
*http://www.icesi.edu.co/jcalonso/images/pdfs/Publicaciones/una_mirada_descriptiva_a_las_comunas_de_cali.pdf

Gallego, J., Gómez, M. y otros (n.d.), 'Pereira imaginada 2009 ? 2014'.
*http://revistas.utp.edu.co/index.php/miradas/article/view/1349/5193

García Ramírez, I. N. et al. (n.d.), Los grandes proyectos urbanos en contextos étnicos. Estudio de caso Macroproyecto de interés Social Nacional Ciudadela San Antonio en su relación con el proyecto Malecón Bahía de la Cruz en Buenaventura–Colombia, Master's thesis, Universidad Nacional de Colombia-Sede Manizales.

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black (n.d.), 'Wc (2000): Análisis multivariante', *Iberia: Prentice Hall*.

Hernandez, L. Gutierrez, S. O. (2014), *Caracterización Socioeconómica De La Comuna Cinco Del Área Doña Luz De La Ciudad De Villavicencio Hacia Una Estrategia De Desarrollo Participativo 2014 -2018*.

Lohr, S. L. y Velasco, O. A. P. (2000), *Muestreo: diseño y análisis*, number 519.52 L64., International Thomson México.

Pérez Juste, R., García Llamas, J. L., Gil Pascual, J. A. y Galán González, A. (2009), 'Estadística aplicada a la educación', *Madrid: UNED-Pearson* .

Rincón Galvis, F. A., Agudelo Vega, A. D., Villate Corredor, Y. L. y Torres López, J. R. (n.d.), 'Análisis situacional en salud municipio de Duitama'.
*http://duitama-boyaca.gov.co/apc-aa-files/62653261643164376130336162613534/ SITUACIONAL_MUNICIPAL_FINAL_SALUD.pdf

ANÁLISIS RESULTADOS PRUEBA SABER 11 DE DUITAMA COLEGIOS CALENDARIO A - 2014

Especialización en Estadística

Lucero Rodíguez López[1,a], Sandra Patricia Cárdenas Ojeda[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadístcia, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

En este artículo se encuentra la descripción y análisis de la prueba SABER 11 de las institucines de Duitama - calendario A año 2014, de la cual se evaluan que relación puede tener los resultados de los componentes en cuanto al desempeño con algunas variables, donde dicha relación se analizo a través de pruebas de independencia y medidas de asociación.

***Palabras clave*:** SABER 11, componentes, desempeño, pruebas de independencia.

Abstract

In this article the description and analysis of the test SABER 11 of the institucines of Duitama is - calendar A 2014, which are evaluated that relationship can have the results of components for performance with some variables , where that relationship I was analyzed through independent tests and measures of association.

***Key words*:** SABER 11, components, performance, tests of independence.

1. Introducción

En este artículo se brinda la información acerca de la prueba SABER 11 de los colegios de Duitama para el año 2014, de calendario A, donde el objetivo principal es identificar las variables sociodemográficas que pueden influir en el nivel de desempeño, las variables sociodemográficas que se consideran son aquellas que el ICFES pregunta al estudiante en el momento de su inscripción y que tienen que ver con aspectos familiares y del colegio donde cursan en ese momento su grado 11.

Es relevante conocer que la Prueba SABER 11 es un examen que el ICFES aplica de manera anual el cual ha sido modificado en el 2014 en su estructura para que sus resultados sean comparables. Esto se ha logrado mediante la reestructuración en torno a la evaluación de competencias genéricas, la introducción de una prueba de competencias ciudadanas, la distinción en la prueba de matemáticas entre lo que es genérico y lo que no lo es y la fusión en diferentes pruebas en torno a las competencias genéricas que evalúan en común: Lenguaje y Filosofía que fusionaron en una prueba de Lectura crítica; Física, Química y Biología se fusionaron en una prueba de Ciencias naturales (que incluye el componente de Ciencia, Tecnología y Sociedad establecido en los Estándares); y las competencias ciudadanas se evaluarán mediante una prueba de Sociales y ciudadanas (Medina N. y Salazar 2015, pág 26).

[a]Estudiante de Especialización en Estadística. E-mail: lucero.rodriguez@uptc.edu.co

[b]Profesor asistente. E-mail: sandra.cardenas@uptc.edu.co

Los Objetivos que de conformidad con el Decreto 869 del 17 de marzo de 2010, en su artículo 1 menciona que la prueba es un instrumento estandarizado para la evaluación externa y que a su vez hace parte del Sistema Nacional de Evaluación(MEN 2010). Bajo ese mismo decreto uno de los objetivos es "comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado úndecimo de la educación media"(MEN 2010, Hoja1)

Las competencias a evaluar en la SABER11 son:

Lecto escritura Se evalúa el identificar y entender los contenidos locales que conforman un texto; comprender cómo se articulan las partes de un texto para darle un sentido global y reflexionar en torno a un texto y evaluar su contenido(ICFES 2014).

Matemáticas: en la prueba de matemáticas se definen tres competencias que recogen los elementos de los procesos de pensamiento descritos en los Estándares básicos de competencias: interpretación y representación; formulación y ejecución; y argumentación(ICFES 2014).

Sociales y ciudadanas: se evalúan tres competencias básicas, pensamiento social, interpretación y análisis de perspectivas y pensamiento reflexivo y sistémico(ICFES 2014).

Naturales: se evalúa el uso comprensivo del conocimiento científico,Explicación de fenómenos y la Indagación.

Inglés: los niveles de desempeño son B+, B1, A2, A1, A-, siendo A- el nivel donde el estudiante promedio no supera las preguntas de menor complejidad de la prueba(ICFES 2014).

Teniendo en cuenta las competencia descritas en los párrafos anteriores, el estudiante según del puntaje (valores en una escala de 0 a 100) obtenido en cada competencia se puede clasificar en un nivel de desempeño, donde se tiene Bajo (puntaje menor a 30), Medio (puntaje entre 30 y 70), Alto (puntaje superior a 70). En esta aplicación, además de describir las variables sociodemográficas que caracterizan los estudiantes que presentan la prueba SABER 11 para el año 2014, también identifica mediante pruebas de independencia si el nivel de desempeño está asociado con alguna variable sociodemográfica.

2. Referente Concenptual

En esta sección se presentan algunos conceptos relacionados con la prueba SABER 11 y sus componentes[1],en el análisis de los resultados de la prueba SABER 11 se tiene en cuenta tablas de contingencia, pruebas de independencia y medidas de asociación [2].

2.1. Tablas de contingencia

Una tabla de contingencia de es un arreglo bidimensional, de una variable fila con $f-$categorías o modalidades y una variable columna con $c-$categorías, donde hay $f \times c$ celdas, las entradas de las celdas son las frecuencias o conteos del número de casos en cada una de las combinaciones de valores de ambas variables. En general, se nota con n_{ij} a la frecuencia de la i-ésima modalidad de la variable fila y j-ésima de la variable columna.

El total por fila o por columna está formado por las frecuencias marginales, y se notan por $n_{i.}$ (donde el punto señala que se suman columnas dentro de la fila i) y $n_{.j}$ (donde el punto señala que se suman filas dentro de la columna j), respectivamente.

La suma de las frecuencias por celda es igual a la suma de las frecuencias marginales e igual al número total de individuos seleccionados y clasificados; se nota por n.

[1]Los conceptos son tomados de (Medina N. y Salazar 2015, pág.24)
[2]Los conceptos son tomados de (Díaz M. y Morales R. 2009, pág.23-42)

De acuerdo con (Díaz M. y Morales R. 2009), la notación general, para una tabla de contingencia de $f-$filas y $c-$columnas, se muestra en la tabla 1.

Filas	Columnas						
	1	2	...	j	...	c	Total($n_{i.}$)
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
f	n_{f1}	n_{f2}	...	n_{fj}	...	$n_{f.}$	n
Total($n_{.j}$)	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..} = n$

Tabla 1: Tabla de contingencia

donde

- La frecuencia de la $i-$ésima modalidad de la variable fila y la modalidad $j-$ésima de la variable columna se escribe como n_{ij}.
- El total de observaciones en la $i-$ésima modalidad de la variable fila se nota por $n_{i.}$, es decir,

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

- El total de observaciones en la $j-$ésima modalidad de la variable columna se nota por $n_{.j}$; es decir,

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{fj} = \sum_{i=1}^{f} n_{ij}$$

- El número total de observaciones en la muestra se escribe con n, y es igual a la suma de los márgenes fila o columna, es decir,

$$n = \sum_{i=1}^{f} \sum_{j=1}^{c} n_{ij}$$

Por otra parte, las frecuencias pueden ser transformadas en proporciones o porcentajes. Un primer porcentaje se obtiene de dividir cada frecuencia n_{ij} por el número total de observaciones n; este porcentaje se escribe como f_{ij}, es decir,

$$f_{ij} = \frac{n_{ij}}{N} \times 100$$

la cantidad f_{ij} corresponde a la proporción o porcentaje de elementos que tienen los atributos i y j.

El segundo porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal fila $n_{i.}$, así:

$$f_{j|i} = \frac{n_{ij}}{ni.} \times 100$$

La cantidad $f_{j|i}$ es la proporción de elementos de cada celda, respecto al total de la fila i. Según (Díaz M. y Morales R. 2009) "La expresión $j|i$ (que se lee: "j dado i") significa "estar en la columna j, a condición de estar en la fila i", es decir, se deja fija la fila i y se recorren sus columnas. Estas frecuencias corresponden al *perfil fila*".

El tercer porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal columna $n_{.j}$:

$$f_{ij} = \frac{n_{ij}}{n} \times 100$$

La cantidad $f_{i|j}$ es la proporción de elementos de cada celda, respecto al total de la columna j. Nuevamente según (Díaz M. y Morales R. 2009) "La expresión $i|j$ (se lee: "i dado j") significa "estar en la fila i, a condición de estar en la columna j", es decir, se deja fija la columna j y se recorren sus filas. Estas frecuencias corresponden al *perfil columna*".

Se obtienen tres tipos de tablas adicionales, la primera hace referencia al porcentaje de cada celda con relación al número total de individuos n; la segunda, al porcentaje de cada celda respecto al total de la respectiva fila (perfil fila) y la tercera, al por- centaje de cada celda con relación al total de la respectiva columna (perfil columna).

2.2. Pruebas de independencia

Al disponer de la información en una tabla de contingencia, es posible indagar si las variables que constituyen dicha tabla son independientes o no.

la hipótesis nula de independencia está dada por:

Ho: La variable fila es independiente de la variable columna

La estadística de prueba que se empleada en el juzgamiento de esta hipótesis es:

$$\chi^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

que bajo la hipótesis nula de independencia, tiene distribución de probabilidad *ji-cuadrado* con $(f-1)\times(c-1)$ grados de libertad.

Se rechaza la hipótesis nula a un nivel α cuando se verifica que $\chi_0^2 > \chi^2_{(f-1)(c-1),\alpha}$

2.3. Medidas de asociación

A continuación algunas medidas relacionadas con la estadística ji-cuadrado, como se ver en (Díaz M. y Morales R. 2009, pág 38), algunas de estas son:

2.3.1. El coeficiente de contingencia

Es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando se tiene información clasificadora acerca de uno o ambos conjuntos de atributos. El grado de asociación entre dos conjuntos de atributos, sean ordinales o no, se puede describir mediante la siguiente fórmula:

$$C = \sqrt{\frac{\chi_0^2}{\chi_0^2 + n}}, \qquad \text{donde} \qquad \chi_0^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

La estadística C toma valores entre 0 y 1. Valores cercanos a cero muestran una baja asociación entre las variables, mientras que valores próximos a 1 indican una posible alta asociación.

2.3.2. El coeficiente (V) de Cramer

Este coeficiente tiene un valor máximo en tablas de contingencia de cualquier tamaño. Se define como

$$V = \sqrt{\frac{\chi_0^2}{nk}} \tag{3}$$

donde k =mín{f-1,c-1} es el menor número de modalidades fila (o columna) menos uno de la tabla de contingencia. Se trata de un coeficiente que toma el valor 1 cuando hay asociación perfecta entre los atributos, cualquiera que sea el tamaño de la tabla de contingencia.

3. Metodología

El desarrollo de esta aplicación empleó el enfoque cuantitativo y tipo de investigación exploratoria. Teniendo como población los estudiantes de los colegios de la ciudad de Duitama en calendario A.

En lo refeerente a la consecución de los datos, de la página del ICFES www.icfes.gov.co, en el módulo investigadores y estudiantesde posgrado, en la opción acceso a bases de datos y siguiendo los pasos según la GUIA DE ACCESO A BASES DE DATOS DEL ICFES(ICFES 2013) se decargó la carpeta SABER11, de ella se utilizaron los diccionarios SB11-2014-2-Diccionario de VARIABLES-V1-0.pdf, y de la carpeta SB11-20142RGSTRO-CLFCCN-V1-0.zip se tomó la base de datos SB11-20142-RGSTRO-CLFCCNNV1.txt, y siguiendo las intrucciones del archivo tutorial(ICFES n.d.) se abrió en hoja de excel la base de datos con un total de 541875 registros, información de los resultados de la prueba SABER11 presentada para el año 2014 segundo semestre en Colombia, de dicha base se seleccionaron los 2070 registros correspondientes a los estudiantes de los colegios de la ciudad de Duitama, calendario A, que presentaron la prueba SABER11.

Una vez se tiene la muestra de 2070 estudiantes, se procedió a identificar las variables y el tipo de variable. Para el caso de las variables cuantitativas, tales como, edad y puntaje en cada competencia (lectura crítica, matemáticas, sociales, naturales e inglés) se calcularon algunas estadísticas de tendencia central, posición, forma y apuntamiento.

Para todas las variables se hace el análisis descriptivo y para el caso del nivel de desempeño en cada competencia, teniendo en cuenta los elementos de evaluación(ICFES 2014) se llevó a cabo la realización de tablas de contingencia entre nivel de desempeño y las variables sociodemográficas género, área residencia, naturaleza del colegio, jornada y estrato de la vivienda, obteniendo un total del 25 tablas de contingencia, para cada tabla se realiza la prueba de independencia y se calculan los coeficientes de asociación Cramer y Contingencia. El procesamiento de los datos se llevó a cabo en el software libre R(R Core Team 2015). Finalmente la organización del documento.

4. Resultados

En esta sección se encuentran los resultados primero en cuanto a variables asociadas con el estudiante, luego a variables asociadas con la familia, seguido de información del colegio y por último lo relacionado con el desempeño en las competencias.

4.1. Variables Sociodemográficas

Se presenta la información respecto a género, edad, si se encuentra trabajando y área de ubicación.

De las pruebas SABER 11, realizadas el segundo semestre de 2014, podemos afirmar que el 54.11 % correspondiente a 1120 estudiantes que presentaron la prueba fueron hombres, mientras que el 45.89 % equivalente a 950 estudiantes fueron mujeres.

Género	Estudiantes	%
M	1120	54,11
F	950	45,89
Total	2070	100

Tabla 2: Clasificación por género.
Fuente la Autora, 2016

Edades	Estudiantes	%
<14	7	0,338
14-16	921	44,493
17-19	912	44,058
20-22	73	3,527
23-25	24	1,159
26-28	30	1,449
29-31	25	1,208
32-34	17	0,821
35-37	18	0,870
>37	43	2,077
Total	2070	100

Tabla 3: Clasificación por edad.
Fuente la Autora, 2016

Mínimo	Máximo	Media	Q1	Q2	Q3	SD	CV	Asimetría	Kurtosis
0	67	18,17	16	17	18	5,7971	31,9034	4,4028	27,4448

Tabla 4: Algunas estadísticas de la variable edad. Fuente la Autora, 2016

En las edades de los estudiantes que presentaron la prueba SABER 11 se observa que el 0.342 % correspondiente a 7 estudiantes diligenciaron mal el año de nacimiento puesto que afirman que nacieron en el año 2014. Es notorio que los estudiantes de edades entre 14 y 19 años conforman la mayor parte de estudiantes que presentaron las pruebas SABER 11 realizadas en el segundo semestre de 2014, ya que corresponde al 88,551 % del total de estudiantes que la presentaron.

Trabaja	Estudiantes	%
No	1807	87,29
< 20 h semana	131	6,33
> 20 h semana	132	6,38
Total	2070	100

Tabla 5: Estado laboral. Fuente la Autora, 2016

Área	Estudiantes	%
Urbana	1762	85,12
Rural	308	14,88
Total	2070	100

Tabla 6: Clasificación por área. Fuente la Autora, 2016

El 87.29 %(1807) de los estudiantes afirmaron que no trabajaban y el 6 % manifestó trabajar más de 20 horas.

Con relación al área donde reside el estudiante, el 85.12 % correspondiente a 1762 estudiantes que presentaron la prueba SABER 11 viven en área urbana, mientras que el 14.88 % correspondiente a 308 estudiantes viven en área rural.

4.2. Variables asociadas a la familia

NIVEL EDUCATIVO PADRES		Ninguno	Pri_Inco	Pri_Com	Sec_In	Sec_Com	Tec_Inc
Padre	Estudiantes	53	288	418	297	463	25
	%	2,56	13,91	20,19	14,35	22,37	1,21
Madre	Estudiantes	25	263	354	313	575	24
	%	1,21	12,71	17,10	15,12	27,78	1,16
NIVEL EDUCATIVO PADRES		Tec_Com	Prof_Inco	Prof_Com	Post	No_sabe	Total
Padre	Estudiantes	98	34	224	68	102	2070
	%	4,73	1,64	10,82	3,29	4,93	100
Madre	Estudiantes	110	35	263	64	44	2070
	%	5,31	1,69	12,71	3,09	2,13	100

Tabla 7: Nivel educativo padres. Fuente la Autora, 2016

El 22.37 % de los estudiantes correspondiente a 463 indicó que sus padres tienen secundaria completa y en cuanto al nivel de formación de las madres el 27.78 % de los estudiantes, equivalente a 575 respondió que ellas tenían secundaria completa.

OCUPACIÓN PADRES		Empre	Peq empre	Direc gere	Niv dire	Tec prof	Aux adm	
Padre	Estudiantes	45	71	54	18	181	42	
	%	2,17	3,43	2,61	0,87	8,74	2,03	
Madre	Estudiantes	14	51	22	32	141	134	
	%	0,68	2,46	1,06	1,55	6,81	6,47	
OCUPACIÓN PADRES		Obre ope	Prof ind	Trab prop	Hogar	Pensi	Otr act	Total
Padre	Estudiantes	591	99	599	17	92	261	2070
	%	28,55	4,78	28,94	0,82	4,44	12,61	100
Madre	Estudiantes	121	73	195	1090	20	177	2070
	%	5,85	3,53	9,42	52,66	0,97	8,55	100

Tabla 8: Ocupación padres. Fuente la Autora, 2016

Los padres de los estudiantes que presentaron la prueba SABER 11 en el segundo semestre del año 2014, 28.94 % correspondiente a 599 son trabajadores por cuenta propia y 0.82 % correspondiente a 17 tiene como ocupación el hogar. Las madres el 52.66 % correspondiente a 1090 tiene como trabajo u oficio el hogar y el 0.68 % correspondiente a 14 son empresarias.

Estrato	Estudiantes	%
1	438	21,16
2	1122	54,20
3	446	21,55
4	55	2,66
5	9	0,43
Total	2070	100

Tabla 9: Clasificación por estrato. Fuente la Autora, 2016

Nivel SISBEN	Estudiantes	%
1	565	27,29
2	568	27,44
3	46	2,22
4	10	0,48
5	881	42,56
Total	2070	100

Tabla 10: Clasificación por SISBEN. Fuente la Autora, 2016

El 96 % de los estudiantes afirman pertenecer a los estratos 1,2 y 3 siendo el 54 % dfel estrato 2.

Se encuentra que la mayoría están en los niveles 1 con 27.29 %(565) y 2 con 27.44 %(568), aunque también llama la atención que 42.56 % correspondiente a 881 estudiantes que reportan estar en nivel 5.

Personas Hogar	Cuartos Hogar				
Nř	1	2	3	4	5 o más
1	7	1	2	0	0
2	22	92	15	3	3
3	17	220	154	12	5
4	7	206	406	81	4
5	3	92	262	113	25
6	2	24	94	57	24
7	0	14	25	19	9
8 o más	1	3	13	12	21

Tabla 11: Personas vs cuartos hogar. Fuente la Autora, 2016

Bienes/Servicios	Si	%	No	%
Teléfono	508	24,54	1562	75,46
Celular	1985	95,89	85	4,11
Televisor	1047	50,58	1023	49,42
Computador	1405	67,87	665	32,13
DVD	1333	64,40	737	35,60
Lavadora	1465	70,77	605	29,23
Microondas	610	29,47	1460	70,53
Nevera	1845	89,13	225	10,87
Horno	784	37,87	1286	62,13
Internet	784	37,87	1286	62,13
Automóvil	676	32,66	1394	67,34

Tabla 12: Posesión de algunos bienes y servicios. Fuente la Autora, 2016

De los 2070 estudiantes que presentaron la prueba SABER 11 en el segundo periodo de 2014 el 75.46 %(1562), dicen no tener teléfono, 95.89 %(1985) afirman tener celular, 50.58 %(1047) tienen televisor, 67.87 %(1405) tienen computador, 64.40 %(1333) tienen DVD, 70.77 %(1465) tienen lavadora, 70.53 %(1460) no tienen microondas, 89.13 %(1845) tienen nevera, 62.13 %(1286) no tiene horno, 62.13 %(1286) dicen no tener servicio de internet en sus hogares, 67.34 %(1394) afirman no contar con automóvil en sus hogares.

Pisos	Estudiantes	%
Tierra	16	0,77
Cemento	470	22,71
Madera burda	165	7,97
Baldosa	1419	68,55
Total	2070	100

Tabla 13: Material de los pisos. Fuente la Autora, 2016

En cuanto al tipo de material de los pisos que predomina en el hogar de los 2070 estudiantes que presentaron la prueba SABER 11 en el segundo semestre de 2014, el 68,55 % correspondiente a 1419 afirman tener piso tipo 4 el cual corresponde a Madera pulida-Baldosa-Tableta-Mármol-Alfombra.

4.3. Variables asociadas al colegio

Naturaleza colegio	Estudiantes	%
Oficial	1447	69,90
No oficial	623	30,10
Total	2070	100

Tabla 14: Naturaleza del colegio. Fuente la Autora, 2016

En colegios oficiales se evaluaron un total de 1447 estudiantes, que equivale al 69,90 %, y en colegios no oficiales un total de 623 estudiantes, que equivale al 30,10 % de total de los estudiantes evaluados. De los colegios oficiales un total de 633 alumnos están sujetos a estudiar en jornada completa, mientras que en colegios no oficiales son 480 estudiantes.

Jornada	Estudiantes	%
Completa	1113	53,77
Mañana	568	27,44
Noche	18	0,87
Sabatina	253	12,22
Tarde	118	5,70
Total	2070	100

Tabla 15: Clasificación por Jornada
Fuente la Autora, 2016

De los 2070 estudiantes que presentaron la prueba SABER 11 en el segundo semestre del año 2014, el 53.77 % correspondiente a 1113 estudiantes estaban en jornada completa u ordinaria, mientras que el 0.87 % correspondiente a 18 estudiantes están en la jornada nocturna siendo la jornada de menos estudiantes, que presentaron dicha prueba en ese año.

JORNADA	NATURALEZA			
	OFICIAL		NO OFICIAL	
	Estudiantes	%	Estudiantes	%
Completa	633	30,58	480	23,19
Mañana	463	22,37	105	5,07
Noche	0	0,00	18	0,87
Sabatina	242	11,69	11	0,53
Tarde	109	5,27	9	0,43

Tabla 16: Naturaleza vs Jornada. Fuente la Autora, 2016

Valor pensión	Estudiantes	%
No paga	1642	79,32
< 87000	160	7,73
87000-120000	122	5,89
120000-150000	41	1,98
150000-250000	99	4,78
>250000	6	0,29
Total	2070	100

Tabla 17: Valor de la pensión. Fuente la Autora, 2016

El 79.32 % correspondiente a 1642 estudiantes pertenecen al valor numero 1 asignado por el ICFES el cual quiere decir, que no tiene ningún costo el valor de la pensión de estos estudiantes y el 0.29 % correspondiente a 6 estudiantes pertenecen al valor 6, lo cual quiere decir que cancelan $250000 o más de pensión.

Valor pensión	NATURALEZA			
	OFICIAL		NO OFICIAL	
	Estudiantes	%	Estudiantes	%
No paga	1445	69,81	197	9,52
< 87000	0	0	160	7,73
87000-120000	1	0,05	121	5,85
120000-150000	0	0	41	1,98
150000-250000	0	0	99	4,78
>250000	1	0,05	5	0,24

Tabla 18: Naturaleza vs Valor Pensión. Fuente la Autora, 2016

SMLV	Estudiantes	%
<1	464	22,42
1 y < 2	944	45,60
2 y <3	398	19,23
3 y < 5	171	8,26
5 y < 7	55	2,66
7 y < 10	18	0,87
> 10	20	0,97
Total	2070	100

Tabla 19: Clasificación por salario. Fuente la Autora, 2016

Componente	Desempeño						
	Alto	%	Medio	%	Bajo	%	Total
LEC_CRÍTICA	108	5,22	1948	94,11	14	0,68	2070
MATEMÁTICAS	255	12,32	1808	87,34	7	0,34	2070
SOCIALES	140	6,76	1918	92,66	12	0,58	2070
NATURALES	200	9,66	1859	89,81	11	0,53	2070

Tabla 20: Desempeño competencias. Fuente la Autora, 2016

Entre las asignaturas con puntajes más altos, encontramos matemáticas con un total de 255 alumnos que equivale a 12,32 % que obtuvieron éste desempeño. Se encuentran deficiencias en cuanto a lectura crítica ya que solamente 108 estudiantes, que equivale al 5,22 % del total de alumnos que presentaron las pruebas se encuentran situados en desempeño alto.

DESEMPEÑO INGLÉS		
Componente	Estudiantes	%
B+	36	1,74
B1	139	6,71
A2	311	15,02
A1	875	42,27
A-	709	34,25
Total	2070	100

Tabla 21: Desempeño en Inglés. Fuente la Autora, 2016

Un total de 36 alumnos se situaron en desempeño alto en el área de inglés, equivalentes al 1,74 % del total de los alumnos que presentaron las pruebas. El 76,52 % del total de los alumnos que presentaron las pruebas se situaron en un nivel de desempeño bajo (A1 y A-).

PUNTAJES										
Areas	Min.	Max	Mediana	Media	Q1	Q3	SD	CV	Asim.	Kurtosis
Lec crítica	0	92	54	54,23	48	61	9,7594	17,99	-0,0864	3,4749
Matemáticas	20	100	55	56,16	48	64	11,3926	20,28	0,4072	3,2034
Sociales	23	93	55	55,07	48	62	10,0810	18,31	0,0363	3,0552
Naturales	19	93	55	55,57	49	63	10,4254	18,76	0,0928	2,9203
Inglés	30	100	52	53,5	47	58	10,2385	19,14	1,4148	5,8750

Tabla 22: Algunas estadísticas para los puntajes en las competencias. Fuente la Autora, 2016

4.4. Puebas de independencia

A continuación se presenta una tabla resumen con los resultados obtenidos de las pruebas de independencia y los coeficientes de asociación Cramer y contingencia, que fueron calculados para las 25 tablas de contingencia, las cuales se hicieron cruzando el nivel de desempeño en cada competencia (Alto, Medio, Bajo) para Lectura Crítica, Matemáticas, Sociales y Naturales con las variables sociodemográficas género, área donde reside(rural o urbana), naturaleza del colegio (oficial o no oficial), jornada(completa, mañana, noche, tarde y sabatina-dominical) y estrato (1,2,3,4,5,6). Cabe mencionar que en estos primeros resultados se utilizan todas las categorías de las variables mencionadas anteriormente, pero se debe realizar el análisis reagrupando categorías tanto de las variables nivel de desempeño, puesto que hay muy pocos estudiantes en nivel bajo y alto; de igual forma en las variables como estrato se presenta frecuencias muy bajas para las categorías 3,4,5 y para la variable jornada hay baja frecuencia en noche, sabatina-domical y tarde.

Observando la tabla 23 para el desempeño en Lectura Crítica la única variable con la que no esta relacionado el desempeño es el género, es decir, ser hombre o mujer no influye en el nivel obtenido en la prueba.

Desempeño	Variable	Valor Chi-cuadrado	p-valor	Coeficientes	
				Cramer	Contingencia
Lectura Crítica	Genero	0.065927	0.9676	0.006	0.006
	Area_Resi	6,7571	0.0341	0.057	0.057
	Naturaleza	44.556	2,113E-10	0.145	0.147
	Jornada	69.721	5,58E-12	0.181	0.13
	Estrat_Vivie	96.641	2.2e-16	0.211	0.153
Matemáticas	Genero	37.235	8,21E-09	0.133	0.134
	Area_Resi	12.685	0.00176	0.078	0.078
	Naturaleza	67.812	1,88E-15	0.178	0.181
	Jornada	149.92	2.2e-16	0.26	0.19
	Estrat_Vivie	105.64	2.2e-16	0.22	0.16
Sociales	Genero	23.794	6,81E-06	0.107	0.107
	Area_Resi	7,5773	0.02263	0.06	0.061
	Naturaleza	39.397	2,79E-09	0.137	0.138
	Jornada	109	2.2e-16	0.224	0.162
	Estrat_Vivie	56.445	2.31e-09	0.163	0.117

Naturales	Genero	23.608	7,47E-06	0.106	0.107
	Area_Resi	12.033	0.002438	0.076	0.076
	Naturaleza	87.911	2.2e-16	0.202	0.206
	Jornada	156.81	2.2e-16	0.265	0.195
	Estrat_Vivie	102.38	2.2e-16	0.217	0.157
Inglés	Genero	9,8454	0.04311	0.069	0.069
	Area_Resi	42.844	1,12E-08	0.142	0.144
	Naturaleza	161.05	2.2e-16	0.269	0.279
	Jornada	393.6	2.2e-16	0.4	0.218
	Estrat_Vivie	294.79	2.2e-16	0.353	0.189

Tabla 23: Resultados prueba de independencia

Al revisar los resultados de las pruebas de independencia de las competencias Matemáticas, Sociales, Naturales e Ingés con las variables sociodemográficas se observa que el nivel de desempeño en cada competencia si está asociadas con dichas variables. Es decir, el género, el área de residencia, la naturaleza del colegio, la jornada y el estrato si influye en el nivel de desempeño de la prueba SABER 11, lo anterior se ratifica con los valores de los coeficientes Cramer y Contingencia. Es importante resaltar que para la competencia de ingés, el nivel de desempeño esta influenciado por las variables sociodemográficas y que hay una fuerte asociación entre el nivel de desempeño con la naturaleza del colegio, es decir, influye y es muy relevante si es oficial o no; a su vez nivel de desempeño con la jornada y el estrato. Y tal como se mencionó con anterioridad es importante reagrupar las cateogorías para algunas variables y observar si se mantienen las conclusiones, además, para las variables ordinales es posible usar otros coeficientes de asociación como Pearson para detectar si hay tendencia lineal.

5. Conclusiones

- En cuanto a las variables sociodemográficas se concluye que la mayoría de los estudiantes son hombres (54 %), además la edad promedio de ellos es 18 años, un alto número de estudiates no trabajan (88 %) y se puede afirmar que en cuanto a su área de residencia predomina la urbana (85 %).

- De las variables asociadas a la familia, se evidencia que el nivel educativo de los padres en gran parte es primaria completa (20 %) y secundaria completa (22 %), para el caso de las madres la primaria completa es para el 17 % y la secuendaria completa el 28 %; en lo relacionado a la ocupación de los padres gran parte tienen trabajo propio (29 %), como también son obreros operarios (29 %), las madres, por su parte, trabajan en el hogar(53 %). La mayoria de los estudiantes pertenecen a los estratos socioeconomicos correspondientes a 1 y 2 (75 %), tienen nivel 1 y 2 de sisben (55 %) y los ingresos de las familias oscilan entre de uno a dos SMLV.

- Gran parte de los estudiantes son de colegios oficiales (70 %), de los cuales, el 31 % esta en jornada completa. El 79 % de los estudiantes manifestó no pagar pensión y respecto al salario de las familias a las que perteneces el 67 % tiene ingresos por debajo de dos salarios mínimos mensuales.

- Entre las asignaturas con puntajes más altos, encontramos matemáticas. Por otro lado, la gran mayoría de los estudiantes se sitúa en desempeño medio en los componentes y tuvieron un nivel bajo en inglés.

- Se puede observar que con respecto a los puntajes, el puntaje promedio está alrededor de 55 puntos, al revisar el valor de la desviación estándar se aprecia que los puntajes en cada competencia son heterogéneos, lo anterior lo anterior se corrobora con los coeficientes de variación, los cuales oscilan entre el 17 % y 20 %. El valor alto de la kurtosis, en especial para la prueba de ingés indica la presencia de valores atípicos.

- Al revisar las pruebas de independencia se observó que en el desempeño de Lectura Crítica la única variable con la que no esta relacionado es el género, es decir, ser hombre o mujer no influye en el nivel obtenido en la prueba. Para las demás competencias se encontró que el nivel de desempeño si esta asociado con el género, naturaleza del colegio, jornada, área donde reside y estrato; que en particular en la competencia de ingés el nivel de desempeño esta altamente relacionado con la jornada, naturaleza del colegio y el estrato del estudiante.

Referencias Bibliográficas

Alonso, J., Casasbuenas, P., Gallo, B. y Torres, G. (2012), 'Bilingüismo en santiago de cali: Análisis de los resultados de las pruebas saber 11 y saber pro'.

Díaz M., Guillermo, L. y Morales R., Mario, A. (2009), *Análisis estadístico de datos categóricos*, primera edn, Universidad Nacional de Colombia.

ICFES (2013), 'Guía de acceso a bases de datos icfes'.

ICFES (2014), *Lineamientos generales 2014 - 2*, ISBN de la versión electrónica: 978-958-11-0630-1.

ICFES (n.d.), 'Turorial carga de bases de datos en excel'.

Medina N., Armando, E. y Salazar, de la Torre, J. (2015), 'Implementación del proyecto pro-saber 11 en el grado 11 del colegio ekklesía para fortalecer las competencia que se evalúan en las pruebas saber 11-2015 y lograr un desempeño medio alto en las pruebas del estado'.

MEN (2010), 'Decreto 869. Bogotá'.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

PLANTEAMIENTO DE UN DISEÑO EXPERIMENTAL DE MEZCLAS PARA LA FABRICACIÓN DE CEMENTO CON PUZOLANA

Especialización en Estadística

David Julián Sotelo López[1,a], Eduardo Dávila Sanabria[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El cemento de uso general producido en cementos ARGOS, planta Sogamoso contiene principalmente Clinker, yeso, escoria y caliza. Para preservar los niveles de adición y de consumo de materias primas, se deben mantener los inventarios de cada uno de sus componentes. En este momento se presenta una escases de la escoria de alto horno, lo cual hace necesario la posibilidad de reemplazar este material; mediante la inclusión de puzolana, la cual reemplazara en proporción a la escoria. Para ello se plantea un diseño experimental aleatorio, donde se le aplicara un análisis de variancia a los datos obtenidos, teniendo como premisa el cumplimiento de las resistencias a la compresión detalladas en la NTC 121 y así formular la mezcla más eficiente frente al consumo de puzolana.

***Palabras clave*:** Análisis de Varianza, Diseño de Experimentos, Escoria, Puzolana..

Abstract

The cement of general use produced in Cementos Argos, Sogamoso plant, mainly contain Clinker, gypsum, slag and limestone. To preserve the addition levels and the consumption of raw materials, it is required to maintain the reserves of these components. At this moment, a scarcity of slag from the top oven is present, then, by inclusion of the pozzolan is expected to reduce in a portion the slag required in the process. For that, it proposes a random experimental design, with the objective to apply a variance analysis of the data, to formulate the most efficient mixture, with the premise of the fulfilment of the normativity of the resistance to compression detailed in the NTC 121.

***Key words*:** Variance Analyse, Experimental design, Slag, Pozzolan..

1. Introducción

En la planta de cementos ARGOS, ubicada en la ciudad de Sogamoso, se produce cemento de uso general, donde sus materias primas son el Clinker, yeso, caliza y escoria de alto horno. La caliza utilizada es proveniente de las minas de cementos ARGOS y la escoria es comprada a la siderúrgica VOTORANTIM , esta adición es producto de la ineficiencia en la producción de acero, lo cual resulta un material que ha futuro disminuirá su producción, por lo tanto el reemplazo de esta materia prima es fundamental, para mantener los niveles de adición que actualmente se mantienen.

[a]Estudiante de Especialización en Estadística. E-mail: david.sotelo@uptc.edu.co
[b]Profesor asistente. E-mail: eduardo.davila@talex.com.co

La disminución de la producción de escoria por parte de VOTORANTIM[1], hace insostenible la adición de escoria en los cementos producidos por cementos ARGOS, planta Sogamoso. La cantidad de escoria necesaria para mantener los tenores de escoria en el año 2016 está estimada en $260.000t$ de escoria húmeda, esta adición está distribuida entre el cemento de uso general con una adición del 40 % y de cemento concretero con una adición del 10 %. Por parte de VOTORANTIM el contrato está pactado por $150.000t$, lo cual implicaría un déficit de 110.000 t de escoria.

En el año anterior este déficit se suplió con escoria importada con un costo aproximado de $224.266 por tonelada, que al realizar la comparación con el costo de la escoria de VOTORANTIM es de $35.791, lo cual aumenta el costo por tonelada en $188.475, y por tal razón el costo del cemento. Por esta razón surge la necesidad de buscar un reemplazo a la escoria, bien sea para mantener las adiciones actuales y cumplir con el presupuesto del año tanto de producción como de costo del cemento.

La búsqueda de una mezcla que satisfaga los objetivos de calidad y costo, es fundamental para la sostenibilidad del negocio, por lo tanto la inclusión de una adición activa como la puzolana dentro del proceso de la fabricación del cemento en la planta Sogamoso de cementos ARGOS, podría ser la mejor opción técnicamente hablando y de fácil adquisición dentro de la región.

El costo de la puzolana es de $\$34.996/t$, por lo cual habría una reducción de costos de $795 por tonelada de puzolana que se reemplace por escoria, y al realizar la comparación con el costo de la escoria importada estaríamos hablado de una reducción de $\$189.270/t$, de igual forma la importancia de esta mezcla, es que se mantengan los consumos de Clinker, en primera medida por su costo el cual es de $\$162.000/t$, ya que al no hacerlo se vería afectada la relación Clinker cemento (t de cemento/ t de Clinker), lo cual implicaría mayor costo del cemento, y mayor emisión de CO2 a la atmósfera, por lo tanto la posibilidad de usar la puzolana en el cemento de uso general, tendiendo a reemplazar en su mayor proporción la escoria de alto horno, cumpliendo con la NTC 121 en lo referente a las resistencias a la compresión, mediante el planteamiento de un diseño experimental y posterior análisis de los datos obtenidos.

2. Referente Conceptual

2.1. Diseño Experimental

En el diseño estadístico se trata de entender de donde proviene la varianza y asegurar su correcta distribución entre las diferentes muestras (Casella 2010). De igual forma los objetivos de un experimento son los de determinar cuáles son las variables que tienen mayor influencia sobre la respuesta y, y cuál es el ajuste de las covariables que tiene mayor influencia para que "Y" esté casi siempre cerca del valor nominal deseado, para que los efectos de las variables no controlables $Z_1, Z_2, ..., Z_q$ sean mínimos.Montgomery (2004)

La aplicación de las técnicas del diseño experimental en las fases iniciales del desarrollo de un proceso puede redundar en mejorar el rendimiento de un proceso, reducir la variabilidad y conformidad más cercana con los requerimientos nominales o proyectados, reducción del tiempo de desarrollo y reducción de los costos globales.

Los tres principios básicos del diseño experimental son la realización de réplicas, la aleatorización y la formación de bloques. Por realización de réplicas se entiende la repetición de la condición experimental o tratamiento, en unidades homogéneas, la aleatorización se entiende que tanto la asignación del material experimental como el orden en que se realizarán las corridas o ensayos individuales del experimento que se determinan al azar; y la formación de bloques es una técnica de diseño que se utiliza para mejorar la precisión de las comparaciones que se hacen entre los factores de interés. Muchas veces la formación de bloques se emplea para reducir o eliminar la variabilidad transmitida por factores perturbadores; es decir, aquellos factores que pueden influir en la respuesta experimental pero en los que no hay un interés específico

La realización de réplicas posee dos propiedades importantes. Primera, permite al experimentador obtener una estimación del error experimental. Esta estimación del error se convierte en una unidad de medición básica para determinar si las diferencias observadas en los datos son en realidad estadísticamente diferentes. Segunda, si se usa la media muestral para estimar el efecto de un factor en el experimento, la realización de réplicas permite al experimentador obtener una estimación más precisa de este efecto.

[1] Acerías Paz Del Rio

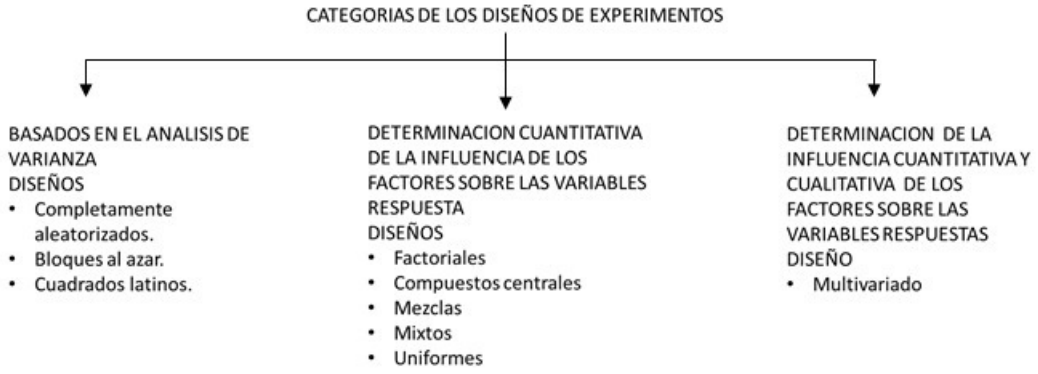
La aleatorización es la piedra angular en la que se fundamenta el uso de los métodos estadísticos en el diseño experimental. Por aleatorización se entiende que tanto la asignación del material experimental como el orden en que se realizarán las corridas o ensayos individuales del experimento se determinan al azar. Uno de los requisitos de los métodos estadísticos es que las observaciones (o los errores) sean variables aleatorias con distribuciones independientes. La aleatorización hace por lo general que este supuesto sea válido. La formación de bloques es una técnica de diseño que se utiliza para mejorar la precisión de las comparaciones que se hacen entre los factores de interés. Muchas veces la formación de bloques se emplea para reducir o eliminar la variabilidad transmitida por factores perturbadores; es decir, aquellos factores que pueden influir en la respuesta experimental pero en los que no hay un interés específico.

De igual forma se debe tener una metodología para plantear un diseño experimental y se deben seguir las diferentes recomendaciones para su realización.

2.2. Pautas generales para el diseño de experimentos

- Identificación y exposición del problema.
- Elección de los factores, los niveles y los rangos.
- Selección de la variable de respuesta.
- Elección del diseño experimental.
- Realización del experimento.
- Análisis estadístico de los datos.
- Conclusiones y recomendaciones.

Los diseños de experimentos han sido agrupados en tres categorías, atendiendo al tipo de problema que se pretende resolver con los mismos. En la Figura 1 se observa que la primera categoría incluye los diseños basados en el análisis de varianza, la segunda incluye aquellos en los que se pretende encontrar un modelo matemático, por lo general polinomial, que conecte los factores con las respuestas, y la tercera categoría incluye solamente a los diseños multivariados, los cuales permiten definir un modelo matemático similar al descrito para la segunda categoría. Además, en este pueden incluirse variables cualitativas siendo llevadas a numéricas empleando la técnica de análisis de componentes principales.



Especialistas en diseños de experimentos han definido 8 reglas para el desarrollo exitoso de estos. A continuación se listan las mismas Pérez y et al. (2008)

1. Definir objetivos.
2. Medir respuestas cuantitativamente.

3. Replicar para amortiguar la variación incontrolable (ruido).
4. Aleatorizar el orden de las corridas.
5. Bloquear las fuentes de variación conocidas.
6. Conocer cuáles efectos pueden estar confundidos.
7. Realizar una secuencia de diseños de experimentos.
8. Confirmar siempre los resultados críticos.

2.3. Análisis de varianza

El análisis de varianza es una técnica estadística para analizar mediciones que dependen de varias clases de efectos que operan simultáneamente, para estimar los efectos y para decidir cuales efectos son importantes. Díaz (2009)

La aplicación del modelo del análisis de varianza simple, se realiza donde solo un factor es el centro de la investigación y debe cumplir con la asignación aleatoria de tratamientos en las unidades experimentales, de igual forma los errores del modelo deben ser variables aleatorias que siguen una distribución normal e independientes con media cero y varianza δ^2.

Se encontrará útil describir las observaciones de un experimento con un modelo. Una manera de escribir este modelo es

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{1}$$

Donde y_{ij} es la observación ij-ésima, μ es la media global, τ_i es un parámetro único del tratamiento al que se le llama el efecto del tratamiento i-ésimo, ε_{ij} es un componente del error aleatorio que incorpora las fuentes de variabilidad del experimento.

El ANOVA es la mejor manera de pensar acerca de los datos y diseños de experimentos. Fisher (1934) lo llamó por primera vez " un método conveniente para la organización de la aritmética ", pero luego fue demostrado rigurosamente por Speed (1987). Las ideas de la repartición de la variación, distribuyendo correctamente los grados de libertad, y la identificación de los términos de error correctamente, son fundamentales para analizar los datos.

En un análisis de la varianza (ANOVA), a los datos recogidos se aplica una repartición de la variabilidad de los datos en partes que pueden atribuirse a diferentes factores.

$$SS_{Total} = SS_{Tratamientos} + SS_E \tag{2}$$

Donde a $SS_{Tratamientos}$ se le llama la suma de cuadrados debida a los tratamientos (es decir, entre los tratamientos), y a SS_E se le llama la suma de cuadrados debida al error (es decir, dentro de los tratamientos).Hay $an = N$ observaciones en total; por lo tanto, SS_{Total} tiene $N - 1$ grados de libertad. Hay a niveles del factor (y medias de a tratamientos), de donde $SS_{Tratamientos}$ tiene $a - 1$ grados de libertad. Por último, dentro de cualquier tratamiento hay n réplicas que proporcionan $n - 1$ grados de libertad con los cuales se estima el error experi-mental. Puesto que hay tratamientos, se tienen $a(n - 1) = an - a = N - a$ grados de libertad para el error.

2.4. Diseños con Mezclas

En las generalidades de los diseños con mezclas es preciso aclarar que estos diseños de experimentos no son una nueva alternativa a los clásicos planes factoriales a dos niveles. Estos últimos trabajan con variables que son totalmente independientes (variables de proceso), mientras los primeros están definidos para problemas de mezclas físicas de componentes donde no existe total independencia entre las variables, que en este caso

son las proporciones de los ingredientes de una formulación dada. La dependencia entre las variables está condicionada por la restricción de unicidad, su representación matemática es la siguiente:

$$0 \leq x_i \leq 1,\ i = 1,\ 2,\ 3, ...,\ q \tag{3}$$

$$\sum_{i=1}^{q} x_i = x_1 + x_2 + x_3 + x_4 + ... + x_q = 1 \tag{4}$$

De esta forma se define que la variable respuesta de los problemas con mezclas depende solamente de la proporción de los ingredientes presentes en la mezcla y no en la cantidad de mezcla. Cornell (2011) Para el análisis de los datos de un diseño de mezclas se deben responder las siguientes preguntas.

1. Los valores de la variable respuesta son similares en las diferentes mezclas planteadas?
2. Si la respuesta es "NO", los valores de la variable respuesta están en función de la de la mezcla?
3. Si su respuesta es "SI" a la pregunta 2, es posible expresar la relación entre la variable respuesta y la composición química de la mezcla en forma de una ecuación que tenga sentido?

Los objetivos potenciales de un experimento de mezcla, son lo de modelar la dependencia de la variable respuesta en las proporciones relativas de los componentes con alguna forma de ecuación matemática. Las fórmulas matemáticas se pueden configurar para tener en cuenta lo siguiente:

1. La influencia en la respuesta de cada componente por separado y en combinación con los otros componentes se pueden medir. Los componentes con el menor efecto o de menor influencia podrían entonces ser "filtrados", dejando sólo los componentes que tienen el mayor efecto en la variable respuesta.
2. Las predicciones de la variable respuesta a cualquier mezcla o combinación de las proporciones de los componentes se pueden hacer.
3. Mezclas de los componentes que producen los valores deseables de la respuesta.

Para desarrollar el análisis de los datos el primer paso en la prueba, es demostrar que los componentes se mezclan de forma no lineal, lo cual es determinar si los valores de las medias son diferentes. Para ello, se denota el verdadero porcentaje medio de cada una de las n mezclas utilizando la letra griega μ y el estado de la hipótesis nula es:

$$\begin{aligned} H_0 &= \mu_1 = \mu_2 = \mu_3 = \mu_n \\ H_1 &= \mu_1 \neq \mu_2 \neq \mu_3 = \mu_n \end{aligned} \tag{5}$$

Mientras que la hipótesis alternativa se plantea donde uno o más de las medias son diferentes; El análisis se inicia mediante la creación de un análisis de varianza (ANOVA); Se calculan, por tanto, $MS_{Tratamientos}$ y MS_E como una medida de las dispersiones comentadas y se comparan mediante una prueba de hipótesis F. Si no existe diferencia estadísticamente significativa entre ellas, la presencia de errores aleatorios será la causa predominante de la discrepancia entre los valores medios. Si, por el contrario, existe algún error sistemático, será $MS_{tratamientos}$ mucho mayor que MS_E, con lo cual el valor calculado de F será mayor que el valor tabulado F_{tab} para el nivel de significación α escogido y los grados de libertad mencionados. Boqué y Maroto (000)

Al rechazar la hipótesis nula se da respuesta a la primera pregunta anteriormente mencionada para el análisis de datos, ahora se debe responder a la pregunta si los los de la mezcla, para comprobar su dependencia se debe comenzar ajustando la ecuación de regresión lineal simple y posteriormente realizar el análisis de varianza al modelo y de igual forma se debe evaluar la bondad de ajuste del modelo.

2.5. Medida de la Bondad de Ajuste

El análisis de la varianza descrito anteriormente nos da un criterio para decidir si alguno de los parámetros es distinto de cero y, por tanto, si las variables regresoras explican significativamente la variabilidad de la variable independiente, sin embargo, no miden el grado de la relación existente entre la dependiente y las regresoras. Una medida descriptiva del grado de la relación existente entre las variables se denomina *Coeficiente de Determinación*, se denota con R^2 y se define como el cociente entre la suma de cuadrados explicada y la suma de cuadrados total:

$$R^2 = \frac{SS_{Tratamiento}}{SS_{Total}} \tag{6}$$

Está acotado entre 0 y 1 y multiplicado por 100 representa el porcentaje de la variabilidad de la variable dependiente explicado por la introducción de las regresoras en el modelo lineal modelo lineal. Para el modelo de regresión simple en el que se dispone de una sola variable regresora, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación de Pearson.

El coeficiente de determinación es sencillo y fácil de interpretar aunque tiene un problema importante, aumenta con el número de variables regresoras, estén o no relacionadas con la dependiente, de forma que es posible conseguir una bondad del ajuste próxima a 1 simplemente introduciendo en el modelo un número elevado de variables. Para evitar este problema se define el *Coeficiente de Determinación Ajustado*, en el que las sumas de cuadrados se dividen por sus correspondientes grados de libertad.

$$R^2 = \frac{SS_{Tratamientos}/(a-1)}{SS_{Total}/(N-1)} \tag{7}$$

3. Diseño Metodológico

3.1. Método de Campo

Se realizara el planteamiento de un experimento de mezclas, en la fabricación de cemento de uso general en la planta de cementos ARGOS, ubicada en el kilómetro 7 vía Sogamoso Corrales, donde se tiene como objetivo la inclusión de la puzolana como adición en cemento teniendo como variable respuesta las resistencias a la compresión dada en MPa cumpliendo la norma NTC 121.

3.2. Sistema de Hipótesis

3.2.1. Hipótesis General

Incluir la puzolana en la fabricación de cemento tipo uso general.

3.2.2. Hipótesis Específica

Es posible usar la puzolana en el cemento de uso general, tendiendo a reemplazar en su mayor proporción la escoria de alto horno, cumpliendo con la NTC 121 en lo referente a las resistencias a la compresión.

3.3. Sistema de Variables

La variable respuesta del experimento son las resistencias a la compresión del cemento producido, este valor será medido en el laboratorio físico de cementos ARGOS, según el procedimiento de la NTC121, planta Sogamoso.

3.3.1. Variables independientes

Las variables intervinientes en el experimento es la proporción de puzolana y escoria.

3.3.2. Variables Intervinientes

En el momento de realizar el experimento se debe tener en cuenta las siguientes variables, las cuales pueden incidir en el resultado final.

El C3S del Clinker utilizado debe estar en los rangos de 59 % a 62 %. La finura del cemento debe estar en un retenido del $7\,\% \pm 0.5$ en malla de 90μ. El SO_3 en el cemento debe estar en un rango de $2.8\,\% \pm 0.1$. La dosificación del aditivo XP203, debe estar en 300 ppm. Si durante la realización del experimento la molienda para por algún motivo este debe detenerse y no se reinicia hasta normalizar el proceso de molienda.

3.3.3. Diseño Experimental

El tipo de diseño experimental a aplicar es de tipo aleatorio, basado en un diseño experimental de mezclas, donde se realizara un análisis de variancia para describir los tratamientos como proporciones que varían entre los dos ingredientes.

4. Resultados

El planteamiento del experimento de mezclas se basa en mantener la adición de caliza en 20 %, Clinker en 37 % y SO_3 en 3 % con las restricciones anteriormente mencionadas; los elemento a variar son la puzolana y la escoria las cuales se muestran en la Tabla 1.

Componentes	Proporciones(%)				
Escoria	0	10	20	30	40
Puzolana	40	30	20	10	0

Tabla 1: Proporciones por componente

El procedimiento para realizar el experimento se denota en la tabla 2.

Etapa	Tiempo [min]	Parámetro medido	Resultado
Preparación de materiales a dosificar		Volumen para la molienda de 5h cemento uso general.	Clinker 116,5t.; Escoria 63t.; Puzolana 63t.; SO_3 9,45t.
Inicio de molienda y ajuste de parámetros	60	Retenido 7 % en malla de 90μ. Escoria 0 %; Puzolana 40 %;	Toma de muestra para análisis de laboratorio.
Mezcla 2	60	Retenido 7 % en malla de 90μ. Escoria 10 %; Puzolana 30 %;	Toma de muestra para análisis de laboratorio.
Mezcla 3	60	Retenido 7 % en malla de 90μ. Escoria 20 %; Puzolana 20 %;	Toma de muestra para análisis de laboratorio.
Mezcla 4	60	Retenido 7 % en malla de 90μ. Escoria 30 %; Puzolana 10 %;	Toma de muestra para análisis de laboratorio.
Mezcla 5	60	Retenido 7 % en malla de 90μ. Escoria 0 %; Puzolana 40 %;	Toma de muestra para análisis de laboratorio.

Tabla 2: Resultados de la aplicación del procedimiento experimental.

Este procedimiento tendrá 4 repeticiones, las cuales se harán con diferencia de 24h.

El planteamiento del análisis de variancia seria de la siguiente forma teniendo como nivel de significancia de 0,05.[2]

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrado medio
Entre mezclas	$SS_{Tratamientos}$	$5-1=4$	$MS_{Tratamientos}$
Dentro de las mezclas	SS_E	$5(4-1)=15$	MS_E
Total	SS_{Total}	$(5*4)-1=19$	

Tabla 3: Análisis estadístico del procedimiento

5. Cronograma

ACTIVIDAD	MES	Junio				Julio			
	SEMANA	1	2	3	4	1	2	3	4
Preparación de materia prima	Programado								
	Ejecutado								
Mezcla 1	Programado								
	Ejecutado								
Mezcla 2	Programado								
	Ejecutado								
Mezcla 3	Programado								
	Ejecutado								
Mezcla 4	Programado								
	Ejecutado								
Mezcla 5	Programado								
	Ejecutado								

6. Conclusiones

Al aplicar un diseño de experimentos podemos contemplar todas las variables que pueden influir en la variable respuesta, y así determinar la forma en que se analizaran los datos para de esta manera obtener el mayor beneficio del experimento, de igual la aplicación de un ajuste de bondad optimo es fundamental para evaluar el modelo de regresión y así realizar las transformaciones necesarias para garantizar la fiabilidad del modelo planteado, de igual forma el protocolo del experimento se formaliza para su ejecución y evaluación.

Referencias Bibliográficas

Boqué, R. y Maroto, A. (2000), *El análisis de la varianza*, Grupo de Quimiometría y Cualimetría. Universitat Rovira i Virgili. Pl. Imperial Tarraco, 1. 43005-Tarragona.

Casella, G. (2010), *Statistical design*, Springer.

Cornell, J. (2011), *A Primer on Experiments with Mixtures*, WILEY.

[2]El estudio estadístico de prueba (F) se obtuvo a partir de la relación $\frac{MS_{Tratamientos}}{MS_E}$

Díaz, A. (2009), *Diseño estadístico de experimentos*, Universidad de Antioquia.

Montgomery, D. (2004), *Diseño y análisis de experimentos*, Limusa Wiley.

Pérez, I. y et al. (2008), 'Diseños de experimentos en tecnología y control de los medicamentos', *Revista Mexicana de Ciencias Farmacéuticas* pp. 28–40.

APLICACIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA CON RESPUESTA POLITÓMICA ORDINAL EN EL ANÁLISIS DEL DESEMPEÑO ACADÉMICO EN MATEMÁTICAS

Proyecto desarrollado con base en los registros institucionales del colegio Mariano Ospina de Bogotá del año 2012

Especialización en Estadística

John Jairo Gonzalez Gonzalez[1,a], Carmen Helena Cepeda Araque[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Un profesor debe usar herramientas que le permitan intervenir en el proceso de enseñanza aprendizaje de sus alumnos para aumentar el índice de efectividad académica. El siguiente artículo hace uso de la regresión logística ordinal para identificar los factores que afectan el desempeño de estudiantes en matemáticas. Se encontró que el promedio total del estudiante, el número de materias perdidas el año anterior, la calificación en comportamiento y el género del estudiante explican la variabilidad en el desempeño en el área. Se estimaron los parámetros mediante el método de máxima verosimilitud, se calculó la calidad del ajuste mediante el coeficiente pseudo-R^2de Mc-Fadden y de Nagelkerke.

***Palabras clave*:** Regresión Logística Ordinal, Desempeño estudiantes.

Abstract

A teacher should use tools that allow it to intervene in the process of learning of their students to increase academic effectiveness index . The following article uses ordinal logistic regression to identify factors that affect the performance of students in mathematics. It was found that the total average student, the number of lost materials last year, qualifying in behavior and gender of the student explain the variability in performance in the area. It was estimated The parameters estimation mere through method of maximum likelihood, it was calculated the quality of the adjustment through the coefficient $R^2$and pseudo-R^2 and Nagelkerke.

***Key words*:** Performance students, ordinal logistic regression..

1. Introducción

Mediante la aplicación de la Regresión Logística Ordinal se da a conocer a los usuarios de la estadística una herramienta de análisis de dependencia en el caso de la variable respuesta politómica ordinal, con este propósito se efectuó el análisis del desempeño académico, año 2012, de los estudiantes de grado sexto del colegio Mariano Ospina de la ciudad de Bogotá.

[a]Estudiante de Especialización en Estadística. E-mail: jhosand01@yahoo.es
[b]Profesor asistente. E-mail: carmen.cepeda@uptc.edu.co

La información que arrojó la aplicación es importante ya que permitió identificar que el promedio total del estudiante, el número de materias perdidas el año anterior, la calificación en comportamiento y el género son factores que influyen en el desempeño académico de un estudiante en el área de matemáticas. Aspecto que contribuye a diseñar planes de mejoramiento del trabajo de aula del profesor.

Para estudiar varias mediciones simultáneamente es útil un modelo matemático para explicar las observaciones y sus relaciones. El modelamiento es la aplicación de una serie de pasos tales como estimación, juzgamiento de hipótesis, diagnosis y replanteamiento para conseguir una explicación apropiada del comportamiento de una variable respuesta (datos) a partir de una función ponderada de una o más variables explicativas modelo (Díaz y Morales, 2012). [1]

El uso de la regresión logística ha ampliado su campo de acción ya que por medio de ella se pueden analizar variables de tipo categórico y por la utilidad de la información que se deriva del análisis del denominado odds ratio (Hosmer y Lemeshow 2000).

En los eventos cuya probabilidad que se desea explicar corresponden a variables dependientes categóricas ordinales, es decir, aquellas cuyos valores no sólo diferencian a los individuos sino que también permiten establecer un orden entre estos Heredia, Rodríguez y Vilalta (2009), es recomendable el empleo de la regresión logística ordinal, ya que esta capta la relación de orden de las diferentes categorías de la variable objeto de estudio a fin de realizar predicciones más confiables. La importancia de mantener una clasificación de la información de los estadios de la variable dependiente según Ponsot. E Shira, Agresti. A y Mccullangh (1990) (Agresti 1990), radica en que sin este orden, no se captaría cabalmente la influencia de las variables explicativas sobre la variable dependiente al no considerar la información acerca de las diferencias de orden entre las categorías de esta última.

En gran parte del territorio Nacional, el desempeño del estudiantes es valorado en determinadas materias mediante una escala ordinal, en ocasiones estas últimas varían junto a sus parámetros de exigencia, comenzando desde, bajo o deficiente hasta las más altas superior o excelente. Para el caso particular de este estudio se tiene que el desempeño en matemáticas se determina por bajo, básico, alto y superior.

2. Referente Conceptual

Un modelo lineal generalizado se origina cuando interesa modelar un experimento en el cual la variable respuesta Y tiene una distribución que pertenece a la familia exponencial de densidades y que está asociada a un conjunto de variables explicativas X_1, X_2, ..., X_P. En esta clase de modelos se distinguen tres componentes denominados componente aleatorio, sistemático y función de enlace. La siguiente es la descripción de cada uno de ellos. De acuerdo a (Díaz Monroy y Morales Rivera 2012, pág.135) para estudiar entre los datos y el modelo existe una discrepancia que se denomina error residual, es decir Datos = modelo + error:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i \tag{11}$$

El propósito del modelamiento estadístico es la "búsqueda del modelo más simple que sea capaz de explicar los datos con el mínimo error posible". Esto equivale a buscar un modelo parsimonioso que se ajuste adecuadamente a los datos (Díaz Monroy y Morales Rivera 2012, pág.136).

Componente aleatorio: Está representada por el conjunto de variables respuesta independientes y_i, $i = 1, 2, ..., n$ cuya distribución para todo i pertenece a la familia exponencial.

Componente sistemático: Está representada por el conjunto de variables explicativas $X_1, X_2, ..., X_P$ y una relación de la forma $\eta_s = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$ equivalente a $\eta = X\beta$ con $s = 1, 2, ..., n$, donde β es el vector de parámetros y X es la matriz del modelo que puede estar asociada a un modelo de regresión múltiple de rango completo si las X_i, son variables cuantitativas o a un modelo de regresión de rango incompleto si el diseño es por ejemplo un diseño factorial, un diseño de bloques u otro diseño experimental con las X_i variables categóricas o de clasificación.

Función de enlace: Es una función g monótona, derivable que asocia o enlaza la componente aleatoria y sistemática por la relación: $g(u_i) = \eta_s$

[1] Este proyecto se ha desarrollado con base en los registros institucionales del colegio Mariano Ospina de Bogotá del año 2012 (Hosmer y Lemeshow 2000)

2.1. Regresión logística ordinal - RLO

En muchas situaciones, las categóricas de la variable respuesta tienen alguna clase de ordenamiento. En estos casos no es apropiado el uso de la regresión logística nominal, porque se podría perder la capacidad de detectar la forma en que la variable respuesta está relacionada con las variables independientes.

Si queremos modelar una variable respuesta categórica, Y, de categorías y1, . . . , yg, con un conjunto de variables explicativas (factores o covariables) X= X_1, ..., X_P, mediante un modelo lineal general, podemos plantearnos las opciones siguientes:

Si se tienen dos categorías en la variable respuesta y no importa el orden utilizamos la regresión logística. Si se tienen tres o más categorías en la variable respuesta y no importa el orden hacemos uso de la regresión logística multinomial. Si se tienen tres o más categorías en la variable respuesta y al mismo tiempo importa el orden, recurrimos a la regresión logística ordinal.

Así, el modelo de regresión ordinal es adecuado para modelar la variable de desempeño académico, y para este caso se tendrán 4 categorías en la variable respuesta.

Siguiendo los planteamientos de Ponsot y goitía (1980), la regresión logística en su forma más simple, es decir, con una respuesta binaria, propone que el logaritmo de la razón de probabilidad (odds según su denominación en inglés se puede acoplar a una distribución Bernoulli entendida normalmente como predictor lineal ya que es similar a una función lineal en los parámetros.

Un aspecto común es la función de enlace ya que tiene la capacidad de relacionar las covariables de forma líneal con la razón de probabilidad acumulada hasta una determinada categoría de una variable ordinal, por lo general, en la RLO se recurre a la función de enlace logit. Sin embargo, también se puede usar la Cloglog y la Probit, Según Agresti. A (1990), no existe un criterio que defina con claridad en qué caso es más adecuada cada una de estas funciones, y cuando existen dudas sobre cuál emplear, generalmente se utilizan ambas y se comparan los resultados para escoger los más satisfactorios.

Así mismo, cada proceso relacionado con la función logit cumple por lo general que para variables o datos de estudio se debe observar una distribución de frecuencias uniforme en todos los niveles, por otro lado Mccullagh (1980) plantea que la función de enlace Logit es más adecuada para analizar datos ordinales cuya distribución de frecuencia es uniforme a lo largo de todas las categorías, mientras que la unión Cloglog es preferible para analizar datos categóricos cuyas categorías de mayor valor son las más probables.

En el caso que nos compete los valores de nuestra variable respuesta son los posibles desempeños en el área de matemáticas por parte de los estudiantes de grado sexto, es decir que los valores de la variable respuesta corresponden a una variable ordinal, y donde los registros institucionales y otros son tomados por la valoración cuantitativa de los docentes en unas áreas específicas. La variable desempeño de los estudiantes puede dividirse en categorías, convirtiendo a esta, la variable respuesta en ordinal, tales categorías son determinadas por la valoración hecha a partir de la aplicación de talleres evaluaciones entre otros y puede mostrar una idea de la evolución del proceso de aprendizaje del estudiante en un determinado espacio de tiempo, estas categorías son BAJO, BASICO, ALTO y SUPERIOR.

Así las cosas, hemos de recurrir al enlace Logit el más adecuado para nuestra situación ya que por lo anterior no se puede establecer que los niveles más altos sean en donde se destaquen la mayoría de estudiantes o en el caso del estudio los más probables.

La expresión de la función Logit para la RLO es la siguiente:

$$\ln(O_i) = \alpha_i + \beta X \tag{1}$$

En esta ecuación, O_i es la razón de probabilidad (odds) asociada a la categoría i de la variable dependiente, siendo la expresión de esta razón:

$$O_i = \frac{p(\text{valor sea} \leq \text{categoria i/valores de X})}{p(\text{valor sea}>\text{categoria i/valores de X})} \tag{2}$$

Lo que es lo mismo que:

$$O_i = \frac{p(\text{valor sea} \leq \text{categoria i/valores de X})}{(1 - p(\text{valor sea}>\text{categoria i/valores de X}))} \tag{3}$$

Con la premisa de tener en cuenta los planteamientos expuestos por Heredia, Rodríguez y Vilalta (2009) en las expresiones (2) y (3) el término valor hace referencia a cualquier valor de la variable dependiente. En cada una de las expresiones las probabilidades son condicionales lo que indica que cada una de ellas está ligada a las posibles situaciones que experimente el estudiante.

Continuando con los planteamientos de Heredia , Rodríguez y Vilalta (2009), en la ecuación (1), $\propto_i$ es el intercepto asociado a la ecuación que modela la razón de probabilidad de la categoría i, y B es vector de coeficientes de la ecuación de regresión. Si existen p variables independientes, existen p coeficientes, y BX corresponde a la combinación lineal $BX_1 + BX_2 + ... + BX_P$. Estos coeficientes cuantifican el efecto de las variables independientes sobre el logaritmo de la razón de probabilidad.

Es de vital tener en cuenta que si la variable respuesta o dependiente posee K categorías, las ecuaciones que se pueden generar serán siempre una categoría menor (k-1), dado que la categoría de mayor desempeño no se asocia a los odds, porque al ser una probabilidad acumulada esta última será igual a 1, dejando a las restantes como probabilidades menores a este.

Se puede afirmar que lo anterior obedece a un modelo conocido como ¿modelo logit acumulado? en donde se aprecia que es construido partiendo de la probabilidad acumulada de la variable respuesta dado un determinado valor(es) de las covariables.

También se puede denominar modelo de razón de probabilidad proporcional esto es porque los coeficiente no tienen relación alguna con las categorías de la variable respuesta, proporcionando las k-1 ecuaciones que se forman para las categorías, según Agresti, A (1990) lo anterior implica asumir que la relación entre las variables explicativas y la variable dependiente ordinal, es independiente de las categorías de esta última, y por tanto que los cambios en las variables explicativas provocan el mismo cambio en la razón de probabilidad acumulada de todas las categorías.

Existen representaciones graficas que permiten comprobar el supuesto citado, en nuestro caso podemos considerar el test de líneas paralelas. Solo se diferencian las k-1 ecuaciones por valor del intercepto, ya que los coeficientes de las ecuaciones son iguales para todas las variables explicativas es decir que las afectan de igual forma. Para estimar los coeficientes de la ecuación de regresión se utilizan diversos procedimientos, siendo la estimación de máxima verosimilitud el más empleado Agresti, A (1990).

Para la estimación de los coeficientes del modelo y de sus errores estándar se utiliza la estimación por máxima verosimilitud. Supongamos que disponemos de una muestra aleatoria de tamaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, X_2, ..., X_n$. Denotemos a cada combinación de valores de las variables explicativas por $x_q = (, x_q1, ., x_qn)^T$ con $x_{q0} = 1 \; \forall \; q = 1, 2, ..., Q$. En cada una de estas combinaciones se tiene una muestra aleatoria de d_q observaciones independientes de la variable de respuesta politomica Y, de entre las cuales denotamos por $Y_{j/q}$ al número de observaciones que caen en la categoría de respuesta $Y_j \; \forall \; j = 1, 2, ..., k$. Así que se verifica que, $\sum_{j=1}^{k} Y_{j/q} = d_q = N$. Los vectores $(y_(1/q), ..., y_(k/q))^T \; \forall \; q = 1, ..., Q$.Siguen una distribución de probabilidad multinomiales independientes $M(d_q; p_{1/q}, , p_{k/q})$ siendo $p_{j/q} = P[\frac{Y/Y_j}{X/x_q}]$ y verificado que $\sum_{q=1}^{k} y_{j/q} = 1$

Por tanto, la función de verosimilitud de los datos viene dada por $V = \prod_{q=1}^{Q} \left[\frac{d_q!}{\prod_{j=1}^{k}(y_{j/q})} \prod_{j=1}^{k} (P_{j/q}^{y_{j/q}}) \right]$. Para obtener los estimadores de máxima verosimilitud hay que resolver $k-1$ sistemas de p ecuaciones no lineales de la forma $\frac{\Delta K}{b_{sj}} = \sum_{q=1}^{Q} y_{jq} x_{qs} - \sum_{q=1}^{Q} n_q x_{qs} \frac{exp \sum_{s=0}^{n} b_{sj} x_{qs}}{\sum_{j=1}^{k} exp \sum_{s=0}^{n} b_{sj} x_{qs}}$. Así que para resolverlo utilizamos se utiliza el método iterativo de Newton-Raphson, con este método obtenemos el estimador de los parámetros b, que es una matriz de dimensión $(p)(k-1)$. la matriz de covarianzas de b, que es la inversa de la matriz de información de Fisher, dada por $cov(\hat{b_j}) = \left[-E \left(\frac{\Delta^2}{\Delta b_{rj} \Delta b_{sj}} \right) \right]^{-1} = \left[X' Diag[d_p p_{j/q}(1 - p_{j/k})] X \right]^{-1}$.

A continuación y después de realizar el proceso de máxima verosimilitud se debe comprobar el aporte o significancia estadística de los coeficientes de regresión del modelo, para conocer dicho aporte se puede recurrir al estadístico de WALD, el estadístico G de razón de verosimilitud y la prueba Score. No solo se debe comprobar el aporte estadístico de los coeficientes, también hay que comprobar la significancia global de la ecuación, lo que quiere decir que debemos evaluar si las covariables tienen un efecto real sobre la variable

respuesta, para ellos aplicamos el estadístico G, el cual se puede calcular al reemplazar en la formula.

$$G = -2\ln\left(\frac{\text{Verosimilitud de la muestra sin variables explicativas}}{\text{verosimilitud con la ecuación que incluye la variables}}\right) \tag{4}$$

Este estadístico sigue distribución chi- cuadrado con tantos grados de libertad como variables independientes existan (McCullagh,1980). Al realizar las predicciones e inferencias de la variable con el modelo que tiene todas las variables independientes estas superan las proyecciones que se realizan sin tenerlas en cuenta; el estadístico G ayuda a concluir que al menos una de las covariables presenta un aporte estadístico a la variables respuesta y por ello la probabilidad que permite calcular esta variable se ve contaminada con al mezclarse con los demás valores de la variables independientes. Releer esta párrafo las ideas están cortadas, revisar nuevamente. si hemos calculado el aporte del coeficiente, la significancia global del modelo, como piezas fundamentales en la generación del modelo óptimo, también debemos darle prioridad a la significancia individual de cada variable independiente. Normalmente se considera el test de Wald, en la cual se obtiene la significancia del coeficiente estimado para cada variable (Agresti. A ,1990). El estadístico utilizado es el siguiente:

$$Z_{Wald} = \left(\frac{b_j}{ES(b_j)}\right) \tag{5}$$

Aplicando el planteamiento de Heredia , Rodríguez y Vilalta (2009). en el cual bj es el coeficiente de regresión estimado para la variable independiente j. Bajo la hipótesis de que el coeficiente poblacional $j = 0$ para la variable j, la razón entre la estimación de este coeficiente (bj) y el error estándar de esta estimación [(ES bj)], sigue una distribución normal estándar. Mayores valores de este estadígrafo indican que el coeficiente j es distinto de cero, y por ende que la variable independiente tiene efecto sobre la probabilidad de ocurrencia de los valores de la variable dependiente. Podemos conocer si las covariables del estudio presentan un aporte sobre la probabilidad de ocurrencia en los valores que puede tomar la variable respuesta, y así determinar si el modelo tiene un buen ajuste, generalmente es necesario para evaluar tal ajuste la construcción de una tabla de contingencia, en donde se aprecien en las columnas las posibles combinaciones de valores de las covariables y en las filas lo valores que puede tomar la variable respuesta. Referencia. El empleo de una prueba de bondad de ajuste para comparar en cada celda de la tabla los valores observados y los valores predichos según el modelo es útil para valorar la calidad de ajuste. Si la frecuencia predicha para las combinaciones según el modelo, difiere significativamente de la frecuencia con la cual ocurren realmente los valores en estas combinaciones, existe evidencia de falta de ajuste.

$$X^2 = \sum_{l=1}^{k}\sum_{i=1}^{m}\frac{(y_{il} - m_i p_{il})^2}{m_i p_{il}(1 - p_{il})} \tag{6}$$

2.2. Tasa de clasificaciones correctas

Para cuantificar la bondad del ajuste global del modelo se dispone también de otra medida como es la tasa de clasificaciones correctas. Es decir, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados-predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción, del mismo modo que se hace en el análisis discriminante . Se define como la proporción de individuos clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta. $Y(Y_1, Y_2, ..., Y_k)$ coincide con su valor estimado por el modelo.

2.2. Calidad del ajuste

Para medir la calidad de ajuste los mas utilizados en regresión logística multinomial se utiliza Mc-Fadden se plantea que si tenemos $\Delta = -2ln(V)$, se identifica que Δ_0 el valor inicial de la función, es decir el mismo Δ bajo el modelo ajustado con todos los parámetros, obtendremos la siguiente expresión del pseudo-R^2 de Mc-Fadden dado por $R^2_{MF} = 1 - \frac{\Delta_f}{\Delta_0}$ conociendo que los valores deben estar comprendidos entre $0 \leq R^2_{MF} \leq 1$

es muy raro que el valor sea aproximado a 1. Es en buen ajuste cuando el valor esta comprendido entre $0.2 \leq R^2_{MF} \leq 0.4$ y excelente para valores superiores. De igual manera se utiliza el coeficiente de pseudo-R^2de Nagelkerke que esta dado por $R^2_N = \frac{R^2_{cs}}{1-V_0^{2/n}} = \frac{1-exp(\frac{\Delta_f - \Delta_0}{N})}{1-exp(\frac{\Delta_0}{N})}$.

El rango debe estar comprendido entre $0 \leq R^2_N \leq 1$, su interpretación es igual al coeficiente de determinación de la regresión lineal clásica, pero es mucho mas difícil que alcance valores muy cercanos a 1. Es decir que para compararlo con el modelo de regresión logistica politomica con diferentes números de variables predictoras suele introducirse coeficientes pseudo-R^2de Mc-Fadden, dado por $Adj = R^2_{MF} = 1 - \frac{0.5\Delta_f + n + 1}{0.5\Delta_f + n}$, siendo n el número de variables predictoras.

En la expresión se forman m combinaciones con los valores de las variables explicativas y se tiene en cuenta que la variable respuesta tiene k niveles o categorías de manera que: $y_i l$ es la frecuencia observada de la i-ésima categoría de la variable dependiente en la l-ésima combinación de valores de las variables explicativas. p_l es la probabilidad estimada con el modelo para la i-ésima categoría de la variable dependiente en la l-ésima combinación de valores de las variables independientes. m_l es la cantidad de elementos en la l-ésima combinación de valores de las variables explicativas. Mientras mayor es el valor del estadístico X^2 mayor sospecha de falta de ajuste. Si finalmente se concluye la existencia de relación entre las variables explicativas y la dependiente, y si la ecuación lograda presenta buen ajuste, entonces se pueden hacer otros análisis, por ejemplo, para obtener la razón de probabilidad acumulada de la categoría i de la variable dependiente para determinados valores de las independientes, se despeja esta razón de la función logarítmica de forma que:

$$\frac{P(\text{Valor sea } \leq \text{ categoria i / valores X})}{P(\text{Valor sea categoria i / valores X})} = \ell^{\alpha+\beta X} \tag{7}$$

Para continuar con el análisis recordemos que término valor en la expresión (7) hace referencia a cualquier valor que pueda tomar la variable dependiente y de la misma expresión podemos concluir que:

$$P(Valor \text{ sea } \leq \text{ categoria i / valores X}) = \frac{\ell^{\alpha+\beta X}}{1+\ell^{\alpha+\beta X}} \tag{8}$$

y de (8) se deduce que:

$$P(Valor \text{ sea} = \text{categiria i / valores x}) = (Valor \text{ sea} \leq \text{categoria i-P(valor sea categoria i-1)}) \tag{9}$$

Lo anterior es de vital pues ayuda para estimar haciendo uso de la ecuación obtenida y dado un conjunto de valores de las variables regresoras, la probabilidad que la dependiente tome cada uno de sus valores. Con ayuda del estadístico puede calcularse también el odd ratio (ratio de la razón de probabilidad), que genera una transformación en cada variable explicativa. El odds ratio de la variable independiente x evalúa la relación entre la razón de probabilidad asociada a la categoría i cuando $x = x2$, y la razón de probabilidad asociada a la categoría i cuando $x = x1$ (Heredia, Rodríguez y Vilalta, 2009).

$$\text{odds ratio} = \frac{p(Y \leq i \vdots X = X_2)/p(Y > i \vdots X = X_2)}{p(Y \leq i \vdots X = X_1)/p(Y > i \vdots X = X_2)} \tag{10}$$

Las categorías de la variable respuesta se ven afectadas de igual forma por una determinada variable explicativa, de esta manera el odds ratio es usado para identificar el efecto de las variables explicativas sobre la variable respuesta. Según los planteamientos de Agresti (1990). Si éste es igual a uno indica que la variable explicativa no tiene efecto. Si es menor que uno, lo cual sucede cuando el coeficiente de la variable regresora es negativo, indica que, si las otras variables explicativas permanecen constante, los cambios en la variable explicativa analizada incrementan la probabilidad de obtener categorías de mayor valor en la variable objeto de estudio (Hosmer y Lemeshow 2000). Valores de odds ratio mayores que uno muestran que las variaciones en la variable independiente disminuyen la probabilidad de obtener categorías de mayor valor de la dependiente.

Dado que la variable respuesta desempeño de los estudiantes en matemáticas (desempmat) presenta cuatro categorías , existen $k-1$ ecuaciones ya que a la categoría mayor no se asocia odds al ser la probabilidad acumulada hasta ésta igual a uno.

Asi mismo el modelo también puede denominarse según Hosmer y Lemeshow (2000), como modelo logit acumulado ya que es construido basándose en las probabilidades acumuladas de la variable respuesta dados los valores de las variables explicativas los coeficientes estimados son independientes de las categorías de la variable dependiente, siendo los mismos en las $k-1$ ecuaciones que se forman para las categorías.

3. Metodología

El tipo de investigación aplicada es de carácter descriptivo, así mismo el enfoque de ésta es cuantitativo. Se contó con el registro de 163 estudiantes de grado sexto del colegio Mariano Ospina en donde fue de interés la construcción del modelo de odds proporcionales [2]en donde Y corresponde al desempeño en matemáticas en el primer periodo del año 2012 y las variables explicativas corresponden al número de materias perdidas en el año 2011, el género, la calificación en español en el primer periodo del 2012 y el promedio total o general de su rendimiento académico en todas sus asignaturas.Las características estudiadas fueron:

Calificación promedio total: Media aritmética de las calificaciones del estudiantes en todas las asignaturas en grado sexto en el primer periodo.

Calificación final en español: La calificación escolar o nota escolar (?) es un método utilizado para evaluar y categorizar el rendimiento escolar de los alumnos. El sistema de calificación varía dependiendo de las necesidades y criterios de cada docente en una institución educativa, teniendo en cuenta diferentes componentes conceptuales, procedimentales y actitudinales propios de la evaluación integral del estudiante en dicha área.

Género del estudiante: Es el conjunto de características sociales, culturales, políticas, psicológicas, jurídicas, económicas asignadas a las personas en forma diferenciada de acuerdo al sexo;Registro que se tiene en la matrícula de cada estudiante de a cuerdo a la clasificación ¿masculino? o ¿femenino?..

Número de materias perdidas: Número de asignaturas en las cuales la calificación final del estudiante fue inferior a 3.0. Interesa conocer el nivel de pérdida del estudiante.

Calificación final en comportamiento: Valoración que se da al estudiante de acuerdo a su convivencia dentro de la institución. Se revisan diferentes aspectos disciplinarios y/o académicos (faltas leves, graves y especialmente graves).

Los desempéños en matematicas son: El sistema de calificación varía dependiendo de las necesidades y criterios de cada docente en una institución educativa, teniendo en cuenta diferentes componentes conceptuales, procedimentales y actitudinales propios de la evaluación integral del estudiante en dicha área.

Desempeño bajo: calificacion obtenida por el estudiante entre 0.0 a 2.9 *Desempeño basico*: calificacion obtenida por el estudiante entre 3.0 a 3.9 *Desempeño alto*: calificacion obtenida por el estudiante entre 4.0 a 4.5 *Desempeño superior*: calificacion obtenida por el estudiante entre 4.6 a 5.0

La información registrada se analizó por medio del programa estadístico R Core Team (2016) [3] que permitió determinar el modelo ajustado que explica la estimación de los parametros mediante el metodo de maxima verosimilitud y se establece el test estadístico adecuado para el modelo. También se calculó la calidad del ajuste mediante los coeficientes de determinación "de pseudo-R^2de Mc-Fadden, pseudo-R^2 de Nagelkerke". los intervalos de confianza de los parámetros además se valida el modelo.

4. Resultados

Con la intención de tener un panorama más claro de las variables de estudio, a continuación en la tabla 1 se describe la distribución univariada de las variables consideradas en el estudio, las unidades en las que se miden o los valores codificados que toman, a fin de tener en cuenta cada una de las ccaracteristicas de la variables , asi como su naturaleza. las variables independientes como la dependiente del estudio son valoradas a continuación.

[2]Este proyecto se ha desarrollado con base en los registros institucionales del colegio Mariano Ospina de Bogotá del año 2012

[3]Equipo Central R (2016). R : Un lenguaje y entorno de estadística informática. R Fundación para la Computación de Estadística, Viena, Austria. URL https:/www.R-project.org/.

Variable	**Unidades / Valores que toma/Codificación**	**Descriptivo**
Desempeño en matemáticas (N=163).	De 0.0 a 2.9 Desempeño Bajo. De 3.0 a 3.9 Desempeño Básico. De 4.0 a 4.5 Desempeño Alto. De 4.6 a 5.0 Desempeño Superior.	ALTO : 10 estudiantes (6 %) BAJO : 75 estudiantes (46 %) BÁSICO: 67 estudiantes (41 %) SUPERIOR: 11 estudiantes (7 %)
CALIFICACIÓN COMPORTAMIENTO en el primer periodo (N=163).	Escala de 0 a 5, donde 0 indica un comportamiento malo y 5 un comportamiento excelente durante el primer periodo del año.	Min. :2.5 Median :4.5 Mean :4.2 Max. :5.0 C.V.: 0.1669043 Apunt: -0.808658129 Curtosis: -0.21273224
GÉNERO del estudiante (N=163).	¿masculino?: M ¿femenino?: F	F= Femenino = 64 Est. (39 %) M=Masculino= 99 Est. (61 %)
PROMEDIO TOTAL DE MATERIAS en el primer periodo (N=163).	Escala 0 a 5, indica nota promedio de las materias en su primer periodo.	Min. :2.3 Median :3.4 Mean :3.4 Max. :4.8 C.V.: 0.1451445 Apunt: 0.308718457 Curtosis: -0.04458927
NÚMERO DE MATERIAS PERDIDAS en el año anterio (N=163).	Número de asignaturas en las cuales el estudiante tuvo un desempeño bajo en el año anterior. DESEMPEÑO <3.0	Min. :0.0 Median :2.0 Mean :2.4 Max. :8.0 C.V.: 0.8728045 Apunt: 0.468983820 Curtosis: - 0.85440237
CALIFICACIÓN EN ESPAÑOL primer periodo (N=163).	Escala 0 a 5, indica la calificación en español en su primer periodo.	Min. :1.1 Median :3.1 Mean :3.2 Max. :4.9 C.V.: 0.2622803 Apunt:0.007922252 Curtosis -0.54703710

Tabla 27: Descripción de las variables estudiadas

Ccon base en el coeficiente de variacion de totas las variables cuantitativas se pudo concluir que los datos son homogéneos ya que los resultados obtenidos fueron cercanos a 1, de igual manera, al revisar la curtosis de las variables cuantitativas se infiere que no se presentan datos atípicos en las variables.

A partir de la tabla se pueden realizar conclusiones de la probacion, como el hecho de que el (87 %) de los estudiantes tiene un desempeño entre básico y alto, no obstante la población muestrada no está compuesta en una misma proporción por estudiantes de mismo género.

Se estableció que ninguna de las variables cumple el test de shapiro wilk, excepto promedio total del primer periodo académico, por lo que utilizaremos un análisis no paramétrico para comprobar el supuesto de que las variables independientes procedan de la misma población a partir del análisis de sus medias.

Con ayuda del paquete R se aplicó el test a las variables cualitativas de la base de datos versus el desempeño en matemáticas. Los resultados se presentan en la Figura 1.

	GENERO			CHI CUADRADO	P-VALUE
DESEMP. MATEM.		F	M	1,2026	0.7524
	BAJO	28	47		
	BASICO	26	41		
	ALTO	4	6		
	SUPERIOR	6	5		

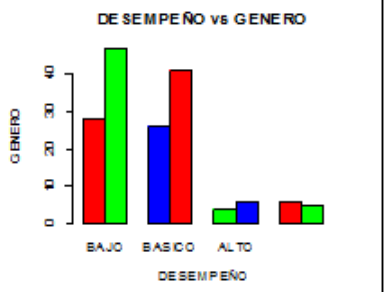


Figura 4: Desempeño Vs. género

Usamos la tabla de doble entrada, que se denomina tabla de contingencia, la cual presenta en cada casilla las frecuencias absolutas o porcentajes de una de las categorías de una variable con una categoría de la otra variable.

Para evaluar el grado de relación y el nivel de significación estadística entre dos variables categóricas se utiliza el test de ji-Cuadrado evidenciando una asociación directa entre el desempeño del estudiante y su género, es notable el comportamiento del desempeño en los estudantes de género femenino y masculino a medida que progresa la variable respuesta.

De la misma forma se realizo un analisis no parametrico a a las variables cuantitativas con relacion al desempeño del estudiante a fin de estudiar el comportaminto de sus medias, estos resultados se presentan en la siguiente tabla.

	Desempeño					
	bajo	Basico	Alto	Superior	chi-squared	p-value
Calific. Comporta.	4.0	4.50	4.95	4.50	18.69	0.0003168
Calific. Español	3.00	3.40	4.15	4.50	47724	2,44E-007
Número. perdidas	4	1	0	0	87689	<2.2e-16
Promedio total	3120	3530	4155	4350	74.08	5,704E-013

Tabla 28: Kruskall Wallis

Podemos concluir que se presentan diferencias de medias estadisticamente significativas dado que $p < 0,05$, respecto a la variable desempeño. Es decir, la calificación en el comportamiento, la calificación en español, el número de materias pérdidas y el promedio total es distinto entre los que obtienen desempeño bajo, básico, alto y superior.

FACTORES ASOCIADOS AL DESEMPEÑO EN MATEMÁTICAS

La selección del modelo óptimo con el método hacia adelante y a través del Criterio de información de Akaike (AIC), indica que las variables que explican el desempeño en matemáticas de los estudiantes son el número de materias perdidas, el promedio de las asignaturas en el año anterior, el género y la calificación en el comportamiento. A continuación se presenta el modelo estimado:

Coefficients:	**Value**	Std, Error	t value
nuperdidas	-0,7663	0,1896	-4,042
promediototal	2,3256	0,7178	3,24
género[T,M]	0,7846	0,3883	2,021
calificcomport	0,4973	0,33	1,507
Intercepts:	Value	Std, Error	t value
BAJO|BASICO	8,3852	2,7956	2,9995
BASICO|ALTO	12,471	2,9577	4,2165
ALTO|SUPERIOR	13,5708	3,0339	4,473
Resid. Deviance	217,2145		
AIC:	231,2145		

Tabla 29: Resumen modelo óptimo

summary(fitted(OrdRegModel.2))							
BAJO		BASICO		ALTO		SUPERIOR	
Min.	0.002098	Min.	0.002442	Min.	2,79E-002	Min.	0.00001
Q1	0.083276	Q1	0.134001	Q1	1,90E+000	Q1	0.00090
Median	0.379355	Median	0.461594	Median	1,77E+001	Median	0.00907
Mean	0.456674	Mean	0.418132	Mean	6,26E+001	Mean	0.06257
Q3	0.855056	Q3	0.678614	Q3	9,81E+001	Q3	0.05800
Max.	0.997516	Max.	0.770268	Max.	2,68E+002	Max.	0.72694

Tabla 30: Resumen modelo óptimo

El ajuste global del modelo a través de razón del test de razón de verosimilitud arrojó una deviance de 216.6081 con 10 parámetros, mientras que la deviance del modelo optimo es 217.2145 con 7 parámetros, podemos calcular los grados de libertad a partir de la diferencia entre los parámetros del ajuste global con el modelo óptimo es decir, 12 grados de libertad, la diferencia de sus deviance corresponde a 0.6064295 con p-valor de 0.8949589.

Al calcular la tasa los valores observados y predichos por el modelo se obtuvo que en el (68 %) de los casos se consiga una predicción correcta lo que nos indica que el modelo es bueno. Lo anterior puede indicar que las calificaciones dependen del número de materias perdidas en el año anterior, el género, calificación comportamiento y del promedio total de las materias en el primer periodo. Para medir la calidad del ajuste del modelo se utilizó los coeficientes de Mc fadden y Nagelkerke.

Realizando los cálculos correspondientes se obtuvo en el coeficiente de mc fadden 0.3806349, como este valor es superior a que 0.2 se puede decir que el modelo presenta un buen ajuste al igual podemos recurrir otras pruebas. Para darle mayor certeza a la anterior inferencia, aplicamos la prueba r2 de Nagelkerke, el cual arrojo un valor de 0.632694 que afianza la idea de que el modelo hallado presenta un buen ajuste ya que su estadístico es cercano a 1. Lo indicado en amarillo ya lo habíamos corregido. En cada una de las graficas se presenta la aplicación del test, especificando que existen k-1 ecuaciones con los mismos coeficientes acompañando a las variables explicativas y que sólo se diferencian en el valor del intercepto.

TEST DE LÍNEAS PARALELAS

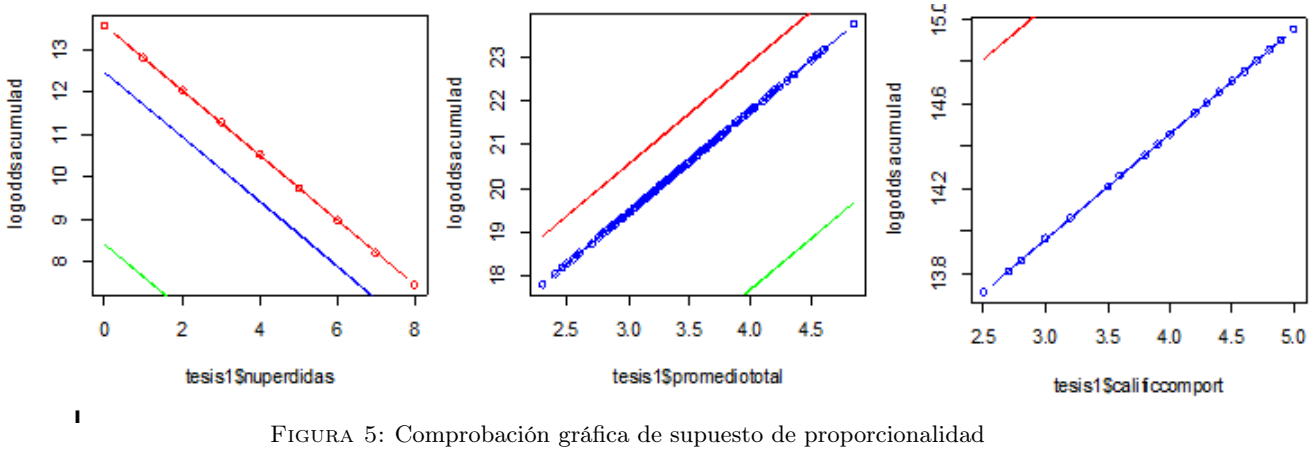


FIGURA 5: Comprobación gráfica de supuesto de proporcionalidad

Podemos observar que se cumple el supuesto de proporcionalidad, lo cual significa en este caso que las variaciones en el número de materias perdidas por los estudiantes de grado sexto en el año 2011, junto al promedio total del estudiante y calificación del comportamiento en el primer periodo académico del año 2012, producen el mismo cambio en la razón de probabilidad acumulada de todas las categorías de la variable respuesta.

INTERPRETACIÓN DEL MODELO

	COEF,	**error**	WALD	P VALOR	OR	INTERVALO DE CONF	
	ESTIMADO					95,00 %	
calificomport	0,4972	0,33	1,506714	0,065941	1,64E+000	0,869766	3,1948132
género[T,M]	0,7845	0,3882	2,020755	0,021652	2,19E+000	1,04E+000	4,774956
numperdidas	-0,7663	0,1895	-4,04208	0,999973	4,65E-001	0,3143925	0,6631879
promediototal	2,3255	0,7178	3,239803	0,000598	1,02E+000	2,5972081	43,903133
BAJO|BASICO	8,3852	2,7956	2,999428	0,001352	4,38E+003		
BASICO|ALTO	12,471	2,9577	4,216452	1,24E-005	2,61E+005		
ALTO|SUPERIOR	13,5708	3,0339	4,473054	3,86E-006	782931,1		

TABLA 31: Resumen modelo óptimo

En la variable numero de materias perdidas se observa que su razon de probabilidad y su intervalo de confanza es menor que 1, podemos concluir entonces que a medida que se aumenta en una materia el numero de materias perdidas del año anterior, se produce una disminucion de la razon de probabilidad acumulada de todos los valores que puede tomar el desempeño.

A medida que aumenta una decima el promedio total del estudiante, se produce un aumento en la razon de probabilidad acumulada de los valores del desempeñno. Es decir es mas probable que ocurran valores mayores de la variable respuesta.

Por otra el intervalo de confanza de la variable género [T,M], sugiere que la probabilidad acumulada aumenta en todos los valores que puede tomar el desempeño si el estudiante es de género masculino respecto a que sea de género femenino, es decir que es más probable que un estudiante de género masculino obtenga valores altos en el desempeño que uno de género femenino.

5. Conclusiones

En este trabajo está presentado un modelo de regresión logística con respuesta politómica ordinal usando un metodo estadísticos para su estimación, la bondad de ajuste, calidad del ajuste, la validación y la selección del modelo más ajustado para que nos de a conocer la asociacion entre el desempeño de los estudiantes y su comportamiento, género, numero de materias perdidas el año anterior, calificación en español y promedio académico total en el primer periodo del año 2012.

Los resultados muestran que en el modelo óptimo es de (68 %) por lo cual es un modelo bueno, es decir se deben tener en cuenta otras variables explicativas. Lo anterior puede indicar que el desemepeño en matematicas no solo depende de las variables estudiadas.

Referencias Bibliográficas

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

Díaz Monroy, L. G. y Morales Rivera, M. A. (2012), *Análisis estadístico de datos categóricos*, Universidad Nacional de Colombia.

Hosmer, D. y Lemeshow, S. (2000), *Applied Logistic Regression*, John Wiley Sons, New York.

Mccullagh, P. (1980), *Sobre la agrupación de niveles del factor explicativo en el modelo logit binario*, Revista Colombiana de Estadística, 32, 157-187.

Ponsot, E., S. S. y Goitía, A. (1980), *Regression models for ordinal data. Journal of the Royal Statistical Society, 42, 109-142.*

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, Vienna, Austria. *https://www.R-project.org/

r comander programa estadistico, Disponible en http://www.duitama.gov.co/dependencias.html (2016).

METODOLOGÍA PARA EVALUAR LA CALIDAD DE LA INFORMACIÓN DEL COMPONENTE DE INSUMOS DEL SIPSA

Especialización en Estadística

Emilcen Rojas Pinzón[1,a], Reinaldo Alarcón Guarín[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Se propone una metodología que permite medir la calidad de la información recolectada en el componente de insumos y factores del Sistema de Información de precios del Sector agropecuario- SIPSA del DANE , utilizando para ello el criterio de calidad de exactitud del precio recolectado versus el precio supervisado, a través de un muestreo bietápico, aplicando un instrumento de verificación desarrollado para tal fin, con el objetivo de construir un indicador de calidad que nos arroja la proporción de precios correctos, versus el número de artículos verificados. La aplicación de esta metodología permite conocer el porcentaje de calidad mes a mes, e identificar falencias y fortalezas en el proceso. Se demuestra cómo su aplicación práctica en la operación estadística para la cuidad de Tunja permitió estimar una proporción de precios correctos de 0,92 y obtener el indicador de calidad, que para el mes de abril cuando se realizó la prueba piloto arrojo un resultado de 92 %.

***Palabras clave*:** calidad, muestreo, información recolectada, proporción, indicador.

Abstract

This work proposes a methodology that permits to measure the quality of the information collected in the component of inputs and factors for SIPSA- Agriculture Area Prices Information System- from DANE Statistics National Department of Colombia. Using for this one the quality approach of the accuracy of the collected prices versus the supervised prices by means of a two stage sample applying a verification instrument developed for this purpose; with the aim to make a quality indicator that shows the correct prices compared to the verified articles.The application of this methodology allows to know the quality percentage month after month, and identify weaknesses and strengths in the process. The pilot test made in Tunja city allowed to estimate a correct prices proportion of 92 % in April, 2016.

***Key words*:** Statistical operation, quality, sampling, data collected, proportion.

1. Introducción

El propósito de este artículo es dar a conocer el proceso que se realizó para el diseño de la metodología que permite estimar la calidad de la información recolectada en el componente de Insumos y factores del SIPSA -DANE, la cual surgió a raíz de que se han evidenciado errores en la información recolectada en los componentes del Sistema. Teniendo en cuenta que el SIPSA es una operación estadística nueva para la entidad, ya

[a]Estudiante de Especialización en Estadística. E-mail: emilcen.rojas@uptc.edu.co
[b]Profesor asistente. E-mail: reinaldo.alarcon@uptc.edu.co

que comenzó a ser operada por el DANE a partir del año 2012 (DANE 2015*a*), hasta el momento no cuenta con una herramienta que permita medir y evaluar eficazmente la calidad de la información recolectada en los diferentes componentes.

El Sistema de Información de precios y Abastecimiento del Sector Agropecuario- SIPSA es el encargado de informar los precios mayoristas de los productos agroalimentarios que se comercializan en el país, así como los precios minoristas de insumos y factores asociados a la producción agrícola y el nivel de abastecimiento de alimentos en las principales ciudades del país.El componente Insumos y factores asociados a la producción agropecuaria genera estadísticas sobre los precios promedio minoristas de los diferentes insumos y factores asociados a la producción agropecuaria; información que es útil en la identificación de los factores y hechos que afectan los precios en la cadena de comercialización y en la toma de decisiones de productores, agroindustriales, comercializadores y determinadores de políticas (DANE 2015*c*).

En cuanto a los antecedentes de la evaluación de calidad en este tipo de operaciones estadísticas, en la revisión de literatura que se realizo acerca del tema no se encontró mucho contenido al respecto, ya que la mayoría de estudios que se conocen sobre evaluación de calidad han sido en el sector de la industria y los servicios, especialmente en educación y salud. En lo concerniente a este tipo de operación estadística, a nivel internacional el caso más similar es el sistema de Precios Coyunturales de Productos Agrícolas del Instituto Nacional Estadístico Español-INE, sin embargo en los textos encontrados mencionan que no existe documentación disponible sobre la evaluación de la calidad de los datos de esta estadística (INE sin año).

A nivel nacional se encuentra la operación estadística del Índice de Precios al Consumidor -IPC, ejecutada por el DANE, en la cual el control de calidad se realiza a través de las diferentes etapas por los que pasa la información (recolector-supervisor-analista- coordinador de campo), se realiza un proceso de seguimiento a los recolectores, por medio de acompañamientos, revisitas y re-entrevistas por parte del supervisor, y coordinador de campo, y se lleva control en los diferentes formatos donde se registran el tipo y la cantidad de errores encontrados con el fin de evaluar la calidad de la información y el desempeño de los encuestadores. Sin embargo, dado que el SIPSA tiene algunas diferencias metodológicas y operativas, el inconveniente radica en que no se pueden aplicar estas mismas herramientas y procesos a esta operación estadística, por lo cual es necesario desarrollar e implementar, herramientas y procesos de control para generar indicadores de calidad específicos.

El presente documento recopila y expone la información conceptual y los criterios necesarios para implementar una metodología que permita evaluar la calidad de la información recolectada en el componente de Insumos. En primer lugar se define la calidad estadística y los criterios que se deben tener en cuenta en la evaluación de calidad de una operación estadística, en segundo lugar se da a conocer el uso de indicadores para medir la calidad de los procesos, y finalmente se realiza una introducción al muestreo bietápico, la cual fue la técnica seleccionada para esta metodología.

2. Referente Conceptual

2.1. Marco conceptual de la calidad de una operación estadística

Dentro del marco del Sistema Estadístico Nacional (SEN), la calidad estadística se entiende como un conjunto de propiedades que deben tener el proceso y el producto estadístico para satisfacer las necesidades de información de los usuarios. Asimismo, en este marco se entiende que la calidad de la información estadística tiene un carácter multidimensional por lo que debe considerarse a partir de la interrelación de atributos como: la pertinencia y relevancia, la continuidad, la exactitud, la oportunidad y puntualidad, la accesibilidad, la interpretabilidad, la coherencia, la comparabilidad y transparencia (DANE 2015*b*).

Precisión o exactitud: grado con que los datos estiman o describen correctamente las cantidades o características que deben medir. La precisión se refiere a la proximidad entre los valores estimados y los valores verdaderos (desconocidos). La precisión tiene muchos atributos, y en la práctica no existe una única medida

agregada o general de la misma (DANE 2015*b*). Por necesidad, estos atributos generalmente se miden o describen en términos del error, o la importancia potencial de error, que se introduce a través de las fuentes individuales.

Indicadores de calidad: Se define un indicador como una expresión cuantitativa observable, que permite describir características, comportamientos o fenómenos de la realidad a través del establecimiento de una relación entre dos o más variables, por tanto un indicador es un cociente que permite analizar rendimientos, también se puede decir que es una expresión del desempeño, que al ser comparada con un nivel de referencia, podrá señalar una desviación. Por lo cual los indicadores son considerados como instrumentos de medición, basados en hechos y datos, que permiten evaluar la calidad de los procesos, productos y servicios para asegurar la satisfacción de los clientes (EUROPARC 2002), miden el nivel de cumplimiento de las especificaciones establecidas para una determinada actividad o proceso empresarial, además permiten tomar medidas preventivas y/o correctivas para asegurar la mejora en el tiempo (DANE 2009).

Componentes de un indicador de calidad (EUROPARC 2002):

Indicador: lo que se quiere medir,

Unidades de medida: Ratios: Monitor/alumno, Tiempo, Porcentaje,

Valores de referencia: nivel mínimo y máximo admisible,

Fuente de los datos: de donde se extraerán los datos (encuestas, informes, etc), responsable de la toma de datos,

Periodicidad: Diario, semanal, mensual, etc.

Metodología: hace referencia al conjunto de procedimientos racionales utilizados para alcanzar el objetivo o los objetivos que rige una investigación científica, una exposición doctrinal o tareas que requieran habilidades, conocimientos o cuidados específicos. Con frecuencia puede definirse la metodología como el estudio o elección de un método pertinente o adecuadamente aplicable a determinado objeto.

Existe un área de control y mejoramiento de la calidad que es el muestro de aceptación, el cual guarda una estrecha relación con la inspección y prueba del producto o servicio, cuyos orígenes se remonta mucho antes que se desarrollara una metodología estadística para el mejoramiento de la calidad. Las operaciones no manufactureras carecen de un sistema de medición natural que permita definir fácilmente la calidad, por lo cual el primer paso a segur es establecer un sistema de medición cuantitativo y objetivo que permita medir la calidad (Montgomery 1991).

2.2. Conceptos de Muestreo y selección de la técnica de muestreo

Con el fin de precisar los métodos utilizados en esta propuesta, es necesario precisar algunos conceptos acerca del muestreo y de las técnicas de muestro aleatorio simple por etapas.

Población objeto de estudio: Todo conjunto de elementos, definido por una o más características, de los que gozan todos los elementos que lo componen, es la totalidad del universo que interesa considerar y que es necesario que esté bien definido para que se sepa en todo momento qué elementos lo componen. Muestra: Subconjunto de la población o colectivo que se investiga, debe ser representativa del conjunto de la población.

Estimador: Se denomina estimador, a cualquier función de n variables, donde después de sustituir en ella los valores muéstrales, el resultado obtenido puede servir como sustituto del valor del parámetro poblacional (León y de Rojas Gómez 2010).

2.2.1. Muestreo Aleatorio Simple (MAS)

Según Cochran (1976), Botero (2001) Gutiérrez (2009) el Muestreo Aleatorio Simple es el procedimiento mediante el cual se eligen n elementos de una población de tamaño N, haciendo la selección con o sin reposición. Se presenta como el prototipo de muestreo por su sencillez y la facilidad para calcular los errores de muestreo. El MAS tiene una propiedad que lo caracteriza, la cual es, que todas las muestras posibles de tamaño n, de una población de tamaño N, tienen la misma probabilidad de ser seleccionadas.

Estimación de proporciones: En el caso que la variable de interés solo pueda tomar dos valores (0 y 1), entonces su media no es más que la proporción de individuos con $y = 1$. Es decir la proporción con la característica de interés (P), que representa la proporción de elementos en la población que poseen el atributo considerado (Galmés y Galmés 1997). Los estimadores de las proporciones poblacionales se estiman a través de la expresión:

$$P = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{A}{N} \tag{1}$$

Donde N_i es el número de elementos en la i-ésima unidad ($i = 1, 2, ...n$) y y_i el número de elementos en la unidad i que poseen la característica de interés (Zapata-Ossa, Cubides-Munévar, López, Pinzón-Gómez, Filigrana-Villegas y Cassiani-Miranda 2011).

Como en la práctica, no se conoce el valor de la proporción poblacional (P), pues el muestreo tiene como propósito estimar dicho valor. Sin embargo, pueden obtenerse estimaciones preliminares de P mediante un estudio anterior de similares características o una muestra piloto cuyo tamaño suele ser entre 30 y 100 elementos.(Hidalgo y Ramírez 2009). Esto se recomienda siempre, porque no se tiene información sobre la variabilidad de las variables que se van a estudiar. La prueba piloto también sirve para diseñar la estrategia del trabajo de campo y para revisar el cuestionario, entre otras tareas(Ojeda, Camacho, Victoria y Landa 2011). Cuando se estima una proporción, la varianza del estimador de la proporción es (Morillas 2007)

$$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}\hat{q}}{(n-1)} \frac{N-n}{N} \tag{2}$$

y el error máximo vendría dado, de acuerdo con lo anterior, por:

$$\varepsilon = |p - \hat{p}| = Z_{1-\alpha/2} \sqrt{\frac{pq}{n}} \tag{3}$$

Por tanto el tamaño de la muestra se obtendrá despejando n de la expresión (3) para poblaciones infinitas

$$n_0 = \frac{P\,(1-P)\,Z^2_{1-\alpha/2}}{\varepsilon^2} \tag{4}$$

y ajustada para poblaciones finitas

$$n_1 = \frac{n_0}{1 + \frac{n_0}{N}} \tag{5}$$

Intervalos de confianza para la proporción: se construirán los correspondientes intervalos de confianza para la proporción, que nos dan una idea de la horquilla en que se mueve el verdadero valor del parámetro(Morillas 2007)

$$\widehat{p} - Z_{1-\alpha/2}\sigma_{\widehat{p}} \leqslant p \leqslant \widehat{p} + Z_{1-\alpha/2}\sigma_{\widehat{p}} \tag{6}$$

2.2.2. Muestreo bietápico (MAS-MAS)

Cochran (1976) define el muestreo bietápico de la siguiente manera:

> Supongamos que cada unidad de la población se puede dividir en cierto número de unidades más pequeñas, o subunidades, si las subunidades contenidas en una unidad seleccionada dan resultados semejantes, no parece económico medirlas todas, una práctica acostumbrada consiste en seleccionar y medir una muestra de subunidades de alguna unidad elegida, lo que es llamado submuestreo, dado que la unidad no se mide completamente, sino que a su vez es objeto de un muestreo. Mahalanobis, le dio el nombre de muestreo en dos etapas, porque la muestra se obtiene en dos pasos

Es una generalización del muestreo por conglomerados en el que se intenta reducir el coste al mínimo. En cada etapa es necesario usar un diseño simple, aunque se pueden realizar diferentes combinaciones. Se comienza seleccionando unas unidades llamadas unidades primarias de muestreo (UPM) que están compuestas, a su vez, por grupos de unidades de menor tamaño. A continuación se extrae una muestra de estas unidades de cada una de las UPM, las que se denominan unidades secundarias de muestreo (USM) o subunidades que son los elementos de la población que van a ser observados (unidades últimas) (Gutiérrez 2009).

Suponiendo que la población está dividida en N_i unidades primarias de muestreo, de las cuales se selecciona una muestra s_i de nI cada unidad $N_i = M$. El sub-muestreo es tal que se selecciona una muestra de exactamente $n_i = m$ u nidades secundarias de muestreo. Por tanto el Tamaño poblacional y muestral está dado por: $N = N_I M$ y $n = n_I m$

Para encontrar los valores óptimos de n_I y m que serán utilizados en la primera y segunda etapa de muestreo de tal forma que dada una función de costo se minimice la varianza del estimador. Por tanto se obtiene: $C = c_1 n_I + c_2 n_I m$

Donde c_1 es el costo del levantamiento del marco de muestreo en cada unidad primaria seleccionada en la muestra s_I y c_2 es el costo de recolectar la información de la característica de interés para los elementos o unidades secundarias de sub-muestreo (Cochran 1976).

Entonces los valores óptimos de n_i y m:

$$n_I = \frac{C}{c_1 + c_2 m} \tag{7}$$

$$m = M\overline{S}^2_{yUi}\sqrt{\frac{c_1/c_2}{S^2_{tyUI} - M\overline{S}^2_{yUi}}} \tag{8}$$

Si la variabilidad de la característica de interés dentro de las unidades primarias es grande, entonces m será grande. Se debe resaltar que los resultados son válidos si la función de costo es correcta (Gutiérrez 2009). Se deben considerar dos fuentes de variabilidad: Entre unidades primarías y entre unidades secundarías, por lo tanto la varianza del estimador del total poblacional se define como:

$$\text{VAR}\left[\hat{Y}_{(2)}\right] = \frac{N\left(N-n\right)}{n} S^2_{UPM} + \frac{N}{n}\sum_{i=1}^{N} M_i\left(M_i - m_i\right)\frac{S_i^2}{m_i} \tag{9}$$

Donde $S^2_{UPM} = \sum_{j=1}^{N} \frac{\left(Y_i - \overline{Y}\right)}{N-1}^2$ es la varianza entre las UPM, y, $S_i^2 = \sum_{j=1}^{M_i} \frac{\left(y_i - \overline{y}\right)}{M_i - 1}^2$ la varianza de las unidades secundarias en la i-ésima unidad primaria (Gutiérrez, 2005).

Cuando las unidades son de tamaño variable existen varias reglas para determinar las fracciones de muestreo y submuestreo y se dispone de varios métodos de estimación. Las ventajas de los diferentes métodos dependen de la naturaleza de la población, de los costos de campo y de los datos adicionales de que disponemos. Uno de los más utilizados y que provee los mejores resultados es el método de selección de unidades con probabilidad proporcional al tamaño, propuesta por Hansen y Hurwitz (1943), la cual provee una media de la muestra que no tiene sesgo y no está sujeta a la inflación de la varianza (Cochran 1976). Por tanto las fórmulas del muestreo para tamaños diferentes se calculan de la siguiente forma: (Särndal y otros autores)

$$f_j = \frac{S_j^2}{G} f_1 = \frac{G}{V_0 + N_I S_I^2}\left[N_I + \frac{1}{G^{1/2}}\sum_{i=1}^{N=1} N_i S_i\right](10) G = S_I^2 - \frac{1}{N_I}\sum_{i=1}^{N} N_i S_i^2 \tag{10}$$

Donde G: es la diferencia de la variabilidad entre unidades primarias y el promedio de la variabilidad dentro de las unidades primarias, es decir las unidades secundarias (artículos) y $V_o = 0{,}0004$

3. Metodología

Para la consecución de la metodología propuesta se siguieron las siguientes etapas, retomando parte de lo propuesto por Piña y de Rojas, 2010 quienes plantean que se debe realizar un plan de muestreo el cual es un procedimiento que abarca varias etapas, las cuales se describen a continuación:

Primera etapa: Definición de los objetivos y criterios de evaluación: Para la realización del diseño metodológico para evaluar la calidad de la información recolectada en el componente de Insumos y factores del SIPSA se definió como criterio de calidad a evaluar la exactitud (precisión) de los precios recolectados, verificándolos y contrastándolos con los precio que suministra la fuente en el proceso de verificación, y de acuerdo al número de aciertos hallar el indicador de calidad. Por tanto el parámetro que se estimó fue la proporción de precios correctos (aciertos) en una muestra aleatoria seleccionada.

Segunda etapa: Determinación de la población objeto de estudio y elaboración del marco muestral: Para este caso la población objeto de estudio fue la totalidad de las fuentes (almacenes) y artículos que se tienen en la cobertura geográfica para el municipio de Tunja. El marco muestral se elaboró a partir del reporte de Excel que genera el aplicativo electrónico de análisis del Sistema dispuesto para ello, el cual nos arroja el listado mensual de fuentes de insumos agrícolas y pecuarios, con su respectiva lista de artículos que se recolectaron en cada fuente. En resumen el marco muestral está compuesto como se observa en la siguiente tabla de la Figura 1.

NOMBRE DE LA FUENTE (UPM)	N_i	**N° DE ARTICULOS RECOLECTADOS (USM)**	M_i
AGRICOLA MERCOSUR	N_1	92	M_1
AMAGRO	N_2	53	M_2
AGROPROTECCION	N_3	81	M_3
CASA DEL AGRO	N_4	53	M_4
INAGROB	N_5	80	M_5
SUPERAGRO	N_6	88	M_6
EL CEBU	N_7	33	M_7
GANAPROTECCION	N_8	61	M_8
SAN JORGE	N_9	74	M_9
TOTAL: 9		**615**	

Figura 1: Resumen población objeto de estudio y marco muestral de Fuentes y Artículos

Tercera etapa: Diseño del formato de verificación a fuentes para el componente de insumos y factores. Para el diseño del instrumento de verificación se tomó como base la planilla de recolección de campo de insumos utilizada antes del año 2012, (cuando se capturaba los datos en medio físico), adicionándole las casillas respectivas: precio supervisado y resultado. Este formato permite visualizar los datos de la fuente que se está verificando, las especificaciones del artículo que se está supervisando, el precio recolectado, el precio supervisado y el resultado de la re-entrevista, donde se registra si el articulo presento o no error, coincide o no. Además de la fecha cuando se aplicó la reentrevista, los datos de quien suministra la información y el responsable (como se observa en el Formato de Verificación- SIPSA, Figura 1).

Cuarta etapa: Implementacion del instrumento de verificación (muestra piloto): El diseño de muestreo utilizado fue un Muestreo Aleatorio Simple en dos etapas. La selección aleatoria sistemática de las "nünidades se realizó a partir del marco muestral de las "N"Fuentes (UPM) y la selección de las "m"subunidades se

realizó a partir del marco muestral de las "M.ªrtículos (UPM).

De acuerdo a lo planteado por PIÑA y de ROJAS, 2010, quienes "sugieren sacar una muestra preliminar con objeto de estimar la varianza de la población y calcular la dimensión de la muestra", se procedió a realizar una prueba piloto, la cual consistió en seleccionar aleatoriamente cuatro de las nueve fuentes y se procedió a realizar un submuestreo intencional del 20 % de artículos por fuente, teniendo en cuenta que dichas fuentes no tienen el mismo número de artículos (tamaño variable), esto con el fin de hallar la proporción estimada de calidad.

El instrumento de verificación se aplicó en el mes de abril mediante re-entrevistas telefónicas a los almacenes donde se indago uno a uno el precio de los artículos que fueron seleccionados registrando en el formato si el precio que informa nuevamente la persona coincide o no con el precio recolectado por el encuestador.

DANE
Para tomar decisiones

FORMATO DE VERIFICACION A FUENTES - SIPSA

VERSIÓN: 01
ELABORÓ: ANALISTA SIPSA
FECHA: 12-04-16

TERRITORIAL O SUBSEDE : TUNJA(BOYACA)

Nombre de la fuente: **GANAPROTECCION** Telefono: 7469435

Direccion: CRA 24-85-15 Cargo de quien responde: MIGUEL CARO (ADMINISTRADOR)

Día: __16_____ Mes: ____04___ Año: ___2016____

ARTCULO	CASA COMERCIAL	REG ICA	PRESENTACION	PRECIO RECOLECTADO	PRECIO SUPERVISADO	RESULTADO
Calfon - Energy	BAYER S.A.	10642	FRASCO|CC|350| 0|0	26400	26.400	0
Calfosgan	NOVARTIS DE COLOMBIA S.A.	2323	BOLSA|CC|250| 0|0	0	ND	0
Calfosvit Se	COMPAÑIA CALIFORNIA S.A.	5117	FRASCO|CC|50| 0|0	26000	26.000	0
Emicina 100	ZOETIS	963	FRASCO|CC|20| 0|0	0	ND	0
Finca Huevos Pequeño Productor	FINCA	8790	BULTO|KILOGRAMO|40| 0|0	46500	46500	0
Finca Novillas Desarrollo	FINCA	4973	BULTO|KILOGRAMO|40| 0|0	47000	47000	0
Flativet	LABORATORIOS PROVET S.A.S.	2224	FRASCO|CC|120| 0|0	19300	19300	0
Inflacor	COLOMBIA S.A.	1948	FRASCO|CC|10| 0|0	19400	19400	0
Ivermectina	VITAGRO LTDA.	5649	FRASCO|CC|100| 0|0	8900	8900	0
Ivermectina	VITAGRO LTDA.	5649	FRASCO|CC|50| 0|0	0	ND	0
Tyloser	S.A.	5726	FRASCO|CC|100| 0|0	24800	24800	0

Responsable: ANALISTA

Convenciones:
ND: producto no disponible 0 Precio coincide 1 Precio no coincide

FIGURA 2: Formato de Verificación a Fuentes SIPSA, Fuente la Autora,2016

Quinta Etapa. Estimación de la proporción y Construcción del indicador de calidad Luego de realizada la prueba piloto se procedió a estimar la proporción en cada una de las fuentes y luego hacer una ponderación de las cuatro proporciones estimadas para hallar la variabilidad entre fuentes y dentro de las fuentes para calcular las fracciones de muestreo óptimas.

Un segundo paso en esta etapa fue la construcción del indicador de calidad, el cual tiene como objetivo establecer el porcentaje de calidad de la información recolectada por recolector y por cuidad, con el propósito de realizar un seguimiento a las inconsistencias que se presentan y así tomar las medidas preventivas y correctivas pertinentes.

Sexta Etapa . Determinación de las fracciones de muestreo para la fase posterior del estudio. A partir de la proporción estimada de precios correctos mediante de la prueba piloto, se procedió a determinar las fracciones de muestreo óptimas para la aplicación posterior de la metodología, con el fin de aplicarla mensualmente y estimar la verdadera proporción de calidad de la información recolectada en el componente. Para los calculos de las fracciones de muestreo se aplicaron las expresiones (9) y (10) de la sección de 2.1 Referente conceptual.

4. Resultados preliminares

Los pasos de la aplicación del Muestreo aleatorio simple en dos etapas en la prueba piloto arrojaron los siguientes resultados:

MUESTRA UPM (n)	FUENTE	ARTICULOS TOTALES (M)	ARTICULOS MUESTREADOS (m)	PRECIOS CORRECTOS	PROPORCION ESTIMADA $\hat{p}$	VARIANZA ENTRE UPM S_j^2
N1	AGROPROTECCION	78	16	16	1,00	
N2	GANAPROTECCION	61	12	12	1,00	
N3	AMAGRO	51	11	9	0,81	
N4	SAN JORGE	74	15	13	0,86	
Total		**264**	**54**	**50**	$\bar{P}$ =0,92	0,0095

Figura 3: Resultados tamaños muestrales y estimación de cada una de las proporciones

En la tabla de la Figura 3, se observa la proporción estimada para cada una de las fuentes y finalmente se observa el promedio ponderado para el total de fuentes, el cual dio un resultado de 0.92. Para el cálculo de la varianza del estimador (varianza dentro de UPM) se utilizó la expresión (2) el cual arrojo un resultado de 0,0013 con un error estándar de 0.036. También se calculó la varianza entre unidades primarias (UPM) es decir entre los cuatro almacenes, la cual dio un resultado de 0,0095, lo que nos dice que hay mayor variabilidad entre almacenes que dentro de los mismos, es decir en los artículos de cada almacén

Construcción del indicador de calidad

Nombre del indicador		**Fórmula**	
Indicador de calidad del proceso de recolección de información por ciudad		$I_{RC} = \frac{N^{\circ}\, precioscorrectos}{N^{\circ}\, preciosmuestreados} * 100\%$	
RANGO DE TOLERANCIA DEL INDICADOR			
NO CUMPLE	**CUMPLE PARCIALMENTE**	**CUMPLE**	**META O TENDENCIA ESPERADA**
MENOR A 90%	ENTRE 90% A 95%	MAYOR A 95%	**100%**
Tipo de indicador	Cuantitativo	Positivo	
Periodicidad:	Mensual		
Fuente:	Formato de verificación a fuentes SIPSA		
Responsable:	Analista SIPSA		

Figura 4: Ficha Indicador de calidad SIPSA, Fuente la Autora, 2016

El resultado obtenido para el mes de abril fue de $I_{RC} = \frac{50}{54} * 100\,\% = 92\,\%$. El nivel de referencia de este indicador de calidad deben tener un resultado del 100 % lo que indica ausencia total de inconsistencias y calidad total en la información (DANE, 2009)

Determinación de las fracciones óptimas de muestreo

(UPM)	Tamaño UPM	n_j	S_i^2 UPM	$N_i S_i^2$	$N_i S_i$	f_j	n_j
AGRICOLA MERCOSUR	92	0,92	8,79121E-06	0,00081	0,27278	0,11381	10
AMAGRO	53	0,92	2,67054E-05	0,00142	0,27389	0,19837	11
AGROPROTECCION	81	0,92	1,1358E-05	0,00092	0,27298	0,12937	10
CASA DEL AGRO	53	0,92	2,67054E-05	0,00142	0,27389	0,19837	11
INAGROB	80	0,92	1,16456E-05	0,00093	0,27300	0,13099	10
SUPERAGRO	88	0,92	9,61338E-06	0,00085	0,27285	0,11902	10
EL CEBU	33	0,92	6,9697E-05	0,00230	0,27550	0,32046	11
GANAPROTECCION	61	0,92	2,01093E-05	0,00123	0,27354	0,17213	11
SAN JORGE	74	0,92	1,36246E-05	0,00101	0,27315	0,14169	10
TOTAL: 9	**615**		2,203E-05	0,01355	2,88642	0,18016	**94**

FIGURA 5: Tabla 3. Cálculos para determinación de las fraccciones óptimas de muestreo

En la tabla de la Figura 5, se observa los cálculos de las fracciones de muestreo para toda la población por lo cual al reemplazar los valores en las expreciones (9) y (10), se obtuvo un resultado de $f_1 = 1,224$, el cual muestra que la fracción de muestreo de fuentes (UPM) es mayor a 1 , e indica en ese caso que para que la muestra sea óptima se deben tomar todas las fuentes, mientras que $f_j = 0,18$ es inferior a 1, indica la fracción de artículos a muestrear que en total nos arroja un n_j de entre 10 y 11 artículos por fuente, para un total de 94 artículos, los cuales serian los tamaños óptimos de muestreo.

5. Conclusiones

Se reconoce la importancia de realizar la prueba piloto antes de implementar la metodología con el fin de estimar la proporción muestral necesaria para calcular las fracciones de muestreo óptimas, así como para identificar fallos en el instrumento de medición o las dificultades que se pueden presentar en la aplicación del mismo.

Como en la prueba piloto se encontró una alta variabilidad en las proporciones entre almacenes, esto conduce a que en etapas posteriores se debe realizar un muestreo bietápico estratificado, dado que el número de unidades primarias es pequeño se deben tomar todas las fuentes.

El resultado del indicador obtenido de acuerdo al nivel de referencia nos estaría indicando que la calidad no es del todo total, aunque a simple vista, esto se debe en a que la muestra preliminar es poco representativa, pero se espera que al aplicar los tamaños óptimos de muestreo esta proporción aumente.

La metodología propuesta permite de forma paulatina, ir evaluando mensualmente la calidad de la información recolectada, e implementando mejoras en el tiempo tanto en el proceso de recolección como en el de evaluación realizando un control o seguimiento exhaustivo del mismo

Referencias Bibliográficas

Cochran, W. (1976), 'Técnicas de muestreo', *Compañía Editorial Continental SA México* .

DANE (2009), 'Guía para la obtención de indicador de calidad en las direcciones territoriales y subsedes índices de precios al consumidor-ipc'.

DANE (2015*a*), 'Ficha metodológica componente insumos y factores'.

DANE (2015*b*), 'Marco de garantía de la calidad del sistema estadístico nacional'.

DANE (2015*c*), 'Metodología General Sistema de Información de Precios y Abastecimiento del Sector Agropecuario Componente de Insumos y Factores Asociados a la Producción Agropecuaria'.

EUROPARC (2002), 'Manual guía para la definición e implantación de un sistema de indicadores de calidad', España.

Galmés, M. y Galmés, M. (1997), Métodos de muestreo, Technical report.

Gutiérrez, H. A. (2009), Estrategias de muestreo diseño de encuestas y estimación de parámetros, Technical report, Universidad Santo Tomás, Bogotá (Colombia).

Hidalgo, F. K. y Ramírez, J. C. R. (2009), 'Aplicación de las técnicas de muestreo en los negocios y la industria', *Ingeniería Industrial* (27), 11–40.

INE (sin año), 'Precios coyunturales de productos agrícolas. informe metodológico estandarizado', España.

León, L. P. y de Rojas Gómez, H. (2010), 'Estadística aplicada a la actividad empresarial (i) técnicas de muestreo y la auditoría (i)', *Economía y Desarrollo* **145**(1-2), 197–220.

Montgomery, D. C. (1991), *Introducción al control estadístico de la calidad*, number 04; TS156, M6.

Morillas, A. (2007), 'Muestreo en poblaciones finitas'.

Ojeda, M. M., Camacho, J. E. D., Victoria, C. A. y Landa, I. T. (2011), *Metodología de diseño estadístico*, Universidad Veracruzana.

Pérez López, C. et al. (2000), *Técnicas de muestreo estadístico: teoría, práctica y aplicaciones informáticas*, number 519.52 P4.

Zapata-Ossa, H. d. J., Cubides-Munévar, A. M., López, M. C., Pinzón-Gómez, E. M., Filigrana-Villegas, P. A. y Cassiani-Miranda, C. A. (2011), 'Muestreo por conglomerados en encuestas poblacionales', *Revista de Salud Pública* **13**(1), 141–151.

METODOLOGÍA PARA EVALUAR ESTADÍSTICAMENTE EL EFECTO DE UN BIOPESTICIDA ELABORADO CON SEMILLAS DE MELIA AZEDARACH
Sobre neoleucinodes elegantalis (guenée), (lepidóptera: crambidae), plaga de las plantas de la familia solanaceae

Especialización en Estadística

Eric Giovanny Osorio Olea[1,a], Luis Guillermo Díaz Monroy[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Este artículo presenta la descripción de la forma de multiplicar el insecto Neolucinodes elegantalis, en condiciones de laboratorio, la forma de preparar un biopesticida en forma artesanal a partir de semillas de Melia azaderach y tras la aplicación del biopesticida al insecto, describir la metodología para evaluar estadísticamente la mortalidad del insecto haciendo uso de un modelo lineal generalizado.

***Palabras clave*:** Modelo lineal generalizado, Distribución binomial, DL50, Neoleucinodes elegantalis, Melia azaderach.

Abstract

This article presents the description of the way to multiply the Neolucinodes elegantalis insect in laboratory conditions, the way to prepare a handcrafted biopesticide since Melia azaderach seeds and after the application of the biopesticide to the insect, to describe the methodology to evaluate statistically the mortality of the insect, using a Generalizated Lineal Model GLM.

***Key words*:** Generalized linear model, Binomial distribution, DL50, Neoleucinodes elegantalis, Melia azedarach.

1. Introducción

Como parte de las actividades del grupo de investigación Agrociencia, de la Carrera de Ingeniería Agronómica de la Universidad de Cundinamarca en Facatativá en su semillero Manejo Integral de Cultivos, énfasis Entomología, los estudiantes del semillero quisieron estudiar los problemas fitosanitarios de los árboles de tomate de árbol sembrados en el vivero experimental. Con el propósito de conseguir información al respecto estudiantes del semillero visitaron, en representación de la Universidad, el municipio de Arbelaez, al suroccidente Cundinamarca. En un momento determinado los agricultores les plantearon la problemática de que sus costos de producción eran muy altos porque debían hacer muchas aplicaciones de insecticidas para proteger sus cultivos de Tomate de árbol del insecto denominado pasador del fruto, y que la efectividad de estos controles era muy discutible. Preguntaron los agricultores a los estudiantes acerca de si ¿no habría una forma

[a]Estudiante de Especialización en Estadística. E-mail: eric.osorio@uptc.edu.co
[b]Profesor asistente. E-mail: lgdmonroy@gmail.com

menos onerosa de hacer control de ese problema?. Tras una lluvia de ideas con los campesinos, surgio la idea de: ¿probar si la elaboración artesanal de una infusión de semillas de árbol de paraíso serviría para controlar los adultos del pasador del fruto? dado que este árbol es común en la zona, y es utilizado como sombrío en los potreros o como cerca viva.

Se elaboró un borrador de las necesidades de investigación para responder las preguntas planteadas haciendo uso por parte de los integrantes del semillero de los laboratorios de la Universidad en Facatativá, y en ese momento surgío la necesidad de saber como manipular estadísticamente los datos de mortalidad y calcular la dosis letal mínima, dado que este fenómeno tiene una distribución binomial y se debe entonces trabajar mediante un modelo lineal generalizado, por tal razón el autor de este artículo, profesor líder del semilero, aprovecho la oportunidad de estar realizando una especialización en estadística en la Universidad Pedagógica y Tecnológica de Colombia en Duitama, Boyacá para con la ayuda del Doctor Luis Guillermo Díaz Bermúdez dar respuesta a este interrogante y plantear una metodología.

El presente trabajo establecerá la metodología estadística que permita valorar la efectividad de tal biopesticida, expresada en mortalidad y establecerá el cálculo de la dosis letal media LD50 mediante un modelo lineal generalizado.

2. Referente Conceptual

El perforador del fruto del tomate de árbol y otras solanáceas, Neoleucinodes elegantalis (Guenée) (Lepidoptera: Crambidae) es un insecto plaga clave distribuido en el continente americano y en dieciocho departamentos de Colombia (Díaz A., 2013). El insecto es limitante en cultivos de solanáceas como el tomate de árbol Solanum betaceum Cav., que para el año 2013 en Colombia es uno de los frutales de clima medio más importantes dado que se produjeron 161.748 Ton, siendo Antioquia el principal departamento productor con 82.390,8 Ton, seguido por Cundinamarca con 42.120,2 Ton, Tolima con 10.905 Ton, Boyacá con 6.543,2 Ton y Huila con 4.307 Ton. (CCI, 2013).

En el cultivo de tomate de árbol los daños ocasionados por N.elegantalis pueden alcanzar al 22 % de pérdidas en la cosecha, provocando un alza en los costos de producción (Colorado, Díaz, Yepez y Rueda 2010). La hembra oviposita en la superficie del fruto y al eclosionar la larva neonata se desplaza hasta encontrar el sitio adecuado para perforar el fruto y desarrollarse dentro de él. La larva permanece alimentándose dentro del fruto y al terminar su desarrollo sale a empupar en el suelo o en la planta. Las larvas de último instar expulsan los excrementos hacia el exterior cuando hacen el orificio de salida (Viáfara, Garcia y Díaz 1999). Las prepupas caen al suelo donde construyen la envoltura pupal sobre las hojas secas (Díaz, Peña, Silva y Trochez 2003).

Al alimentarse del endospermo y mesocarpio del fruto causa o la maduración incompleta del fruto y su abcisión anticipada de la planta, o sus galerías con residuos, lo que en ambos casos hacen el fruto invendible. La larva neonata en la superficie del fruto es vulnerable a enemigos naturales y control químico pero no permanece en el fruto más de un promedio de 2,3 horas antes de perforarlo. Debido a que la larva permanece dentro del fruto, el control químico no es eficiente (Sandoval y Manzano 2012). Es importante encontrar alternativas de reducción de población enfocadas al adulto, sin los inconvenientes de selección de poblaciones resistentes.

Melia azedarach, llamado popularmente cinamomo, agriaz, lila, paraíso sombrilla o árbol del paraíso, es un árbol mediano, de hoja caduca, de la familia de las meliáceas. El fruto seco y pulverizado sirve como insecticida y para defenderse de los piojos. La toxicidad de los frutos afecta al ser humano y otros mamíferos, aunque no a las aves. Contiene neurotoxinas, en especial tetranortriterpeno (Mishra, Jawla y Srivastava n.d.).

Los biopesticidas generados a partir de partes de árboles pertenecientes a la familia Meliácea, como Melia azedarach, han sido una estrategia que ha tenido uso diverso, por ejemplo, la elaboración artesanal de un biopesticida ha sido útil para proteger comunidades en Etiopia contra Aedes aegipty vector de la malaria (Trudel y Bomblies 2011). Fueron utilizados utilizados extractos en etanol de Melia azedarach como biopla-

guicida en almacenamiento de frijol caupí Vigna unguiculata L. para defenderlo del gorgojo Callosobruchus maculatus (F) (Kosma, Bakop, Djile, Abdou y Goudoum 2014).

La azaridactina es el principal terpenoide que se manifiesta tener acción insecticida contra lepidópteros entre otros artrópodos, además del efecto insecticida tiene un efecto de anticenosis como el referido a los factores inhibidores de alimentación, de apareamiento o de oviposición, así como el posible efecto insecticida lo que ha sido probado en el pasado a través de diferentes tipos de experimentos. Como por ejemplo para evaluar la antibiosis de Tecia Solanivora (Polilla Guatemalteca) sobre cinco variedades de papa, se realizaron experimentos para estimar el efecto los porcentajes de pupamiento y emergencia de adultos de T. solanivora con el fin de comparar cual variedad induce menor desarrollo poblacional (Ch, Rosero y Bacca 2012). Se han establecido las propiedades entomotóxicas contra otros lepidópteros como Spodoptera exigua Hübner, Sesamia nonagroides (Delgado Barreto, García-Mateos, Ybarra-Moncada, Luna-Morales, Martínez-Damián et al. 2012); (Valladares, Garbin, Defagó, Carpinella y Palacios 2003); (Riba i Viladot, Torra y Martí Martí 1996).

De Melia azaderach se han identificado algunos de sus metabolitos y procedimientos de extracción de sus biocomponentes (Aoudia, Oomah, Zaidi, Zaidi-Yahiaoui, Drover y Harrison 2013). Se plantea con este trabajo inferir a partir de los experimentos realizados con N. elegantalis (Salinas et al, 1993) y otros insectos tales como los ya mencionados S. nonagroides o T. solanivora, lepidópteros, la metodología para la multiplicación de N. elegantalis, Y la forma de elaboración del pesticida en forma artesanal de M. azedarach para ser usado contra sobre esta plaga.

3. Resultados

FASE 1 Establecimiento de la metodología para la multiplicación de los insectos:

Se recolectarán frutos de Tomate de árbol infestados de larvas de N. elegantalis en cultivos de Arbeláez Cundinamarca, Colombia ubicados a 1.486 m.s.n.m. Los frutos serán transportados al Laboratorio de Entomología y Acarología de la Universidad de Cundinamarca en Facatativá, y dejados en bandejas de plástico hasta la recuperación de pupa.

Después de la emergencia, hembras copuladas serán expuestas a ovipositar diariamente en tomates, dentro de casa de malla de medidas 1 metros de largo y ancho por 0.70 de alto. Los frutos con huevos serán retirados diariamente y transportados a vasos plásticos con papel picado que imita la hojarasca en la cual el último instar larval construye su capullo, de esta forma se obtendrá la segunda generación de pupas. Con los adultos de la segunda generación se realizará el experimento de control con el biopesticida.

FASE 2 Propuesta metodológica para preparación del bioextracto:

Con las semillas secas, se debe proceder a molerlas en una licuadora utilizándose este material para las suspensiones en agua destilada, los extractos serán preparados mezclando 1, 2, 3,4 y 5 del polvo de semillas de Paraíso, separadamente en recipientes con 100 ml de agua destilada. Las mezclas serán dejadas en reposo durante 24 horas, almacenadas en condiciones de oscuridad para la extracción de los metabolitos secundarios hidrosolubles, en especial los tetranortriterpenos, siendo posteriormente filtradas con filtro fino, descartándose la parte sólida y utilizándose solo la parte líquida. Se procede así para demostrar que es posible fabricar un biopesticida en forma artesanal, sin incluir procedimientos complejos como extracción en etanol u otros solventes, una preparación semejante fue utilizada contra Aedes aegypti, Díptera: Culicidae pero por inmersión (Trudel y Bomblies 2011).

FASE 3 Diseño estadístico del experimento y análisis de los datos:

Unidad Experimental. Cada unidad experimental estará conformada por un recipiente que consiste en dos vasos plásticos unidos por su boca y horadados 3 mm para el ingreso de oxígeno y permitir el ingreso del orificio de salida del asperjador manual. En el interior de la unidad experimental se colocará un fruto de tomate. Dentro de cada unidad manualmente se colocaran tres (3) parejas de adultos de Neoleucinodes elegantalis, con 5 repeticiones por tratamiento totalizando treinta (30) adultos por tratamiento. O sea 180 adultos para todo el experimento.

Figura 1: Adulto de Neulocinodes elegantalis ampliado y en casa de malla

Tratamientos. Los tratamientos estarán constituidos por los gramos del extracto/milímetros de agua (g/100ml) en el estudio comparativo de la eficacia de los extractos de Paraíso para el control de Neoleucinodes elegantalis, serán como sigue:

Estructura de los datos

Tratamiento	Respuesta: Toxicidad por contacto directo
Agua	Mortalidad
Suspensión de Polvo de semillas 1 g/100 ml	Mortalidad
Suspensión de Polvo de semillas 2 g/100 ml	Mortalidad
Suspensión de Polvo de semillas 3 g/100 ml	Mortalidad
Suspensión de Polvo de semillas 4 g/100 ml	Mortalidad
Suspensión de Polvo de semillas 5 g/100 ml	Mortalidad

Tipo de diseño: El diseño experimental utilizado será Factorial completamente aleatorizado en donde el factor dosis de biopesticida tendrá seis niveles y cinco repeticiones para cada uno. El factor madurez del fruto tendrá tres niveles. Se presentarán cinco repeticiones para cada uno.

Variable a medir: La variable a medir es la mortalidad de los adultos.

Toxicidad por contacto directo con solución biopesticida: Estos ensayos se llevarán a cabo para los adultos. Las pruebas toxicológicas se realizarán con cohortes de hembras y machos adultos sexados de acuerdo a su morfología antenal y colocados al azar en las unidades experimentales y con menos de 48 horas de emergidos de las pupas. Los bioensayos se realizarán bajo condiciones de oscuridad. Los adultos serán alimentados antes de los bioensayos toxicológicos con una solución de miel de abejas empapada en algodón ubicada dentro de la caja de malla donde se encuentran todos los adultos. Los insectos serán capturados con la mano en condiciones de luz y transferidos uno a uno a la unidad experimental. Para cada uno de los tratamientos se utilizarán 3 individuos por 5 repeticiones. Las concentraciones acuosas de los extractos de semilla de paraíso disueltos en agua destilada y el agua destilada se aplicarán en hora crepuscular, esparciendo aproximadamente 25 ml promedio de la solución biopesticida por unidad experimental (según ensayo previo) y posteriormente las unidades experimentales se mantendrán en condiciones de oscuridad utilizando telas oscuras que recubran el área experimental. Los individuos se considerarán muertos cuando al cabo de un

tiempo aproximado de 3.5 horas los líquidos dentro de la unidad se evaporen y los adultos dentro de la misma no se posen sobre las paredes del recipiente con sus patas y se encuentren en el fondo del recipiente con las patas hacia arriba, durante 10 s de observación al microscopio estereoscopio, habiendo estado previamente en oscuridad. El tratamiento control consistirá en agua destilada. Se harán revisiones cada 2 horas las primeras 24 horas y cada 6 horas el segundo día.

FASE 4 Análisis de los datos:

Inicialmente se realizará un análisis descriptivo exploratorio de las observaciones para verificar los supuestos del modelo, es decir la presencia o no de normalidad del error, errores no correlacionados, presencia de observaciones inusuales e influyentes.

Con los datos de mortalidad se establecerá el cálculo de la dosis letal media DL50 a partir de un modelo lineal generalizado en la siguiente forma: Según (García 2002) Un modelo lineal generalizado involucra:

- Una variable respuesta, que hace parte del componente aleatorio del modelo que tiene una distribución perteneciente a la familia exponencial.
- Unas variables explicativas que hacen parte del componente sistemático del modelo.
- Una función de enlace que une los componentes aleatorios y sistemático

Los ensayos de tipo dosis respuesta son aquellos en donde una droga es administrada en$k$ diferentes dosis d, a individuos quienes cambian de estado por la ocurrencia de un suceso como la muerte. En la situación contemplada el modelo sugerido es el modelo binomial que se expresa:

$$Y = Bin(\pi, m) \tag{1}$$

Siendo: π: probabilidad de muerte, y m: número de insectos que recibirán una dosis de biopesticida.

El objetivo de este tipo de experimento es determinar las dosis efectivas necesarias para eliminar la mitad de la población, dato útil para comparar la potencia de diferentes productos.

Como la gráfica de tal situación toma aspecto sigmoidal el problema, entonces, consiste en encontrar una curva sigmóide que se ajuste bien a los datos y a partir de ella obtener la dosis letal media DL50. Es necesario transformar esa curva sigmóide en una recta con el fin de estimar los parámetros mediante los procedimientos comunes de la regresión, lo que puede graficarse generando una curva de la siguiente forma:

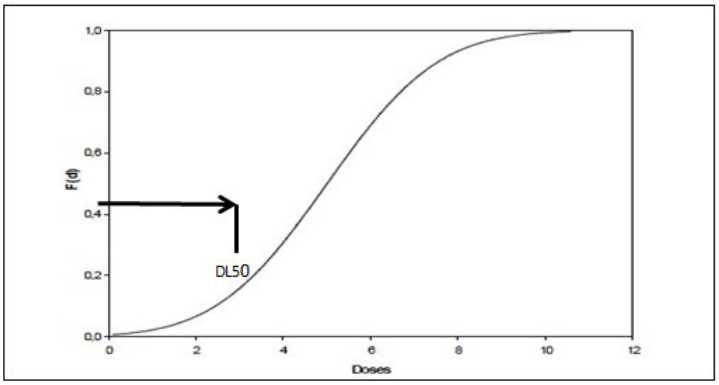


Figura 2: Curva respuesta Mortalidad y variable independiente Dosis

Modelo probit (Probability unit:) El modelo Probit asume que los datos tienen una distribución normal con media μ y desviación σ^2 Osea:

$$f_u(U;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(u-\mu)}{2\sigma^2}\right]$$

Y entonces el modelo binomial se transforma en una función de la forma:

$$Probit(\pi) = \phi^{-1}(\pi) = \beta 1 + \beta 2di$$

Valoración de la eficacia La eficacia de los extractos será calculada contabilizándose los adultos muertos por cada unidad experimental, corrigiéndose con la fórmula de Abbott (1925), a expresarse originalmente en porcentaje, considerándose eficaces los tratamientos que superan el 80 % de mortalidad corregida. Mtr - Mte MC = x 100 100 ? Mte Donde: MC: porcentaje de mortalidad corregida; Mtr: número de individuos muertos en el tratamiento con el producto; Mte: número de individuos muertos en el testigo absoluto. (Caballero y Mena 2013)

4. Conclusiones

Los resultados esperados deberán demostrar si la fabricación artesanal de biopesticida es efectiva y en que grado controla el defoliador, lo que sería un avance en el control de la plaga a un costo razonable.

El cálculo de la dosis letal 50 permite cuantificar el costo de una aplicación de biopesticida incluida la fabricación y ser comparada con los insecticidas de síntesis para establecer entonces si económicamente tiene ventajas el fabricar el biopesticida con subproductos del arbol del paraíso así ambientalmente sea evidente que es mejor.

Referencias Bibliográficas

Aoudia, H., Oomah, B., Zaidi, F., Zaidi-Yahiaoui, R., Drover, J. C. y Harrison, J. (2013), 'Phenolics, antioxidant and anti-inflammatory activities of melia azedarach extracts', *International Journal of Applied Research in Natural Products* **6**(2), 19–29.

Caballero, V. L. T. y Mena, E. F. G. (2013), 'Acción insecticida y repelente del neem sobre adultos de callosobruchus maculatus f.(coleoptera: Bruchidae) en granos de poroto (vigna unguiculata)', *Investigación Agraria* **13**(2), 107–111.

Ch, M. F. O., Rosero, J. F. y Bacca, T. (2012), 'Resistencia de cinco variedades de (solanum spp., solanaceae) al ataque de tecia solanivora (lepidoptera: Gelechiidae)', *Boletín Científico. Centro de Museos. Museo de Historia Natural* **16**(1), 108–119.

Colorado, W., DÍAz, A., Yepez, F. y Rueda, J. (2010), Evaluación de la feromona sexual de neoleucinodes elegantalis neoelegantol®(guenée)(lepidoptera: Crambidae) en solanáceas cultivadas y silvestres, *in* 'Resúmenes XXXVII Congreso de la Sociedad Colombiana de Entomología SOCOLEN, Bogotá, Colombia', Vol. 36.

Díaz, A. (2013), 'Manejo integrado del gusano perforador del fruto del lulo y tomate de árbol.', Fondo regional de tecnología agropecuaria. Fontagro. Corpoica C.i La selva, Ronegro Antioquía, Colombia.

Delgado Barreto, E., García-Mateos, M., Ybarra-Moncada, M., Luna-Morales, C., Martínez-Damián, M. et al. (2012), 'Propiedades entomotóxicas de los extractos vegetales de azaradichta indica, piper auritum y petiveria alliacea para el control de spodoptera exigua hübner', *Revista Chapingo. Serie horticultura* **18**(1), 55–69.

Diaz, A., Peña, J., Silva, J. y Trochez, A. (2003), 'Control biológico y mecánico del perforador del fruto de tomate de mesa, neoleucinodes elegantalis', *Revista Regional Novedades Técnicas* **4**, 22–26.

García, C. (2002), *Modelos lineales generalizados en investigaciòn agronómica*, Escuela de agricultura Luis de Quiroz, Universidad de San Pablo. Piracicaba.

Kosma, P., Bakop, R., Djile, B., Abdou, B. y Goudoum, A. (2014), 'Bioefficacity of the powder of melia azedarach seeds and leaves against callosobruchus maculatus, on cowpea seeds (vigna unguiculata) in storage', *Journ. Agric. Res. Dev* **5**(4), 72–78.

Mishra, G., Jawla, S. y Srivastava, V. (n.d.), 'Medicinal chemistry & analysis'.

Riba i Viladot, M., Torra, E. y Martí Martí, J. (1996), 'Bioactividad de extractos de melia azedarach l. sobre el taladro del maíz sesamia nonagrioides lef.', *Boletín de sanidad vegetal. Plagas, 1996, vol. 22, núm. 2, p. 261-276* .

Sandoval, S. F. M. y Manzano, M. R. (2012), 'Neoleucinodes elegantalis (lepidoptera: Crambidae) plaga de solanum quitoense ¿es vulnerable al control el primer estadio larval?', *Acta Agronómica* **61**(5), 61.

Trudel, R. E. y Bomblies, A. (2011), 'Larvicidal effects of chinaberry (melia azederach) powder on anopheles arabiensis in ethiopia', *Parasit vectors* **4**, 72.

Valladares, G., Garbin, L., Defagó, M. T., Carpinella, C. y Palacios, S. (2003), 'Actividad antialimentaria e insecticida de un extracto de hojas senescentes de melia azedarach (meliaceae)', *Revista de la Sociedad Entomológica Argentina* **62**(1-2), 53–61.

Viáfara, H., Garcia, F. y Díaz, A. (1999), 'Parasitismo natural de neoleucinodes elegantalis (guénee) (lepidoptera: Pyralidae) en algunas zonas productoras de solanáceas del Cauca y Valle del Cauca colombia', *Revista Colombiana de Entomología* **25**(3-4), 151–159.

DISEÑO Y ANÁLISIS DE UN EXPERIMENTO PARA TUTORADO EN ARVEJA BAJO PRESENCIA DE SOBREDISPERSIÓN RESPECTO AL MODELO POISSON

Especialización en Estadística

María Eliana Díaz Sosa[1,a], Eduardo Dávila S.[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El objetivo del presente artículo es plantear un diseño de experimentos y analizar los datos para comparar el rendimiento de la arveja de seis variedades de semilla con dos sistemas de tutorado. El experimento es trazado para condiciones climáticas del municipio de Tota del departamento de Boyacá, en el que se plantea un diseño en bloques completamente al azar, donde se evaluan tres sistemas de tutorado (horizontal, vertical y sin tutorado) en el cultivo. Donde las variables respuesta (Y) son: número de vainas por planta, número de granos por vaina y peso de 100 semillas. Para el análisis de los datos se propone los modelos de análisis de varianza (ANOVA), modelo de regresión poisson (M.R.P) y modelo de regresión binomial negativa (M.R.B.N). La variable $Y \sim Poisson$, y se supone que presenta sobredispersión, es decir, $V(Y) > E(Y)$, esto debido a la variabilidad de Y, concluyendo según (Krzanowski 1998) que cuando existe exceso de varianza en los estimadores y no se tiene en cuenta en el análisis se llegar a inferencias inapropiadas.

***Palabras clave*:** Sobredispesión, Poisson, Diseño de bloques completamente aleatorizados, Tutorado.

Abstract

The objective of this article is to propose a design of experiments and analyzing data to compare efficiency peas six varieties of seeds with two backing systems. It is used to design in a randomized complete block, where three backing systems (horizontal, vertical and without backing) are evaluated in the crop. Besides, for data analysis is proposed models analysis of variance (ANOVA), poisson regression model (M.R.P) and negative binomial regression model (M.R.B.N). The variable $Y \sim Poisson$, and is supposed to present overdispersion.

***Key words*:** Overdispersion, Poisson, Design in a randomized complete block, Backing.

1. Introducción

En Colombia, el cultivo de la arveja, es el segundo en importancia en cuanto a leguminosas después del fríjol, existiendo dos sistemas de producción. El primero, y de mayor cobertura, es el "tutorado" para la producción de arveja en vaina o verde, comúnmente denominada "colgada". El segundo, el rastrero sin tutorado, se usa especialmente para la producción de semilla (DANE 2015).

[a]Estudiante de Especialización en Estadística. E-mail: mariaeliana.diaz@uptc.edu.co
[b]Profesor asistente. E-mail: eduardo.davila@talex.edu.co

La investigación en arveja ha sido escasa en el país y ha estado enfocada hacia la obtención de nuevos materiales para el uso en tutores y al manejo del cultivo en los departamentos de Cundinamarca y Nariño. En el departamento de Boyacá se llevó a cabo en los municipios de Samacá, Viracachá, Ciénega y Ramiriquí, el proyecto innovación y desarrollo tecnológico y participativo para la agricultura sostenible del cultivo de arveja. Los logros evidenciados fueron (Reyes y Garcia n.d.): los productores adoptaron prácticas de selección de semillas, manejo integrado del cultivo (manejo integrado de plagas y enfermedades y utilización de abonos orgánicos); Sin embargo, para la provincia del Sugamuxi no hay investigación reportada.

En arveja se han formulado modelos que permiten predecir el comportamiento de caracteres de rendimiento, calidad de grano y precocidad y su interacción con el ambiente; sin embargo, estos modelos han sido propuestos para ser utilizados en zonas donde existen estaciones (González 2001). Por lo tanto, se requiere plantear el diseño de experimentos y un modelo que permita explicar el rendimiento de la producción de arveja según el tipo de tutorado, para variable respuesta de conteo y en presencia de sobredispersión con respecto al modelo Poisson.

2. Referente Conceptual

Esta sección se divide en tres partes. Una primera parte en la que se presenta las generalidades del cultivo de arveja y los tipos de tutorado. Una segunda parte donde se muestran los diseños de experimentos y sus principios. En la tercera parte se revisan algunos modelos para análisis de datos experimentales.

2.1. Cultivo de Arveja

A continuación se presentan las generalidades de las variedades de la arveja y los sistemas de tutorado:

2.1.1. Generalidades

Las variedades existentes de arveja presentan características según tipo de suelo y al mejoramiento genético que se ha desarrollado en cada región, en Colombia, se siembran nueve variedades de arveja, dentro de las cuales se encuentran las que se adaptan bien entre 2.200 y 3.000 m.s.n.m que son:Alcalá, ICA Tominé, Andina, Santa Isabel, Sureña y San Isidro (DANE 2015).

2.1.2. Sistema de Tutorado

En Colombia existen diversos sistemas de siembra de arveja en monocultivo, entre los que (Castro Restrepo 1995) destaca los siguientes:

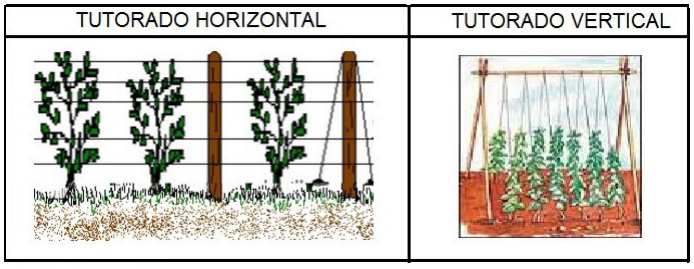


Figura 1: Sistemas de tutorado horizontal y vertical. Fuente Ecoagricultor, 2016.

El uso de tutorado en la arveja consiste en colocar palos, cañas o estacas cerca de las plantas para que éstas puedan crecer en forma vertical y no tumbadas sobre el suelo, dañándose al ser pisadas, así también se ahorra espacio en el huerto. Es una práctica imprescindible para mantener la planta erguida, ya que algunos tallos, se parten con facilidad.

Para (Arévalo Vela y Ortega Alvizuri 1995) con la utilización de tutores en el cultivo de arveja se obtiene un mayor rendimiento por hectárea y una mejor calidad de frutos. El mayor rendimiento se debe a que los tutores permiten aprovechar mejor el espacio aéreo, disponiendo de mayor terreno para sembrar más plantas. La mejor calidad de los frutos se debe a más luminosidad que recibe el cultivo, favoreciendo un mejor llenado de las vainas; además, la posición vertical de las plantas contribuye a un control más eficiente de plagas, enfermedades y daños por pájaros.

2.2. Diseño de experimentos

En esta sección se hace una revisión de las generalidades, principios y directrices para realizar un diseño de experimentos.

2.2.1. Conceptos básicos de diseños de experimentos

En (Gutiérrez Pulido y De la Vara Salazar 2004) se establecen la definiciones de algunos términos y conceptos a aplicar en el diseño de un experimento:

1. Variable respuesta: A través de esta variable se conoce el efecto de los resultados de cada prueba experimental.
2. Unidad experimental: La unidad experimental es la unidad (sujeto, planta, matera, animal) que se asigna al azar a un tratamiento.
3. Factores controlables: Son variables de proceso o características de los materiales experimentales que se pueden fijar en un nivel dado. Algunos de estos son los que se controlan durante la operación normal del proceso.
4. Niveles y tratamiento: los diferentes valores que se asignan a cada factor estudiado en un diseño experimental se llaman niveles. Una combinación de niveles de todos los factores estudiados se llama tratamiento.
5. Factores estudiados: son las variables que se investigan en el experimento, respecto de cómo influyen o afectan a la variable de respuesta.
6. Factores no controlables o de ruido: Son variables o características de materiales y métodos que no se pueden controlar durante el experimento o la operación normal del proceso.
7. Error aleatorio y error experimental: siempre que se realiza un estudio experimental, parte de la variabilidad observada en la respuesta no se podrá explicar por los factores estudiados. Esto es, siempre habrá un remanente de incertidumbre que se debe a causas comunes o aleatorias, que generan la variabilidad natural del proceso, la cual es conocida como error aleatorio. Sin embargo, el error aleatorio también absorberá todos los errores que el experimentador comete durante los experimentos, y si estos son graves, más que el error aleatorio se hablará de error experimental.

2.2.2. Principios del diseño de experimentos

Los tres principios básicos del diseño de experimentos son (Casella 2008):

- Aleatorización: Tal vez el principio más fundamental del diseño es la aleatorizaciõn, es decir,la obtención de las observaciones (o, más precisamente, las unidades experimentales) en una forma aleatoria que es tan libre de sesgod como sea posible .
- Replicación: es la repetición de la situación experimental para replicar la unidad experimental.
- Control local o bloqueo: es una técnica de diseño utilizada para mejorar la precisión con la que las comparaciones entre los factores de interés están hechos.

2.2.3. Diseño de bloques completamente aleatorizados

El diseño de bloques completos aleatorizados es utilizado para controlar y reducir el error experimental, en él las unidades experimentales quedan estratificadas en bloques de unidades homogéneas, cada tratamiento se asigna al azar a un número igual (por lo general uno) de unidades experimentales en cada bloque y es posible hacer comparaciones más precisas entre los tratamientos dentro del conjunto homogéneo de unidades experimentales en un bloque. A continuación se presenta un modelo estadístico para el respectivo diseño y los supuestos: La respuesta de la unidad con el i-ésimo tratamiento en el j-ésimo bloque se escribe como:

$$y_{ij} = \mu + \tau_i + \rho_j + e_{ij} \; i = 1, 2, 3, ...j = 1, 2, 3, ...$$

Donde μ es la media general,τ_i es el efecto del tratamiento y $e_{ij} \sim N\left(0, \sigma^2\right)$ es el error experimental. El efecto del bloque ρ_j representa la desviación promedio de las unidades en el bloque j a partir de la media general (Kuehl 2001).

2.3. Algunos modelos para análisis de datos experimentales

Los modelos de análisis de datos experimentales son varios y no se van a desarrollar todos. Los que se van a presentar son, quizás, los más representativos.

2.3.1. Modelos Anova clásicos

La técnica del Análisis de la Varianza (ANOVA) es la más utilizada en los análisis de los datos de los diseños experimentales. Se utiliza cuando se quiere contrastar más de dos medias, por lo que puede verse como una extensión de la prueba t para diferencias de dos medias. Básicamente es un procedimiento que permite dividir la varianza de la variable dependiente en dos o más componentes, cada uno de los cuales puede ser atribuido a una fuente (variable o factor) identificable. El ANOVA parte de algunos supuestos que han de cumplirse:

- Y debe medirse al menos a nivel de intervalo.
- Independencia de las observaciones.
- $e_{ij} \sim N\left(0, \sigma^2\right)$.
- Homoscedasticidad.

Existen tres tipos de modelos:
Modelo I. Modelo de efectos fijos: Si se toman K niveles de un factor, a cada uno se asignan las muestras y las inferencias se refieren exclusivamente a los K niveles y no a otros que podrían haber sido incluidos, el ANOVA se llama de efectos fijos, sistemático o paramétrico, es decir, que el experimentador ha considerado para el factor todos los posibles valores que éste puede tomar.

Modelo II. Modelos de efectos aleatorios: Cuando los niveles son varios y se seleccionan al azar K niveles, pero las inferencias se desean hacer respecto al total de niveles, el análisis de varianza se denomina de efectos aleatorios.

Modelos mixtos. Cuando se utilizan dos factores, cada uno con varios niveles, uno de efectos fijos y otro de efectos aleatorios, el análisis de varianza es mixto.

Una observacióón individual se representa como:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \text{ Modelo I}$$
$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \text{ Modelo II}$$
$$Y_{ij} = \mu + \alpha_i + A_j + \varepsilon_{ij} \text{ Modelo III}$$

Donde μ es la media global, α_i es el factor fijo, A_j es el factor aleatorio ε_{ij} es la variable aleatoria residual o error, donde los residuos son independientes y $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$

Definición 1: Transformación de Box y Cox

La familia de transformaciones más utilizada para resolver los problemas de falta de normalidad y de heterocedasticidad es la familia de Box-Cox, cuya definición es la siguiente:
Se desea transformar la variable Y, cuyos valores muestrales se suponen positivos, en caso contrario se suma una cantidad fija M tal que $Y + M > 0$. La transformación de Box-Cox depende de un parámetro λ por determinar y viene dada por:

$$Z(\lambda) \begin{cases} \frac{y^{\lambda}-1}{\lambda} si \lambda \neq 0 \\ \log(y) si \lambda = 0 \end{cases}$$

2.3.2. Modelos lineales generalizados y extensiones

En muchas situaciones los modelos de regresión lineal clásicos no se pueden aplicar directamente, como en casos donde se pretende modelar respuestas no normales como conteos o proporciones. Esta situación motivó el desarrollo de los modelos lineales generalizados, con los que se pueden modelar respuestas categóricas, binarias, de proporciones y de conteo, entre otras. Un modelo lineal generalizado (MLG) se define mediante la especificación de tres componentes:

1. Componente aleatorio: Un vector $Y_1, ..., Y_n$ un conjunto de variables respuesta, caracterizado por los parámetros θ_i y ϕ, que pertenece a la familia exponencial, con la forma:

$$f(y_i; \theta, \phi) = \exp\left\{ \frac{1}{a(\phi)} [y\theta - b(\theta) + c(y; \phi)] \right\}$$

2. Un componente sistemático: Especifica las variables explicativas $x_i = (x_{i1}, ..., x_{ip})^t$ que ingresen en forma de efectos fijos de un modelo lineal, y se relacionan como:

$$n_i = \beta_0 + \sum_{i=1}^{p} \beta_j x_{ij}$$

   Donde β_j es el j-ésimo coeficiente de regresión y x_{ij} es el j-ésimo predictor, en el i-ésimo individuo, para $i = 1, ..., n$ y $j = 1, ..., p$

3. Función Enlace: Los dos componentes son combinados en el modelo mediante la elección de un enlace denotado como $g(.)$, de manera que relaciona μ_i con el predictor lineal η_i, a través de la función:

$$g(\mu_i) = \beta^t x_i \text{ con } i = 1, 2, ..., n$$

Definición 2: Modelo de regresión poisson (M.R.P)

Como caso especial esta el M.R.P que es el modelo de referencia en estudios de variables de recuento (Cameron 1998). Los tres componentes del MRP son:

1. Componente aleatoria: Dado $Y_i, ..., Y_n$ un vector de variable respuesta positiva $Y \sim Poisson(\mu)$ con $Var(Y) = \mu$

2. Componente sistemática: El predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho es: $\eta_i = \beta^t x_i$

3. Función de enlace: aquella que relaciona η con μ es: $g(\mu_i) = \log(\mu_i)$ la cual es la más utilizada, sobre la función identidad y raíz cuadrada.

Definición 3: Sobredispersión

La sobredispersión con respecto al modelo poisson ocurre cuando $V(Y) > E(Y)$.
Entre las diversas causas de la sobredispersión se puede mencionar (Winkelmann 2000):

- Alta variabilidad en los datos.
- Los datos no provienen de una distribución Poisson.
- Los eventos no ocurren independientemente a través del tiempo.
- Falta de estabilidad, es decir, la probabilidad de ocurrencia de un evento puede ser independiente de la ocurrencia de un evento previo pero no es constante.
- Errores de especificación de la media μ como omitir variables explicativas o que entran al modelo a través de alguna transformación en lugar de linealmente.

Definición 4: Modelo de regresión binomial negativa (M.R.B.N)

Una forma de debilitar el supuesto de equidispersión es especificar una distribución que permita un modelado más flexible de la varianza, como lo es con la binomial negativa (Vives Brosa 2002).

Binomial negativa como familia exponencial
Por tanto para definir el modelo alternativo se asume con $\Theta_i = \Gamma(\alpha, \lambda_i)$ donde la variable respuesta Y_{ij} con $i = 1, ..., p$ y $j = 1, ..., r_i$ tiene distribución binomial negativa con media $E(Y_{ij}) = \mu_i$ y varianza $Var(Y_{ij}) = \mu_i\left(1 + \frac{\mu_i}{\alpha}\right)$
De este modo para α fijo la función de distribución es

$$f(y_{ij}) = \frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha)\, y_{ij}!}\left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_{ij}}\left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha}$$

con parámetros μ_i y α, con $\mu_i > 0$ y $\alpha > 0$, que pertenece a la familia exponencial; es decir que se puede escribir de la forma

$$f(y;\theta) = \exp\left\{y_{ij}\ln\left(\frac{\mu_i}{\mu_i + \alpha}\right) + \ln\left(\frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha)\, y_{ij}!}\right) + \alpha\ln\left(\frac{\alpha}{\mu_i + \alpha}\right)\right\}$$

Donde $\theta = \frac{\mu_i}{\mu_i + \alpha}$, $b(\theta) = \alpha\ln\left(\frac{\alpha}{\mu_i + \alpha}\right)$ y $a(\phi) = 1$.

Modelo
Se dice que una variable Y_i sigue el M.R.B.N, si cumple que

$$Y_i = BN(\mu_i, \alpha) \text{ con } g(\mu_i) = x_i^t\beta$$

Elementos del M.R.B.N

1. Componente aleatoria: Dado $Y_1, ..., Y_n$ una variable aleatoria independiente que indica el número de sucesos necesarios para obtener r-éxitos. Es decir, el número de éxito está predeterminado y la aleatoriedad es el número de sucesos, de modo que: $Y_i = BN(\mu_i, \alpha)$ con $Y_i \in \{0, 1, ...\}$
2. Componente sistemática: Dado μ_i y el llamado predictor lineal simbolizado por:$\eta_i = x_i^t\beta$
3. Función de enlace: Los dos componentes son combinados en el modelo, mediante la elección de la función enlace: $g(\mu_i) = \eta_i$, la cual utiliza las mismas de M.R.P.

3. Diseño metodológico

La siguiente sección se divide en cuatro partes: el método de campo, sistema de hipótesis, el sistema de variables y el diseño experimental.

3.1. Reconocimiento y planteamiento del experimento

El diseño se plantea para la vereda Tota del municipio de Tota, el cual se encuentra localizado sobre la cordillera Oriental; a una altura de 2.870 m.s.n.m, con una temperatura promedio de 12 $°C$, presenta un clima frío y húmedo, dada su cercanía a la Laguna de Tota. Presenta una precipitación de alrededor de 903 mm (*Nuestro municipio-Información general* 2015). Condiciones climáticas que hacen posible plantear el experimento in situ.
Recolectar la información de la unidad experimental de la siguiente forma: La unidad experimental está constituida por un área a sembrar de 30 m^2 (6 m x 5 m). Con cinco surcos, en los cuales se siembran 2 semillas por golpe con una densidad de siembra 0.2 m entre planta y 1 m entre surco, 30 plantas por surco y 150 plantas en los cinco surcos. El área útil a tener en cuenta en la parcela para cada una de las unidades experimentales, se calcula descartando los surcos bordes y las plantas extremas de los surcos (Vaca 2011).

Características de las unidades experimentales:

• Unidad experimental neta (3 surcos)	12 m^2
• Número de plantas del surco	20 plantas
• Número de plantas por parcela	60 plantas
• Separación entre bloques	1 m
• Área total del ensayo	450 m^2
• Número de unidades experimentales	15

Figura 2: Características de la unidades experimentales. Fuente Díaz E.

El rendimiento se mide al cosechar las vainas de cada unidad experimental, en el cual se procede a determinar los datos de los componentes de rendimiento y registrarlos.

3.2. Sistema de hipótesis

Se plantean las siguientes hipótesis
Hipótesis general: Los dos sistemas de tutorado tienen la misma producción de arveja

$$H_0 : \mu_{t_1} = \mu_{t_2} = \mu_{t_3}.vs.H_A : \mu_{t_1} \neq \mu_{t_2} \neq \mu_{t_3}$$

Hipótesis especifica: Mediante el montaje del experimento se puede determinar cuál es el sistema de tutorado más eficiente.
Hipótesis M.L.G
Sea $\Omega = \beta_0 + \varepsilon$ y $\omega = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ con

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_A : \beta_1 \neq \beta_2 \neq 0$$

3.3. Variable respuesta y factores

Para el planteamiento del sistema de variables se tiene:

TIPO DE VARIABLE	SIMBOLO	VARIABLE	DEFINICION CONCEPTUAL
VARIABLE RESPUESTA	Y_1	Número de vaina por planta (VP)	Se obtiene contando el número de vainas por planta, en 10 plantas de la parcela útil tomadas al azar y registrando
	Y_2	Número de granos por vaina (GV)	Se obtiene contando los granos en 20 vainas tomadas al azar de la parcela útil y registrando
	Y_3	Peso de 100 semillas (PS)	Se registra el peso de 100 granos en gramos
VARIABLES INDEPENDIENTES	x_1	Tutorado	T1: Sin tutorado
			T2: Tutorado Horizontal
			T3: Tutorado Vertical
	x_2	Variedad de arveja	I. Alcalá
			II. Santa Isabel
			III. Sureña
			IV. Andina
			V. San Isidro
			VI. ICA-Tominé
VARIABLES INTERVINIENTES	Incidencia de plagas		Por hongos y enfermedades
	Clima		Cambios climáticos extremos

Figura 3: Sistema de variables. Fuente Díaz E.

3.4. Diseño experimental

A continuación se presenta el diseño seleccionado para realizar el experimento y los modelos de análisis.

3.4.1. Diseño de bloques completamente aleatorizados

Como el rendimiento de arveja depende de la variabilidad en la variedad de la semilla, se trabajará con cinco bloques, donde B1=Alcala, B2=Sureña, B3=Andina, B4=Santa Isabel, B5=San Isidro, B6=Tominé, de modo que cualquier diferencia en el rendimiento, causadas por la variedad de la semilla incluida en el modelo. La distribución de las parcelas experimentales en el campo se muestra en la siguiente figura.

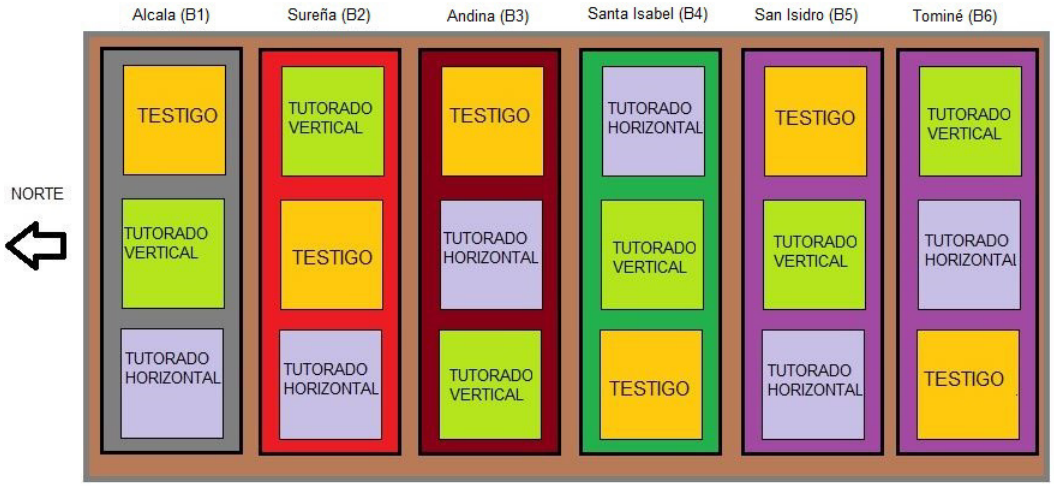


Figura 4: Croquis de distribución de tratamientos y repeticiones en el campo por bloques completamente al azar. Fuente Díaz E.

3.4.2. Métodos de análisis

En seguida se presenta algunos de los modelos a utilizar para realizar el análisis de los datos recolectados en el experimento.

ANOVA

: La respuesta de la unidad con el i-ésimo tratamiento en el j-ésimo bloque se escribe como:

$$y_{ij} = \mu + \tau_i + \rho_j + e_{ij} \; i = 1, 2, 3, j = 1, 2, 3, 4, 5, 6$$

Donde μ es la media general,τ_i es el efecto del tratamiento y $e_{ij} \sim N\left(0, \sigma^2\right)$ es el error experimental. El efecto del bloque ρ_j representa la desviación promedio de las unidades en el bloque j a partir de la media general. En la siguiente figura se puede observar el análisis de varianza para el diseño en bloques aleatoriazados

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F
Bloques	5	SCTr	CMTr	$\frac{CMTr}{CMR}$
Tratamientos	2	SCBl	CMBl	$\frac{CMBl}{CMR}$
Error	10	SCR	CMR	
Total	17	SCT	CMT	

Figura 5: Análisis de varianza para el diseño en bloques aleatorizados. Fuente Díaz E.

Si el Anova detecta efectos significativos para tratamientos, se procede a hacer pruebas de comparación múltiple.

Prueba de Tukey

Para un grupo de k medias de tratamiento, se calcula la diferencia honestamente significativa como (DHS):

$$DHS\left(k, \alpha_E\right) = q_{\alpha,k,v}\sqrt{\frac{CMR}{r}}$$

Donde $q_{\alpha,k,v}$ es el estadístico estandarizado de Student para un grupo de $k = 3$ medias de tratamiento en un arreglo ordenado. Los valores críticos de la tasa de error con respecto al experimento, $\alpha_E = 0,05$, los $v = 15$ grados de libertad y r=6. Prueba que permite comparar las medias de los niveles de un factor. Se establece que dos medias de tratamientos no son iguales, $\mu_i - \mu_j \neq 0$ si:

$$|\bar{y}_i - \bar{y}_j| > DHS\left(k, \alpha_E\right)$$

Función de desvío para M.R.P

El desvío se puede usar para probar la bondad de ajuste de un modelo y comparar dos modelos anidados

$$d^2\left(y_i, \hat{\mu}_i\right) = \begin{cases} 2\left\{y_i \log\left(\frac{y_i}{\mu_i}\right) - \left(y_i - \hat{\mu}_i\right)\right\}, si \\ 2\hat{\mu}_i, si, y_i = 0 \end{cases} \quad , y_i > 0$$

Donde $\mu_i = g^{-1}\left(x_i^t\beta\right)$ y $d^2 \sim x^2_{n-p}$

Función de desvío para M.R.B.N

Si se asume α fijo, la función de desvío esta dada por:

$$D^*\left(y; \hat{\mu}\right) = 2\sum_{i=1}^{n}\left[\alpha \log\left\{\frac{\hat{\mu}_i + \alpha}{y_i + \alpha}\right\} + y_i \log\left\{\frac{y_i\left(\hat{\mu}_i + \alpha\right)}{\hat{\mu}_i\left(y_i + \alpha\right)}\right\}\right]$$

Donde $\mu_i = g^{-1}\left(x_i^t\beta\right)$ y $D^* \sim x^2_{n-p}$

4. Resultados esperados

- Montaje de un experimento diseñado estadisticamente acorde a las hipótesis planteadas y al sistema de variables.
- Comparación de las inferencias de los modelos, determinando las inferencias a las que lleva el tipo de modelo escogido.
- Recomendaciones reales a las condiciones de la zona sobre el tutorado y las diferentes variedades de semilla de arveja a utilizar.

Referencias Bibliográficas

Arévalo Vela, C. y Ortega Alvizuri, V. (1995), 'Uso de tutores en el cultivo de arveja'.

Cameron, A. y Trivedi, P. (1998), *Regression Analysis of Count Data*.

Casella, G. (2008), *Statistical Design*.

Castro Restrepo, D. (1995), 'Comportamiento y rendimiento de plantas de mora (rubus glaucus benth) producidas in vitro en tres sistemas de tutorado.', *Investigaciones-Universidad Católica de Oriente (Colombia)*.

DANE (2015), 'Insumos y factores asociados a la producción agropecuaria', *Boletín mensual del DANE* (33). *https://www.dane.gov.co/files/investigaciones/agropecuario/sipsa/Bol_Insumos31_mar_2015.pdf

Díaz, L. y Morales, M. (2016), Análisis estadístico de datos categóricos.

González, M. (2001), Interacción genotipo por ambiente en guisante proteaginoso (pisum sativum l.), Master's thesis, Universidad de Valladolid.

Gutiérrez Pulido, H. y De la Vara Salazar, R. (2004), *Análisis y diseño de experimentos*, México DF.

Krzanowski, W. J. (1998), *An Introdution to Statistical Modelling*.

Kuehl, R. O. (2001), *Diseño de experimentos. Principios estadísticos de diseño y análisis de investigación*.

Ligarreto, G. A. y Ospina, A. R. (2009), 'Análisis de parámetros heredables asociados al rendimiento y precocidad en arveja voluble (pisum sativum l.) tipo santa isabel', *Agronomía Colombiana* **27**(3), 333–339.

Montgomery, D. C. (2008), *Design and analysis of experiments*.

Nuestro municipio-Información general (2015). *http://www.tota-boyaca.gov.co/informacion_general.shtml

Reyes, Y. y Garcia, D. (n.d.), 'Producción sostenible de arveja en boyacá'. *http://www.corporacionpba.org/portal/novevades/produccion-sostenible-de-arveja-en-boyaca

Sánchez, E. A., y. M. T. (2006), 'Establecimiento de una metodología para la inducción de regenerantes de arveja (pisum sativum) variedad santa isabel.', *Agronomía Colombiana* (24), 17–27.

Vaca, R. (2011), 'Evaluación de tres bioestimulantes con tres dosis en el cultivo de arveja en santa martha de cuba-carchi', *Universidad Técnica del Norte.* .

Vives Brosa, J., . L. V. J. M. (2002), 'El diagnóstico de la sobredispersión en modelos de análisis de datos de recuento'.

Winkelmann, R. (2000), *Econometric Analysis of Count Data*.

DISEÑO Y VALIDACIÓN DE UN CUESTIONARIO PARA MEDIR LAS CARACTERÍSTICAS DEL VISITANTE DEL ANILLO TURÍSTICO DEL LAGO TOTA

Especialización en Estadística

Camilo Ernesto Caicedo Eslava[1,a], Dairo Sigifredo Gil Gil[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Se pretende diseñar y validar un instrumento que permita determinar las características del visitante del Anillo Turístico del Lago de Tota en el departamento de Boyacá Colombia. Se trata de elaborar un cuestionario estructurado de tal manera que permita formularle preguntas pertinentes solo a los visitantes del Anillo que vengan con fines turísticos. La estructura del cuestionario consta de cuatro secciones: la primera va enfocada a darle información al correspondiente sobre los objetivos y el derecho de la privacidad. La segunda permite determinar si el visitante realmente viene con fines turísticos, caso en el cual se le hacen las preguntas pertinentes; de lo contrario solo se le aplica la última sección ya que no se le pregunta a quien no sabe del tema objeto de la investigación, pero tampoco se descarta porque ello conduciría a modificar el tamaño muestral y las consecuentes implicaciones en la significancia práctica y estadística y todo lo concerniente al efecto tamaño y al tamaño del efecto. La tercera sección temática contiene las preguntas objeto de la investigación: Motivacionales, socioeconómicas, sociodemográficas y actitudinales. La cuarta sección contiene todas aquellas preguntas que permitan realizar tablas de clasificación cruzada y complementar la caracterización del visitante.

***Palabras clave*:** Escala tipo Likert, fiabilidad, panel de expertos, validez concurrente, aparente y de constructo.

Abstract

This research pretends to design and to validate a tool that allows determining the characteristics presented by the visitor of Totaťs lake touristic ring in Boyacá state of Colombia. It is proposed to make a structural survey, which lets to ask some relevant questions just for foreign people, who are taking holidays in the Totaťs lake ring. The mean feature of this survey consists of 4 stages: the first one is focused on giving some information to the user about the objectives and the privacy statements. The second part lets the investigator to observe if the visitor is really a tourist, if the answer is positive he or she may continue with the relevant questions. On the contrary, he or she may only answer the last and fourth segment because he or she doesnťt know about the topic of this research. However, this investigation canťt discard the answers due to it could modify the sampling size and the implications given by the practical and statistical meaning and everything about the size effect and effect of the size. The third is called the thematic section that contains the meaning questions of this project: motivational, socio-economical, socio-demographical and attitudinal. The fourth area has the questions that allow making cross classification tables and complementing the visitors profile

***Key words*:** Likert Scale, structural, survey, focused, privacy statements, focused, user, discard, modify, sampling, size effect, profile, concurrent validity.

[a]Estudiante de Especialización en Estadística. E-mail: kmilokicedo117@gmail.com

[b]Profesor asistente. E-mail: dsigifre@gmail.com

1. Introducción

El Lago de Tota es de los parajes de Boyacá con más influencia turística en el país, abarca los Municipios de Aquitania 108 km^2, Cuítiva 25 km^2 y Tota 7.4 km^2, es el segundo en importancia en Sudamérica1, después del Titicaca, cuenta con un área de aproximadamente 60 kilómetros cuadrados y profundidad máxima de 67 metros.

En el Lago, famoso por sus truchas arco iris, se pueden practicar deportes acuáticos. Hay hoteles y restaurantes en el área y es un lugar ideal para realizar paseos y caminatas. En las cercanías del Lago de Tota se presentan fuertes fluctuaciones de temperatura, la cual puede oscilar entre 0^o C y 22^o C. Ubicada sobre los 3.015 m.s.n.m. tiene una temperatura promedio de 12^o C. ocupando una depresión de alta montaña.

El turismo se ha convertido en una de las actividades más dinámicas en las economías actuales debido, entre otros factores, a los cambios en el consumidor turístico. Por ello, para fortalecer las relaciones con los consumidores potenciales y mejorar la prestación de servicios, es necesario conocer y analizar las características de la demanda turística en el lago de Tota. Con este fin, conocer las características de los visitantes de este destino turístico se ha convertido en objeto de estudio de la presente investigación.

Por todo lo anterior se hace necesario diseñar y validar una herramienta que permita medir las características del Visitante del Anillo Turístico del Lago de Tota. Para este estudio se Implementa un Cuestionario estructurado el cual está diseñado para obtener información específica. En él se realizan variedad de preguntas en cuanto a conducta, intensiones, actitudes, conocimiento, motivaciones, características demográficas y de estilos de vida de los visitantes, toda esta información es necesaria a la hora de determinar estas características en el visitante y facilita la obtención de los resultados del estudio.

El objetivo principal de este trabajo es construir un instrumento que permita medir las características del visitante del Anillo Turístico del Lago de Tota, analizando previamente la fiabilidad del instrumento mediante el Alpha de Cronbach (González y D'Ancona 1997*a*), así mismo se realizó un análisis de la validez del instrumento mediante consulta a un panel de expertos los cuales validaron el instrumento e hicieron algunas observaciones para mejorar el mismo las cuales fueron acatadas previamente a la aplicación.

2. Metodología

Para el diseño de cualquier instrumento que se requiera con el fin de medir todo tipo de variables estadísticas se debe basar en cualquier tipo de muestreo, esta investigación se basa en los muestreos estructurales (Gil 2015), estos se caracterizan porque los individuos que componen la muestra son seleccionados en virtud a sus posiciones sociales, situación en una red sociométrica, en una cadena de comunicación, en una jerarquía de dominación, etc.

Es decir, las muestras estructurales tienen como unidad muestral aquellos elementos que están conectados por una relación especifica.

La estructura del cuestionario (Gil 2015) se basa en el diagrama de flujo para un cuestionario estructurado, donde la primera sección es protocolaria allí se explica el principal objetivo y se deja claro la confidencialidad del instrumento, la segunda sección corresponde a las preguntas de introducción las cuales permiten identificar si el entrevistado es el idóneo para brindar la información que se necesita en el estudio, la tercera sección corresponde a las preguntas temáticas las cuales contienen la mayoría de la información objeto de estudio, por último se encuentra la sección de clasificación de donde se extrae la información económica, social y demográfica de los entrevistados, esta sección puede ser molesta para ellos por el tipo de preguntas que se les realiza, por ello se deja para el final del cuestionario.

Las preguntas del cuestionario diseñado son de tipo Likert5 las cuales consisten en un conjunto de ítems bajo la forma de afirmaciones o juicios ante los cuales se solicita la reacción (favorable o desfavorable, positiva o negativa) de los individuos, por ejemplo: Muy de Acuerdo, De Acuerdo, Ni de acuerdo ni en desacuerdo, En desacuerdo, Muy en desacuerdo.

Para corroborar que el instrumento diseñado sea confiable se realiza un análisis de fiabilidad (González y D'Ancona 1997*b*), esta se define como la capacidad de conseguir resultados firmes en mediciones sucesivas de mismo fenómeno, los resultados logrados en mediciones repetidas han de ser iguales para que la medición sea fiable. Existen cuatro métodos que permiten medir la fiabilidad integrando dos conceptos básicos como

son la estabilidad y la consistencia, dichos métodos son el método test-retest, método alternativo, método de las dos mitades, y por último el método utilizado para evaluar la viabilidad del instrumento en esta investigación, **Alpha de Cronbach** el cual permite estimar la fiabilidad de un instrumento de medida a través de un conjunto de ítems que se espera que midan el mismo constructo o dimensión teórica. La validez de un instrumento se refiere al grado en que el instrumento mide aquello que pretende medir. Y la fiabilidad de la consistencia interna del instrumento se puede estimar con el alfa de Cronbach. La medida de la fiabilidad mediante el alfa de Cronbach asume que los ítems (medidos en escala tipo Likert) miden un mismo constructo y que están altamente correlacionados. Cuanto más cerca se encuentre el valor del alfa a 1 mayor es la consistencia interna de los ítems analizados. La fiabilidad de la escala debe obtenerse siempre con los datos de cada muestra para garantizar la medida fiable del constructo en la muestra concreta de investigación. Como criterio general, se sugieren las recomendaciones siguientes para evaluar los coeficientes de alfa de Cronbach:

Coeficiente alfa >.9 es excelente - Coeficiente alfa >.8 es bueno -Coeficiente alfa >.7 es aceptable - Coeficiente alfa >.6 es cuestionable - Coeficiente alfa >.5 es pobre - Coeficiente alfa <.5 es inaceptable.

La fórmula utilizada para hallar el coeficiente de Alpha de Cronbach7 es :

$$\alpha = \frac{\mathrm{k}}{\mathrm{k\text{-}1}} \left[1\text{-}\frac{\sum \mathrm{S_i^2}}{\mathrm{S_T^2}} \right]$$

Donde:

- k : *El número de items.*
- S_i^2**:** *Sumatoria de varianzas de los items.*
- S_T^2**:** *Varianza de la suma de los items.*
- α**:** *Alpha de Cronbach.*

También, antes de confiable, el instrumento debe ser válido según el tipo de investigación que se lleve a cabo, la validez (Escobar-Pérez y Cuervo-Martínez 2008, Gil 2015, González y D'Ancona 1997*b*) debe demostrar que el cuestionario proporcione la información necesaria y adecuada para poder desarrollar el objetivo principal de la investigación, básicamente para medir la validez de un instrumento existen tres modalidades de medición, la validez de criterio, validez de contenido y validez de constructo.

Esta investigación se basa en la **validez de contenido**, la cual consiste en que tan adecuado es el muestreo que hace una prueba del universo de posibles conductas, de acuerdo con lo que se pretende medir; los miembros de dicho universo U pueden denominarse reactivos o ítems, este tipo de validez de contenido es un componente importante de la estimación de la validez de inferencias derivadas de los puntajes de las pruebas, ya que brinda evidencia acerca de la validez de constructo y provee una base para la construcción de formas paralelas de una prueba en la evaluación a gran escala. Este tipo de validez generalmente se evalúa a través de un panel o juicio de expertos, El juicio de expertos es un procedimiento que nace de la necesidad de estimar la validez de contenido de una prueba y este se define como una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en este, y que pueden dar información, evidencia, juicios y valoraciones. La identificación de las personas que formaran parte del juicio de expertos se debe realizar bajo los siguientes criterios de selección: (a) Experiencia en la realización de juicios y toma de decisiones basada en evidencia o experticia (grados, investigaciones, publicaciones, posición, experiencia y premios entre otras), (b) reputación en la comunidad, (c) disponibilidad y motivación para participar, y (d) imparcialidad y cualidades inherentes como confianza en sí mismo y adaptabilidad. También plantean que los expertos pueden estar relacionados por educación similar, entrenamiento, experiencia, entre otros.

Para que los expertos realicen la validez del instrumento se debe seguir los siguientes pasos:

1. Definir el objetivo del juicio de expertos. En este apartado los investigadores deben tener clara la finalidad del juicio, ya que puede utilizarse con diferentes objetivos: (a) Establecer la equivalencia semántica de una prueba que se encuentra validada en otro idioma, (b) evaluar la adaptación cultural, es decir, el objetivo de los jueces es evaluar si los ítems de la prueba miden el mismo constructo en una cultura distinta;

así por ejemplo, los ítems que midan agresividad en una prueba validada en el Tíbet, pueden no estar midiendo lo mismo en Alemania, y (c) validar contenido en una prueba diseñada por un grupo de investigadores.

2. Selección de los jueces. Para ello han de tomarse en cuenta los criterios especificados anteriormente para la selección, considerando la formación académica de los expertos, su experiencia y reconocimiento en la comunidad. Se propone un mínimo de cinco jueces, dos de los cuales deben ser expertos en medición y evaluación, y para el caso de traducciones y adaptaciones de pruebas, se requiere por lo menos un experto en lingüística.

3. Explicitar tanto las dimensiones como los indicadores que está midiendo cada uno de los ítems de la prueba. Esto le permitirá al juez evaluar la relevancia, la suficiencia y la pertinencia del ítem. No hay que dar por sentado que el juez únicamente con la descripción del constructo a medir pueda identificarlo claramente, ya que como se mencionó anteriormente, es posible que existan diferentes definiciones de un mismo constructo.

4. Especificar el objetivo de la prueba. El autor debe proporcionar a los jueces la información relacionada con el uso de la prueba, es decir, para que van a ser utilizados los puntajes obtenidos a partir de esta. Esto aumenta la contextualización del juez respecto a la prueba, incrementando a su vez el nivel de especificidad de la evaluación; ya que la validez de los ítems está directamente relacionada con su utilización, por ejemplo, para hacer un diagnóstico o un tamizaje, o evaluar desempeño, entre otros.

5. Diseño de planillas. La planilla se debe diseñar de acuerdo con los objetivos de la evaluación.

6. Calcular la concordancia entre jueces, se deben establecer los criterios de evaluación y calificación.

Cuando el panel de expertos avala el cuestionario y se aplican las observaciones dadas por ellos, se ajusta el instrumento y se procede a aplicar la prueba piloto.

3. Resultados

Para esta investigación se diseña un Cuestionario Estructurado de 25 preguntas, con preguntas tipo Likert y escalas aditivas bipolares con cinco o más opciones de respuesta de tipo impar dándole la oportunidad al entrevistado de declararse indiferente al ítem que se le pregunta, consta de cuatro secciones:

1. **Protocolo:** Se explica el objeto del instrumento y se asegura la confidencialidad del mismo.

2. **Sección de introducción:** (pregunta 1) Permite clasificar al entrevistado en la intensidad del viaje, si lo hace con fines turísticos o no, esta sección se les aplica a los individuos que realizan el viaje o visita con fines turísticos, de lo contrario se les aplica la sección de clasificación, con el fin de no disminuir el tamaño de la muestra.

3. **Sección Temática:** (preguntas 2-19) Esta sección se divide en tres subsecciones. A) Motivacional: Definir el motivo de la visita al Anillo Turístico del Lago de Tota y evaluación de su experiencia respecto a los servicios utilizados. B) Socioeconómica: Dirigida a definir la composición social del visitante y sus características económicas. C) Sociodemográfica: Define los aspectos sociales y demográficos de los turistas.

4. **Sección de Clasificación:** (preguntas 20-25) Recoge toda aquella información propia de la persona visitante no ligada directamente al turismo y el objetivo principal es realizar las tablas de contingencia que permitan definir la caracterización del visitante del Anillo Turístico del Lago de Tota.

A continuación, se evidencian las Secciones y las preguntas del instrumento con sus objetivos.

I. MOTIVACIONAL.

Tabla 1. Objetivos de las preguntas que clasifican la motivación del visitante.

Pregunta	**OBJETIVO**
2	Determinar si el visitante es habitual o no que visite el anillo turístico del lago de Tota
3	Establecer si el visitante quiere conocer el entorno del anillo turístico
4	Determinar si conoce los alrededores del anillo turístico del lago de Tota
5	Clasificar al visitante según los motivos de su visita.
6	Determinar si el visitante hace turismo habitual o no.
7	Determinar el tiempo promedio de permanencia de la visita
8	Establecer la proporción de grupos que vista el anillo turístico de lago de Tota ”DEMANDA”
9	Determinar el medio de comunicación que lo motivo para su visita
10	Determinar el grado de importancia de permanecer hospedado en el anillo turístico

Tabla 1: Fuente: El autor

II. SOCIOECONÓMICA.

Tabla 2. Objetivos de las preguntas que clasifican el nivel socioeconómico de los encuestados.

Pregunta	**OBJETIVO**
7	Determinar el tiempo promedio de permanencia de la visita
8	Establecer la proporción de grupos que vista el anillo turístico de lago de Tota ”DEMANDA”
10	Determinar el grado de importancia de permanecer hospedado en el anillo turístico
11	Clasificar el tipo de alojamiento que desea el visitante.
12	Establecer los tipos de restaurante que utiliza para su visita
13	Determinar el medio de transporte que utiliza para su visita
14	Establecer la capacidad económica del visitante
15	Establecer la capacidad económica diaria del visitante para los aspectos relacionados con su visita

Tabla 2: Fuente: El autor

III. PERCEPCIÓN.

Tabla 3. Objetivos de las preguntas que clasifican el nivel de percepción de los encuestados.

Pregunta	**OBJETIVO**
16	Clasificar la opinión del visitante sobre los aspectos que tuvo en su visita
17	Percepción del visitante según los aspectos de la vista
18	percepción del visitante según grado de satisfacción de la visita si recomienda o no el anillo turístico
19	Sugerencias del visitante según la percepción de la calidad de los servicios.

Tabla 3: Fuente: El autor

IV. SOCIODEMOGRAFICA.

Esta encuesta quiere determinar los factores sociodemográficos de las personas que visitan el anillo turístico del lago de tota, mediante la clasificación de la encuesta.

FIABILIDAD DEL INSTRUMENTO

El análisis de la fiabilidad se realizó mediante el Alpha de Cronbach para cada una de las secciones mencionadas, obteniéndose un valor de Alpha 0,656 y un poco más.

En síntesis, se puede afirmar que el instrumento tiene alta validez en todos los aspectos, incluyendo la predictiva. Utilizando la prueba piloto se puede hacer una aproximación a la caracterización del visitante del Anillo Turístico del Lago de Tota, ya que se tienen 50 observaciones de una población que es dinámica, tiene mucha variabilidad longitudinal a través del tiempo.

VALIDEZ DEL INSTRUMENTO

A la hora de evaluar un diseño de investigación existen varios criterios a seguir, tal vez el más fundamental es que el diseño se adecue a los objetivos principales de la investigación y mida realmente o que se pretende.

Por lo anterior todo instrumento dirigido a medir características de la población debe ser válido antes de su aplicación definitiva, para lograr validar el documento diseñado en esta investigación se recurre a un panel de expertos bastante calificado el cuál se relaciona a continuación. A cada uno de ellos se les presentó

PANEL DE EXPERTOS	
NOMBRE	**ESPECIALIDAD**
Carmenza Pérez	Gerencia del talento humano, Admón. en Informática Educativa.
Héctor García	PhD en Educación
Edgar Caicedo	Gerencia en Informática Educativa
Ángela Páez Solano	Magister en Literatura.
Esneider Agudelo	PhD en Educación, Sociólogo

Tabla 4: Fuente: El autor.

el instrumento a validar junto con el formato de Evaluación Panel de Expertos donde se encuentran todas las especificaciones de calificación de la Validez del mismo teniendo los siguientes criterios:

1. **Pertinente:** Si corresponde o no al tema y objetivo.
2. **Suficiente:** Si basta para el tema y el objetivo que se pretende evaluar.
3. **Coherente:** Si tiene conexión lógica con el tema y el objetivo.
4. **Relevante:** Si el ítem es importante, si se debe tener en cuenta.
5. **Sintaxis:** Si la ordenación de las palabras y la relación mutua entre las mismas en la construcción de las oraciones es adecuada al objetivo.
6. **Semántica:** Si las palabras empleadas son adecuadas, en cuanto al significado en cada frase del instrumento.

CALIFICACIÓN DEL PANEL DE EXPERTOS

En promedio, el panel de expertos le dio al instrumento una calificación promedio de 4.4 lo que quiere decir que el instrumento se conserva y se deben atender algunas observaciones.

OBSERVACIONES DEL PANEL DE EXPERTOS

1. Los expertos recomiendan ajustes de tipo semántico y sintáctico para algunas preguntas.

2. Anexar criterios faltantes en las preguntas.

3. Ajustar los cuadros al tamaño de la hoja, y acomodar según las márgenes.

4. Conceptos positivos sobre los constructos, lo que afirma que las preguntas miden lo que el instrumento desea evaluar.

5. Los grupos de preguntas son congruentes al objetivo, el instrumento tiene validez concurrente y validez aparente.

PRUEBA PILOTO

La población de estudio para la prueba piloto fueron 50 turistas que se encontraban visitando el Anillo Turístico del Lago de Tota. Donde se miden las actitudes sobre los gustos generales, los aspectos de preferencia en alojamientos, gastronomía, clima, paisaje y otros aspectos de calidad y confort en su estadía, al igual que las características socioeconómicas y socio demográficas.

OBSERVACIONES DE LOS ENCUESTADOS

Al aplicar el instrumento los encuestados proporcionan las siguientes observaciones.

1. Intervalo promedio por entrevista: 11 - 14 minutos

2. Las preguntas 16 y 17 el entrevistado las percibe como equivalentes.

3. Cuando se llega a la sección de percepción se nota fatiga en los entrevistados.

4. En la sección de clasificación hicieron falta opciones de algunas ocupaciones.

5. El lenguaje utilizado en las preguntas fue fácil de entender por parte de los entrevistados.

4. Conclusiones

Al verificar los resultados de la prueba piloto, se evidencia que las preguntas son coherentes con los objetivos, ya que al realizar algunas tablas de contingencia su coeficiente es mayor a 0,7 lo que muestra una moderada asociación entre las variables, si aumentamos el tamaño de la muestra la asociación de las variables aumentará considerablemente al tamaño de la muestra.

Para la validez del instrumento se recurre a un panel de expertos lo cual muestra que es idóneo y cumple con la metodología para ello, es un método de validación de bajo costo lo que es pertinente para esta investigación.

La implementación del cuestionario estructurado se realizó con la finalidad de tener preguntas tipo Likert en su mayoría para poder medir las actitudes de los visitantes del Anillo Turístico del Lago de Tota.

Los encuestados comprendieron fácilmente el contenido de cada una de las preguntas.

El instrumento permitió medir eficientemente el objetivo a evaluar.

5. Discusión de resultados

El cuestionario consta de veinticinco ítems; para la evaluación de contenido, concurrente, aparente y de constructo se recurrió a la opinión de un panel de expertos. Para medir la validez predictiva y la consistencia interna se aplicó un encuesta piloto a n=50 visitantes del Anillo. El panel de expertos hizo algunas observaciones especialmente de forma, quedando el instrumento definitivo compuesto de 24 ítems que fueron los que se probaron con la encuesta piloto.

En síntesis, el tiempo máximo de respuesta por cuestionario fue de diez y ocho minutos durante los cuales se observaron algunas actitudes sintomáticas de posible fatiga y dificultades para comprender algunas de las preguntas. Se revisaron una a una las preguntas del cuestionario para tratar de corregir las dificultades observadas durante la aplicación de la prueba piloto y se envió con carta a cada uno de los miembros del panel de expertos quienes conceptuaron que los constructos apuntaron a los objetivos, que tiene relativamente buena validez aparente, concurrente y de contenido.

Referencias Bibliográficas

Escobar-Pérez, J. y Cuervo-Martínez, A. (2008), 'Validez de contenido y juicio de expertos: una aproximación a su utilización', *Avances en medición* **6**, 27–36.

Gil, D. (2015), *Módulo de muestreo y Diseño de Experimentos*, Escuela de Licenciatura en Matemáticas y Estadística Facultad Seccional Duitama.

González, R. A. y D'Ancona, M. A. C. (1997*a*), 'Metodología cuantitativa. estrategias y técnicas de investigación social'.

González, R. A. y D'Ancona, M. A. C. (1997*b*), 'Metodología cuantitativa. estrategias y técnicas de investigación social'.

Malave, N. (2007), 'Trabajo modelo para enfoques de investigación acción participativa programas nacionales de formación. escala tipo likert', *Caracas: Universidad Politécnica Experimental de Paria. Facultad de Ingeniería. Modalidad MBA* .

APLICACIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA CON RESPUESTA POLITÓMICA NOMINAL EN EL ANÁLISIS DE PREFERENCIAS ALIMENTARIAS DE AVES
El escenario del proyecto fue la Sabana inundable del municipio de Paz de Ariporo Casanare

Especialización en Estadística

Jorge Alberto Chaparro Pesca[1,a], Carmen Helena Cepeda Araque[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El presente estudio expone la aplicación de un modelo de regresión logística multinomial con respuesta politomica, herramienta que se utilizada en investigaciones biológicas, entre otras áreas. A partir de este modelo estadístico se explica la relación que existe entre las preferencias alimentarias de las aves y su variables morfométria y ecológicas, datos tomados del estudio previo *"Preferencias alimentarias de aves asociadas a bosques riparios de sabana inundable en Paz de Ariporo, Casanare"* (Ardila, 2009). Se obtuvo como modelo óptimo el que relaciona las preferencias alimentarias con el alto del pico, ancho del pico, parche y peso. Se estimaron los parámetros mediante el método de máxima verosimilitud, se calculó la calidad del ajuste mediante el coeficiente pseudo-R^2de Mc-Fadden y de Nagelkerke.

***Palabras clave*:** Regresión Logística Multinomial, Preferencias alimentarias de aves.

Abstract

This study me explored the application of multinomial logistic regression model with politomic exits important tool in biological research. From this statistic model, it was explained the relationships that exists between the alimentary preferences of birds and their morphometric and field ecological variables. The mentioned analysis was based on data taken from the previous study *"Alimentary preferences of the birds associated to the riparian forests of the floodable savannah in Paz de Ariporo, Casanare"*(Ardila, 2009). It was obtained as an optimum model relating the alimentary preferences with the high and width of beak, width of the beak, nidification patch and weight. It was estimated The parameters estimation mere through method of maximum likelihood, it was calculated the quality of the adjustment through the coefficient $R^2$and pseudo-R^2 and Nagelkerke.

***Key words*:** Multinomial logistic regression model, Alimentary preferences of birds.

1. Introducción

La regresión logística multinomial es un método estadístico multivariado usado para el análisis de datos, aplicable a variables predictoras discretas o continuas. Este tipo de modelos con variables de tipo nominal

[a]Estudiante de Especialización en Estadística. E-mail: chapis_teto@yahoo.es
[b]Profesor asistente. E-mail: carmen.cepeda@uptc.edu.co

con más de dos categorias(politómicas) es una extension multivariante de la regresión logistica binaria clásica (Fernández y Fernández 2004). El propósito de este artículo es aplicar un modelo de regresión logístico con respuesta politómica nominal para determinar los factores morfométricos que inciden en las preferencias alimentarias de las aves de los bosque riparios de la sabana inundable del municipio Paz de Ariporo - Casanare[1] Las caracteristicas alimentarias se determinaron por las heces que permitieron clasificar las aves en frugívoras, insectívoras, granívoras, frugívoras-insectivoras y omnivoras, en proporción de peso seco. Determinar los factores morfométricos que influyen en las preferencias alimentarias es util ya que a partir de la forma en que las aves utilizan los recursos alimenticios disponibles, se pueden establecer los patrones de uso de hábitat, la conducta de las especies y la dinámica de su población (Avila 2000). Además, en el departamento de Casanare, la información referente a las aves es poca, por tal motivo, la aplicación constituye un gran aporte al conocimiento de la avifauna departamental. A nivel internacional las investigaciones en torno a las preferencias alimentarias de las aves en la región tropical son varias, por ejemplo en los complejos espacios de la Orinoquia, el uso y transformación de los ambientes ha llevado a la disminución de los bosques primarios y al establecimiento y expansión de áreas disturbadas o bosques secundarios en algún grado de sucesión. Dichos bosques se convierten en fuente de recursos estacionales para las aves, en especial a partir del aprovechamiento de los frutos (Loiselle & Blake 1990) o de las flores (Stiles 1979), a su vez que la respuesta de las aves ante dichas condiciones, en especial en las especies polinizadoras y las dispersoras de semillas, son de vital importancia en el funcionamiento y en la recuperación de los sistemas alterados (Stiles 1985). Por lo tanto se han realizado múltiples estudios en el uso de los recursos vegetales por parte de las aves, donde los esfuerzos se han centrado en el análisis de contenidos estomacales (Rosenberg 1990), lavados estomacales (Montalti & Coria 1993), egragópilas o bolos de regurgitación (Pardinas & Cirignoli 2002), observación directa al recurso (De la Peña & Pensiero 2003) o residuos de las materias fecales (Rouges & Blake 2001) como es el caso del presente estudio. El porque se eligieron las muestras de heces radica en que según Loiselle & Blake (1990) este es un método no invasivo efectivo para recolectar información sobre las dietas de las aves capturadas con redes de niebla. Aunque pueden presentarse sesgos relacionados con el paso por el aparato digestivo, se ha observado una fuerte correspondencia entre heces y contenidos estomacales, sin encontrarse sesgos para ítems pequeños o de cuerpo blando (Ralph 1985).

2. Referente Conceptual

Para estudiar varias mediciones simultáneamente es útil un modelo matemático para explicar las observaciones y sus relaciones. El modelamiento es la aplicación de una serie de pasos tales como estimación, juzgamiento de hipótesis, evaluación y la selección del modelo para conseguir una explicación apropiada del comportamiento de una variable respuesta (datos) a partir de una función ponderada de una o más variables explicativas modelo.

De acuerdo a (Díaz Monroy y Morales Rivera 2012, pág.135) entre los datos y el modelo existe una discrepancia que se denomina error residual, es decir Datos = modelo + error. Uno de los ejemplos clásicos de modelos estadísticos es la regresión lineal múltiple en la cual una variable respuesta Y_i es "explicada" a través de las variables X_1, X_2, ..., X_P mediante el modelo estadístico:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i \tag{1}$$

El propósito del modelamiento estadístico es la "búsqueda del modelo más simple que sea capaz de explicar los datos con el mínimo error posible". Esto equivale a buscar un modelo parsimonioso que se ajuste adecuadamente a los datos (Díaz Monroy y Morales Rivera 2012, pág.136).

Un modelo lineal generalizado se origina cuando interesa modelar un fenómeno en el cual la variable respuesta Y tiene una distribución que pertenece a la familia exponencial de densidades y está asociada a un conjunto de variables explicativas $X_1, X_2, ..., X_P$. En esta clase de modelos se distinguen tres componentes, el aleatorio, sistemático y la función de enlace.

[1]Los datos proceden del estudio Preferencias alimentarias de aves asociadas a bosques riparios de sabana inundable en Paz de Ariporo, Casanare realizado por Jennifer Ardila en el 2009 como trabajo de grado en el programa de biología, el cual fue dirigido por el doctor Oscar Rodríguez Fandiño, Biologo de la Fundación Universitaria Internacional del Tropico Americano.

Componente aleatorio: Está representado por el conjunto de variables respuesta independientes Y_i, $i = 1, 2, ..., n$ cuya distribución para todo i pertenece a la familia exponencial.

Componente sistemático: Está representada por el conjunto de variables explicativas $X_1, X_2, ..., X_P$ y una relación de la forma $\eta_s = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$ equivalente a $\eta = X\beta$ con $s = 1, 2, ..., n$, donde β es el vector de parámetros y X es la matriz del modelo que puede estar asociada a un modelo de regresión múltiple de rango completo si las X_i, son variables cuantitativas o a un modelo de regresión de rango incompleto si el diseño es por ejemplo un diseño factorial, un diseño de bloques u otro diseño experimental con las X_i variables categóricas o de clasificación.

Función de enlace: Es una función g monótona, derivable que asocia o enlaza la componente aleatoria y sistemática por la relación: $g(u_i) = \eta_s$

2.1. Modelo de respuesta multinomial

Hosmer & Lemeshow (2013) (citado por Pando & San Martín, 2004), indican que la regresión logística multinomial es utilizada en modelos con variables dependientes de tipo nominal con más de dos categorías (politómicas) y es una extensión multivariante de la regresión logística binaria clásica.

Se Considerar una variable de respuesta politómica Y con más de dos categorías de respuesta que denotaremos por $Y_1, Y_2, ..., Y_K$.

Se pretende explicar la probabilidad de cada categoría de respuesta en función de un conjunto de covariables $X = [X_1, X_2, ..., X_p]$ observadas. Es decir, ajustar un modelo de la forma $P_j(x) = P(Y = Y_j/X = x) = f_j(x)\ \forall\ j = 1, 2, ..., k$. Para cada vector X de valores observados de las variables explicativas.

Así que para obtener un modelo lineal, obtendremos $\binom{k}{r}$ transformaciones "logit" para comparar cada par de categorías de la variable respuesta, que sería de este tipo.

$$ln\left[\frac{\frac{p_i(x)}{p_i(x)+p_j(x)}}{\frac{p_j(x)}{p_i(x)+p_j(x)}}\right] = \left[\frac{p_i(x)}{p_j(x)}\right], \forall i, j = 1, 2, ..., k(i \neq j) \tag{2}$$

Las cuales representan el logaritmo "odds" de respuesta Y_i frente a Y_j condicionado a las observaciones de las variables independientes que caen en uno de ambos niveles. Para las probabilidades de respuesta, podemos escribir el modelo de la siguiente forma $p_j(x) = \frac{exp(\eta_s)}{1+\sum_{j=1}^{k=1} exp(\eta_s)}$, $s = 1, 2, ..., n$. La interpretación de los parámetros del modelo, depende del tipo de variables explicativas, cuantitativas o cualitativas.

Más de una variable predictora cuantitativa: Para el modelo multinomial múltiple, la razón de odds se definen incrementando una de las variables y controlando fijas las demás.

$$\theta_j(\Delta X_r = \frac{1}{X_s} = x_s, s \neq r) = \frac{\frac{P[Y=\frac{Y_j}{X_r}=x_r+1, X_s=x_s\neq r]}{P[Y=\frac{Y_k}{X_r}=x_r+1, X_s=x_s\neq r]}}{\frac{P[Y=\frac{Y_j}{X_r}=x_r+1, X_s=x_s\neq r]}{P[Y=\frac{Y_j}{X_r}=x_r+1, X_s=x_s\neq r]}} = exp(b_{rj})\forall j = 1, 2, ..., k-1 \tag{3}$$

Siendo $\theta_j(\Delta X_r = \frac{1}{X_s} = x_s, s \neq r)$ el cociente de los odds de respuesta Y_j frente a la última categoría, Y_k, cuando aumenta en una unidad la variable X_r, y las demás se mantienen fijas.

Para la estimación de los coeficientes del modelo y de sus errores estándar se utiliza la estimación por máxima verosimilitud. Supongamos que disponemos de una muestra aleatoria de tamaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, X_2, ..., X_n$. Denotemos a cada combinación de valores de las variables explicativas por $x_q = (, x_q1, ., x_qn)^T$ con $x_{q0} = 1\ \forall\ q = 1, 2, ..., Q$. En cada una de estas combinaciones se tiene una muestra aleatoria de d_q observaciones independientes de la variable de respuesta politomica Y, de entre las cuales denotamos por $Y_{j/q}$ al número de observaciones que caen en la categoría de respuesta $Y_j\ \forall\ j = 1, 2, ..., k$. Así que se verifica que, $\sum_{j=1}^{k} Y_{j/q} = d_q = N$. Los vectores $(y_(1/q), ..., y_(k/q))^T$ $\forall\ q = 1, ..., Q$ siguen una distribución de probabilidad multinomiales $M(d_q; p_{1/q}, , p_{k/q})$ siendo $p_{j/q} = P[\frac{Y/Y_j}{X/x_q}]$ y verificando que $\sum_{q=1}^{k} y_{j/q} = 1$

Por tanto, la función de verosimilitud de los datos viene dada por $V = \prod_{q=1}^{Q}\left[\frac{d_q!}{\prod_{j=1}^{k}(y_{j/q})}\prod_{j=1}^{k}(P_{j/q}^{y_{j/q}})\right]$. Para obtener los estimadores de máxima verosimilitud hay que resolver $k-1$ sistemas de p ecuaciones no lineales de la forma $\frac{\Delta K}{b_{sj}} = \sum_{q=1}^{Q} y_{jq}x_{qs} - \sum_{q=1}^{Q} n_q x_{qs}\frac{exp\sum_{s=0}^{n} b_{sj}x_{qs}}{\sum_{j=1}^{k} exp\sum_{s=0}^{n} b_{sj}x_{qs}}$. Así que para resolverlo se utiliza el método iterativo de Newton-Raphson, con este método obtenemos el estimador de los parámetros b, que es una matriz de dimensión $(p)(k-1)$. la matriz de covarianzas de b, que es la inversa de la matriz de información de Fisher, dada por $cov(\hat{b_j}) = \left[-E\left(\frac{\Delta^2}{\Delta b_{rj}\Delta b_{sj}}\right)\right]^{-1} = \left[X'Diag[d_p p_{j/q}(1-p_{j/k})]X\right]^{-1}$.

El sistema de hipótesis para juzgar la bondad de ajuste global del modelo viene dado por

$$H_0 : p_{j/q} = \frac{exp(\sum_{s=0}^{n} b_{sj}x_{qs})}{1+exp(\sum_{s=0}^{n} b_{sj}x_{qs})} \forall q = 1,2,...,Q; \forall j = 1,2,...,k \tag{4}$$

$$H_1 : p_{j/q} \neq \frac{exp(\sum_{s=0}^{n} b_{sj}x_{qs})}{1+exp(\sum_{s=0}^{n} b_{sj}x_{qs})} \text{ para algún } q \text{ y } j \tag{5}$$

El cual se puede juzgar con *Test chi-cuadrado de Pearson*, estadístico tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre el número de parámetros $p_{j/q}$ y el número de parámetros independientes en el modelo, $q - p*(k-1)$. Es decir, $\chi^2(M)_{dp\to\infty}^{d\to\chi^2_{q-p*(k-1)}}$. Así que se rechaza la hipótesis nula con un nivel de significancia α, cuando $\chi^2(M)_{obs} \geq \chi^2_{q-p*(k-1)}$, o equivalentemente podemos definir el $p-valor$ del contraste como la probabilidad acumulada a la derecha del valor observado: $p - valor = P[\chi^2(M) \geq \chi^2(M)_{obs}]$, se rechaza la hipótesis nula cuando $p - valor \leq \alpha$.

El estadístico Wilks de razón de verosimilitudes: Para el contraste de bondad de ajuste del modelo de regresión logística multinomial M corresponde a $G^2(M) = 2\left[\sum_{q=1}^{Q}\sum_{j=1}^{k} y_{j/q} ln\left(\frac{y_{j/q}}{\hat{m}_{j/q}}\right)\right]$. Este estadístico tiene distribución asintonica chi-cuadrado con grados de libertad corresponde a la diferencia entre la dimensión del espacio paramétrico y la dimensión de este espacio bajo la hipótesis nula. Para un modelo de regresión logística multinomial los grados de libertad es la diferencia entre el número de parámetros $p_{j/q}$ y el número de parámetros b_{sj} bajo el modelo, es decir, $q-p*(k-1)$ grados de libertad, $G^2(M)_{dp\to\infty}^{d\to\chi^2_{q-p*(k-1)}}$. Así que se rechaza la hipótesis nula con un nivel de significación a cuando $G^2(M)_{obs} \geq \chi^2_{Q-(n+1)*(k-1);\alpha}$. O equivalentemente cuando $p - valor = P[G^2(M) \geq G^2(M)_{obs}] \leq \alpha$. Al estadístico de Wilks, $G^2(M)$, se le denomina devianza.

Tasa de clasificaciones correctas: Para cuantificar la bondad del ajuste global del modelo se dispone también de otra medida como es la tasa de clasificaciones correctas. Es decir, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados-predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción, del mismo modo que se hace en el análisis discriminante. Se define como la proporción de individuos clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta, Y_i, coincide con su valor estimado por el modelo.

Calidad del ajuste: Para medir la calidad de ajuste los más utilizados en regresión logística multinomial son R^2 del Mc-Fadden y Nagelkerke. Plantea que si tenemos $\Delta = -2ln?(V)$, se identifica que Δ_0 el valor inicial de la función, es decir el mismo Δ bajo el modelo ajustado con todos los parámetros, obtendremos la siguiente expresión del pseudo-R^2 de Mc-Fadden dado por $R^2_{MF} = 1 - \frac{\Delta_f}{\Delta_0}$ conociendo que los valores deben estar comprendidos entre $0 \leq R^2_{MF} \leq 1$ es muy raro que el valor se aproxime a 1. Es en buen ajuste cuando el valor esta comprendido entre $0.2 \leq R^2_{MF} \leq 0.4$ y excelente para valores superiores. De igual manera se utiliza el coeficiente de pseudo-R^2de Nagelkerke que esta dado por $R^2_N = \frac{R^2_{cs}}{1-V_0^{2/n}} = \frac{1-exp(\frac{\Delta_f-\Delta_0}{N})}{1-exp(\frac{\Delta_0}{N})}$. El rango debe estar comprendido entre $0 \leq R^2_N \leq 1$, su interpretación es igual al coeficiente de determinación de la regresión lineal clásica, pero es mucho mas difícil que alcance valores muy cercanos a 1. Es decir que para compararlo con el modelo de regresión logistica politomica con diferentes números de variables predictoras suele introducirse coeficientes pseudo-R^2de Mc-Fadden, dado por $Adj = R^2_{MF} = 1 - \frac{0.5\Delta_f+n+1}{0.5\Delta_f+n}$, siendo n el número de variables predictoras.

2.2. Contraste sobre los parámetros del modelo

Para construir el modelo , ajustarlo y estimarlo debemos comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo. Para ello se pueden emplear básicamente dos métodos: el estadístico de Wald y el estadístico condicional de razón de verosimilitud. Así que plantea contrastar si un subconjunto de los parámetros del modelo de regresión logística multinomial, que denotaremos por $b = (b_1, b_2, ..., b_r)$, es nulo. El sistema de hipótesis corresponde a:

$$H_0 : b = 0 \tag{6}$$

$$H_1 : b \neq 0 \tag{7}$$

El cual se puede juzgar a través del estadístico de Wald y el de razón de verosimilitud.

Contraste de Wald: Se basan en la normalidad asintótica de los estimadores de máxima verosimilitud. Así que se rechaza la hipótesis nula al nivel de significación α cuando el valor observado de este estadístico sea mayor o igual que el cuantil de orden $(1 - \alpha)$ de la distribución χ^2_r. Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. Es decir si se quiere contrastar:

$$H_0 : b_{sj} = 0 \tag{8}$$

$$H_1 : b_{sj} \neq 0 \tag{9}$$

el estadístico será $W = \frac{\hat{b}^2_{sj}}{\hat{\sigma}^2(\hat{b}_{sj})}$, que tiene distribución chi-cuadrado asintótica con un grado de libertad. Así que se rechaza la hipótesis nula con nivel de confianza $(1 - \alpha)$, si $W_{obs} \geq \chi^2_1; \alpha$. Es decir, la obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación ademas no es recomendable su uso si se están empleando variables de diseño, en estos casos se recomienda el uso del test de razón de verosimilitudes.

Contrastes condicionales de razón de verosimilitud: Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo. Supongamos que tenemos un modelo de regresión logística multinomial Mg que se ajusta bien y se desea contrastar si un subconjunto de parámetros, $b = (b_1, b_2, ..., b_r)$, son nulos. Sea M_p el modelo con ese subconjunto de parámetros ceros. Así que M_p está anidado en el modelo general M_G. Así que planteamos el contraste:

$$H_0 : b = 0 \ (M_p \text{ se verifica}) \tag{10}$$

$$H_1 : b \neq 0 \ (\text{asumiendo cierto } M_G) \tag{11}$$

Si asumimos que M_G se verifica, el estadístico del test de razón de verosimilitudes para contrastar si M_p se verifica es $G^2(\frac{M_p}{M_G}) = -2(L_p - L_G) = G^2(M_p) - G^2(M_G)$, siendo L_p y L_G, siendo los máximos de la log-verosimilitud bajo la suposición de que se verifican los modelos saturados, M_p y M_G, respectivamente. Es decir, el test de razón de verosimilitudes para contrastar dos modelos anidados es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo.

$$G^2(\frac{M_p}{M_G}) = -2(L_p - L_G) = G^2(M_p) - G^2(M_G), \ L_p \text{ y } L_G \tag{12}$$

Así que se rechaza la hipótesis nula al nivel de significación α cuando $G^2_{obs}(\frac{M_p}{M_G}) \geq \chi^2_1; \alpha$.

Selección del modelo: Una vez conocido el procedimiento de ajuste de modelos de regresión logística multinomial, el siguiente paso es el desarrollo de estrategias para seleccionar las variables que mejor explican a la variable de respuesta. Para ello se adoptará el principio de parsimonia que consiste en seleccionar el modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla en términos de cocientes de ventajas. Siempre que se incluya o se excluya una de estas variables, todas las demás categorías deben ser incluidas o excluidas en bloque. Si no se tiene en cuenta esta consideración, implicaría que se habría recodificado la variable y por tanto, la interpretación de la misma no sería la correcta. Además hay que tener en cuenta la significación que pudiera tener cada variable dummy. No siempre todas las categorías de una covariable son significativas, o todas no significativas. Por lo que, cuando ocurra esta situación es recomendable contrastar el modelo completo frente al modelo sin la covariable mediante la prueba de razón de verosimilitud, decidiendo incluir o excluir la covariable dependiendo del resultado de la prueba y del interés de la covariable(Rodríguez 2013).

La evaluación del modelo puede hacerse a través de los residuales de pearson que esta dado por $r_{j/q} = \frac{y_{j/q} - d_q \hat{p}_{j/q}}{[d_q \hat{p}_{j/q}]^{1/2}}$. Con esta expresión, podemos definir el estadístico chi-cuadrado de Pearson como

$\chi^2 = \sum_{q=1}^{Q} \sum_{j=1}^{k} r^2_{j/q} r_{j/q} =$. Para contrastar la significación estadística de los residuos planteamos el las siguientes hipótesis:

$$H_0 : r_{j/q} = 0 \tag{13}$$

$$H_1 : r_{j/q} \neq 0 \tag{14}$$

Bajo la hipótesis nula $r_{j/q}$ tiene una distribución asintótica normal con media cero y varianza estimada $\hat{\sigma}^2(r_{j/q}) < 1$, es decir que los residuos tienen menor variabilidad que una variable aleatoria estándar, pero suelen ser tratados como normales estándar, considerándose significativos cuando sus valores absolutos son mayores que dos (falta de ajuste). Para evitar este problema se definen los residuos de Pearson ajustados que presentan distribuciones asintóticas normales estándar y vienen dados por $r^2_{j/q} = \frac{r_q}{\hat{\sigma}^2(r_{j/q})}$.

3. Metodología

La investigación privilegió un enfoque cuantitativo descriptivo, el cual nos permitió establecer la relacion de las preferencias alimentarias con sus características morfométricas a través de un modelo de regresión logística politómica concretamente el interés se centra en la construcción del modelo $p_j(x) = \frac{exp(\eta_s)}{1+\sum_{j=1}^{k=1} exp(\eta_s)}$, $s = 1, 2, ..., n$; en donde Y corresponde a la preferencia alimentaria y X a la matriz de las variables morfométricas de las aves. La base de datos que se analizó para este trabajo es tomada de una investigación que se realizó en la reserva natural la esperanza ubicada al Oeste, perteneciente a la Vereda Caño Chiquito del Municipio de Paz de Ariporo, Casanare[2]En la base de datos se conto con un total de 251 aves capturadas con sus características taxonómicas, familia, genero, especie.(Ardila Ayala 2009). Las características morfométricas estudiadas fueron:

Peso: Medida de la masa corporal en gramos, tomada para cada ejemplar por medio de un dinamometro o una balanza.

Sexo del espécimen: Se determina a partir de patrones de coloración del plumaje aunque muchas especies no presentan dimorfismos sexuales de coloración. Para determinar con certeza el sexo de un ejemplar es necesario mirar las gónadas lo que se hace al preparar una piel de estudio. También se pueden utilizar técnicas más avanzadas como la laparotomía. Puede ser a) macho; b) hembra; c) desconocido.

Largo del pico: Longitud del pico o culmen total se mide desde el comienzo de la parte córnea del pico en la parte frontal del cráneo, en línea recta hasta su punta.

[2]La reserva se encuentra localizada en lo que se denomina sabana inundable o hiperestacional de los departamentos de Arauca y Casanare, el área se encuentra aproximadamente a los 120 m.s.n.m y según el sistema de Holdridge se puede clasificar dentro del clima cálido húmedo o cálido seco.

Alto del pico: La altura del pico se mide desde la parte inferior de la mandíbula hasta la parte superior de la maxila a nivel de las narinas.

Ancho del pico: La anchura del pico o rictus ("gape") se mide la distancia entre las comisuras de la boca o pico (es medir la sonrisa del ave).

Longitud del tarzo: La longitud del tarso se mide desde la parte inferior al comienzo del tarso, antes de la saliente ósea parecida al tobillo, hasta la parte frontal de la última escama completa que da la vuelta al tarso, justo antes del comienzo de la mano y dedos.

Longitud del ala: Se mide desde la "muñeca" (es decir, donde nacen las plumas primarias y se detecta una pequeña saliente) hasta la punta de la pluma primaria más larga con el ala cerrada.

Longitud de la Cola: Se mide desde el nacimiento de las dos plumas centrales de la cola, justo debajo de la glándula uropigial, hasta la punta de la pluma rectriz más larga con la cola cerrada.

Parche: Es el estado reproductivo del ejemplar determinado a partir de la presencia o ausencia del parche de incubación. Este atributo puede tomar tres valores: a) parche de incubación presente: se evidencia por la ausencia de plumas en el abdomen e incluso en la parte central del pecho, justo en la quilla y sus alrededores. En muchas ocasiones, también se observa un aumento de tamaño de las venas de la región abdominal y un engrosamiento de la piel, el vientre puede presentar también una bolsa llena de líquido fluido; b) parche de incubación aparente: se presentan sólo algunas de éstas características o la piel en el abdomen se encuentra arrugada y retraída; y c) parche de incubación ausente: cuando la piel del abdomen y el vientre no muestran ninguna de estas características.

Grasa incuba: Es la cantidad de grasa en la fúrcula y flancos, medida relativa de la abundancia de grasa subcutánea a nivel de la fúrcula y los flancos del ejemplar. Este atributo puede tomar valores de abundancia entre 0 y 5. Se dice que la grasa es 0 cuando no está presente; 1, cuando hay grasa visible solamente en la base de unión de las fúrculas; 2, cuando además presenta líneas delgadas de grasa a lo largo de ambas fúrculas; 3, se presentan también trazas de grasa en la parte media sin ser abultada; 4, cuando la grasa cubre totalmente el área entre las fúrculas y está abultada y 5, cuando además presenta grasa a nivel de los flancos.

Muda del plumaje: Se considera que existe muda cuando se están desarrollando nuevas plumas y por lo tanto es observable la presencia de cañones o remanentes de cubierta quitinosa que envaina la pluma mientras aún se extiende completamente. Este atributo puede tener diferentes valores de acuerdo con su presencia y ubicación: a) ausente: cuando no se observa muda en ninguna pluma; b) cuerpo: cuando se presenta muda en cualquier parte diferente de alas y cola; c) alas: es necesario que la muda sea pareada, es decir que tanto las plumas del ala izquierda como las de la derecha la presenten; d) cola: al igual que en las alas es necesario que sea pareada, es decir, que se presente en las plumas del lado derecho e izquierdo de la cola. e) accidental: cuando se presenta solo una pluma en muda en alas y/o cola sin que sea pareada, esta se considera accidental y debe anotarse en qué parte se presentó. Estado del plumaje: Medida subjetiva acerca del desgaste del plumaje. Puede tomar tres valores: a) fresco: plumaje brillante y sin muescas, ni partes en mal estado; b) gastado: plumaje que se observa opaco, incluso con coloración dispareja y plumas con bordes desgatados o muescas; c) regular: categoría intermedia entre las dos anteriores, a medida que el plumaje se desgasta y comienza a perder brillo y opacarse debido a la abrasión con el medio y/o acción de bacterias y ectoparásitos.(Mauricio Álvarez 2004, pág.108)

Las preferencias alimentarias son: *Frugívoras*: aves cuyas heces están compuestas principalmente por restos de materiales vegetales. *Insectívoras*: aves cuyas heces están compuestas principalmente por insectos. *Granívoras*: aves cuyas heces están compuestas principalmente por semillas. *Frugívoras-insectívoras*: aves cuyas heces están compuestas principalmente por restos de material vegetal e insectos. *Omnivoras*: aves cuyas heces están compuestas principalmente por resto de material vegetal, animal, semillas, insectos y material sin identificar.

La información registrada se analizó por medio del programa estadístico(R Core Team 2016)[3].

[3]Equipo Central R (2016). R : Un lenguaje y entorno de estadística informática. R Fundación para la Computación de Estadística, Viena, Austria. URL https:/www.R-project.org/.

4. Resultados

Variable	Unidades/valores que toma/codificación	Descriptivo
Familia	T=Tyranndae; Th=Thraupidae; C=Columbidae; Ta=Thamnophilidae; G=Galbulidae; Tu=Turdidae; O=Otras	T=111 (41.73 %); Th=67 (25 %); C=20(7.46 %); Ta=13(4.85 %); G=8(2.98 %); Tu=8(2.98 %); 0=40(14.92 %)
Género	E=Empidonax; El=Elaenia; C=Columbina; R=Ramphocelus; T=Tangara ; O=Otras	E=40(14.92 %); El=35(13.05 %); C=20(7.46 %); R=20(11.56 %) ; T=11(4.10 %); O=11(4.47 %)
Especie	S.p; Ca=Carbo; P=Parvirostris; Cy=Cayana; Ep=Episcopus; Ru=Ruficaudo; O=Otras	S.p=42(15.67 %); Ca=29(25 %); P=28(10.82 %); Cy=17(6.34 %); Ep=9(3.35 %); Ru=8(52.98 %); O=133(4.47 %)
Peso	Gramos	Media=23.15; Mediana=22 Min=7; Max=81; No pesados=15; Cv= 0.5238 ;As= 1.6430; K=3.3597
Sexo del Especimen	O=No identificados; M=Machos; H=Hembra	O=226(84.96 %); M=19(7.14 %) ; H=21(7.89 %)
Largo del Pico	milimetros	Media=14.38; Mediana=12.65; Min=5.90; Max=53.80; Cv= 0.5529;As= 3.3597; K=12.7387
Alto del Pico	milímetros	Media=6.265; Mediana=5.200; Min=3.000; Max=64.000; Cv= 0.8603;As=7.3067; K=66.4445
Ancho del Pico	milímetros	Media=11.35; Mediana=10.50; Min=7.60; Max=95; No medidos=2; Cv= 0.6686 ;As=9.21; K=95.7134
Longitud del Tarso	milímetros	Media=20.77; Mediana=20.65; Min=3.90; Max=38.10; Cv= 0.2661;As=0.2665; K=2.2681
Longitud del Ala	milímetros	Media=76.46; Mediana=74; Min=0.00; Max=180.00; No medidos=1; Cv= 0.2144;As=0.1541; K=2.2681
longitud de la Cola	milímetros	Media=66.56; Mediana=64.00; Min=0; Max=147.00; No medidos=1; Cv=0.2917;As=0.6291; K=2.6578
~~Parche~~	O=Ausencia; 1=Presencia; No=No identificado	O=163(61.56 %); 1=102(38.49 %); 2=3(1.11 %
Grasa Incuba	O=Ninguno; 1=poco; 2=Bajo; 3=Medio; 4=Alto; 5=Cobertura Total; 6= No identificados	O=78 (29.43 %); 1=52 (19.52 %); 2=56(21.13 %); 3=57(21.51 %); 4=16(6.04 %); 5=7(0.03 %)
Muda del Plumaje	O=Ausencia; CAT=Cuerpo+Alas+Timonera; C=Cuerpo; T=Cola(Timonera)	O=138(52.08 %); CAT=75(23.77 %); C=13(11.71 %); T=14(4.10 %)
Estado del plumaje	F=Fresco; G=Gastado; R=Regular	F=137(52.09 %); G=113(42.97 %); R=13(4.94 %)
Preferencias Alimentarias	Fru=Frugívoro; Ins=Insectívoro; Fru-Ins=Frugívoro-Insectívoro; Gra=Granívoro Om=Omnívoro	Fru=102 (38.35 %); Ins=80 (30.08 %); Fru-Ins=56(21.05 %); Gra=19(7.14 %); Om=9(3.85 %)

Tabla 1: Descripción de las variables estudiadas.

En la Tabla 1 se describe la distribución univariada de las variables consideradas en el estudio, las unidades en las que se miden o los valores codificados que toman. La mayoria de las variables cualitativas estan codificadas númericamente, para tratarlas en el estudio del modelo de regresión logistica multinomial.

En la reserva natural la Esperanza en el municipio de Paz de Ariporo, Casanare predominaron las aves de las familia tyrannidae 41.73 % y Thraupidae 25 % lo anterior es fácil de explicarse si tenemos encuenta que ambas familias se caracterizan por presentar una gran cantidad de especies que esta asociadas a los bosques riparios que se presenta en la orinoquia. Razón por la cual al 84.9 % no se les pudo determinar el sexo por su poca edad y tambien por tener dimorfismo sexual (no tiene organos sexuales visibles), se tiene que el 62 % hay ausencia de parche aspecto que esta asociado al ciclo de reproducciòn, no se encontraba en esa epoca y 52 % con plumaje fresco esto debido a que la probabilidad de cambio de plumaje durante todo el año es la misma. A partir de los coeficientes de variación se determina que las medidas de longitudes como peso, ancho, largo y alto del pico, que vienen determinadas más por la especie y familia a la que pertenece el ave, la variabilidad relativa es muy grande. Es decir, las aves presentan bastante heterogeneidad en estas variables. Caso contrario pasa en el tarso, ala y cola. Lo anterior es consistente y se corresponde con la distribución asimétrica o simétrica de las variables. Los coeficientes de kurtosis son altos para el alto, largo y ancho del pico, lo que puede estar indicando la presencia de datos atípicos. En las variables peso, tarso, ala y cola se descarta presencia de este tipo de datos.

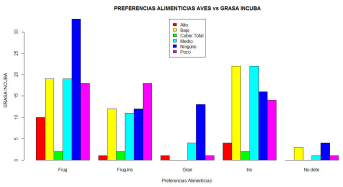
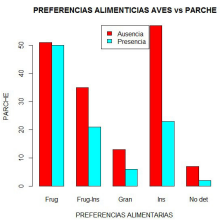
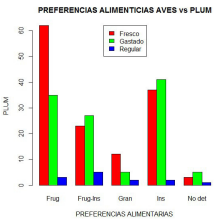
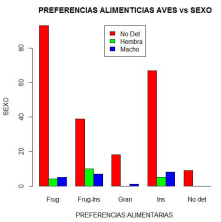
		Frug	Frug-Inc	Gran	Insec	Omn	χ^2	P-valor	Gráfica
Grasa incuba	Alto	10	1	1	4	0	36.308	0.01415	
	Bajo	19	12	0	22	3			
	Cober total	2	2	0	2	0			
	Medio	19	11	4	22	0			
	Ninguno	33	12	13	16	4			
	Poco	18	18	1	14	1			
Parche	Ausencia	51	35	13	57	50	9.7942	0.04404	
	Presencia	50	21	6	23	2			
Plumaje	Fresco	62	21	9	3	37	14.906	0.06101	
	Castado	35	27	5	41	5			
	Regular	3	5	2	2	1			
Sexo	No det	93	3	18	67	9	19.036	0.01467	
	Hembra	4	10	0	5	0			
	Macho	5	7	1	8	0			

Tabla 2: Relación entre las preferencias alimentarias ylas variables cualitativas

En la Tabla 2 se observa la relación entre las variables cualitativas, usamos la tabla de doble entrada, que se denomina tabla de contingencia, la cual presenta en cada casilla las frecuencias absolutas o porcentajes de una de las categorías de una variable con una categoría de la otra variable. Para evaluar el grado de relación y el nivel de significación estadística entre dos variables categóricas se utiliza el test de ji-Cuadrado.

Se observa que existe asociación estadística entre las variables grasa, parche, pluma, sexo y las preferencias alimentarias. Con el fin de establecer si existen diferencias significativas respecto a las preferencias alimentarias y las variables cuantitativas se aplicó el test de kruskal-wallis debido a que luego de aplicar el test de normalidad de shapiro wilk se determinó que el supuesto se cumple.

	frug	Frug-Ins	Gran	Ins	Omn	χ^2	P
Alto del Pico	4.5	7.2	4.1	5.5	7.4	23.439	0.0001035
Longitud del Ala	74.0	76.0	76.0	72.5	80.0	11.557	0.02097
Ancho del Pico	10.35	12.00	6.60	10.55	12.10	47.94	$9.712\times^{-10}$
Longitud de la Cola	64.0	71.5	53.0	61.5	75.0	11.215	0.02425
Largo del Pico	10.5	13.4	11.9	14.0	13.0	22.815	0.0001379
Peso	16	22	36	21	27	29.469	$6.277\times^{-6}$
Longitud del Tarso	20.40	22.25	17.20	21.6	16.1	31.523	$2.394\times^{-6}$

Tabla 3: Relación entre las preferencias alimentarias de las aves y las variables cuantitativas

A partir de los resultados de la Tabla 3 se determina que existe diferencia significativa en todas las variables. Por lo que podemos afirmar que existen diferencias en la media de las preferencias respecto a Alto del Pico, Longitud del Ala, Ancho del Pico, Longitud de la Cola; Largo del Pico, Peso, Longitud del Tarso nos explica la relación que existe entre lo que comen y ecoevolución de las aves.

FACTORES ASOCIADOS A LAS PREFERENCIAS ALIMENTARIAS

El modelo inicial contempló las variables Peso, Sexo del espécimen, Largo del pico, Alto del pico, Ancho del pico, Longitud del Tarso, Longitud del Ala, Longitud de la Cola, Parche, Grasa Incuba, Muda del Plumaje, Estado del plumaje. Para la selección del modelo se utilizó el método hacia adelante comenzando con el modelo inicial que contiene sólo la constante, en cada paso se analizó la inclusión o no de alguna de las variables mediante contrastes de razón de verosimilitudes, se encontró que el modelo óptimo queda determinado por Alto del pico, Ancho del Pico, Parche y Peso. A continuación en la Tabla 4 se presenta el resumen del modelo.

Coeficiente	(Intercepto)	Alto del Pico	Ancho del Pico	Parche(T.Presencia)	Peso
Frug.-Insec	-14.274.872	0.1418912	-0.004955250	-0.4737023	0.01111216
Granívoros	38.923.934	-0.1041174	-0.907058013	-0.7768076	0.12168586
Insectívoro	-0.3270395	0.1268980	0.007662909	-0.7914572	-0.01817625
Omnívoro	-21.859.234	0.1378881	-0.203958229	-15.614.018	0.00595263
Std	Errores:				
Frug-Insec	0.5591492	0.07999382	0.02952459	0.3589122	0.01784538
Granívoros	16.155.397	0.23984523	0.20030544	0.8792831	0.03238572
Insectívoro	0.4890901	0.07961102	0.01852520	0.3247021	0.01850331
Omnívoro	17.375.734	0.09002601	0.17237122	11.141.157	0.02819228
Residual	Deviance	5.745.282			
AIC	6.147.282				

Tabla 4: Resumen del modelo ajustado

Preferencias Alimentarias	Var. Independientes	Est. de coeficientes	(E.E)Errores Estandar	Test de Wald	OR	IC 95 % OR	P-v
Frug.-Insec.	Intercepto	-1.427	0.559	-2.553	0.24	(0.080,0.717)	0.005
	Altura del pico	0.141	0.079	1.785	1.151	(0.985,1.348)	0.037
	Anchura del pico	-0.004	0.029	-0.138	0.996	(0.939,1.054)	0.445
	Parche (presencia)	-0.473	0.358	-1.321	0.623	(0.308,1.258)	0.093
	Peso	0.011	0.017	0.647	1.011	(0.976,1.047)	0.258
Granívoros.	Intercepto	3.892	1.616	2.408	49.008	(2.067,1163.04)	0.008
	Altura del pico	-0.104	0.239	-0.435	0.901	(0.563,1.441)	0.331
	Anchura del pico	-0.907	0.2	-4.535	0.404	(0.272,0.598)	0.001
	Parche (presencia)	-0.777	0.879	-0.884	0-46	(0.082,2.576)	0,188
	Peso	0.122	0.032	3.813	1.129	(1.059,1.203)	0.001
Insectívoro.	Intercepto	-0.327	0.49	-0.667	0.721	(0.274,1.880)	0.252
	Altura del pico	0.127	0.08	1.588	1.136	(0.971,1.327)	0.056
	Anchura del pico	0.008	0.019	0.421	1.008	(0.971,1.044)	0.336
	Parche (presencia)	-0.791	0.325	-2.434	0.453	(0.239,0.856)	0.093
	Peso	-0.018	0.019	-0.947	0.982	(0.947,1.018)	0.258
Omnívoro	Intercepto	-2.186	1.738	-1.258	0.112	(0.003,3.386)	0.104
	Altura del pico	0.138	0.09	1.533	1.147	(0.962,1.369)	0.062
	Anchura del pico	-0.204	0.172	-1.186	0.815	(0.581,1.143)	0.117
	Parche (presencia)	-1.561	1.114	-1.401	0.209	(0.023,1.863)	0.08
	Peso	0.056	0.028	2.000	1.057	(1.000,1.117)	0.022

Tabla 5: Factores asociados a las preferencias alimentarias de las aves. Modelo de regresión logística politómica.

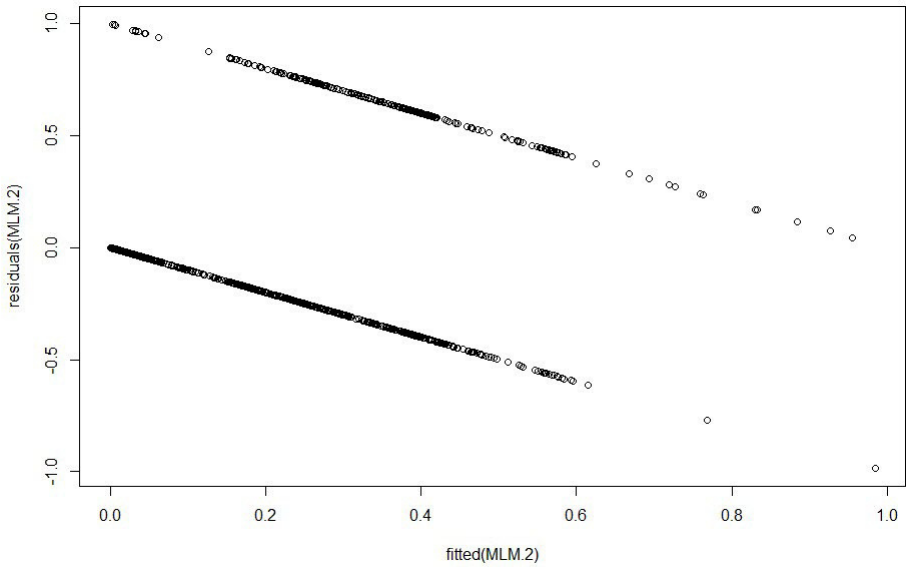


Figura 1: Residuales del modelo óptimo

En primer lugar se realiza la evaluación del modelo mediante los residuos de la devianza (ver tabla 5), considerando que los residuos que indican una falta de ajuste global son aquellos cuyo valor absoluto mayor que 2. Estos residuos se calculan mediante la función "residuals" y se presenta en la tabla 5.

Frugívoros		Frugívoros-Insectívoros		Granívoros		Insectívoros		Omnívoros	
Min	0.0002090	Min	0.004563	Min	0.000000	Min	0.001696	Min	0.00000
Q1	0.3125026	Q1	0.169764	Q1	0.001953	Q1	0.252008	Q1	0.00968
Mediana	0.3848421	Mediana	0.207027	Mediana	0.006938	Mediana	0.333995	Mediana	0.02147
Media	0.3784882	Media	0.211150	Media	0.071713	Media	0.310706	Media	0.02789
Q3	0.4672832	Q3	0.250306	Q3	0.021692	Q3	0.389808	Q3	0.03309
Max	0.6151045	Max	0.568511	Max	0.984087	Max	0.594805	Max	0.45907

Tabla 6: Resumen de los residuales modelo ajustado

A partir de la información anterior se evidencia que no se presentan datos atípicos que el modelo se ajusta bien los datos(ver figura 1). El ajuste global del modelo a través del test de razón de verosimilitud indica que la diferencia de la devianza entre el modelo saturado (64 parametros) y el modelo óptimo(20 parametros), corresponde 65.94451 con un p-valor de 0.9822925, lo cual indica que el modelo propuesto ajusta bien a los datos. Al calcular la tasa de clasificación correcta con los valores observados y predichos por el modelo se obtuvó que en el 48 % de los casos se consigue una predición correcta. Lo que nos indica que el modelo es regular. Lo anterior puede indicar que las preferencias alimentarias no sólo dependen de los factores morfometricos si no de otros factores como las familias, genero, especie y características de las garras.

Para medir la calidad de ajuste se utilizó el pseudo-R^2de Mc-Fadden a traves del cual se obtuvó un valor 0.1507487, como este valor es menor que 0.2, se puede decir que el modelo no presenta un buen ajuste. Se afirma lo mismo con el pseudo-R^2de Nagelkerke el cual arrojó un valor de 0,3581507.

La tabla 6 presenta la información del modelo Multinomial determinado para el estudio. Estadístico de WALD indica que con un valor de significacia del 5 % que ninguna de las variables explican que un ave sea frugívora-insectívora; La anchura del pico explica que sea granívora; la presencia de parche explica que sea insectívora y que sea omnívora queda explicada por el peso. A partir de la coeficientes de ODDS se puede indicar que: A medida que disminuya un milímetro el ancho del pico y aumente su peso un gramo el riesgo de que el ave sea granívora frente a frugívora se multiplica por 0.404, y 1.129 respectivamente. El riesgo de ser insectívoro disminuye frente a ser frugívoro queda multiplicado por 0.453 si el ave tiene parche con respecto a no tenerlo. A medida que aumente un gramo el peso del ave el riesgo que sea omnívora frente a frugívora queda multiplicado por 1.057, es decir aumenta el riesgo.

5. Conclusiones

Se pude afirmar que las preferencias alimentarias de las aves y su morfométria, este modelo permitio que las preferencias alimentarias están determinadas Alto, Ancho del Pico, Parche y Peso. Los resultados muestran que a través de estas variables el modelo óptimo es de 48 % es decir se deben tener en cuenta otras variables explicativas. Lo anterior puede indicar que las preferencias alimentarias no sólo dependen de los factores morfométricos si no de otros factores como las familias, genero, especie y la medición de sus garras, que se deben tener en cuenta para futuros estudios. El estudio determino que lo que explica que un ave sea granívora es la anchura del pico y su peso; que sea insectívora queda explicado por la ausencia de parche y que sea omnívora queda explicado por su peso. Finalmente, se puede decir que los modelos de regresión logística son muy útiles en estudios de tipo biológico tal como se ha puesto de manifiesto. El modelo construido puede indicar los aspectos a tener encuenta a la hora de implementar planes de conservación del hábitat en esta región del pais.

Referencias Bibliográficas

Ardila Ayala, J. (2009), *Preferencias alimentarias de aves asociadas a bosques riparios de la sabana inundable en Paz de Ariporo, Casanare*, Yopal,Casanare.

Díaz Monroy, L. G. y Morales Rivera, M. A. (2012), *Análisis estadístico de datos categóricos*, Universidad Nacional de Colombia.

Fernández, V. P. y Fernández, R. S. M. (2004), *Regresión logística multinomial*, Departamento de estadística e investigación operaria.

Mauricio Álvarez, S. C. y. o. (2004), *Manual de métodos para el desarrollo de inventarios de biodiversidad Bogotá D. C*, Impreso en Bogotá D. C.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, Vienna, Austria. *https://www.R-project.org/

Rodríguez, M. A. D. (2013), *Modelo de respuesta discretas en R y aplicación con datos reales*, Universidad De Granada.

RESULTADOS SABER PRO 2015 Y SU RELACIÓN CON VARIABLES SOCIODEMOGRÁFICAS [1]

Especialización en Estadística

Yuri Carolina Niño Castillo[1,a], Sandra Patricia Cárdenas Ojeda[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Normalmente las personas que optan por un proceso de formación académica que no requiera de una asistencia presencial constante, lo hacen porque buscan que dicho proceso no afecte el cumplimiento de otro tipo de obligaciones. Se podría pensar que cuando el proceso se lleva a cabo a distancia, las limitaciones que se tuvieron en cuenta para elegir un programa presencial, también influyen en el rendimiento acaémico de los estudiantes, evidenciado de alguna manera en los resultados de la prueba saber pro. Es así como se buscó determinar si el desempeño en cada una de las competencias que dicha prueba evalúa, depende de algunas de las variables sociodemográficas que el Icfes tiene en cuenta a la hora de inscripción. Encontrando a nivel general que la relación no es tan fuerte como se esperaría.

***Palabras clave*:** Educación a distancia, saber pro, competencias genéricas, pruebas de independencia.

Abstract

People who choose an academic development process that does not require a constant site support, normally prefer them because they look for that process not to affect the fulfillment of other type of obligations. It is possible to think that when the process is carried out at distance, the limitations that were taken into account to choose a permanent attendance program, as well influence in the students? academic performance, and it is evidenced somehow in the results of the Saber Pro exam. For this reason, I sought to determine if the performance in each competence that this exam evaluates, depends on any of the sociodemographic variables that the Icfes considers in the inscription step. The relationship in general was not as strong as it would be expected.

***Key words*:** distance education, saber pro, general skills, test of independence.

1. Introducción

En el proceso de evolución que ha tenido el ser humano a lo largo de los años se ha visto la necesidad de hacer algunas mejoras que le permitan desenvolverse con más facilidad en el medio que lo rodea; y de la misma forma, brindar herramientas que fortalezcan las habilidades en futuras generaciones. Para esto se recurría a un proceso con el que se pudieran transmitir las experiencias que dejaban enseñanza o lograron

[1]Caso: Licenciatura en Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana; Universidad Pedagógica y Tecnológica de Colombia, Facultad de Estudios A Distancia, sede Tunja; además las variables que se tomaron para verificar la relación fueron: género, estrato, estado civil, trabajaba, situación de su hogar, No de personas a cargo, título de bachillerato y cabeza de familia.

[a]Estudiante de Especialización en Estadística. E-mail: yuricarolina.nino@uptc.edu.co

[b]Profesor asistente. E-mail: sandra.cardenas@uptc.edu.co

cambios en las costumbres que se tenían, es decir, se trataba de educar a los descendientes, y esto es según Platón, dar al cuerpo y alma toda la perfección que son capaces de soportar.[1]

Al igual que cualquier proceso, en ocasiones necesitan pulirse ciertos detalles para brindar resultados más óptimos. En Colombia, el modelo educativo tuvo un cambio que se pensó con el propósito de corregir dichos detalles; siendo más específicos, se hizo con miras a facilitar el ingreso de los ciudadanos a programas e instituciones educativas de todos los niveles, en especial el superior; haciéndolos participantes activos en el proceso de educación que obviamente conlleva a una evolución social muy positiva.

Teniendo en cuenta que todos los ciudadanos tienen derecho a la educación, se diseñó una propuesta que les abriera todas las puertas posibles a este proceso, entonces nació la Educación Abierta y a Distancia, la cual se plantea como una muy buena opción para aquellas personas que no disponen del tiempo suficiente para participar de un proceso presencial; por lo cual su proceso ideal debía reducir eficazmente los obstáculos relacionados con el tiempo y espacio (Alfonso Sánchez 2003).

Es así como surge la educación a distancia, la cual según Alfonso Sánchez (2003), "Es un conjunto de procedimientos e interacciones de mediación que se establece entre educandos y profesores en el desarrollo del proceso enseñanza-aprendizaje mediante la utilización racional de recursos tecnológicos informáticos". Este proceso, hoy en día ha tenido bastante aceptación, por las facilidades que ofrece.

Para garantizar que la Educación impartida, realmente tenga un efecto positivo, se ejercen ciertos criterios que evalúan su calidad, término que se asocia con lo bueno, con el hecho de alcanzar unas metas planteadas. Sin embargo, su concepto puede variar dependiendo del contexto en el que se aplique, así que para centrar la temática, se especificará lo que es el concepto de calidad aplicado a la educación.

Si se partiera del hecho de que la calidad hace referencia a un producto final, solo se estaría viendo al docente como un "obrero", el cual hace uso de materiales que le llegan prefabricados y por lo que la calidad vendría a medirse por hechos aislados que surgen en el "producto final"(Aguerrondo 1993).

Según Aguerrondo (1993), este concepto se debe replantear, ya que al evaluar únicamente el producto final, no se está teniendo en cuenta el hecho de que otros factores como la calidad del docente, de los procesos, de la infraestructura, entre otros; estén afectando los resultados.

Es importante evaluar dicha calidad, teniendo en cuenta que cuando se culmina el proceso, estas personas saldrán a desempeñar un papel social, que los afectará tanto a ellos como al entorno, dependiendo de la formación que recibieron, además, se considera que si el sistema no transmite conocimiento socialmente válido, no es de calidad (Aguerrondo 1993), por eso se debe estar en un proceso permanente de evaluación, ya que el sector educativo, al igual que el comercial, de salud, etc., están en la obligación de responder a las diferentes necesidades que tiene la sociedad, las cuales se van viendo reformadas con el paso del tiempo

Debido a la necesidad de estar evaluando constantemente la educación, ya que en ella se fundamenta la base de toda sociedad y por lo cual se pretende que sea impartida con calidad, se implementan una serie de procesos que permiten llevar a cabo inspección y vigilancia, para garantizar que los esfuerzos tanto económicos como humanos, surtirán efectos positivos.

Entonces fue implementada por el Instituto Colombiano para la Evaluación de la Educación - Icfes, la prueba saber pro, precisamente para evaluar la calidad de la educación superior.Dicha prueba se viene aplicando desde el 2003, su presentación es de carácter obligatorio y la pueden realizar estudiantes de programas tecnológicos y profesionales universitarios que hayan aprobado por lo menos el 75 % de los créditos académicos del programa que están cursando. Anteriormente la prueba se hacá dos veces en el año, pero a partir del 2014 se realiza solo una vez, en el mes de noviembre (Alvear 2014).

Esta prueba tiene tres objetivos: comprobar el desarrollo de competencias en los estudiantes, proporcionar información para la comparación entre programas e instituciones y recoger datos para construir indicadores de evaluación (*Para qué sirven las pruebas saber pro* 2012); a su vez, se divide en dos módulos, uno de competencias genéricas (comunicación escrita, inglés, razonamiento cuantitativo, lectura crítica, competencias ciuadadanas), que califica los conocimientos que deben tener todos los estudiantes sin importar el programa de educación superior que están cursando, y otro de competencias específicas, que evalúa el conocimiento de acuerdo al programa o carrera escogidos por el estudiante.

[1]Concepto tomado de Dimas Márques (2014)

Respecto a las competencias genéricas, se hace a continuación una pequeña descripción de cada una de ellas, ya que son las de interés en este caso

Comunicación escrita: este módulo evalúa la habilidad de trasnmitir ideas referentes a un tema dado, por escrito. Estos temas son de dominio público por lo que el estudiante no requiere de conocimientos especialiados al respecto, y todos tienen la misma oportunidad de producir un texto sobre estos. La calificación de los escritos tiene en cuenta los siguientes aspectos: planteamiento que se hace del texto (perspectivas innovadoras), organización (secuencia de las ideas) y la expresión (lenguaje usado). De igual forma, el desempeño en esta competencia, se designa con número de 1 a 8, siendo 1 el más bajo. Cuando no aparece calificación se da por hecho que el estudiante no respondió o el texto que escribió no fue legible.[2]

Inglés: evalúa la habilidad para comunicarse de forma efectiva en este idioma, y su nivel de desempeño está clasificado en A1, A2, B1 y B2, siendo A1 el nivel más bajo; en todos se tiene en cuenta la facilidad con la que el estudiante comprende un escrito, sabe comunicar ideas y mantener una conversación.[3]

Razonamiento Cuantitativo: este módulo se relaciona con las habilidades matemáticas que todo ciudadano debe tener, que le son de utilidad en situaciones cotidianas y que no necesariamente se involucran con su profesión. Las competencias que se tienen en cuenta y su porcentaje en el módulo para el nivel profesional (se manejan diferentes para los niveles técnico y tecnológico) son la interpretación y representación (33 %), Formulación y ejecución (33 %) y Argumentación (34 %); de acuerdo a la calificación obtenida en cada uno de estos aspectos el estudiante estará en el nivel 1, 2 o 3, siendo 1 el nivel más bajo.[4]

Lectura crítica: en este módulo se tienen en cuenta la capacidad de entender, interpretar y evaluar textos que se pueden encontrar en ámbitos no especializados. Esta prueba evalúa tres competencias, como son identificar y entender los contenidos de un texto, comprender cómo se articulan sus partes para darle sentido global y reflexionar a partir del texto evaluando su contenido. De acuerdo a esto, el estudiante se puede encontrar en el nivel 1 (muy poca comprensión y no alcanza a cumplir con lo exigido en el nivel 2), 2 (comprende los contenidos generales del texto y su propósito) ó 3 (contextualiza adecuadamente el texto tomando una posición crítica respecto a este), [5]

Competencias ciudadanas: esta prueba evalúa las habilidades que conducen a una mejor comprensión del entorno, aportando positivamente a este. Los componentes a tener en cuenta para la prueba son los conocimientos mostrados, la valoración de argumentos, el multiperspectivismo (capacidad de ver la problemática desde varias perspectivas), y el pensamiento sistémico (construir sistemas, relaciones entre las diferentes dimensiones de la problemática)[6]

De acuerdo al tema que se está manejando, se habla entonces de las diferentes investigaciones que lo han abordado. Aguerrondo (1993), en su escrito “La educación a distancia", habla de la implementación de herramientas como el internet, las cuales permiten un mejor aprendizaje y ayudan a los estudiantes a cumplir sus expectativas en este proceso. Ya que "podemos enseñar a robar y no estamos educando", recalca que en este se debe tener muy clara la diferencia entre enseñar y educar, para dar nitidez a lo que se está logrando.

También Rodriguez Albor, Gómez Lorduy y Ariza Dau (2014), realizaron un trabajo denominado ¿Calidad de la Educación Superior a Distancia y Virtual, un análisis de desempeño académico en Colombia?, en el cual se buscó determinar si habían diferencias estadísticamente significativas entre la educación tradicional y la no tradicional, que incluye tanto educación a distancia como virtual, esto en términos de calidad, para lo que se tomaron y analizaron los resultados de las pruebas saber pro en programas como Administración, Licenciaturas, Ingenierías, Contaduría y Psicología de instituciones de Educación Superior a nivel nacional, teniendo en cuenta características sociodemográficas que el Instituto Colombiano para la Evaluación de la Educación-Icfes, solicita a la hora de inscribirse para presentar dicha prueba.

[2]Guía de orientación, Módulo de Comunicación escrita, saber pro 2015-2-ICFES
[3]Guía de orientación, Módulo de Inglés, saber pro 2015-2-ICFES
[4]Guía de orientación, Módulo de Razonamiento Cuantitativo, saber pro 2015-2-ICFES
[5]Guía de orientación, Módulo de lectura crítica, saber pro 2015-2-ICFES
[6]Guía de orientación, Módulo de Competencias Ciudadanas, saber pro 2015-2-ICFES

A nivel nacional son bastantes las instituciones que ofrecen programas en la modalidad a distancia, pero en esta ocasión, se tomará el caso de la Universidad Pedagógica y Tecnológica de Colombia-UPTC, siendo más específicos el programa de Licenciatura en Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana, que es ofrecido por la Facultad de Estudios A Distancia-FESAD. Contemplando la posiblidad de que existan algunos factores que influyan en el rendimiento académico de los estudiantes, quizá parecidos a los que les impiden llevar a cabo el proceso de manera presencial.

Es así como se pretende determinar si ciertos aspectos de la vida personal del estudiante tienen influencia en su rendimiento académico, evidenciado de alguna manera en los resultados de las pruebas saber pro; teniendo en cuenta que en la educación a distancia no se tiene el mismo contacto con el docente que cuando se es partícipe de un proceso presencial, además el hecho de no estar conviviendo en el ambiente estudiantil que se tiene en este proceso, posiblemente afecte el rendimiento académico del estudiante.

2. Referente Conceptual

En el siguiente apartado se dan a conocer algunos conceptos relacionados con la estadística, como son las tablas de contingencia, las pruebas de independencia y algunos coeficientes de asociación.

2.1. Tablas de contingencia

Una tabla de contingencia de es un arreglo bidimensional, de una variable fila con $f-$categorías o modalidades y una variable columna con $c-$categorías, donde hay $f \times c$ celdas, las entradas de las celdas son las frecuencias o conteos del número de casos en cada una de las combinaciones de valores de ambas variables. En general, se nota con n_{ij} a la frecuencia de la i-ésima modalidad de la variable fila y j-ésima de la variable columna.

El total por fila o por columna está formado por las frecuencias marginales, y se notan por $n_{i.}$ (donde el punto señala que se suman columnas dentro de la fila i) y $n_{.j}$ (donde el punto señala que se suman filas dentro de la columna j), respectivamente.

La suma de las frecuencias por celda es igual a la suma de las frecuencias marginales e igual al número total de individuos seleccionados y clasificados; se nota por n.

De acuerdo con Díaz M. y Morales R. (2009), la notación general, para una tabla de contingencia de $f-$filas y $c-$columnas, se muestra en la tabla 1.

Filas	Columnas						
	1	2	...	j	...	c	Total($n_{i.}$)
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
f	n_{f1}	n_{f2}	...	n_{fj}	...	$n_{f.}$	n
Total($n_{.j}$)	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..} = n$

Tabla 1: Tabla de contingencia

donde

- La frecuencia de la $i-$ésima modalidad de la variable fila y la modalidad $j-$ésima de la variable columna se escribe como n_{ij}.
- El total de observaciones en la $i-$ésima modalidad de la variable fila se nota por $n_{i.}$, es decir,

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

- El total de observaciones en la $j-$ésima modalidad de la variable columna se nota por $n_{.j}$; es decir,

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{fj} = \sum_{i=1}^{f} n_{ij}$$

- El número total de observaciones en la muestra se escribe con n, y es igual a la suma de los márgenes fila o columna, es decir,

$$n = \sum_{i=1}^{f} \sum_{j=1}^{c} n_{ij}$$

Por otra parte, las frecuencias pueden ser transformadas en proporciones o porcentajes. Un primer porcentaje se obtiene de dividir cada frecuencia n_{ij} por el número total de observaciones n; este porcentaje se escribe como f_{ij}, es decir,

$$f_{ij} = \frac{n_{ij}}{N} \times 100$$

la cantidad f_{ij} corresponde a la proporción o porcentaje de elementos que tienen los atributos i y j.

El segundo porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal fila $n_{i.}$, así:

$$f_{j|i} = \frac{n_{ij}}{ni.} \times 100$$

La cantidad $f_{j|i}$ es la proporción de elementos de cada celda, respecto al total de la fila i. Según Díaz M. y Morales R. (2009) "La expresión $j|i$ (que se lee: "j dado i") significa "estar en la columna j, a condición de estar en la fila i", es decir, se deja fija la fila i y se recorren sus columnas. Estas frecuencias corresponden al *perfil fila*".

El tercer porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal columna $n_{.j}$:

$$f_{ij} = \frac{n_{ij}}{n} \times 100$$

La cantidad $f_{i|j}$ es la proporción de elementos de cada celda, respecto al total de la columna j. Nuevamente según Díaz M. y Morales R. (2009) "La expresión $i|j$ (se lee: "i dado j") significa "estar en la fila i, a condición de estar en la columna j", es decir, se deja fija la columna j y se recorren sus filas. Estas frecuencias corresponden al *perfil columna*".

Se obtienen tres tipos de tablas adicionales, la primera hace referencia al porcentaje de cada celda con relación al número total de individuos n; la segunda, al porcentaje de cada celda respecto al total de la respectiva fila (perfil fila) y la tercera, al porcentaje de cada celda con relación al total de la respectiva columna (perfil columna).

2.2. Pruebas de independencia

Al disponer de la información en una tabla de contingencia, es posible indagar si las variables que constituyen dicha tabla son independientes o no.

la hipótesis nula de independencia está dada por:

Ho: La variable fila es independiente de la variable columna

La estadística de prueba que se empleada en el juzgamiento de esta hipótesis es:

$$\chi^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

que bajo la hipótesis nula de independencia, tiene distribución de probabilidad *ji-cuadrado* con $(f-1)\times(c-1)$ grados de libertad.

Se rechaza la hipótesis nula a un nivel α cuando se verifica que $\chi_0^2 > \chi^2_{(f-1)(c-1),\alpha}$

2.3. Medidas de asociación

A continuación algunas medidas relacionadas con la estadística ji-cuadrado, como se ver en Díaz M. y Morales R. (2009, pág 38), algunas de estas son:

2.3.1. El coeficiente de contingencia

Es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando se tiene información clasificadora acerca de uno o ambos conjuntos de atributos. El grado de asociación entre dos conjuntos de atributos, sean ordinales o no, se puede describir mediante la siguiente fórmula:

$$C = \sqrt{\frac{\chi_0^2}{\chi_0^2 + n}}, \qquad \text{donde} \qquad \chi_0^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

La estadística C toma valores entre 0 y 1. Valores cercanos a cero muestran una baja asociación entre las variables, mientras que valores próximos a 1 indican una posible alta asociación.

Este coeficiente tiene un valor máximo en tablas de contingencia de cualquier tamaño. Se define como

$$V = \sqrt{\frac{\chi_0^2}{nk}} \tag{3}$$

donde k =mín{f-1,c-1} es el menor número de modalidades fila (o columna) menos uno de la tabla de contingencia. Se trata de un coeficiente que toma el valor 1 cuando hay asociación perfecta entre los atributos, cualquiera que sea el tamaño de la tabla de contingencia.

3. Metodología

Para el desarrollo de esta aplicación se utilizó el enfoque cuatitativo y tipo de investigación exploratoria. Teniendo como población los estudiantes de la Licenciatura en Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana.

Los datos fueron tomados de la página del ICFES www.icfes.gov.co, en el módulo investigadores y estudiantes de posgrado, en la opción acceso a bases de datos y siguiendo los pasos según la GUIA DE ACCESO A BASES DE DATOS DEL ICFES (ICFES 2013) se decargó la carpeta saberpro, de ella se utilizaron los diccionarios SBRPRO-Diccionario de datos componentes-v1-1.pdf, SBPRO Diccionario de variables v1-0.pdf y de la carpeta SBPRO 20153 RGSTRO CLFCCNNU GEN.zip se tomó la base de datos SBPRO 20153NRGSTRO CLFCCNNU GEN.txt, y siguiendo las intrucciones del archivo tutorial (ICFES n.d.) se abrió en hoja de excel la base de datos con un total de 324849 registros, información de los resultados de la prueba SABERPRO para el año 2015 en Colombia, de dicha base se seleccionaron los 570 registros correspondientes a los estudiantes de la Licenciatura.

Las etapas que se llevaron a cabo para el desarrollo fueron la siguientes:

Primera etapa: Una vez tomada la bse de la Licenciatura en Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana, la cual es ofrecida por la FESAD, UPTC se identificaron 88 columnas (variables) y 570 filas (registros-estudiantes).

La segunda etapa: se llevó a cabo la exploración de la base de datos, con lo cual se identificó que al momento de inscripción, el Icfes indaga al estudiante sobre aspectos sociodemográficos como el municipio de residencia, género, estado civil, si se es cabeza de familia, edad, bienes con los que cuenta su vivienda y el estrato de esta,nivel educativo alcanzado por los padres; y académicos como el tipo de bachillerato del que es egresado, semestre y créditos aprobados, si tomó algún curso de preparación para el examen; datos que se trataron como variables cualitativas. Además relacionado con estos aspectos (académicos), se encuentran los puntajes obtenidos por los estudiantes en cada competencia junto con sus respectivos desempeños.

Tercera etapa: con ayuda del software estadístico R project (R Core Team 2015), se realizó un análisis descriptivo univariado de los datos, es decir, se realizaron tablas de distribución de frecuencia tanto para variables cualitativas como para cuantitativas (puntajes), a estas últimas también se les hallaron medidas de tendencia central(moda, mediana, media), de posición (cuartiles), de variabilidad (desviación estándar, coeficiente de variación) de forma y apuntamiento (asimetría y kurtosis). Además, se detectaron datos atípicos mediante los puntajes estandarizados y el diagrama de caja.

Cuarta etapa: se procedió a realizar el análisis bivariado de los datos, a traveés de tablas de contingencia entre los desempeños de las diferentes competencias y algunas variables sociodemográficas, pruebas de independencia y coeficientes de asociación; y el análisis multivariado se llevó a cabo hallando las matrices de varianzas y covarianzas y de correlación entre los puntajes obtenidos en las distintas competencias junto con la edad de los estudiantes.

4. Resultados

Al revisar los municipios de residencia de los estudiantes que presentaron el examen Saber Pro 2015, se observa que en su mayoría residen en municipios como Bogotá, Chiquinquirá, Tunja, Sogamoso, Ubate, Fusagasuga, Tulúa, Yopal, Paipa, Barrancabermeja, Saboyá, Villavicencio y Acacías; los cuales al totalizarlos se obtienen 358; de los demás, las mayorías se encuentran en departamentos como Boyacá, Cundinamarca y Valle. Teniendo en cuenta los departamentos de residencia, se dice que Boyacá es el lugar donde más residen estudiantes de la Licenciatura en Básica con énfasis en Matemáticas, Humanidades y Lengua Castellana, con un 46.84 %. Se observa un comportamiento similar en lo relacionado con el municipio y departamento en el que el estudiante presenta el examen en cuestión.

Género	Estado Civil					Total
	Soltero(1)	Casado(2)	Viudo(3)	Separado(4)	Unión Libre(5)	
F	264	120	1	14	107	506
%	46,32	21,05	0,18	2,46	18,77	88,77
M	35	17	0	2	10	64
%	6,14	2,98	0,00	0,35	1,75	11,23

Tabla 2: Clasificación estudiantes por estado civil. Fuente la Autora, 2016

Género	Cabeza de Familia				Total
	NO		SI		
	estudiantes	%	estudiantes	%	
F	360	63,16	146	25,61	506
M	23	4,04	41	7,19	64
TOTAL	383	67,19	187	32,81	570

Tabla 3: Clasificación estudiantes cabeza de famila por género. Fuente la Autora, 2016

Ahora, en cuanto al género de los estudiantes del programa, se evidencia que predominan las mujeres 88 %; además, en lo que respecta al estado civil, tanto hombres como mujeres en su mayoría, son solteros; seguidos por los que están casados y los que viven en unión libre (Tabla 2); lo cual, por el lado de las mujeres coincide en lo relacionado a la pregunta si son cabeza de familia o no, ya que el 71,15 % de ellas respondió que no; pero en el caso de los hombres, el 64,06 % respondió que sí, entonces en los hombres, es mayor la proporción de los que son cabezas de familia (Tabla 3).

También, el Icfes indaga sobre el número de personas de las que se encuentra a cargo, dando como posibilidades de respuesta: 1, 2, 3, 4 y 5; en este caso, el mayor porcentaje se presenta para los que no tienen hijos 32.28 % (184), y en seguida están los que tienen 1 y 2, con un porcentaje aproximadamente igual 25 %, y en el caso de las mujeres predominan las que no tienen hijos (169), para los hombres el mayor número de estudiantes respondieron que tienen 2 hijos (18).

Para la edad de los aspirantes en el momento de presentar el examen, se observa que la gran mayoría tienen entre 20 y 30 años, sin embargo, también se muestra una cantidad considerable de estudiantes que tienen entre 30 y 50 años, además se evidencia la presencia de datos atípicos, (edad 11), ya que considerando el nivel de educación que se está abordando, no suelen encontrarse estudiantes con dicha edad.

Bien o Recurso	SI		NO		Total	Bien o Recurso	SI		NO		Total
	Estud	%	Estud	%			Estud	%	Estud	%	
Celular	557	97,72	13	2,28	570	Microondas	105	18,42	465	81,58	570
Internet	311	54,56	259	45,44	570	Nevera	507	88,95	63	11,05	570
Servicio TV	366	64,21	204	35,79	570	Computador	480	84,21	90	15,79	570
Telefonía	150	26,32	420	73,68	570	Automóvil	98	17,19	472	82,81	570
Lavadora	368	64,56	202	35,44	570	Repr. DVD	271	47,54	299	52,46	570
Horno	122	21,40	448	78,60	570						

Tabla 4: Bienes con los que cuenta la vivienda. Fuente la Autora, 2016

Al estudiante se le pregunta si su vivienda cuenta con ciertos bienes, se esperaría que por la condición de ser educación a distancia, los estudiantes cuenten con las principales herramientas que necesitan para poder llevar a cabo el proceso de manera cómoda y de algún modo satisfactoria, como computador e internet; en el caso del primero, la gran mayoría cuenta con este, pero la cuestión de internet no es tan mayoritaria, la cantidad de estudiantes que cuenta con este servicio supera el 50 % de la totalidad de estudiantes, pero quizá no es la que se esperaría.

Con respecto a los demás bienes, el que predomina es el celular como principal medio de comunicación, seguido de la nevera, luego de la lavadora y en seguida el servicio de televisión. Bienes como el horno y automóvil son los que menos predominan en las viviendas de los estudiantes.

En lo que respecta al número de dormitorios con los que cuenta cada estudiante en su vivienda, el Icfes da posibilidades de respuesta entre 1 y 10; en este caso, el 42.63 % (243) cuenta con 2 dormitorios, le siguen los que tienen 3 con un porcentaje de 35.08 (200), para las demás respuestas, de las cuales, la cantidad máxima de dormitorios es 7, se presentan cantidades pequeñas de estudiantes que las eligieron.

A la pregunta sobre el material de los pisos que predominan en las viviendas dando como opciones de respuesta 1 (tierra, arena), 2 (cemento, gravilla), 3 (Madera burda, Tabla o tablón), 4 (baldosa, tableta, ladrillo, vinilo), 5 (mármol, madera pulida, alfombra o tapete de pared a pared); la respuesta que predomina entre los estudiantes con un porcentaje de 67.36 (384) es la 4, seguida de la 3 con 22.98 % (131); relacionado con los aspectos de la vivienda se habla también del estrato de la residencia, el 61.22 % (349) dijeron estar en el estrato 2 y el 20 % (114) en el 1; se observa un comportamiento similar en lo que tiene que ver con el nivel de sisben en el que están clasificadas las respectivas familias, anotando que en el nivel 5 (o reporta) se encuentra una cantidad considerable de estudiantes (101).

Para el número de personas que conforman el hogar, el rango de respuesta está entre 1 y 12, y las respuestas que predominan son 3, 4 y 5 con porcentajes de 27.71 % (158), 25.78 % (147) y 18.59 % (106) respectivamente, cabe anotar que no hubo respuestas superiores a 9, por lo que se dice que las familias de los estudiantes del programa en cuestión no son numerosas.

Nivel	Madre		Padre		Nivel	Madre		Padre	
	Estud	%	Estud	%		Estud	%	Estud	%
(0) Ninguno	15	2,63	35	6,14	(14) Téc comp	17	2,98	9	1,58
(9) Pria incomp	175	30,70	212	37,19	(15) Profesión incomp	3	0,53	4	0,70
(10) Pria comp	148	25,96	143	25,09	(16) Profesión comp	24	4,21	15	2,63
(11) Sec incomp	95	16,67	75	13,16	(17) Posgr	9	1,58	7	1,23
(12) Sec comp	77	13,51	62	10,88	(99) No sabe	1	0,18	0	0
(13) Téc incomp	6	1,05	8	1,40	Total	570	100	570	100

Tabla 5: Nivel educativo de los padres. Fuente la Autora, 2016

Respecto al nivel educativo alcanzado por padre y madre de los estudiantes, se observa en la Tabla 5, que la mayoría tanto de madres como padres tienen la primaria incompleta, seguida de la primaria completa, luego secundaria incompleta y finalmente secundaria completa, por lo cual se dice que se evidencia una clara voluntad de querer salir adelante, porque en gran parte sus padres tienen un nivel educativo bastante bajo. El Icfes pregunta a los aspirantes sobre la situación de su hogar actual, haciendo esto referencia a 1 (si es habitual o permanente) y 2 (temporal por razones de estudio u otra razón), en este caso el 82.98 % (473) respondió que su hogar es permanente, esto a raíz de que la FESAD cuenta con los denominados CREADS (Centros Regionales de Educación A Distancia) y así le brinda a sus estudiantes la posibilidad de acceder a sus programas sin tener que desplazarse de su lugar de residencia.

Sobre el tipo de bachillerato del que es egresado el estudiante, donde las opciones de respuesta son A (Académico), N (Normalista superior) y T (Técnico), el porcentaje más alto se presenta en la respuesta A, este es de 44.21 % (252), seguido por N y luego por T. Lo cual asociándolo con el género, se observa que la mayoría de las mujeres con un porcentaje de 39.64 % obtuvieron un título técnico (226), pero en el caso de los hombres, hay una cantidad igual en académico y normalista (26).

En lo relacionado al semestre que se está cursando y el porcentaje de créditos que tiene aprobados a la fecha de inscripción el estudiante, en este caso el 50.35 % (287) respondió estar en el décimo semestre, seguido por el 40.35 % (230) de noveno, los demás estudiantes están en menores cantidades en el semestre 7, 8, 11 y 12. En lo relacionado al porcentaje de créditos se dan las siguientes opciones: 0 (no sigue sistema de créditos), 1 (menos del 75 %), 2 (entre el 76 % y 80 %), 3 (entre el 81 % y 90 %) y 4 (el 90 % o más); para lo cual el 55.08 % (314) de los estudiantes respondió 4, seguido por 3 y 2; se esperaría que los estudiantes que siguen el sistema de créditos opten por alguna de estas tres, teniendo en cuenta la reglamentación para la presentación de la prueba, sin embargo hay estudiantes que optaron por 1.

Respecto a si tomaron un curso para presentar la SABERPRO, el 97.89 % (558) de los estudiantes respondió que no, y en lo relacionado a otra actividad de preparación, el 69.82 % (398) respondió que repasó por cuenta propia.

En las respuestas a la pregunta que se plantea sobre el costo de la matrícula el año anterior, teniendo en cuenta que las opciones eran: 0 (no pagó matrícula), 1 (menos de 500 mil), 2 (entre 500 mil y menos de un 1 millón), 3 (entre 1 y 3 millones), 4 (entre 3 y 5 millones) 5 (más de 5 millones); el 70.17 % (400) de los estudiantes, respondió 2, seguido por los que respondieron 3, que fue el 26.84 %(153). A pesar de que la Licenciatura en básica con énfasis en Matemáticas, Humanidades y Lengua Castellana tiene un costo de 1

SMMLV en Boyacá y Casanare, y de 1.5 SMMLV fuera de Boyacá, hubo 2 estudiantes que optaron cada uno, por la respuesta 4 y 5.

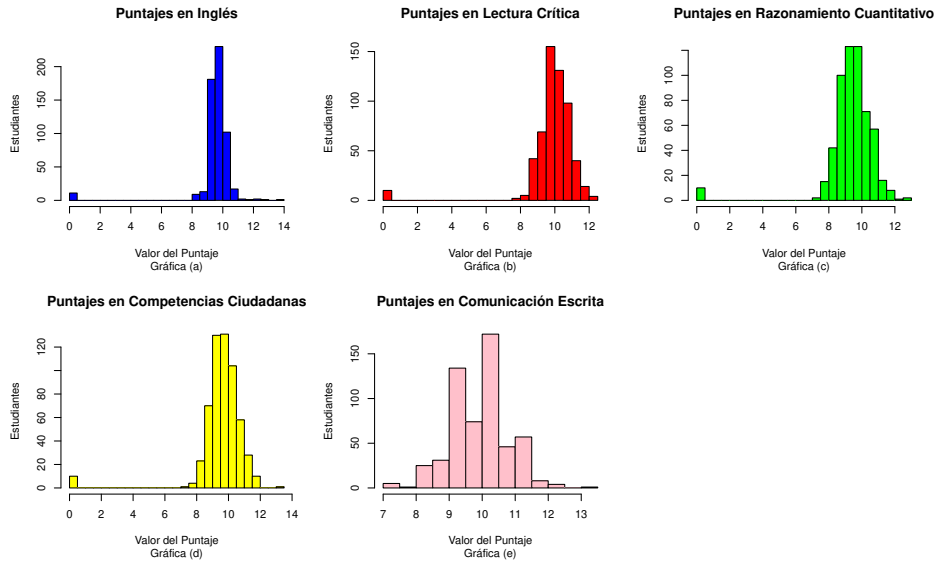
Ahora, el 67.54 % (385) de los estudiantes respondieron que se financiaron el costo de la matrícula, por lo cual se pensaría que son personas que para lograrlo deben trabajar paralelamente al desarrollo de sus actividades académicas, sin embargo, el 62.92 % (353) de los estudiantes respondieron que trabajan pero sin ninguna remuneración y tan solo el 18.77 % (107) trabajan, ya sea para adquirir experiencia laboral o para sus gastos personales. Además, el 47.19 % (269) de los estudiantes dijo que trabajan 20 horas a la semana sin remuneración, lo cual de alguna forma puede afectar sus ingresos; sin embargo, el 33.50 % (191) de los estudiantes, que trabajan sin remuneración, tienen ingresos familiares entre 1 y menos de 2 salarios mínimos, lo que llevaría a pensar que claramente son personas que aún conviven con sus padres o tienen un apoyo económico de algún tipo.

A continuación se presentan algunas estadísticas para la variable puntaje en cada una de las competencias.

Competencia	Min	Q1	Q2	Mean	Q3.	Max	SD	CV	Asimetria	Kurtosis
Comunicación escrita	7,4	9,4	10,1	10,01	10,4	13,2	0,848	0,085	0,0346	3,36
Inglés	0	9,36	9,61	9,526	9,99	13,6	1,432	0,15	-5,56	38,144
Razonamiento Cuantitativo	0	9	9,5	9,433	10,1	12,8	1,523	0,161	-3,996	26,27
Lectura Crítica	0	9,6	10,1	9,9	10,6	12,4	1,526	0,154	-4,745	31,823
Competencias Ciudadanas	0	9,2	9,7	9,6	10,3	13,2	1,517	0,157	-4,318	28,705

Tabla 6: Algunas estadísticas de los puntajes obtenidos en cada competencia. Fuente la Autora, 2016

Con respecto a la comunicación escrita, la desviación estándar es la más pequeña de todas, lo que indica una mayor concentración de los puntajes alrededor de la media, lo cual se ratifica con el coeficiente de variación de los puntajes en dicha competencia, ya que este está indicando poca variabilidad con respecto a la media, y esto también lo están evidenciando los cuartiles presentados en la Tabla 6. La desviación estándar y el coeficiente de variación de las otras competencias, no son tan grandes comparados a los de la comunicación escrita, pero en lo que tiene que ver con la asimetría y la kurtosis, las diferencias son demasiado grandes, y de acuerdo a estos valores correspondientes a dichas competencias, indican la presencia de datos atípicos (Tabla 6), lo cual se verificó a través de los puntajes estandarizados; se encontró que la presencia de puntajes iguales a cero, son los que ayudan a generar dichos valores de asimetría y kurtosis.



Se observa en los histogramas que hay presencia de datos aípicos el los puntajes de cada competencia, lo cual se confirma con el diagrama de caja.

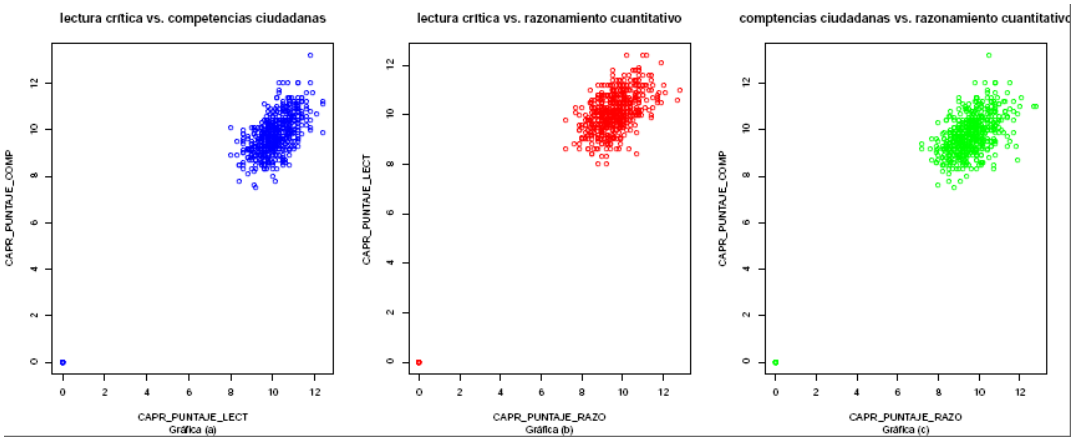
	Com. Escrita	Inglés	Raz. Cuanti.	Lect. Crítica	Comp. ciudad.	Edad
Com. Escrita	0.720	0.033	0.150	0.126	0.155	-0.344
Inglés	0.033	0.436	0.105	0.079	0.110	-0.014
Raz. Cuanti.	0.150	0.105	0.741	0.324	0.340	-0.105
Lect. Crítica	0.126	0.079	0.324	0.571	0.358	-0.012
Comp.Ciudad.	0.155	0.110	0.340	0.358	0.661	-0.183
Edad	-0.344	-0.014	-0.105	-0.012	-0.183	40.04

Tabla 7: Matriz de varianzas y covarianzas para puntajes

	Com. Escrita	Inglés	Raz. Cuanti.	Lect. Crítica	Comp. ciudad.	Edad
Com. Escrita	1	0.060	0.205	0.197	0.224	-0.064
Inglés	0.060	1	0.185	0.159	0.204	-0.003
Raz. Cuanti.	0.205	0.185	1	0.498	0.486	-0.019
Lect. Crítica	0.197	0.159	0.498	1	0.582	-0.002
Comp.Ciudad.	0.224	0.204	0.486	0.582	1	-0.035
Edad	-0.064	-0.003	-0.019	-0.002	-0.035	1

Tabla 8: Matriz de correlación para puntajes

Por otra parte, se revisó si podía existir asociación lineal entre los puntajes obtenidos en cada competencia (Tabla 7), encontrando para el caso de las varianzas, que los valores más altos están entre Lectura Crítica y Comptencias Ciudadanas, Razonamiento cuantitativo y competencias ciudadanas, Razonamiento Cuantitativo y Lectura Crítica, y por último Comunicación Escrita y Edad, y para las demás se encontraron valores cercanos a cero lo que indica que no existe asociación, esto llevaría a pensar que existe asociación entre las primeras variables mencionadas(valores positivos asociación directa; valores negativos, inversa) sin embargo, al hallar la matriz de correlación, (Tabla 8) teniendo en cuenta que los valores mayores a 0.6 indican asociación, se encontró que esto solo se cumple para los tres primeros casos mencionados. Lo cual se muestra en los diagramas de dispersión.



Cabe señalar que tanto los valores de la matriz de varianzas y covarianzas fueron calculados con todos los datos, es decir, involucrando la información de aquellos estudiantes que tanto en las competencias de lectura crítica, razonamiento cuantitativo y compentencias ciudadanas obtuvieron un puntaje de cero.

Como uno de los propóstitos de esta aplicación es identificar si hay algunas variables sociodemográficas que estén relacionadas con el el nivel de desempeño en cada competencia, se procedió a contruir tablas de contingencia, donde se cruzó nivel de desempeño en cada competencia con las variables géro, estrato, estado civil, trabaja, personas a cargo, hogar actual, título de bachillerato y cabeza de familia, obteniendo un total de 40 tablas de contingencia.

Desempeño	Variable	Valor Chi-cuadrado	p-valor	coeficientes	
				Cramer	Contingencia
Comunicación escrita	Género	1,585	0.662	0.053	0.053
	Estrato	5,301	0.505	0.069	0.097
	Estado civil	1,893	0.929	0.041	0.058
	Trabaja	4,51	0.341	0.064	0.09
	Personas a cargo	9,363	0.404	0.075	0.128
	Hogar actual	3,63	0.304	0.081	0.08
	Titulo bachillerato	3,237	0.778	0.054	0.076
	Cabeza de flia	1,176	0.758	0.046	0.046
Inglés	Género	4,923	0.085	0.086	0.086
	Estrato	2,381	0.666	0.046	0.065
	Estado civil	5,614	0.467	0.07	0.099
	Trabaja	1,179	0.881	0.032	0.045
	Personas a cargo	6,744	0.564	0.077	0.108
	Hogar actual	0.830	0.660	0.038	0.038
	Titulo bachillerato	5,114	0.275	0.067	0.094
	Cabeza de flia	10,246	0.016	0.134	0.133
	Género	17,598	0.0001	0.176	0.173
Razonamiento Cuantitativo	Estrato	7,503	0.111	0.081	0.114
	Estado civil	2,17	0.903	0.044	0.062
	Trabaja	7,997	0.091	0.084	0.118
	Personas a cargo	6,978	0.539	0.078	0.11
	Hogar actual	2,247	0.325	0.063	0.063
	Titulo bachillerato	3,706	0.447	0.057	0.08
	Cabeza de flia	7,125	0.028	0.112	0.111

Tabla 9: Estadísticas relacionadas con la pruebas de independencia. Fuente la Autora, 2016

Adicionalmente, en las Tablas 9 y 10 aparece los valores de la estadística de prueba Chi-cuadrado, el p-valor y los coeficientes de Cramer y Contingencia.

Para la realización de las Tablas 9 y 10, se tomó como hipótesis el hecho de que el desempeño en cada competencia era independiente de cada una de las variables con las que se cruzó. En el caso de la comunicación escrita esta no depende de ninguna de las variables de la derecha, ya que el p-valor en todos los casos es mayor que los niveles de significancia (0.01, 0.05, 0.10), por lo que no hay evidencia estadística para rechazar la hipótesis de independencia; esto se verifica con los coeficientes de asociación, porque estos valores son menores que 0.3, lo cual indica que no hay asociación entre las variables. Se observa una situación similar con el desempeño en la prueba de inglés, lectura crítica y competencias ciudadanas.

Desempeño	Variable	Valor Chi-cuadrado	p-valor	coeficientes	
				Cramer	Contingencia
Lectura Crítica	Género	2,981	0.225	0.072	0.072
	Estrato	5,061	0.281	0.067	0.094
	Estado civil	4,999	0.543	0.066	0.093
	Trabaja	1,349	0.853	0.034	0.049
	Personas a cargo	9,985	0.266	0.094	0.131
	Hogar actual	0.985	0.611	0.042	0.042
	Titulo bachillerato	2,619	0.623	0.048	0.068
	Cabeza de flia	5,738	0.056	0.1	0.1
Competencias ciudadanas	Género	2,981	0.225	0.072	0.072
	Estrato	5,061	0.281	0.067	0.094
	Estado civil	4,999	0.543	0.066	0.093
	Trabaja	1,559	0.816	0.037	0.052
	Personas a cargo	9,985	0.266	0.094	0.131
	Hogar actual	0.985	0.611	0.042	0.042
	Titulo bachillerato	2,619	0.623	0.048	0.068
	Cabeza de flia	5,738	0.056	0.1	0.1

Tabla 10: Continuación estadísticas relacionadas con la pruebas de independencia. Fuente la Autora, 2016

Para el caso de razonamiento cuantitativo, se evidencia que con la variable género, se dió un p-valor bastante pequeño, y a la vez los coeficientes de asociación son de 0.17, medianamente cercanos a 0.3, lo que lleva a pensar que el desempeño en esta competencia si depende del género de quien la responde.

5. Conclusiones

Se podría pensar que las personas que optan por un proceso de formación a distancia, tienen algún tipo de obligación por la cual requieren de trabajar para subsistir y de igual forma estudiar; sin embargo en este caso, los estudiantes de la Licenciatura en Básica con énfasis en matemáticas, humanidades y lengua castellana son en gran parte personas relativamente jóvenes entre 20 y 30 años, el 52.14 % son solteros, de igual forma la gran mayoría no tiene responsabiliades de familia y a su vez una buena parte de los estudiantes trabajan sin obtener una remuneración, lo que indica que dependen económicamente de sus padres o cuentan con otra fuente de ingresos.

El 61.22 % de las familias de los estudiantes se encuentran en el estrato 2 (socioeconómico), a su vez, el 67.33 % manifestaron que los pisos predominantes en sus viviendas eran baldosa, tableta, vinilo o ladrillo, y el 54.73 % respondió que los ingresos familiares son entre 1 y 2 salarios mínimos, sin embargo, la gran mayoría cuenta con las herramientas indispensables en la modalidad de educación a distancia, que son el computador e internet; lo cual ratifica la intención de quere salir adelante, teniendo en cuenta que más del 30 % de los estudiantes (porporción más alta comparada con las demás) manifestaron que el nivel educativo alcanzado por sus padres fue la primaria incompleta.

A pesar de que la prueba se considera importante porque de alguna manera refleja lo aprendido en el proceso de formación, hay estudiantes que quizá no le dan la importancia suficiente, ya que no respondieron en competencias como Inglés, Razonamiento Cuantitativo , Lectura Crítica y Competencias Ciudadanas, lo que generó puntajes iguales a cero.

No existe asociación lineal entre los puntajes obtenidos en cada comptencia, incluso los que posiblemente se podrían relacionar, no tienen un nivel de asociación lo suficientemente fuerte.

El género no influye en el desempeño de las pruebas de comunicación escrita, inglés, lectura crítica y competencias ciudadanas, pero en lo que tiene que ver con el razonamiento cuantitativo, el desempeño en esta prueba si depende del género de quien la está presentando.

Referencias Bibliográficas

Aguerrondo, I. (1993), 'La calidad de la educación: ejes para su definición y evaluaci ón', *Revista Interamericana de desarrollo educativo* **37**(116), 561.
*https://www.researchgate.net/profile/Ines_Aguerrondo2/publication/44818477_La _Calidad_de_la _educacin_ejes_para_su_definicin_y _evaluacin/links/53f518c90cf2fceacc6f2e70.pdf

Alfonso Sánchez, I. (2003), 'Educación a distancia', *Acimed* .

Alvear, M. A. (2014), 'Saber 11 y saber pro: dos pruebas estresantes', *El Universal* .
*http://www.eluniversal.com.co/educacion/saber-11-y-saber-pro-dos-pruebas-estresantes-177367

Díaz M., Guillermo, L. y Morales R., M. (2009), *Análisis estadístico de datos categóricos*, primera edn, Universidad Nacional de Colombia.

Dimas Márques, S. S. (2014), 'Concepto educación'.
*http://repository.uaeh.edu.mx/bitstream/handle/123456789/16615

Hernández R., F. C. y Baptista, L. (2010), *Metodología de la investigación*, Mc Graw Hill.

ICFES (2013), 'Guía de acceso a bases de datos icfes'.

ICFES (2015*a*), 'Guía de orientación. módulo de competencias ciudadanas. saber pro 2015-2'.

ICFES (2015*b*), 'Guía de orientación. módulo de lectura crítica. saber pro 2015-2'.

ICFES (2015*c*), 'Guías. módulo de comubicación escrita. saber pro 2015-2'.

ICFES (2015*d*), 'Guías. módulo de inglés. saber pro 2015-2'.

ICFES (n.d.), 'Turorial carga de bases de datos en excel'.

Para qué sirven las pruebas saber pro (2012), Technical report.
*http://www.elespectador.com/noticias/actualidad/vivir/sirven-pruebas-saber-pro-articulo-352085

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Rodriguez Albor, G., Gómez Lorduy, V. y Ariza Dau, M. (2014), 'Calidad de la educación superior a distancia y virtual: un análisis de desempeño académico en colombia', *Investigación y desarrollo* . *http://biblio.uptc.edu.co:2086/ehost/detail/detail?vid=21/sid=5ed22875-5dfe-4ebf-aa6d-e8bd9650825-40sessionmgr115-hid=107-bdata-Jmxhbmc9ZX-Mmc2l0ZT1laG9zdC1saXZl-AN-95806031-db-aph

SERIES DE PRECIPITACIÓN PLUVIOMÉTRICA EN EL MUNICIPIO DE TOTA

Especialización en Estadística

Wilmer Antonio Martinez Suancha[1,a], Álvaro Calvache Archila[2,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia

Resumen

El estudio de las series de precipitación pluviómetrica es muy importante para tomar decisiones y tomar acciones en el manejo de las políticas del recurso hidríco; Para hacer cualquier estudio a dichas series, se requiere de la creación y reconstrucción de las series, seguido de un análisis de control de calidad de los registros y observaciones, junto con pruebas de homogeneidad. Luego se hizo un llenados de datos faltantes por medio del método de cociente normal, posteriormente un análisis descriptivo de las series.

***Palabras clave*:** Escorrentía, precipitación, homogeneidad, sequía, serie de tiempo.

Abstract

The study of the series of precipitation is very important to take decisions and to take actions in the managing of the policies of the resource hidríco; to do any study to the above mentioned series, it is needed of the creation and reconstruction of the series, followed by an analysis of quality control of the records and observations, together with tests of homogeneity. Then there were done fillings of lacking information by means of the method of normal quotient, later a descriptive analysis of the series.

***Key words*:** surface run-off, precipitation, homogeneity, Sequa, Series of time.

1. Introducción

Los eventos de sequía, han sido devastadores en los últimos años, en algunas regiones de Colombia se han perdido centenares de hectáreas de bosque por incendios ocasionados, algunos por las fuertes oleadas de calor y falta de lluvias; la disminución en la escorrentía por aumento en la demanda hidríca (Aguas 2013) y la sequía, que en departamentos como Boyacá, se llevó a declarar calamidad publica en los primeros meses de 2016, además de que muchas personas no cuentan con acceso a agua potable,[1] estos problemas para los que se estima un panorama muy seguramente peor en los próximos años (IANAS 2012), preocupan y hacen urgente tomar acciones que prevengan y busquen dar solución. Como señala el Instituto de Hidrología, Metereología y Estudios Ambientales - IDEAM, "Los crecimientos en el consumo, la deforestación y la escasa gestión sobre las cuencas y los acuíferos u otros recursos naturales, en conjunto con la ausencia casi total de tratamiento de aguas residuales están causando problemas serios de disponibilidad, limitaciones por calidad, desabastecimiento y racionamiento en un número cada vez mayor de municipios del país. Estos efectos adversos

[a]Estudiante de Especialización en Estadística. E-mail: wilmer.martinez@uptc.edu.co

[b]Profesor Titular. E-mail: acalvachea@gmail.com

[1]La Organización de las Naciones Unidas - ONU afirma que: "A pesar de los importantes logros alcanzados bajo los Objetivos de Desarrollo del Milenio, 800 millones de personas aún no tienen acceso a una fuente de agua mejorada, mientras que millones más carecen de acceso a agua limpia y segura. Se calcula que hasta 1.800 millones de personas viven en países que enfrentarán escasez de agua para 2025"

sobre la calidad de vida y las actividades económicas, requerirán de lineamientos de políticas y estrategias para la planificación del manejo integral y sostenible del recurso hídrico (superficial y subterráneo)" (Aguas 2013).

"El agua se encuentra disponible en diferentes fuentes: la humedad del suelo, aguas subterráneas, en forma de nieve, corrientes superficiales o reservoriós. Los impactos ocasionados por eventos de sequía, se pueden manifestar en la disminución de la oferta hídrica" (*IDEAM, Estudio Nacional del Agua 2014. Bogotá, D.C., 496 páginas.* n.d.).

Conocer el comportamiento de ocurrencia y variación de las fuentes hídricas, es de los primeros pasos en búsqueda de las soluciones. Por lo anterior, en este estudio interesa identificar el comportamiento en cuanto a ocurrencia y variación mensual, así como la tendencia anual, de las precipitaciones pluviométricas en el municipio de Tota, para los últimos treinta años. Estos resultados serán necesarios para ser utilizados en un próximo estudio del balance hidríco, de las aguas subterráneas en el municipio de Tota, las cuales se recargan principalmente de las precipitaciones por medio de la escorrentía (Michael 2003). Este estudio se requiere para la búsqueda de nuevas fuentes de abastecimiento hídrico, ya que "Las aguas subterráneas constituyen importantes reservas de agua dulce con una menor susceptibilidad a procesos de contaminación y degradación en comparación con las fuentes superficiales. De ahí la importancia de conocer su ocurrencia, distribución y principales características hidráulicas, hidrológicas e hidrogeoquímicas para una gestión adecuada y sostenible del recurso" (*IDEAM, Estudio Nacional del Agua 2014. Bogotá, D.C., 496 páginas.* n.d.):

> La precipitación pluvial, en adelante precipitación, es una variable climatológica de gran importancia en la hidrología, ya que es insumo fundamental para el cálculo de balances hídricos; la generación de alertas tempranas por riesgo de inundación o sequía; el estudio del régimen ambiental de caudales; y el diseño de sistemas de acueducto y alcantarillado, por mencionar las más importantes. La medición de la precipitación tiene como propósito conocer su distribución en el espacio y en el tiempo. En la escala temporal, el registro de precipitación se estudia como una serie de tiempo horaria, diaria, mensual o anual, y su utilidad dependerá, en gran medida, de su completud (Organización Meteorológica Mundial, citado por: (León 2015)

La posible acentuación de periodos de poca precipitación al lado del aumento en la temperatura, puede traer consigo grandes impactos en el medio ambiente y consecuencias fuertes en las actividades socioeconómicas, especialmente en la agricultura (Víctor, Arturo y Ramón. 2014), actividad muy importante en la economía de muchos municipios de Colombia, como es el caso de Tota, donde la agricultura es de las principales fuentes económicas, si no la más importante.

En Tota, las principales fuentes de abastecimiento hidríco son, la laguna de Tota y las quebradas(como se indica en la sección 1), las cuales se abastecen de la escorrentía provocada por las precipitaciones y descargas de aguas subterráneas (Michael 2003).

Existen estudios sobre la precipitación en la zona andina como (Cortés Moya Diego Ernesto, na Ocampo William Alexander y Fernando 2012), donde indica que en la zona central de Colombia se presenta un patrón bimodal de lluvias, que se caracteriza por tener dos periodos lluviosos en el año, los cuales están intercalados por un periodo seco, en (IDEAM y de Modelación Subdirección de Hidrologia 2014) también se indica de este comportamiento y más precisamente, en la cuenca del lago de Tota. Se hace un análisis de algunas series de precipitación para un estudio hidrológico del año 2014 y se concluye que en los meses de junio y agosto se presentan las precipitaciones máximas. También en (CORPOBOYACA y PUJ 2005), se encuentra un análisis de las precipitaciones de la cuenca, para el año 2004 que utilizan para hacer un diagnóstico del recurso hídrico en la cuenca de la laguna de Tota, se evidencia para ese año, que en los meses de julio y agosto se presentan la precipitaciones máximas, pero que contrario a lo que dice el estudio de Cortés Moya Diego Ernesto, na Ocampo William Alexander y Fernando (2012), para el año 2004 se encuentra un comportamiento monomodal de las precipitaciones. (Ver Figura 1). Lo que puede sugerir que, no en todas las estaciones de la cuenca de la laguna de Tota, se presenta el mismo comportamiento de las precipitaciones, punto muy interesante en identificar para las precipitaciones en el municipio de Tota.

Los objetivos que nos ocupan son entonces:

- Recopilar y reconstruir las series con datos, lo más completo posible, de precipitaciones para la región del municipio de Tota, que permitan el estudio de la ocurrencia y distribución, mensual y anual; así como la variabilidad y tendencia multianual de dichas series de precipitación.

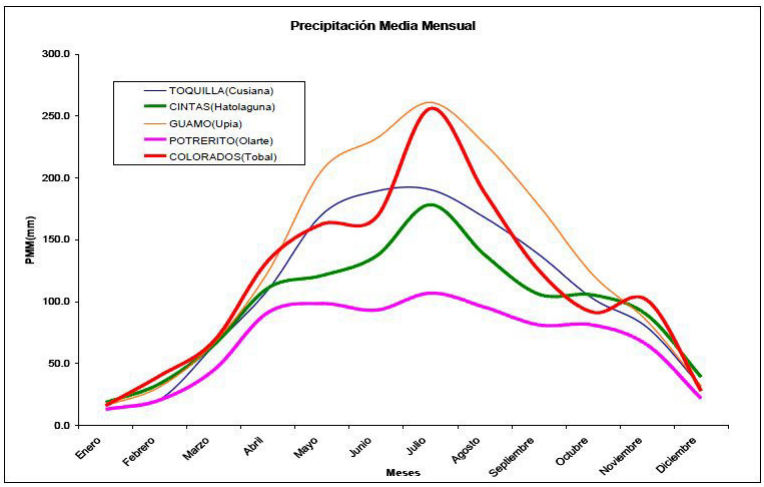


FIGURA 1: Precipitaciones medias mensuales, en estaciones de la cuenca hidrográfica de la Laguna de Tota

- Identificar la distribución mensual de las precipitaciones en el municipio de Tota, para cada una de las series y así conocer los meses en que se presentan los periodos de invierno.
- Identificar la variabilidad y tendencia multianual, de las precipitaciones en cada serie que tiene influencia en el municipio de Tota, para conocer si han aumentado, disminuido o si por el contrario no hay tendencia, en las precipitaciones en los últimos treinta años.
- Identificar el comportamiento medio mensual y anual de las precipitaciones de la región del municipio de Tota.

2. Referente Conceptual

2.1. Sequía

Es de importancia aclarar que no existe un consenso en la definición de la palabra "sequía", dado que el clima seco afecta a las personas de forma distinta.[2]

2.2. Precipitación

Este fenómeno se define como la "fase del ciclo hidrológico que da origen a corrientes de aguas superficiales y profundas. La cantidad de precipitación, depende de variables como la altura, la humedad del aire y la velocidad vertical del mism"(Maderey, 2005 citado por: Cortés Moya Diego Ernesto, na Ocampo William Alexander y Fernando (2012))

2.3. Escorrentía

Según Michael (2003): En hidrología la escorrentía hace referencia a la lámina de agua que circula sobre la superficie en una cuenca de drenaje.

[2]El Servicio Metereológico Británico define sequía como un periodo con al menos quince días consecutivos sin lluvia medible; para su propósito, "medible" significa más de 0.25 milímetros (mms). En México, la comisión Nacional del Plan Hidráulico define una sequía como "un fenómeno metereológico que ocurre cuando la precipitación o el escurrimiento de agua natural de un periodo es menor que su valor normal y cuando esta deficiencia es lo suficientemente grande y prolongada como para dañar las actividades humanas".

2.4. Series de tiempo

Perpiñán Lamigueiro (2014) señalan que:

> Una serie de tiempo es una secuencia de observaciones certificadas en instantes de tiempo consecutivos. Cuando estos instantes de tiempo uniformemente son espaciados, llaman la distancia entre ellos el intervalo de muestreo. La visualización de serie de tiempo es requerida para revelar los cambios de una o varias variables cuantitativas durante el tiempo, y mostrar las relaciones entre las variables y su evolución durante el tiempo.

2.4.1. Clasificaciones de las series temporales

Una serie temporal puede ser discreta o continua dependiendo de cómo sean las observaciones.

Si se pueden predecir exactamente los valores, se dice que las series son determinísticas.

Si el futuro sólo se puede determinar de modo parcial, por las observaciones pasadas y no se pueden determinar exactamente, se considera que los futuros valores tienen una distribución de probabilidad que está condicionada a los valores pasados. Las series son así estocásticas.

Las componentes principales de una serie temporal, son:

a. Si los datos presentan forma creciente o decreciente (tendencia).

b. Si existe influencia de ciertos periodos, de cualquier unidad de tiempo (estacionalidad).

c. si los datos presentan variación aleatoria a corto plazo(Irregular)

El estudio descriptivo de series temporales se basa en la idea de descomponer la variación de una serie en varias componentes básicas. Este enfoque no siempre resulta ser el más adecuado, pero es interesante cuando en la serie se observa cierta tendencia o cierta periodicidad. Hay que resaltar que esta descomposición no es en general única.

Este enfoque descriptivo consiste en encontrar componentes que correspondan a una tendencia a largo plazo, un comportamiento estacional y una parte aleatoria.

Según Alvaro (2007), las componentes o fuentes de variación que se consideran habitualmente son las siguientes:

a. Tendencia: Se puede definir como un cambio a largo plazo que se produce en relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

b. Efecto Estacional: Muchas series temporales presentan cierta periodicidad o dicho de otro modo, variación de cierto periodo (anual, mensual ...). Por ejemplo, el paro laboral aumenta en general en invierno y disminuye en verano. Estos tipos de efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar del conjunto de los datos, desestacionalizando la serie original.

c. Componente Aleatoria: Una vez identificados los componentes anteriores y después de haberlos eliminado, persisten unos valores que son aleatorios. Se pretende estudiar qué tipo de comportamiento aleatorio presentan estos residuos, utilizando algún tipo de modelo probabilístico que los describa.

Las dos primeras componentes son determinísticas y la tercera es de tipo aleatorio. Un modelo de serie de tiempo se puede modelar con la siguiente fórmula:

$$X_t = T_t + E_t + I_t \tag{1}$$

donde T_t es la tendencia, E_t es la componente estacional, que constituyen la señal o parte determinística, e I_t es el ruido o parte aleatoria. Es necesario aislar de alguna manera la componente aleatoria y estudiar qué modelo probabilístico, es el más adecuado. Conocido este, podremos conocer el comportamiento de la serie a largo plazo. Esto será motivo de estudio en Inferencia Estadística. El aislamiento de la componente aleatoria se suele abordar de dos maneras:

- Enfoque descriptivo: Se estima T_t y E_t y se obtiene I_t como:

$$I_t = X_t - T_t - E_t \tag{2}$$

- Enfoque de Box-Jenkins: Se elimina de X_t la tendencia y la parte estacional (mediante transformaciones o filtros) y queda sólo la parte probabilística. A esta última parte se le ajustan modelos paramétricos.

3. Metodología

En el desarrollo de esta aplicación se utilizó el enfoque cuantitativo; la información recopilada en cuanto a operaciones e indicadores propios (Precipitación pluvial).

3.1. Área de estudio

El municipio de Tota, se encuentra localizado sobre la cordillera Oriental; a una altura de 2870 msnm, con una temperatura promedio de 12 grados Centigrados, presenta un clima frío y húmedo, dada su cercanía a la Laguna de Tota.

Su extensión territorial es de 314 Km^2 repartidos en 10 veredas (Tota, Ranchería, Toquechá, Romero, Sunguvita, Corrales, La Puerta, Guáquira, El Tobal y Daisy). Sus fuentes hídricas son la Laguna de Tota, río Tota, las quebradas de Aguaná, Guacható, Ochiná, el Común, quebrada Verde, El Caimán y Tota,(Ver Figura 2).

FIGURA 2: Mapa fisico del municipio de Tota, incluye la zonas donde se encuentran las estaciones meteorológicas mas cercanas al municipio. Fuente Googlemaps.

3.2. Bases de datos

Se recopilaron las series diarias de precipitación pluvial en estaciones más cercanas a la región en donde se encuentra el municipio de Tota, procedentes de registros en formato digital suministrados por el IDEAM; se obtuvieron en un comienzo registros de seis estaciones meteorológicas, de las cuales se descartaron tres, debido a que se encuentran suspendidas desde finales de la década de los 90. Con las tres estaciones restantes que se tuvieron en cuenta, se trabajaron observaciones desde enero de 1984 hasta diciembre de 2014, periodo

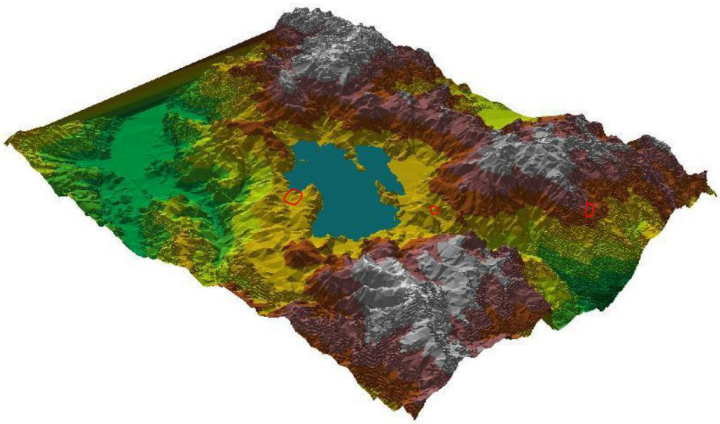


FIGURA 3: Modelo digital de elevación, en rojo la localización de las estaciones en uso. Fuente:Diagnostico del Recurso Hídrico-PLAN DE ORDENACIÓN Y MANEJO DE LA CUENCA DEL LAGO DE TOTA

para el cual, las estaciones seleccionadas tienen registros, (Ver Figura3). Lo anterior se hizo, teniendo en cuenta que: "La Organización Meteorológica Mundial (OMM, World Meteorological Organization, WMO) establece, de forma general, un periodo mínimo de 30 años como la longitud recomendable que han de tener los registros meteorológicos para que su media y demás índices estadísticos tengan significación climática" (Martín-Vide, 2003 citado por:Víctor, Arturo y Ramón. (2014, pág.84))

Sin embargo se debe reconstruir la base de datos debido a que para algunos periodos de tiempo hay datos faltantes en las series, lo cual es normal y se dan en general por falta de lectura debido a diversas causas (León 2015), "si el emplazamiento de las estaciones meteorológicas es próximo, las diferencias en las cantidades de precipitación mensual son normalmente muy pequeñas, salvo excepciones debidas a un terreno muy complejo" (Luna et al., 2012-citado por: Víctor, Arturo y Ramón. (2014, pág.84)).

ID	Estación	Huecos %	Longitud	Latitud	Altitud(msnm)
1	Potrerito	6.25	-72.949	5.477583	3047
2	Tunel	16.93	-72.948	5.574914	3047
3	Guamo	9.64	-72.917	5.366667	2575

TABLA 1: Información básica de las tres series de precipitación reconstruidas: ID, Nombre, porcentaje de faltantes, coordenadas geográficas (en grados) y altitud.

Una vez se obtiene la base de datos depurada y organizada, de las 3 series de precipitación descritas anteriormente, se hace la comprobación de la calidad y homogeneidad de la series.

3.3. Control de calidad

Se construyeron las tres series bajo las condiciones descritas anteriormente, luego del control de calidad y llenado de datos faltantes, se tuvieron 11.566 registros diarios por cada una de las series. Al organizarlas mensualmente, se obtuvieron 372 observaciones por cada una de ellas (Tabla 1).

El control de la calidad en el análisis de las series climatológicas es un aspecto muy importante y consiste en generar criterios y/o filtros para ayudar a identificar datos no razonables y/o erróneos, como sugieren (Lizeth y David 2014).

Posteriormente se realiza una validación para las diferentes estaciones en estudio, la cual consta del cálculo de varios criterios que ayudan a identificar datos atípicos y/o erróneos para su posterior corrección. Los datos identificados en esta sección son reemplazados por NA's. La última columna de la Tabla 2, indica el % total

de datos faltantes, que serán llenados en la sección datos faltantes. Se recomienda que estos, no superen el 20 % en cada una de las estaciones.

Estación	% Datos atípicos	% Datos fuera de rango	% Total datos NA
Potrerito	2.43	0	6.25
Guamo	2.1	0	9.64
Tunel	1.94	0.01	16.93

Tabla 2: Resumen control de calidad para las series de precipitación.

Se encuentra que la base de datos efectivamente tiene datos faltantes y algunos atípicos. Ninguna de las estaciones supera el 20 % de faltantes, por lo que los podemos imputar y se escoge el método de cociente normal, cuya fórmula es la siguiente:

$$D_j = \frac{a_j p_A + b_j p_B}{2}, \tag{3}$$

donde:

$$\begin{aligned} p_A =& \frac{\text{Precipitacion anual estación dato faltante}}{\text{Precipitacion anual estación referencia A}}, \\ p_B =& \frac{\text{Precipitacion anual estación dato faltante}}{\text{Precipitacion anual estación referencia B}}, \\ D_j =& \text{Precipitación estimada para el día j.} \\ a_j, b_j =& \text{Precipitación registrada en las dos estaciones de referencia el día j.} \end{aligned}$$

Aunque existen diversos métodos para la imputación y estimación de datos faltantes en series, escogimos el de cociente normal guiados por el trabajo de DARÍO (2008), en donde se hace una revisión de los diferentes métodos. Para utilizar este método es necesario tener datos de varias estaciones para el mismo periodo y con buena correlación, lo cual en este estudio se cumplen dichas condiciones.

Estación	Potrerito	Guamo	Tunel
Potrerito	1	0.706	0.7629
Guamo	0.706	1	0.481
Tunel	0.7629	0.481	1

Tabla 3: Correlacion entre las series

De acuerdo con Víctor, Arturo y Ramón. (2014), una serie de tiempo es homogénea cuando sus variaciones son naturales y no han tenido intervención humana, ya que para una serie de datos no homogénea, no será fiable ningún resultado de su análisis. Entonces, el "primer paso necesario para analizar las bases de datos mensuales de las series de precipitación reconstruidas debe consistir en evaluar su calidad y homogeneidad temporal" (González Hidalgo et al., 2002 citado por: Víctor, Arturo y Ramón. (2014, pág.87)).

3.4. Series y análisis de las series

Con la ayuda del software R, se calcula el total de precipitación anual y mensual en mms, así también estadísticos como la media, mediana, cuartiles, valores máximos y mínimos para cada serie.

4. Resultados

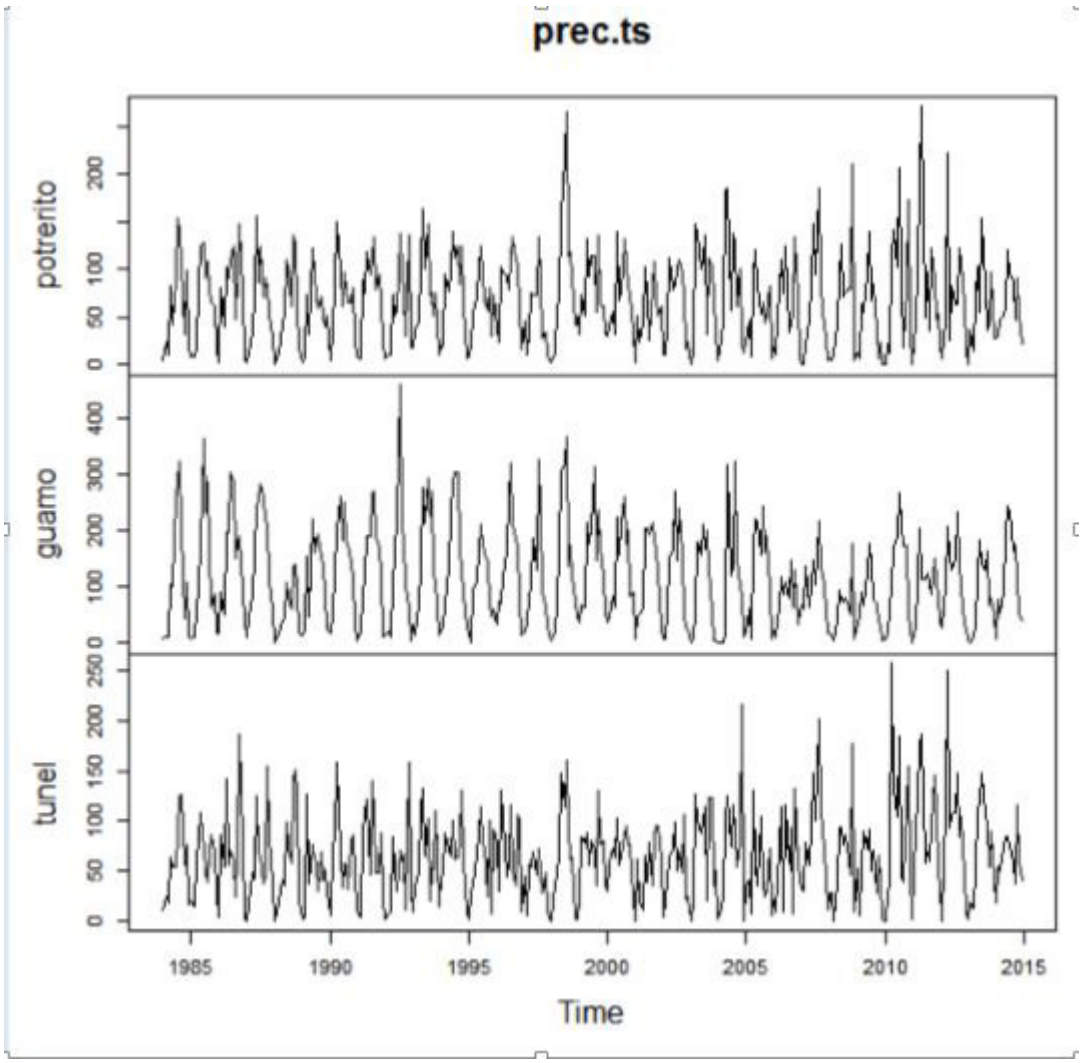


Figura 4: Series de precipitación mensuales, para las tres estaciones seleccionadas

El promedio de precipitaciones anuales para la estacion potrerito es de 833.54 mms, para la estacion tunel es de 785.17 mms y para la estacion guamo es de 1421.26 mms.

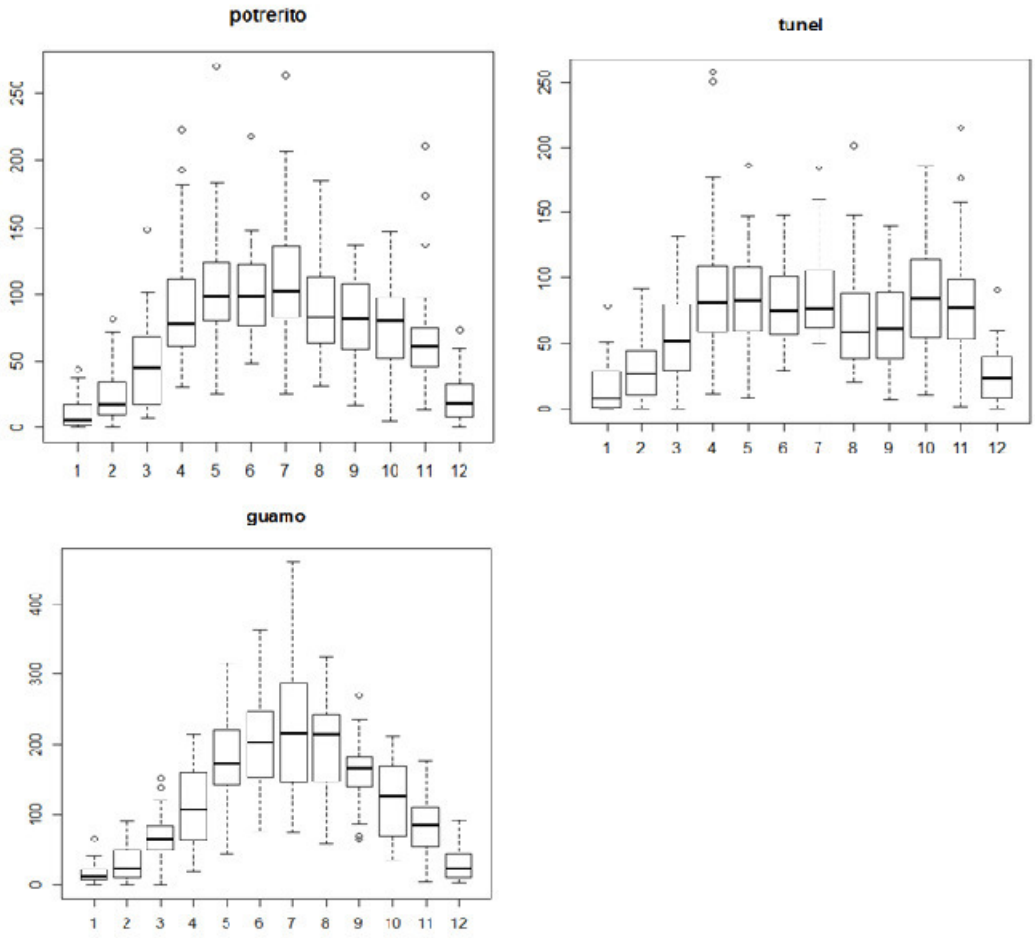


FIGURA 5: Distribución promedio mensual de precipitaciones en las 3 estaciones.

Estación	Mínimo	Máximo	Medio
Potrerito	591.9	1199.75	833.54
Guamo	700.38	1945.9	1421.26
Tunel	486.4	1164.75	785.17

TABLA 4: Valores promedio de precipitación anual en mms

Se evidencia que la distribución de las precipitaciones no es igual para las 3 series (Ver Figura 5), los cual sugiere que el comportamiento no es uniforme en la región del municipio de Tota. la serie de estación potrerito no presenta un patrón monomodal muy marcado, pero definitivamente en acuerdo con el estudio de (CORPOBOYACA y PUJ 2005), la estación potrerito presenta sus precipitaciones máximas para los meses de mayo y julio, ver figura: 1, además coincide también la distribución media mensual, lo cual es muy importante y nos indica que nuestros datos son fiables.

En la estación túnel, observamos un patrón bimodal en la ocurrencia de las precipitaciones altas, las cuales se presentan por un lado en los meses de abril y mayo; luego se vuelven a presentan grandes lluvias en octubre y comienzos de noviembre, cuando finalmente se presenta un periodo de verano muy marcado.

Para la estación guamo, se presenta un indudable patrón monomodal, muy marcado de precipitaciones intensas, especialmente en los meses de junio, julio y parte de agosto. Si vemos las tres distribuciones y las

relacionamos, notamos que a pesar de tener patrones diferentes, las lluvias mas fuertes se concentran entre los meses de abril a octubre.

La serie de la estación potrerito es la que menos presenta sospechas de in-homogeneidad, y la que tiene mas datos originales, ademas es la estación mas cercana al municipio de Tota, como podemos observar en la figura

5. Conclusiones

La preocupación por la escasez de abastecimiento hidríco en algunas zonas del municipio de Tota, implica la necesidad analizar de forma concisa la variabilidad mensual y la tendencia de las precipitaciones. Con ese fin se recopilo y reconstruyo una bases de datos completa, con 3 series de precipitaciones diarias desde el año 1984 hasta 2014, procedentes de registros en formato plano que suministro el IDEAM.

Según el análisis de las series de precipitación, la media anual de precipitaciones en Tota, no supera los 1430 mms, lo cual coloca a Tota en el 22 % del territorio Colombiano que registra lluvias anuales inferiores a 2000 mms, ya que según (Ingeominas 2002) cerca del 88 % del territorio tiene precipitaciones por encima de los 2000 mms, con promedio anual cercano a los 3000 mms.

Los meses en los cuales es mas probable que se recarguen las aguas subterráneas, estarían entre abril a octubre, ya que para esos meses se encuentra estacionalidad de las series, cuando tiene sus valores mas altos de precipitaciones, es decir seria la época de fuertes lluvias.

No se encuentra una tendencia en aumento o disminución de las series de precipitación, para el municipio de Tota, lo cual sugiere que a pesar del cambio climático, este no han influido en el comportamiento de las precipitaciones en Tota. Sin embargo es recomendable hacer un análisis mas minucioso a los componentes de las series.

Al parecer el comportamiento de las precipitaciones, no es la causa de que las fuentes de agua superficial en Tota, vengan con una disminución del caudal para los últimos años.

Aun que las precipitaciones promedio anuales en Tota, son bajas en comparación con otras regiones del país, estas tienen picos altos para algunos meses, lo cual puede ser causa de deslizamientos en épocas de lluvia.

Se debe hacer un estudio que compare el comportamiento de las series de precipitación con las temperaturas máximas, mínimas, y con la evaporatranspiración para el Tota, de esa manera tener un mejor panorama del comportamiento de las lluvias en esa región

Referencias Bibliográficas

Aguas, I. (2013), *Aguas Subterráneas en Colombia: una Visión General*, IDEAM Bogotá D.C., 2013. 284 páginas.

Alvaro, M. G. (2007), *Series de Tiempo*.

CORPOBOYACA y PUJ (2005), 'Diagnostico del recurso hídrico-plan de ordenación y manejo de la cuenca del lago de tota'.

Cortés Moya Diego Ernesto, ., na Ocampo William Alexander, P. y Fernando, S. G. L. (2012), 'Modelamiento espaciotemporal de la precipitación total mensual, para el año 2007, en la zona andina colombiana'.

DARÍO, M. R. R. (2008), 'Estimación estadística de valores faltantes en series históricas de lluvia'.

IANAS (2012), *DIAGNÓSTICO DEL AGUA EN LAS AMÉRICAS*, Derechos Reservados FCCyT, ISBN: 978-607-9217-04-4. 448 páginas. México.

IDEAM y de Modelación Subdirección de Hidrologia, G. (2014), 'Informe batimetria lago de tota'.

IDEAM, Estudio Nacional del Agua 2014. Bogotá, D.C., 496 páginas. (n.d.).

Ingeominas (2002), *GUÍA METODOLÓGICA PARA FORMULAR PROYECTOS DE PROTECCIÓN INTEGRADA DE AGUAS SUBTERRÁNEAS*.

León, G. R. P. (2015), 'Imputación de datos en series de precipitación diaria caso de estudio cuenca del río quindío'.

Lizeth, L. H. y David, A. (2014), *Clima y Sector Agropecuario Colombiano, Adaptación para la sostenibilidad productiva*, MinAgricultura, convenio: CIAT-MADR.

Michael, P. (2003), *Agua Subterránea*, Primera Edición México.

Perpiñán Lamigueiro, O. (2014), 'Displaying time series, spatial, and space-time data with r oscar perpiñan'.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

S.P.Paul, Cowpertwait y Metcalfe, A. V. (2011), *Introductory Time Series with R*.

Víctor, R. A., Arturo, S.-L. y Ramón., G. M. (2014), 'Creación de una base de datos con series largas de precipitación en la región de murcia y análisis temporal de la serie media anual, 1914-2013'.

Apéndice A

mes	Mínimo	Máximo	Medio
1	0	78.5	16.59
2	0	92.7	31.77
3	0	131.4	58.67
4	12	257.2	93.33
5	9.23	186.3	86.37
6	30.3	147.5	78.59
7	49.7	184.1	87.0
8	20.9	201.2	68.78
9	6.83	139.8	66.6
10	11.1	186	89.29
11	2.2	215.7	81.24
12	0	91.25	26.08

Tabla 5: Valores de precipitación mensual en mms, para la series Tunel

mes	Mínimo	Máximo	Medio
1	0	44.8	11.6
2	0	81.6	23.72
3	7.5	148.5	49.07
4	31.7	222.7	91.78
5	26	170.6	106.59
6	48.7	218.1	100.32
7	26.2	264.75	109.63
8	32.5		68.78
9	6.83	139.8	66.6
10	11.1	186	89.29
11	2.2	215.7	81.24
12	0	91.25	26.08

Tabla 6: Valores de precipitación mensual en mms, para la series Potrerito

APLICACIÓN DE UN MODELO DE SOBREVIDA PARA TIEMPOS DE FALLA DE GENERADORES ELÉCTRICOS

Especialización en Estadística

Maryluz Castro Moreno[1,a], Carmen Helena Cepeda Araque[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Este trabajo determina los factores asociados a los tiempos de falla de generadores eléctricos y la función de sobrevida. Se utilizó el modelo semiparamétrico de Cox y el estimador de Kaplan- Meier, utilizando los datos suministrados por el proveedor aggreko correspondientes al primer trimestre de 2016. Se identifica una razón de riesgo consistente con la revisión teórica realizada.

***Palabras clave*:** Modelo de sobrevida, estimador Kaplan- Meier, modelo de Cox, tiempos de falla de generadores eléctricos..

Abstract

This work determines the times associated with failure of electrical generators and factors survival function. The semi-parametric Cox model and Kaplan-Meier estimator was used.

***Key words*:** Survival model,Kaplan- Meier estimator, Cox model, electric generators.

1. Introducción

El uso del modelo de sobrevida para modelar tiempos de falla en componentes eléctricos y los factores asociados a que se presente la falla, es un método estadístico que permite tener en cuenta la presencia de la censura en los datos. Los modelos de sobrevida se han utilizado principalmente en estudios clínicos.

El objetivo es determinar a través de un modelo de sobrevida los factores que inciden en el tiempo hasta que falla un generador eléctrico en el bloque Pacific Rubiales Energy departamento del Meta. Se abordó el modelamiento estadístico, el modelo de sobrevida donde se presenta la función de sobrevida, la función de densidad de probabilidad, la función de riesgo y la vida media residual; de igual forma el estimador de Kaplan-Meier para la función de sobrevida, el modelo de Cox y el juzgamiento, selección y evaluación del mismo. El propósito de la aplicación es dar a conocer a los usuarios de la estadística una herramienta de análisis de dependencia en el caso donde la variable respuesta es el tiempo hasta que ocurre un evento. Además, la información obtenida a través de la aplicación modelo permitá realizar los mantenimientos preventivos de forma y de esta forma el proveedor prestará el servicio de forma más eficiente.

[a]Estudiante de Especialización en Estadística. E-mail: maryluz.castro@uptc.edu.co
[b]Profesor asistente. E-mail: carmen.cepeda@uptc.edu.co

2. Referente Conceptual

El objetivo básico del modelamiento estadístico está en que a partir de la variabilidad observada, se construye un modelo donde la variabilidad sistemática sea mayor a la variabilidad aleatoria, de este modo:

$$\textit{Variable respuesta} = \textit{Componente sistemático} + \textit{Componente aletorio}$$

McCullagh & Nelder(1989), presentan el modelo lineal generalizado (GLM), caracterizado por:

La *variable respuesta*: la cual se asume que tiene la misma forma y pertenece a la familia exponencial de densidades.

El *predictor lineal*: un conjunto de variables explicativas y un vector de parámetros, denotados por $\eta = X\beta$.

Una función de *enlace* $g(.)$ monótona y diferenciable tal que: $\eta_i = g(\mu_i) = {X_i}^T\beta$.

Para ampliar la información sobre modelos lineales generalizados veáse Dobson (2008) & García (2002).

2.1. Modelos de sobrevida

Un tipo importante de datos es el tiempo desde un punto de partida bien definido hasta que se produce algún evento, denominado "falla". En el análisis de sobrevida, el interés se centra en un grupo o varios grupos de individuos para cada uno de los cuales (o del cual) hay un evento puntual definido, llamado falla, que ocurre después de un tiempo, llamado tiempo de falla. Por ejemplo, el tiempo de vida de componentes de máquinas en confiabilidad industrial, la duración de huelgas o periodos de desempleo en economía, los tiempos que toman los sujetos para completar tareas específicas en experimentación psicológica y comúnmente a los tiempos de sobrevida de pacientes en un ensayo clínico.

Para determinar el tiempo de falla de forma precisa, se requiere un tiempo origen que debe ser definido sin ambigüedad, una escala para medir el paso del tiempo que debe ser acorde a las necesidades del estudio y finalmente, el significado de falla debe ser totalmente claro (Godoy 2009, pág. 8).

El análisis de estos datos se centra en resumir las principales características de la distribución, tales como mediana u otros percentiles de tiempo hasta que se presenta la falla, y al examinar los efectos de las variables explicativas. Los datos sobre los tiempos hasta que se presenta la falla o tiempos de sobrevida, presentan dos características importantes:

1. Los tiempos no son negativos y por lo general tienen distribuciones asimétricas con colas largas.
2. Algunas de las observaciones pueden sobrevivir más allá del período de estudio por lo que su tiempo real de falla no se conoce; en este caso, y otros casos en que el tiempo de falla no se conocen completamente, se dice censura y/o truncamiento.

Censura: La censura ocurre cuando se sabe que algunos tiempos de falla han ocurrido solamente dentro de ciertos intervalos y el resto de los tiempos de vida son conocidos exactamente. Hay varias categorías de censura, principalmente censura por la derecha, censura por la izquierda y censura por intervalo.

Censura por la derecha (C_r): el evento es observado solo si este ocurre antes de un tiempo predeterminado. Sea X el tiempo de vida y C_r la censura a derecha, el tiempo exacto de vida de un componente es conocido si $X \leq C_r$ y si $X > C_r$ el componente ha sobrevivido al tiempo y su tiempo de vida es censurado en C_r.

Censura por la izquierda (C_l): un tiempo de vida X asociado con una observación específica en el estudio, se considera censurado por la izquierda, si éste es menor que un tiempo de censura C_l.

Censura por intervalo: se presenta cuando se tiene un estudio longitudinal donde el seguimiento del estado de los individuos se realiza periódicamente y, por tanto la falla sólo puede conocerse entre dos períodos de revisión, generando un intervalo (L_i, R_i).

Truncamiento: condición que presentan ciertas observaciones en el estudio y el investigador no puede considerar su existencia.

Truncamiento por la izquierda: los individuos entran al estudio a una edad particular y son observados desde este tiempo. Sí Y es el momento de ocurrencia del evento que trunca a los individuos, entonces para muestras truncadas por la izquierda, solo los individuos tales que $X \geq Y$ serán consideradas en el estudio. *Truncamiento por la derecha*: los individuos que han presentado el evento de interés son incluidos en la muestra, los que no lo han presentado no se tienen en cuenta.

La Figura 1, ilustra diversas formas de censura. La línea horizontal representa los tiempos de sobrevida de las observaciones. TO y TC son el principio y fin del período de estudio, respectivamente. D significa "falla" y A significa "vivos al final del estudio". L indica que el sujeto era conocido por estar vivo en el momento del estudio, pero se perdió información sobre el mismo al terminar el estudio (Dobson, 2008, pág. 187).

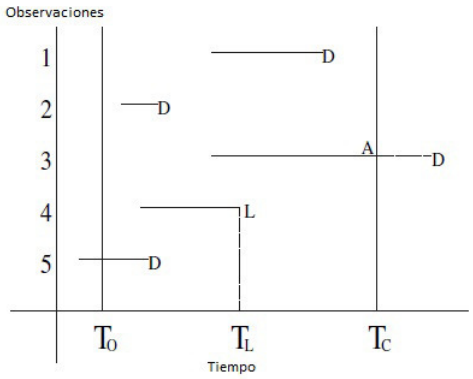


FIGURA 1: Tipos de censura para tiempos de sobrevida

Un modelo de sobrevida se caracteriza por variables aleatorias no negativas, de modo que la variable aletoria T será tomada para denotar el tiempo de falla. La distribución de la variable aleatoria T se expresa a través de la función de sobrevida, la función de densidad de probabilidad, la función de riesgo y la vida media residual.

2.1.1. Función de sobrevida

La función de sobrevida se define como la probabilidad de que una observación sobreviva (no le ocurra el evento de interés) al menos hasta el tiempo t.

Definición: Sea T una variable aleatoria positiva (o no negativa) con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$. La función de sobrevida $S(t)$ es:

$$S(t) = P(T \geqslant t) \tag{1}$$

Donde $S(t)$ es una función monótona no creciente tal que:

$$S(0) = 1 \text{ y } S(t) = 0 \text{ cuando } t \to \infty$$

2.1.2. Función de densidad de probabilidad

El tiempo de sobrevida T tiene una función de densidad de probabilidad, para el caso continuo corresponde a:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du \tag{2}$$

Donde $f(t) = -\frac{dS(t)}{dt}$, que de manera aproximada es la probabilidad de que un evento pueda ocurrir al tiempo t y $f(t)$ una función no negativa con área bajo $f(t)$ igual a uno.

2.1.3. Función de riesgo

La función de razón de riesgos o tasa instantánea de falla $h(t)$ se define como el cociente entre la función de densidad de probabilidad y la función de sobrevida, así:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\ln[S(t)]}{dt} \tag{3}$$

La función de riesgo se interpreta como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt dado que ha sobrevivido hasta el tiempo t. Si $\Delta t \to 0$, $h(t)$ expresa el riesgo instaántaneo de que ocurra el evento en el tiempo t.

2.1.4. Vida media residual

En el análisis de sobrevida la función de *vida media residual* al tiempo t se denota por *mrl(t)*. Para los sujetos de edad t esta función mide la esperanza de tiempo de vida restante, o el tiempo esperado antes de la ocurrencia del evento de interés. Está definida por:

$$mrl(t) = E(T - t\,|T > t\,)$$

Para el caso continuo, por definición de esperanza condicional se tiene que:

$$\begin{aligned} \mathrm{E}(T - t\,|T > t) &= \int_0^{\infty} (u-t) f(u\,|u > t\,)du, \\ &= \int_0^{\infty} (u-t)\frac{f(u)}{S(t)} I_{(t,\infty)}(u)d(u), \\ &= \int_0^{\infty} \frac{(u-t)\,f\,(u)}{S\,(t)}du. \end{aligned}$$

Por lo cual la *vida media residual* al tiempo t queda definida por:

$$mrl(t) = E(T - t\,|T > t) = \int_0^{\infty} \frac{(u-t)\,f(u)}{S(t)}du \tag{4}$$

Se puede apreciar que la *vida media residual* es el área bajo la curva de sobrevida a la derecha de t dividida entre $S(t)$.

De tal modo que la vida media $\mu = E(T) = E(T - 0\,|T > 0\,) = mrl(0)$ es el área total de la curva de sobrevida, es decir:

$$\mu = E\,(T) = \int_0^{\infty} f\,(t)\,dt = \int_0^{\infty} S\,(t)\,dt$$

La relación entre las funciones básicas de sobrevida para la variable aleatoria T en el caso continuo se presenta a continuación(Klein, 1950, pág. 35):

$$S(t) = \int_t^{\infty} f(t)\, dt, \qquad f(t) = \frac{d}{dt} S(t) = h(t)\, S(t),$$

$$h(t) = \frac{d}{dt} \ln [S(t)] = \frac{f(t)}{S(t)}, \qquad mrl(t) = \frac{\int_t^{\infty} S(u)\, du}{S(t)} = \frac{\int_t^{\infty} S(t) dt}{(u-t)\, f(u)}.$$

2.2. Estimador de Kaplan- Meier

Existen modelos paramétricos, es decir, se requiere especificar la distribución de la variable respuesta, para el desarrollo de esta aplicación se aborda el modelo semiparamétrico de Cox, de tal forma que la función de sobrevida es obtenida frecuentemente a través del estimador de Kaplan & Meier (1958) (Colosimo y Giolo 2004, pág. 28), definida por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{5}$$

donde $t_1 < t_2 ... < t_k$ son k tiempos distintos y ordenados de falla, d_j número de fallas en t_j con $j = 1, ..., k$ y n_j número de individuos en riesgo en t_j.

2.3. Modelo de Cox

El procedimiento más utilizado para modelar la relación entre las covariables y la función de sobrevida a datos censurados es el modelo de regresión de Cox (1972). En este modelo, el riesgo para el i-ésimo individuo se define mediante:

$$\lambda_i(t) = \lambda_0(t)\, e^{X_i(t)\beta} \tag{6}$$

donde X_i: Vector de covariables para el i- ésimo individuo en el tiempo t. $\lambda_0(t)$: Función no negativa, arbitraria no especificada. β: Vector de parámetros de la regresión de tamaño $p \times 1$.

Se dice que es un modelo semiparamétrico debido a que incluye una parte paramétrica y otra no paramétrica: La parte paramétrica es $r_i = e^{X_i(t)\beta}$, llamada puntaje de riesgo (risk score). La parte no paramétrica es $\lambda_0(t)$, llamada función de riesgo base. El modelo de regresión de Cox es también llamado modelo de riesgos proporcionales debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante sobre el tiempo. Luego, la razón de riesgo para dos observaciones con vector de covariables X_i y X_j corresponde a:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\, e^{X_i\beta}}{\lambda_0(t)\, e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} \tag{7}$$

El vector de parámetros para un modelo de sobrevida se estima a partir de la función de máxima verosimilitud parcial:

$$L_i(\beta) = \frac{Y_i(t)\, r_i(t)}{\sum_j Y_j(t)\, r_j(t)}$$

donde $Y_i(t) = 0$ para datos censurados ó $Y_i(t) = 1$ cuando el evento es observado.

El proceso de estimación por máxima verosimilitud parcial conduce a solucionar iterativamente las ecuaciones mediante el procedimiento de Newton& Raphson, para ampliar el procedimiento ver (Therneau y Grambsch 2000).

Al ajustar un modelo de Cox, para datos de sobrevida es posible realizar el *juzgamiento de hipótesis* sobre β a través del test de razón de verosimilitud, el test de Wald y el test de puntajes (Therneau y Grambsch 2000, pág. 53). De tal forma que la hipótesis a juzgar esta dada por $H_0 : \beta = \hat{\beta}$.

El test razón de verosimilitudes corresponde a $2\left[\log\left(L\left(\beta_0\right)\right) - \log\left(L\left(\hat{\beta}\right)\right)\right]$, donde β_0 corresponde al estimador máximo verisímil para el modelo nulo y $\hat{\beta}$ al estimador máximo verosímil del modelo de interés.

El test de Wald corresponde a $\left(\beta - \hat{\beta}\right)^t I\left(\beta\right)\left(\beta - \hat{\beta}\right)$ donde $I\left(\beta\right)$ corresponde a la matriz de varianzas y covarianzas.

El test de puntajes corresponde a $U\left(\beta\right) = \left(U_1\left(\beta\right), U_2\left(\beta\right), ..., U_p\left(\beta\right)\right)^t$, donde $U_i\left(\beta\right)$ es la derivada parcial de la función de log verosimilitud parcial con respecto a β_i con $i = 1, 2, ..., p$. Para muestras grandes $U\left(\beta\right)$ tiene distribución asintótica normal con el vector cero por media y matriz de varianzas y covarianzas dada por $I\left(\beta\right)$ cuando H_0 no se rechaza. Para la hipótesis nula dada por $H_0 : \beta = \hat{\beta}$, el estadístico de prueba corresponde a $U\left(\hat{\beta}\right)^t I^{-1}\left(\hat{\beta}\right) U\left(\hat{\beta}\right)$.

Los test de razón de verosimilitud, Wald y puntajes siguen una distribución $ji - cuadrada$ con p grados de libertad.

La *selección del modelo* se puede hacer hacia adelante, es decir, se construye inicialmente teniendo en cuenta el modelo nulo y luego se van incluyendo en el modelo aquellas que aportan a la explicación de la variable respuesta.

La evaluación del modelo de Cox puede realizarse a través del análisis de residuales, así:

Residuos de Cox- Snell: evalúan el ajuste del modelo. Se obtienen al construir la gráfica de residuales. *Residuales de Martingala*: son utilizados para verificar la forma funcional de un predictor continuo, definidos como:

$$\hat{M}_i\left(t\right) = N_i\left(t\right) - \hat{E}\left(t\right) = N_i\left(t\right) - \int_0^t Y_i\left(u\right) e^{X_i(u)\beta} d\hat{\Lambda}_0\left(\beta, u\right),$$

donde $\hat{\Lambda}_0\left(\beta, u\right)$ es el estimador de riesgo de base Breslow definido como:

$$\hat{\Lambda}_0\left(\beta, u\right) = \int_0^t \frac{\sum_{i=1}^n dN_i\left(u\right)}{\sum_{i=1}^n Y_i\left(u\right) e^{X_i(u)\beta}},$$

y están basados en la martingala de un proceso de conteo para la $i-ésima$ observación, $\hat{M}_i\left(t\right) = N_i\left(t\right) - \hat{E}\left(t\right)$ definida mediante:

$$\hat{M}_i\left(t\right) = N_i\left(t\right) - \hat{E}\left(t\right) = N_i\left(t\right) - \int_0^t Y_i\left(u\right) e^{X_i(u)\beta} \lambda_0\left(u\right) du.$$

Los residuos de martingala son muy asimétricos y con colas muy largas a la derecha.

Residuos de desvío: son utilizados para la detección de valores atípicos y se obtienen mediante una transformación de normalización de los desvíos de martingala. Si todas las covariables son fijas en el tiempo, los residuos Dde desvío toman la forma:

$$d_i = signo\left(\hat{M}_i\right) * \sqrt{-\hat{M}_i - \hat{N}_i \log\left(\left(N_i - \hat{M}_i\right)/N_i\right)}.$$

La expansión de Taylor para un término muestra que:

$$d_i \approx \frac{N_i - \hat{E}_i}{\sqrt{\hat{E}_i}},$$

equivalentes a los residuos de Pearson de los modelos lineales generalizados.

Residuos de puntaje: utilizados para verificar la influencia individual y para la estimación robusta de la varianza, se definen como:

$$U_{ij} = U_{ij}\left(\hat{\beta}, \infty\right),$$

donde $U_{ij}(\beta, t), j = 1, 2, \ldots, p$ son las componentes del vector fila de longitud p obtenido a través del proceso de puntaje para la $i - ésima$ observación:

$$U_i(\beta) = \int_0^t \left[X_i(t) - \bar{X}(\beta, t)\right] dN_i(t).$$

Residuos de Schoenfeld: utilizados para la verificación del supuesto de riesgos proporcionales, definidos como:

$$s_{ij}(\beta) = X_{ij}(t) - \bar{X}_j(\beta, t_i),$$

con una fila por falla y una columna por covariable, donde i y t_i son los individuos y el tiempo de ocurrencia del evento respectivamente.

Hay casos, donde alguna de las covariables no cumple el supuesto de riesgos proporcionales. Dado el caso, puede ser posible estratificar esa covariable y utilizar el modelo de riesgos proporcionales dentro de cada estrato y considerando las otras covariables. En este caso, los sujetos en el estrato $j - ésimo$ tienen una función de riesgo base arbitraria $\lambda_{oj}(t)$ y el efecto de otras covariables explicativas sobre la función de riesgo puede ser representado por un modelo de riesgos proporcionales en ese estrato de la forma(López, 2011, pág. 20):

$$\lambda_j(j; X) = \lambda_{0j}(t)\, e^{\beta^T X} \tag{8}$$

En este modelo, los coeficientes de regresión se supone que son los mismos en todos los estratos, aunque las funciones de riesgo base pueden ser diferentes y no relacionadas en absoluto.

3. Metodología

La siguiente aplicación se desarrolla a través de un enfoque cuantitativo, con una investigación de tipo exploratorio- descriptivo, la población a las mediciones realizadas sobre los generadores eléctricos que provee el distribuidor aggreko, y la recolección de la información la realizan los técnicos de mantenimiento diariamente. Las fases metodoógicas correspondieron a revisión documental, depuración de los datos, descripción univariada y bivariada de la información, construcción de la función de sobrevida y del modelo de Cox, finalmete se analizó la información obtenida y se elaboraron las conclusiones. Los generadores eléctricos son utilizados para transformar energía mecánica en energía eléctrica, estos pueden presentar fallas, donde los tiempos de falla son el interés del presente estudio, es decir, el generador empieza a trabajar y se registra el tiempo hasta que se presenta la primera falla. El objetivo es estimar la función de sobrevida $S(t)$ y ajustar el modelo $\lambda_i(t) = \lambda_0(t) \exp(X_i(\mathrm{t})\,\beta)$ en función de la antigüedad, la frecuencia, la presión del aceite, la temperatura del agua y la carga. Se analizaron 43 generadores eléctricos tipo diesel, entre enero y marzo de 2016, a continuación se describen las variables:

Tiempo de falla (tiempo): corresponde al tiempo en que se produce una parada en el generador.

Antigüedad (ant): número de horas en que el motor del generador, generalmente eléctrico o mecánico ha funcionado desde la última vez que se ha inicializado el dispositivo. Se mide a través del horómetro.

Frecuencia (frec): tiempo necesario para completar un ciclo vibratorio. Se usan los hercios (HZ) como unidad de medida.

Presión del aceite (pac): corresponde a la viscosidad de lubricante. Se usa el barómetro (Bar) como unidad de medida.

Temperatura del agua (tag): corresponde a la temperatura de agua en el motor del generador. Los grados centígrados se usan como unidad de medida.

Carga en kilovatios (ckw): consumo de energía de un dispositivo eléctrico, medida en kilovatios(kw).

Los datos se modelan utilizando el sotfware estadístico R versión 3.3.0, mediante el paquete *Survival* a través de: la función *Surv* apropiada para los datos que presentan censura, la función $survfit$ que permite obtener estimación de la función de sobrevida utilizando el método de Kaplan-Meier.

4. Resultados

En este apartado primero se presenta el resumen descriptivo de las variables, segundo, la metodología de Kaplan- Meier para estimar la función de sobrevida, con la cual se analiza la evolución de la probabilidad de falla con su respectivo intervalo de confianza. Tercero, se construye el modelo de regresión de Cox para estimar el efecto de las variables de estudio sobre los tiempos de sobrevivencia a la falla del generador. Las características del estudio se describen en la Tabla 1.

La unidad sobre la cual se registra el evento		Generadores eléctricos presentes en el bloque Pacific Rubiales Energy durante el primer trimestre de 2016
Evento de interés		Falla del generador
Variable respuesta		Tiempo hasta la falla: Cuya escala de medición es de razón de tipo continuo, ya que se mide las horas transcurridas hasta que presenta el fenómeno de estudio
Tiempo	Inicial del Estudio	01 enero de 2016
	Origen del Evento	01 enero de 2016
	Final del Estudio	31 marzo de 2016
La escala de medida del tiempo hasta el evento		Horas
Censura	Tipo	Tipo I (a derecha)
	Tiempo	720 horas

Tabla 1: Consideraciones para el modelo de sobrevida tiempo de falla

La tabla 2 presenta algunas medidas de la distribución univariante para cada variable de estudio, a partir de la cual se puede inferir que todas las coovariables tienen poca variabilidad. Contrario a la variable tiempo de falla del generador que tiene una alta heterogeneidad (78.2 %). La carga y la antigüedad son bastante simétricas, en cambio, el coeficiente de asimetría para los tiempos de falla (As=1.16) indica una alta concentración de los datos en los valores bajos de la variable. La frecuencia, presión del aceite y temperatura del

agua evidencian una concentración de sus valores en los valores altos.

La kurtosis de antigüedad, carga y presión indica la presencia de dos tipos de generadores en la muestra y para la frecuencia y temperatura se evidencia algunos valores atípicos grandes. El 50 % de los generadores eléctricos tienen una frecuencia de 59 hz, una carga de 262 kw, la presión del aceite igual 80 bar y, la temperatura del agua de 80.

	Mínimo	Q1	Mediana	Media	Q3	Máximo	CV	Asimetría	Kurtosis
Tiempo	4.0	476.0	714.0	859.5	856.5	2160.0	0.782	1.16	0.17
Antigüedad	5429	9866	11210	11840	13290	18310	0.225	0.445	0.020
Frecuencia	52	59	59	59	60	61	0.031	-2.37	7.745
Pres. aceite	3.800	4.200	4.300	4.247	4.300	4.400	0.031	-1.106	1.582
Temp. agua	75.00	79.00	80.00	79.72	80.00	81.00	0.0137	-1.78	6.90
Carga	260.0	260.0	262.0	261.4	262.0	263.0	0.004	0.007	-1.41

Tabla 2: Resumen de las variables

También se presenta una alta correlación entre la frecuencia y la presión del aceite ($\rho = 0.7978$), por lo que se puede concluir que un cambio en la presión del aceite afecta directamente la frecuencia. De igual forma se observa que la frecuencia y la temperatura del agua se encuentran correlacionadas ($\rho = 0.7853$). Mientras que se evidencia una correlación baja entre la antigüedad del generador y la temperatura del agua ($\rho = 0.027$).

A partir de la tabla 3 que presenta la estimación de la función de sobrevida, utilizando el estimador de Kaplan- Meier, se encontró que el número de horas que transcurren para que la probabilidad de que el generador falle sea del 50 % corresponde a 714 horas.

Para el proveedor es importante que el tiempo de falla del generador (si está se presenta) sea cercana a las 720 horas debido a que entre menos horas de falla se presente durante el mes tanto proveedor como cliente quedan conformes con el servio que prestan y que reciben, respectivamente, sin acarrear pérdidas económicas.

N	Evento	Mediana	0.95LCL	0.95UCL
43	24	714	584	NA

Tabla 3: Estimación de Kaplan- Meier

El modelo óptimo de Cox, obtenido a través del método de selección hacia adelante, determina que las variables que explican la variablidad en los tiempos de falla de los generadores son la frecuencia y la presión del aceite, debido a que la bomba del aceite debe garantizar un caudal y una presión de trabajo variable debido a que esta trabaja en función de las revoluciones del motor. La Tabla 4 muestra el resumen del modelo:

	coef	exp(coef)	se(coef)	Z	P
Frecuencia	-0.915	0.401	0.315	-2.90	0.0037
Pres. aceite	-4.345	0.013	2.741	-1.59	0.1129

Tabla 4: Modelo de Cox

A continuación se procede a verificar los supuestos del modelo, Tabla 5.

	rho	chisq	P
Frecuencia	-0.36	2.74	0.098
Pres. aceite	0.60	10.58	0.0011
GLOBAL	NA	10.67	0.0048

Tabla 5: Verificación de riesgos proporcionales en el modelo

Se encuentra que el supuesto de riesgos proporcionales no se cumple en la presión de aceite, esto significa que con $\alpha = 0.05$, el modelo propuesto no cumple el supuesto de que existe una relación proporcional entre las funciones de riesgo de tiempo de fallo correspondientes a diferentes generadores.

En otras palabras, si consideramos dos generadores, sus funciones de riesgo no son constantes en el tiempo para la variable presión del aceite.

Se decide entonces construir un modelo de riesgos estratificado por la variable presión del aceite, se asignan dos estratos (bajo, alto), lo anterior debido a clasificaciones técnicas. El modelo obtenido corresponde a:

```
              coef  exp(coef) se(coef)      z Pr(>|z|)
Frecuencia -1.2650    0.2822   0.3065 -4.127 3.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
Frec.    0.2822      3.543    0.1548    0.5147

Concordance= 0.795  (se = 0.086 )
Rsquare= 0.505   (max possible= 0.949 )
Likelihood ratio test= 30.22  on 1 df,   p=3.849e-08
Wald test            = 17.03  on 1 df,   p=3.676e-05
Score (logrank) test = 39.01  on 1 df,   p=4.207e-10
```

El modelo es aceptable para cualquiera de los tres criterios (razón de verosimilitudes, test de wald y de puntajes). Estos coeficientes se consideran significativos por sus valores p.

El supuesto de riesgos proporcionales con $\alpha = 0.05$ para el modelo de riesgos estratificado se satisface, Tabla 6:

	rho	chisq	P
Frecuencia	-0.255	1.01	0.315

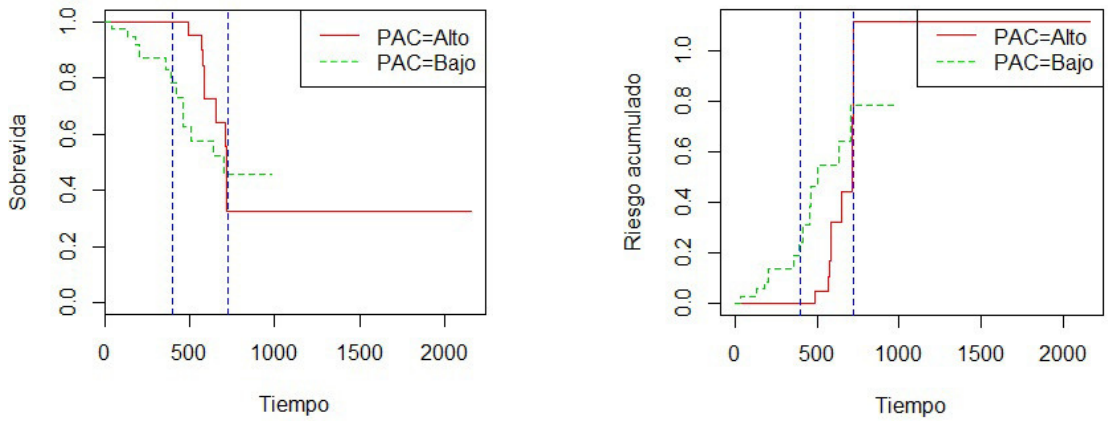
Tabla 6: Supuesto riesgos proporcionales modelo estratificado

A partir del valor del $ODDS$ de la frecuencia (0.282) se tiene que el incremento en 1 Hz en la frecuencia disminuye el riesgo de que falle el generador en 26.5 %. El valor de $R^2 = 0.50$ indica que el modelo propuesto tiene un buen ajuste.

A continuación se presenta la función de sobrevida para el modelo óptimo Tabla 7 y Figura 2.

Alto						
time	n.risk	n.event	survival	std.err	lower 95 % CI	upper 95 % CI
490	24	1	0.953	0.0464	0.866	1.000
569	23	1	0.899	0.0687	0.774	1.000
578	22	1	0.845	0.0837	0.696	1.000
584	21	2	0.725	0.1074	0.543	0.970
653	19	1	0.641	0.1231	0.440	0.934
710	18	1	0.556	0.1327	0.348	0.888
714	17	1	0.480	0.1350	0.276	0.833
718	16	2	0.327	0.1302	0.150	0.714
Bajo						
4	19	1	1.000	0.000168	1.000	1.000
24	18	1	1.000	0.000482	0.999	1.000
35	17	1	0.973	0.027529	0.921	1.000
130	16	1	0.945	0.041596	0.867	1.000
181	15	1	0.916	0.052986	0.818	1.000
205	14	1	0.872	0.069138	0.747	1.000
354	13	1	0.828	0.081425	0.683	1.000
386	12	1	0.784	0.091418	0.623	0.985
416	11	1	0.732	0.101833	0.557	0.961
458	10	1	0.680	0.110089	0.495	0.934
462	9	1	0.629	0.116645	0.437	0.904
507	8	1	0.577	0.121775	0.381	0.872
636	7	1	0.525	0.125655	0.329	0.839
703	6	1	0.456	0.130375	0.261	0.799

Tabla 7: Resumen de K- M tiempos de falla



(a) Curvas de sobrevidafig:label:a

(b) Curvas de riesgo acumuladofig:label:b

Figura 2: Función de sobrevida

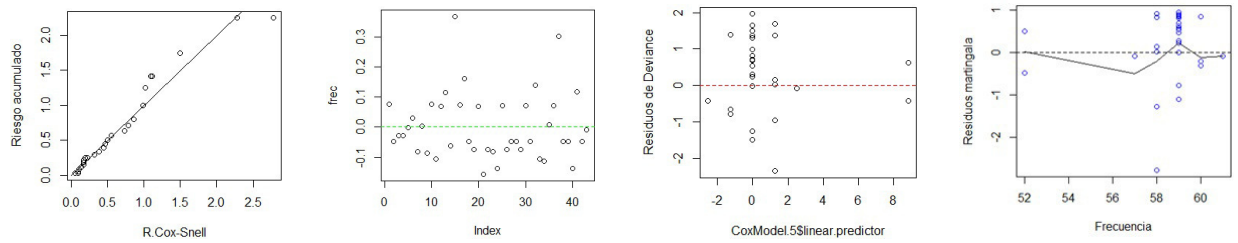
Nótese que es más rápido el decrecimiento antes de las 718 horas y tiende a estabilizarse gradualmente a partir de las 718 horas de funcionamiento después de haberse inicializado el sistema, con una presión del aceite alto donde está alrededor de 32,7 %. Al estimar la función de sobrevida, teniendo en cuenta el efecto de las covariables presentes en el modelo propuesto, Tabla 8, es posible sugerir un mantenimiento preventivo entre las 600 a 700 horas de funcionamiento del generador ya que la probabilidad decrece notoriamente.

Nótese que es más rápido el decrecimiento antes de las 718 horas y tiende a estabilizarse gradualmente a partir de las 718 horas de funcionamiento después de haberse inicializado el sistema, con una presión del aceite alto donde está alrededor de 32,7 %.

Estrato	25	50	75
PAC.Alto	584	718	NA
PAC.Bajo	35	354	462

Tabla 8: Estimación de Kaplan- Meier modelo propuesto

El análisis de residuales para el modelo óptimo evidencia que, este modelo ajusta bien a los datos Figura **??**. Como vemos en la Figura **??** los residuos *dfbeta* están centrados con respecto al origen, no presentan patrones definidos y no se aprecia ninguna irregularidad en la gráfica. La Figura **??** muestra los residuos de deviance estratificados para los dos tipos de presión de aceite, no apreciamos patrones definidos ni tampoco apreciamos residuos alejados del origen.



(a) Residuos Cox- Snell- fig:cox (b) Residuos dfbetafig:beta (c) Residuos de desvío- fig:desvio (d) Residuos de Martingala

Figura 3: Residuales

Los residuos de martingala para la frecuencia, en la que podemos ver una tendencia lineal, estos residuos presentan claramente una forma funcional definida.

5. Conclusiones

A través del análisis de sobrevida, y sin asumir ninguna distribución de probabilidad para los tiempos de falla, se identificó que la frecuencia es un factor que influye en la disminución del riesgo de falla del generador.

A partir de la información que arroja la función de sobrevida se puede diseñar un nuevo esquema de mantenimiento preventivo de los generadores eléctricos.

A partir de la información que arroja el modelo de Cox estratificado se puede diseñar el esquema de mantenimiento predictivo, estableciendo rutinas de inspección y seguimiento a las mediciones asociadas a la presión del aceite y la frecuencia del generador.

Referencias Bibliográficas

Colosimo, E. y Giolo, S. (2004), *Análisis de sobrevivencia*, segunda edición edn, Limusa, Brasil.

Dobson, A. y Barnett, G. (2008), *An Introduction to Generalized Linears Models*, tercera edición edn, Chapman & Hall, Londres.

Enriquez, G. (2005), *El libro práctico de los generadores, transformadores y motores eléctricos*, segunda edición edn, Limusa, Brasil.

Godoy, M. (2009), Introducción al análisis de supervivencia con R, Master's thesis, Universidad Nacional Autónoma de México.

Klein, J. y Moeschberger, L. (1997), *Survival analysis : techniques for censored and truncated data*, segunda edición edn, Springer–Verlag, New York.

McCullagh, P. y Nelder, J. (1989), *Generalized Linear Models*, segunda edición edn, Chapman & Hall, London.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Therneau, T. M. y Grambsch, P. M. (2000), *Modeling survival data: extending the Cox model*, Springer Science & Business Media.

ANÁLISIS MULTIVARIADO PARA EL DIAGNÓSTICO DE HABILIDADES MATEMÁTICAS

Especialización en Estadística

DAYSY MAITE SÁNCHEZ BAREÑO [1,a], EDGAR FELIPE RUIZ ROBERTO [1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

La matemática es una de las áreas del conocimiento de mayor importancia e influencia debido a sus amplias aplicaciones, de ahí que se hiciera un análisis multivariado que permitió la identificación de cuatro grupos del desarrollo matemático de niños en etapa preoperatoria, a través de la aplicación de una prueba de reconocida validez y confiabilidad, "Test de habilidades básicas para la iniciación del cálculo" (TIC), aplicado a niños de 4 a 7 años; El índice de fiabilidad calculado mediante el coeficiente Alpha de Cronbach fue significativa con un p valor de 0.01. Se realizó la clasificación no jerárquica (K-medias), la cual reveló que los niños en etapa preoperatoria poseen algunas habilidades ya innatas, por esto el 50 % y 29 % de la población se ubicó en el cluster 3 y 4 respectivamente. Finalmente, se presentaron los índices de dificultad de los ítems, indicando que a estas edades la habilidad de conservación es la de mayor dificultad.

***Palabras clave*:** Habilidad, k-medias, Etapa preoperacional, Análisis multivariado.

Abstract

Mathematics is one of the areas of knowledge of greater importance and influence because of its broad applications, hence a multivariate analysis allowed the identification of four groups of mathematical development of children in preoperative stage was made through the application of recognized test validity and reliability, "test on basic abilities for calculus introduction (TIC) applied to children 4 to 7 years; The reliability index calculated by Cronbach Alpha was significant with a p-value of 0.01. non-hierarchical classification (K-means) was performed, which revealed that children in pre-operative stage and have some innate abilities, so 50 % and 29 % of the population was located in the cluster 3 and 4 respectively. Finally, the indices of difficulty of the items were presented, indicating that at this age the ability of conservation is the most difficult.

***Key words*:** Abilities, k- means, Preoperational stage, Multivariate analysis.

1. Introducción

El estudio de habilidades lógico matemáticas a nivel escolar es una temática fundamental en el ámbito educativo, debido a que el razonamiento es una de las cualidades del conocimiento más significativas en la formación del individuo. Es importante el diagnóstico de las habilidades matemáticas en los niños en etapa preoperatoria, porque este es el momento oportuno para que los docentes reconozcan las habilidades de sus estudiantes y las falencias que estos poseen, guiándolos en el momento propicio, lo que contribuye a mejorar

[a]Estudiante de Especialización en Estadística. E-mail: daysymaite.sanchez@uptc.edu.co
[b]Profesor asistente. E-mail: feruro42@gmail.com

el aprendizaje en el aula (Cardoso y Cerecedo 2008).

De ahí que se viera la necesidad de un estudio sobre el reconocimiento de las habilidades matemáticas, para así estimular el aprendizaje en los niños oportunamente, siendo este el momento en el cual se crean una idea de lo que será su vida estudiantil. Se toma como objeto de estudio niños y niñas de 4 a 7 años, donde, según Piaget (1955,32), se encuentran en etapa preoperatoria, debido que, una operación mental requiere pensamientos lógicos y los niños no tienen la capacidad de pensar lógicamente, pero en esta etapa desarrollan la capacidad para manejar el mundo haciendo uso de símbolos para representar objetos, lugares y personas, es decir, adquieren un lenguaje y aprenden a manejarlo.

Por tal razón y con base en lo enunciado anteriormente se formulo la siguiente pregunta, ¿Cómo diagnosticar habilidades matemáticas presentes en los niños en etapa preoperatoria del Colegio de la Presentación de Tunja, a partir de las operaciones lógicas elementales?, pero, para responder a esta pregunta es necesario primero hacer una caracterización de la población, describir las habilidades de los niños en etapa preoperatoria y establecer diferencias y similitudes en los estudiantes de preescolar del Colegio de la Presentación.

Para el desarrollo de la propuesta fue necesaria la aplicación de un test, constituido por 32 actividades dirigidas a niños en etapa pre operacional, test tomado del artículo "Test de habilidades básicas para la iniciación al cálculo" (del Solar 2003). La implementación del test como instrumento busca integrar una herramienta útil para que los educadores realicen un buen diagnóstico de sus educandos, como también adecuar los procesos de enseñanza y de aprendizaje a las realidades observadas, permitiendo identificar las habilidades matemáticas presentes en los niños en etapa preoperatoria, para verificar procesos cognitivos existentes en los infantes e identificar la presencia y desarrollo de las operaciones lógicas elementales en que se encuentran los niños, mediante la aplicación de las actividades propuestas por el test.

El estudio se centró, específicamente, en las habilidades cognitivas básicas señaladas por Piaget (1955). Clasificación, seriación, conservación, expresión de juicio lógico y función simbólica. El proceso de formación y desarrollo de habilidades no es espontáneo, ya que conlleva a la adecuada organización y planificación por parte del docente a fin de lograr que el estudiante se apropie adecuadamente de los procedimientos deseados (Álvarez Yero, Ríos Barrios y Velásquez Peña 2014).

Otra investigación importante para este trabajo es la realizada por un equipo de psicólogos encabezado por Libertus, Feigenson y Halberda (2011), indicando que la capacidad para las matemáticas en niños de edad preescolar está fuertemente ligada a su innato y primitivo "sentido numérico", llamado "sistema aproximado de número". Este sentido está presente en todas las actividades que se realizaron a diario como estimar el número de asientos que hay en un cine o el número de personas que se encuentran en un cierto establecimiento etc.

2. Referente conceptual

Para abordar esta temática se hace necesario conceptualizar algunos términos, para lo cual se toman algunos autores como referencia.

2.1. Habilidades matemáticas

Las habilidades matemáticas son reconocidas por muchos autores, entre ellos (H. Hernández, H. González(1999)) quienes la definen como aquellas que se forman durante la ejecución de las acciones y operaciones que tienen un carácter esencial matemático; es decir que estos autores se enfocan en definir la habilidad matemática como un "proceso de experiencias" donde el niño debe tener interacción maestro-alumno para poder desarrollar dichas habilidades.

2.2. Alfa de Cronbach

La validez de un instrumento se refiere al grado en que el instrumento mide aquello que pretende medir. Y la fiabilidad de la consistencia interna del instrumento se puede estimar con el alfa de Cronbach. La medida de la fiabilidad mediante el alfa de Cronbach asume que los ítems (medidos en escala tipo Likert) miden un mismo constructo y que están altamente correlacionados (Welch & Comer, 1988). El coeficiente alfa de Cronbach es la forma más sencilla y conocida de medir la consistencia interna y es la primera aproximación a la validación del constructo de una escala, debe entenderse como una medida de la correlación de los ítems. Está indicada la determinación del alfa de Cronbach en escalas unidimensionales que tengan entre tres y veinte ítems y siempre se debe informar este valor en la población específica donde se empleó la escala (Oviedo y Arias 2005). La fiabilidad de la escala debe obtenerse siempre con los datos de cada muestra para garantizar la medida fiable del constructo en la muestra concreta de investigación.

El alfa de Cronbach es un coeficiente que toma valores entre 0 y 1.Cuanto más cerca se encuentre el valor del alfa a 1 mayor es la consistencia interna de los ítems analizados. El resultado negativo denota un alto grado de inconsistencia interna del test hasta tal punto que no se justifica el cálculo de alfa (Soler Cárdenas y Soler Pons 2012). La popularización del coeficiente alfa de Cronbach se debe a la practicidad de su uso, ya que requiere una sola administración de la prueba. Además, tiene la ventaja de corresponder a la media de todos los posibles resultados de la comparación que se hace en el proceso de dividir en mitades una escala (Kwok y Sharp 1998).

2.3. Análisis univariado

Por su parte el análisis univariado consiste en el análisis de cada una de las variables estudiadas por separado, es decir, el análisis está basado en una sola variable. Las técnicas más frecuentes de análisis univariado son la distribución de frecuencias para una tabla univariada y el análisis de las medidas de tendencia central de la variable. Se utiliza únicamente en aquellas variables que se midieron a nivel de intervalo o de razón (Therese L. Baker, 1997). La distribución de frecuencias de la variable indica cómo están distribuidas las categorías de la variable, pudiendo presentarse en función del número de casos o en términos porcentuales.

3. Conglomerado

Un conglomerado o análisis cluster es un conjunto de técnicas destinadas a agrupar observaciones por afinidad. Cada observación consistirá en p valores numéricos correspondientes a la medición de sendas variables y, por lo tanto, constituirán puntos de R^p. Esa es la razón por la que esta técnica haya sido considerada tradicionalmente como parte de la Estadística Multivariante.

Se analizan las variables que caracterizan a cada individuo, mediante el análisis cluster, grupos diferentes estadísticamente significativos. En primer lugar se realizan técnicas de análisis jerárquico y, una vez decidido el número de grupos a tener en cuenta, técnicas de análisis no jerárquico (o de k medias) se prueba con distintos números de grupos. Empleando como medida de distancia la euclídea al cuadrado (Albarrán Lozano y Alonso González 2006):

$$d_{ij}^2 = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2$$

Según Díaz (2012) "los conglomerados son objetos que poseen características similares, conglomerado es la traducción del termino en ingles cluster"(Díaz Monroy y Morales 2012). Un tipo de cluster o de partición es el k medias permitiendo la clasificación de los niños según sus habilidades se hará con la ayuda del k medias, según Eliseo Martínez H (2003) "La finalidad de la formación de cluster o conglomerados es agrupar elementos en grupos homogéneos en función de "alguna similitud o similaridades" entre ellos. Como podemos ver esta finalidad parece auto-referente. Puesto que si tenemos elementos homogéneos ya por sí estarían formados los cluster. En rigor, quien realiza una formación de cluster lo hace mediante una técnica de homogeneidad, y posteriormente encontrado los cluster o conglomerados realiza el reconocimiento de patrones

de formación. Tanto es así que a menudo al análisis de cluster se le llama métodos de clasificación automática no supervisada".

4. K medias

El k medias también conocido como quick-cluster, se utiliza para agrupar los datos en un número k de conglomerados determinado a priori. La elección de k puede basarse en argumentos formales, como los que se mencionan en la tercera sección, o bien en argumentos gráficos.

La técnica consiste en aglomerar todos los datos en torno a k puntos en función de la proximidad a estos, según la distancia considerada. En ocasiones, estas semillas son establecidas de antemano en función de conocimientos previos, en cuyo caso el método es trivial. Si queremos formar k conglomerados pero no contamos con semillas, puede procederse de la siguiente forma: se seleccionan k datos, bien aleatoriamente o bien los k primeros, que serán las semillas iniciales. Los datos restantes se irán aglomerando en torno a ellos. No obstante, si la semilla mas cercana a un dato dista del mismo más que que la semilla mas cercana a esta, dicho dato reemplaza como semilla a la más cercana y usurpa en lo sucesivo, por así decirlo, su conglomerado. Al final del proceso, se reconstruyen las semillas como centroides de los conglomerados finales y el procedimiento se repite sucesivamente hasta conseguir suficiente estabilidad en los centroides finales.

Se asume que entre los individuos se puede establecer una distancia euclidiana. La idea central de estos métodos es la selección de alguna partición inicial de los objetos para luego modificar su configuración hasta obtener la "mejor" partición en términos de una función objetivo. Varios algoritmos propuestos para estos procedimientos difieren respecto al criterio de optimización (la "mejor" partición). Estos algoritmos son semejantes al de optimización, conocido como el mayor descenso, los cuales empiezan con un punto inicial y generan una serie de movimientos desde un punto a otro, calculando en cada paso el valor de una función objetivo, hasta que se encuentra un óptimo local.

El procedimiento de agrupamiento de K-medias consiste en particionar un conjunto de n individuos en k grupos, se nota la partición por P(n,k), con el siguiente criterio: primero se escogen los centroides de los grupos que minimicen la distancia de cada individuo a ellos, luego se asigna cada individuo al grupo cuyo centroide esté más cercano a dicho individuo. La ventaja del método de k-medias radica en que su algoritmo es más rápido (especialmente con muestras de gran tamaño) y se ve menos afectado ante la presencia de valores atípicos. Su desventaja estriba en la elección de las semillas iniciales, que puede tener una gran trascendencia en el resultado final y, sin embargo, son frecuentemente designadas según criterios bastante arbitrarios.

5. Metodología

El enfoque que privilegia esta investigación es la cuantitativa de tipo inferencial debido a que la investigación cuantitativa recoge y analiza datos sobre variables, los estudiantes que participaron en esta investigación fueron 70 estudiantes, 14 niños y 56 niñas, todos en la etapa preoperatoria, con edades comprendidas entre los 4 y los 7 años que cursan Pre jardín, Jardín y Transición en el Colegio de la Presentación de Tunja.

La prueba utilizada fue el Test De Habilidades Básicas Para La Iniciación Al Cálculo "TIC" validado por Riquelme (2003) y los aportes hechos por Piaget (1975). El test está constituido por 32 ítems organizados en las cinco habilidades principales que son Clasificación, Seriación, Conservación, Expresión de Juicio Lógico y Función Simbólica.

Para la validación del test se calcula el alfa de cronbach permitiendo verificar su confiabilidad, luego se completa una base de datos teniendo en cuenta los resultados arrojados del test y algunas otras variables como edad, género, estrato, grados cursados, estado civil de los padres y nivel de educación de los padres. Se hizo el análisis descriptivo univariado y multivariado de la base de datos lo cual permite determinar si

existen diferencias significativas entre las diversas habilidades de los niños acorde a las variables.

Por último se hace un análisis de conglomerados donde se aplica el método de k- medias con ayuda del programa R. Se hicieron clasificaciones en grupos internamente homogéneos y mutuamente heterogéneos con respecto a las variables estudiadas. Posteriormente se realizó el cruce de cada cluster obtenido con las variables analizadas las características más sobresalientes de los niños pertenecientes a cada uno de ellos. El análisis situó a los niños de acuerdo con la similitud de sus habilidades (Cordón García, Fernández Gómez, Pinto Molina, Alonso Arévalo y Alonso Berrocal 2011).

6. Resultados

Cada uno de los ítems fueron sometido a juicio de expertos: Educadoras de Preescolar en ejercicio, Psicólogas y docentes de matemáticas, en total 10 profesionales. El criterio de selección fue el 99 % de acuerdo entre los jueces para dirimir si el ítem pertenecía o no a la habilidad o dimensión señalada.

Con la finalidad de conocer el grado de confiabilidad que tenía el instrumento se determinó el grado de consistencia interna, aplicando un alfa de Cronbach cuyo valor fue de 0.737. Valor que representa un grado de confiabilidad significativo a un p valor de 0,01.

Resumen de procesamiento de casos

		N	%
Casos	Válido	70	100,0
	Excluido[a]	0	,0
	Total	70	100,0

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,737	6

Figura 1: Se evidencia que la confiabilidad del instrumento fue significativa

Para explorar los datos se identificaron aspectos significativos de estos como lo son el mínimo, el máximo y el promedio de cada una de las variables con el fin de encontrar irregularidades en los datos de cada variable, se hace un resumen de las variables y su caracterización, la cual se presenta en la tabla 2, con ayuda de R.

```
> summary(matematicas1)
      EDAD       GENERO    ESTRATO      GRADOS.CURSADOS  ESTADO.CIVIL.PADRES NIVEL.EDUCACION.MADRE NIVEL.EDUCACION.PADRE
 Min.   :4.000   F:56   Min.   :2.000   Min.   :1.000   CASADOS    :21       MEDIA   :12           MEDIA   : 9
 1st Qu.:4.615   M:14   1st Qu.:3.000   1st Qu.:2.000   SOLTEROS   :29       POSGRADO: 3           POSGRADO:12
 Median :5.050          Median :4.000   Median :3.000   UNION LIBRE:20       SUPERIOR:55           SUPERIOR:49
 Mean   :5.243          Mean   :3.757   Mean   :2.514
 3rd Qu.:5.820          3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :6.830          Max.   :5.000   Max.   :4.000
 PUNTAJE.CLASIFICACION PUNTAJE.SERIACION PUNTAJE.CONSERVACION PUNTAJE.EXPRESION.JUICIO.LOGICO PUNTAJE.FUNCION.SIMBOLICA
 Min.   :2.000         Min.   :2.000     Min.   :1.000        Min.   :1.000                  Min.   :2.0
 1st Qu.:6.000         1st Qu.:5.000     1st Qu.:3.000        1st Qu.:5.000                  1st Qu.:5.0
 Median :8.000         Median :5.500     Median :5.000        Median :5.000                  Median :6.0
 Mean   :6.871         Mean   :5.157     Mean   :4.614        Mean   :4.643                  Mean   :5.4
 3rd Qu.:8.000         3rd Qu.:6.000     3rd Qu.:6.000        3rd Qu.:5.000                  3rd Qu.:6.0
 Max.   :8.000         Max.   :6.000     Max.   :7.000        Max.   :5.000                  Max.   :6.0
 PUNTAJE.TOTAL.TEST
 Min.   :12.00
 1st Qu.:25.00
 Median :28.00
 Mean   :26.73
 3rd Qu.:30.00
 Max.   :31.00
```

Figura 2: Se caracteriza la población describiendo cada variable.

El análisis refleja que no hay irregularidades ademas el puntaje total del test aplicado a los estudiantes de preescolar oscila entre 12 y 32 puntos, con una media de 26,73, ademas se verifica que las edades están entre los 4 y los 7 años y que los puntajes corresponden a los ítem.

Se analizó la correlación existente entre las variables presentada en la tabla 3 evidenciando que las variables PUNTAJE.FUNCION.SIMBOLICA y PUNTAJE.TOTAL.TEST están altamente correlacionadas y la variable fuera del test que mas se relaciona con el puntaje obtenido en este es la edad viendo que a mayor edad mayor puntaje.

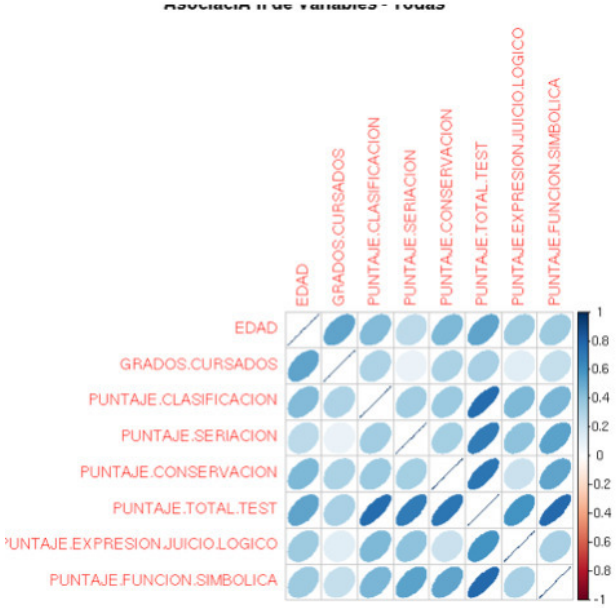


Figura 3: Correlación entre todas las variables.

Vemos también la correlación con respecto al género en la tabla 4 (femenino-masculino), donde se evidencia que la correlación entre la clasificación y la expresión de juicio lógico es mayor en las niñas que en los niños.

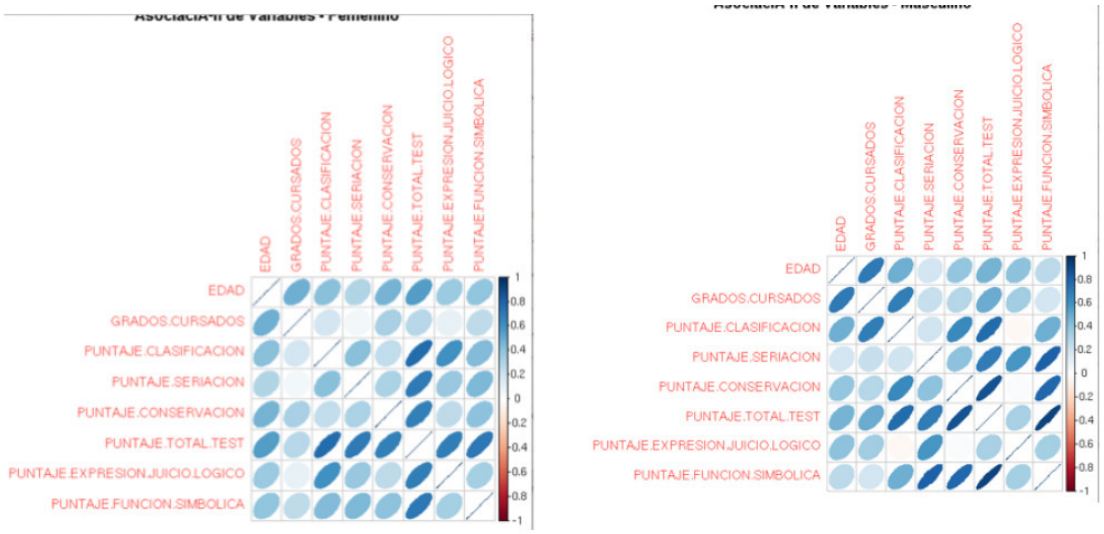


Figura 4: Correlación entre todas las variables según el género.

Para determinar si existían diferencias significativas entre las medias de los niños se hizo un análisis de conglomerados (k medias). Tras analizar el dendrograma se realizaron varias pruebas utilizando el algoritmo

k medias para finalmente seleccionar la clasificación correspondiente a cuatro agrupaciones que se pueden visualizar en el gráfico:

- Cada cluster muestra similitudes entre las habilidades de los niños 4. El niño posee todas las habilidades; 3. El niño posee la mayoría de las habilidades;
- El niño solo posee algunas habilidades;
- El niño solo posee una habilidad.

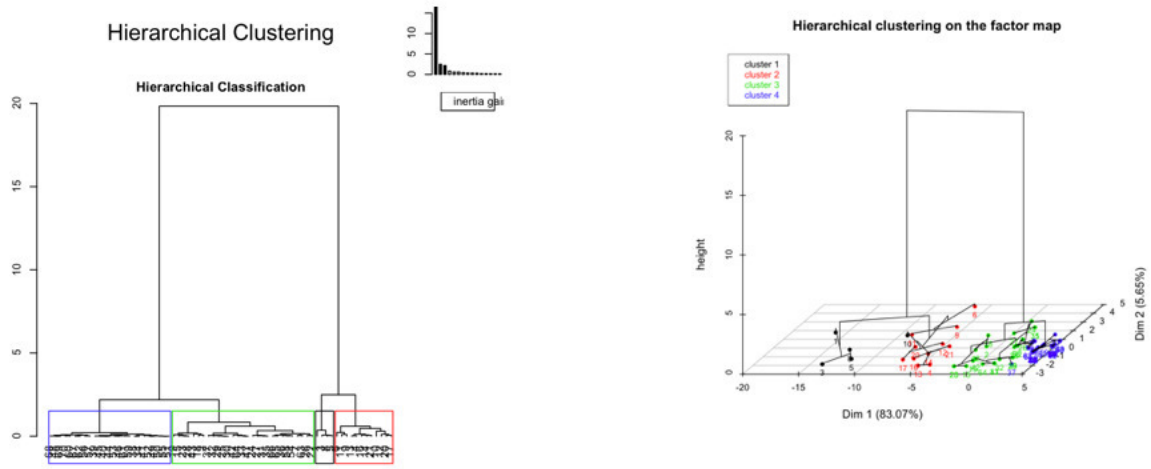


Figura 5: Clasificación mediante el k medias.

Se puede evidenciar que cada grupo tiene diferentes elementos y ahí unos grupos mas grandes que otros como en el casi del grupo 3. Conocidos los centros de los conglomerados, es necesario conocer el grado de diferencia entre ellos considerando la distancia entre los centroides. El método K-medias utiliza la distancia euclídea para calcular las distancias.

Cluster	Frecuencia	Porcentaje	Hablidades
1	5	7%	Clasificación
2	10	14%	Clasificación, seriación y Función simbolica
3	35	50%	Clasificación, seriación, Juicio lógico y Función simbolica
4	20	29%	Todas

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
EDAD	3,153	3	,499	66	6,313	,001
PUNTAJE.CLASIFICACIÓN	3,743	3	,146	66	25,623	,000
PUNTAJE.SERIACIÓN	9,270	3	,293	66	31,645	,000
PUNTAJE.FUNCIÓN. SIMBOLICA	2,318	3	,065	66	35,439	,000
PUNTAJE. CONSERVACIÓN	29,911	3	,485	66	61,619	,000
PUNTAJE.TOTAL.TEST	114,140	3	,669	66	170,622	,000
GRADOS.CURSADOS	,607	3	,662	66	,917	,438

Figura 6: Se identifican las habilidades de cada cluster y el porcentaje de niños que está en cada grupo

En la Figura 6 se evidencian cada uno de los 4 grupos verificando que la mayoría de los estudiantes obtuvo un puntaje que permite afirmar que van en proceso de adquisición de la habilidad (cluster 3).

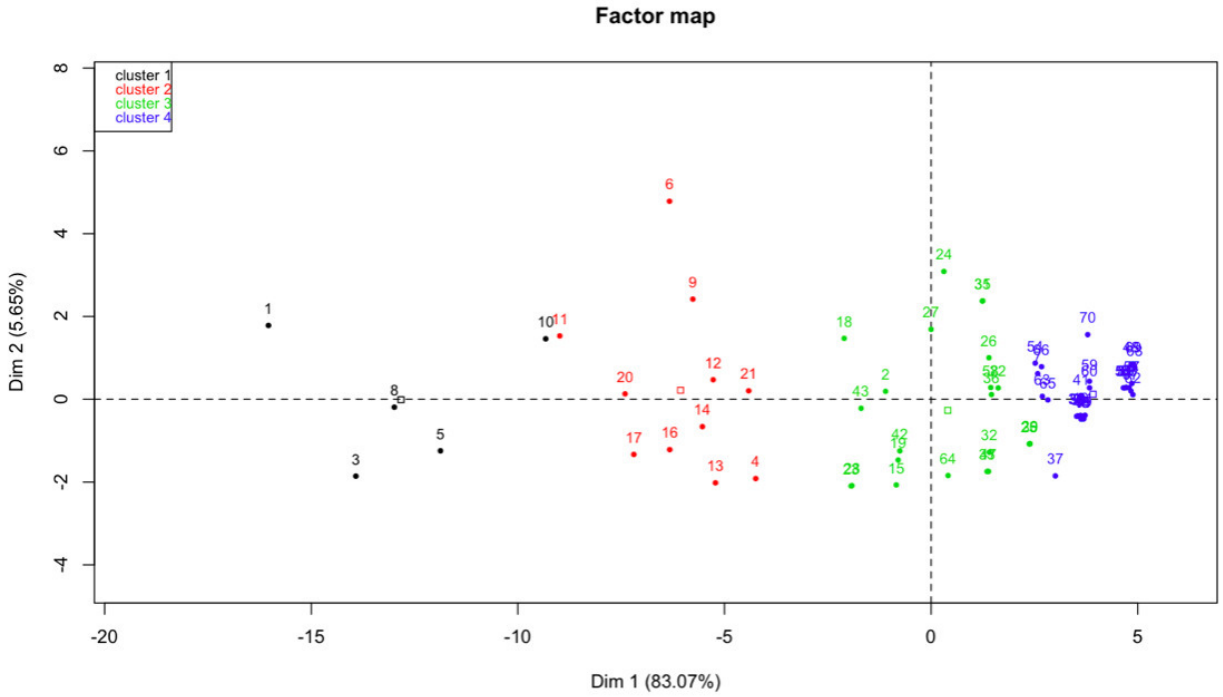


FIGURA 7: Se verifican las observaciones de cada cluster.

7. Conclusiones

- La habilidad de mayor dificultad para los niños es la de conservación.
- El test es fiable ya que se valida la consistencia interna del instrumento.
- Los niños del colegio de la presentación muestran un gran desarrollo en las habilidades de clasificación y seriación.
- Los niños que más han cursados grados obtuvieron los mejores resultados en el test.
- El 50 % de los estudiantes se ubica en el cluster 3 por tanto se puede inferir que la mayoría de los estudiantes llegan a la escuela con todas las habilidades básicas enunciadas por Piaget menos la de conservación.
- Solo el 7 % de los estudiantes se ubica en el cluster 1 destacando solo la habilidad de clasificación.
- El K medias permitió clasificar a los niños según sus habilidades esto con el fin de identificar diferencias y similitudes entre los mismo.

Referencias Bibliográficas

Albarrán Lozano, I. y Alonso González, P. (2006), 'Clasificación de las personas dependientes a partir de la encuesta de discapacidades, deficiencias y estado de salud de 1999', *Revista Española de Salud Pública* **80**(4), 349–360.

Álvarez Yero, J. C., Ríos Barrios, I. y Velásquez Peña, E. A. (2014), 'Requerimientos teórico-metodológicos para desarrollar habilidades en la obtención de información científica en estudiantes universitarios', *Humanidades Médicas* **14**(1), 109–126.

Cardoso, E. y Cerecedo, M. (2008), 'El desarrollo de las competencias matemáticas en la primera infancia', *Revista Iberoamericana de Educación* **47**(5), 1–11.

Cordón García, J. A., Fernández Gómez, M. J., Pinto Molina, M., Alonso Arévalo, J. y Alonso Berrocal, J. L. (2011), 'Las monografías en la edición científica y los perfiles de autoría y productividad en las universidades andaluzas', *Acimed* **22**(4), 317–336.

del Solar, G. R. (2003), 'Test de habilidades básicas para la iniciación al cálculo"tic', *Revista Enfoques Educacionales* **5**(1), 137–156.

Díaz Monroy, L. G. y Morales, M. (2012), 'Análisis estadístico de datos multivariados', *Bogotá: Universidad Nacional de Colombia* .

Kwok, W. C. C. y Sharp, D. J. (1998), 'A review of construct measurement issues in behavioral accounting research', *Journal of Accounting Literature* **17**, 137.

Libertus, M. E., Feigenson, L. y Halberda, J. (2011), 'Preschool acuity of the approximate number system correlates with school math ability', *Developmental science* **14**(6), 1292–1300.

Oviedo, H. C. y Arias, A. C. (2005), 'Aproximación al uso del coeficiente alfa de cronbach', *Revista colombiana de psiquiatría* **34**(4), 572–580.

Soler Cárdenas, S. F. y Soler Pons, L. (2012), 'Usos del coeficiente alfa de cronbach en el análisis de instrumentos escritos', *Revista Médica Electrónica* **34**(1), 01–06.

DISEÑO Y VALIDACIÓN DE UN CUESTIONARIO PARA MEDIR EL GRADO DE CONCIENCIA AMBIENTAL

Especialización en Estadística

Ingrith Marcela Quintero Fuentes[1,a], Reinaldo Alarcón Guarín[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El deterioro de nuestro planeta está en continuo avance y aunque ya no se puede detener este problema, si es posible realizar acciones que conlleven a respetar más el entorno y a cuidar la naturaleza. En este sentido, se percibe la necesidad de diseñar un instrumento que logre evaluar la Conciencia Ambiental (CA) en el ámbito universitario debido a que este escenario es clave en los procesos de transformación de la sociedad a partir de las cuatro dimensiones que conforman este concepto: afectiva, cognitiva, conativa y activa. Se seleccionaron 62 ítems con un formato de escala tipo Likert que tienen relación con la escala NEP , quedando 44 ítems después de hallar su fiabilidad por medio del alfa de Cronbach. Este proyecto trabaja con estudiantes de pregrado la Uptc sede central y en éste se dejará el tamaño de muestra adecuado para aplicar el cuestionario en esta universidad.

***Palabras clave*:** Conciencia Ambiental, Escala NEP, cuestionario.

Abstract

The deterioration of our planet is in continuous advance and although no longer this problem can be stopped, if it is possible to conduct battles that entail to respect plus the surroundings and to take care of the nature. In this sense, the necessity is perceived to design an instrument that manages to evaluate Environmental Awareness (AE) in the university scope because this scene is key in the processes of transformation of the society from the four dimensions that conform this concept: affective, mental, conative and it activates. They selected to 62 items with a scale format Likert type that have relation with scale NEP, being 44 items after finding his reliability by means of the alpha of Cronbach. This project works with undergraduate course students the Uptc headquarters and in this one the suitable sample will be let to apply the questionnaire in this university.

***Key words*:** Environmental Awareness, Scale NEP, Questionnaire .

1. Introducción

La actitud con respecto a los cambios que presenta el medio ambiente está teniendo cada vez más impacto en la sociedad actual debido a que el deterioro a nivel mundial en los últimos años se ha agudizado a través del uso indiscriminado de los recursos naturales y el poco interés por parte de los países que ocasionan mayores daños ambientales, lo que afecta tanto a la naturaleza como a los seres humanos y es por esto, que el fortalecimiento de la CA es esencial.

[a]Estudiante de Especialización en Estadística. E-mail: ingrith.quintero@uptc.edu.co
[b]Profesor asistente. E-mail: reinaldo.alarcon@uptc.edu.co

Ahora, como la universidad es un espacio generador de actitud crítica constructiva y social, es allí donde se debe incentivar a los estudiantes por medio de diversas actividades para tener conocimiento sobre la situación ambiental actual, sus causas y sus consecuencias y así, tratar de frenar el aumento en los índices de contaminación. De igual manera, en la universidad es donde se construye y se transmiten conocimientos y por tanto es responsable de inculcar en sus estudiantes actitudes que contribuyan con el mejoramiento del entorno.

Por lo anterior, se realiza la siguiente pregunta de investigación: ¿Es el cuestionario un instrumento adecuado para medir el grado de CA en los estudiantes universitarios?

Diversas encuestas y escalas se han realizado para medir actitudes pro ambientales, pero las escalas frecuentemente más usadas según Berenger, Corraliza, Moreno y Rodríguez (2002) son la de Preocupación ambiental (Weigel y Weigel 1978) y Nuevo Paradigma Ambiental (Dunlap y Van Liere 1978). La Primera, hace referencia fundamentalmente a creencias ambientales referidas a temas ambientales concretos y la segunda, pretende abarcar la visión de la relación del ser humano-naturaleza evaluando el conjunto de creencias que explican cómo funciona el mundo y la biosfera y cómo esta es afectada por las conductas humanas.

Entre los estudios que utilizaron estas escalas para construir encuestas con el fin de evaluar las actitudes ambientales caben resaltar: (Castanedo 1995) el cual tenía como objeto de estudio elaborar un instrumento eficaz que mida las actitudes pro ambientales de estudiantes universitarios. La escala de actitudes que utiliza contiene ítems de las escalas de Weigel y Weigel (1978) y Van Liere y Dunlap (1981), aunque fueron modificados para adaptarlos al contexto y además utilizaron el método summativo de Likert. Para este proyecto, la construcción de los ítems de la Escala de actitudes pro ambientales siguió la tesis planteada por Maloney y Ward (1973, 1975) al afirmar que: "la crisis ecológica no es un problema tecnológico sino que es una crisis de conducta inadecuada e inadaptada". Otra investigación a nombrar es la de (Berenger, Corraliza, Moreno y Rodríguez 2002) cuyo objeto de estudio es el mismo de Castanedo pero con un enfoque diferente, en este trabajo, igualmente, se diseñó un instrumento de evaluación de actitudes ambientales pero basado en tres premisas: la primera es la necesidad de identificar y diferenciar los contenidos de la evaluación en actitudes ambientales a nivel personal y contextual, la segunda es la necesidad de contemplar la evaluación de la actitud ambiental a nivel general y la tercera premisa es la diferenciación propuesta por Dunlap y Van Liere en cuanto a los temas relevantes en el comportamiento ambiental. Según esto, concluyeron que la problemática ya no se centra en el hecho de la sensibilización ambiental de la sociedad, sino en cómo comprender esta sensibilidad social. De igual manera, Jiménez y Jiménez (2006) manejan la definición de conciencia ambiental bajo las cuatro dimensiones que se estudiarán en este proyecto y analizan dicho concepto basándose en los resultados obtenidos en el Ecobárometro de andaluz, interpretando los resultados analizándolos de manera que puedan ayudar y mejorar en la medición de encuestas donde se desee saber el nivel de conciencia ambiental de los ciudadanos y finalmente, en el trabajo realizado por Gómera, Villamandos y Vaquero (2012) se tomó una muestra de 1082 estudiantes y se implementó una herramienta para la medición de la conciencia ambiental (CA). Así, se diseñó el cuestionario "Conciencia ambiental en los centros universitarios", donde se estudió la CA del alumnado universitario a partir de cuatro dimensiones: cognitiva, afectiva, conativa y activa. Dando como resultado la clasificación de ésta, en tres categorías: CA mayor, CA menor y no condicionados por la CA. Estas tres categorías guardan relación con variables como el género, la edad y el perfil académico, llegando a la conclusión que la Universidad de Córdoba no implementaba aún procesos para la ambientalización curricular. Así, se puede ver la diferencia con Berenger , ya que él solo se centra en las actitudes que conllevan al deterioro del medio ambiente mientras que en Gómera y Lafuente se emplea el término de conciencia ambiental a partir de las cuatro dimensiones.

Así, el presente trabajo tiene como finalidad presentar el diseño y validación de un cuestionario que permita evaluar la CA en sus cuatro dimensiones (afectiva, cognitiva, conativa y activa)

2. Referente Conceptual

En el siguiente apartado se dan a conocer algunos conceptos relacionados con la CA, sus escalas de medición, fiabilidad de un cuestionario y algunas definiciones referentes a muestreo.

2.1. Conciencia Ambiental

Según (Jiménez y Jiménez 2006), la conciencia ambiental es entendida como el conjunto de percepciones, opiniones y conocimientos acerca del medio ambiente, así como de disposiciones y acciones (individuales y colectivas) relacionadas con la protección y mejora de los problemas ambientales. Se trata de un concepto multidimensional en el que, desde una perspectiva analítica, podemos distinguir cuatro dimensiones: afectiva, cognitiva, conativa y activa.

2.2. Dimensiones de la CA

Como se observa, la CA es una integración de varios aspectos actitudinales y comportamentales que están relacionados, por tanto, se clasifica en cuatro dimensiones:

2.2.1. Dimensión Afectiva

Aquella referida a los sentimientos de preocupación por el estado del medio ambiente y el grado de adhesión a valores culturales favorables a la protección de la naturaleza. Se trata de las emociones.

2.2.2. Dimensión Cognitiva

Se refiere al grado de información y conocimiento acerca de las problemáticas ambientales, así como de los organismos responsables en materia ambiental y de sus actuaciones. Se trata de ideas.

2.2.3. Dimensión Conativa

Se refiere a la disposición a actuar personalmente con criterios ecológicos y a aceptar los costes personales asociados a intervenciones gubernamentales en materia del medio ambiente. Se trata de actitudes

2.2.4. Dimensión Activa

La planificación estadística estratégica de entidades gubernamentales tiene como objeto organizar y priorizar la información estadística que se produce dentro de las entidades del orden nacional.

2.3. Escala NEP

La escala "Nuevo Paradigma Ecológico (NEP)" es una herramienta ampliamente utilizada a la hora de medir actitudes pro ambientales. Además, es una escala de creencias generales de la relación del ser humano con el Medio Ambiente, y se relaciona con una serie de valores y de intenciones de conducta pro ambiental. El éxito de este instrumento radica en que es un aparato teórico y empírico capaz de medir con relativa fiabilidad el grado de adhesión de la población a los valores pro ambientales. (Irurtia 2012)

2.4. Escala Likert

La escala tipo Likert es un instrumento de medición o recolección de datos cuantitativos utilizado dentro de la investigación, se basa en medir actitudes. Para (Maldonado 2007), es un tipo de escala aditiva que corresponde a un nivel de medición ordinal; consiste en una serie de ítems o juicios a modo de afirmaciones ante los cuales se solicita la reacción del sujeto y las respuestas son solicitadas en términos de grados de acuerdo o desacuerdo que el sujeto tenga con la sentencia en particular. Son cinco el número de opciones de respuesta más usado, donde a cada categoría se le asigna un valor numérico que llevará al sujeto a una puntuación total producto de las puntuaciones de todos los ítems. Dicha puntuación final indica la posición del sujeto dentro de la escala.

2.5. Confiabilidad de un instrumento

Grado en que un instrumento produce resultados consistentes y coherentes. Es decir en que su aplicación repetida al mismo sujeto u objeto produce resultados iguales. Para calcular la confiabilidad de un instrumento se emplea el Alfa de Cronbach.

2.6. Alfa de Cronbach

Evalúa la magnitud en que los ítems de un instrumento están correlacionados. Está dado por:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum S_i^2}{S_T^2}\right]$$

donde:
K= Número de ítems
S_i^2= Sumatoria de las varianzas de los ítems
S_T^2= Varianza de la suma de los ítems
α= Coeficiente de alfa Cronbach

2.7. Muestreo

El muestreo tiene como objetivo determinar que parte de la población debe examinarse para a partir de ella realizar la inferencia deseada. Existen diferentes tipos de clasificación de muestreos pero básicamente se dividen en dos: muestreos probabilísticos y muestreos no probabilísticos.

2.7.1. Muestreo Probabilístico

Para Sarndal, Swensson y Wietman (1992) es un enfoque para seleccionar una muestra que satisface ciertas condiciones: se puede definir el conjunto de todas las muestras posibles, se conoce la probabilidad de selección de cada muestra p(s), cada elemento de la población tiene una probabilidad de selección distinta de cero y en la selección de la muestra aleatoria cada muestra tiene la probabilidad p(s). Dentro de los muestreos probabilísticos los utilizados con más frecuencia son: Muestreo aleatorio sistemático, Muestreo aleatorio estratificado, muestreo aleatorio por conglomerados y muestreo aleatorio simple.

2.7.2. Muestreo no Probabilístico

Son aquellos que se utilizan cuando el muestreo probabilístico es excesivamente costoso, aunque este tipo de muestreo no se utiliza para realizar inferencias debido a que no se tiene la certeza de que la muestra extraída de la población sea representativa ya que todos los sujetos de la población no tienen la misma probabilidad de ser elegidos. Dentro de este tipo de muestreo se mencionan los más comunes a la hora de realizar investigaciones: Muestreo por cuotas, muestreo por conveniencia y la bola de nieve.

2.7.3. Muestreo Aleatorio Simple

Según Cochran (1996) “Es el método de selección de n unidades de una población de tamaño N de tal modo que cada una de las muestras posibles tenga la misma oportunidad de ser elegida”.

2.8. Estimación

Es un procedimiento de la estadística inferencial mediante el cual se realizan cálculos con los datos de una muestra para obtener valores o resultados que describan las características de la población (Martínez 2012).

2.8.1. Estimación de una proporción poblacional mediante un intervalo de confianza

Un estimador de intervalo de confianza de la proporción en la población p es un intervalo, calculado a partir de los datos de la muestra, en el cual nosotros ?confiamos? se encuentra la proporción de la población. Está dado por:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

2.8.2. Error estándar y tamaño de la muestra para estimar una proporción cuando la población en finita

En el caso de que se tenga una proporción finita y un muestreo sin reemplazo, el error de estimación se convierte en:

$$e = \frac{Z_{\alpha/2}pq}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

despejando n se tiene:

$$n = \frac{Z_{\alpha/2}pqN}{e^2(N-1) + Z^2_{\alpha/2}pq}$$

donde:
N= Total de la Población
p= Proporción esperada
q= 1-p
e=Precisión
n= Tamaño de la muestra

3. Metodología

Para el desarrollo de este proyecto se empleó un enfoque cuantitativo y un tipo de investigación descriptivo. Las etapas que se llevaron a cabo para el desarrollo del proyecto fueron las siguientes:

Primera Etapa: Revisión bibliográfica: Esta etapa consistió en consultar, analizar y resumir los estudios realizados acerca del objeto de investigación.

Segunda Etapa: Diseño del cuestionario: Esta etapa consistió en elaborar un cuestionario de tipo mixto, contemplando en su mayoría un tipo de respuesta Likert, aunque también se manejaron semicerradas y múltiples con el fin de evaluar las cuatro dimensiones de la conciencia ambiental y para esto se tomaron como referencias las encuestas realizadas por Jiménez y Jiménez (2006) y el de Gómera, Villamandos y Vaquero (2012). Además, se utilizaron algunos ítems de la escala NEP de Dunlap, Van Liere, Mertig y Jones (2000) y Van Liere y los de dichos autores con la modificación e implementación de algunos ítems.

Tercera Etapa: Validación del cuestionario: En esta etapa se realizó una prueba piloto con 30 estudiantes de la UPTC sede central con fin de validar el instrumento y encontrar sus posibles fallas. Además, se identificó la confiabilidad del instrumento hallando el alfa de Cronbach para cada una de las dimensiones de la CA

Cuarta Etapa: Población, tipo de muestreo y tamaño de muestra: Para esta investigación se plantea realizar un muestreo aleatorio simple, tomando como población el total de estudiantes de pregrado de la UPTC que son 25521 y utilizando la fórmula para tamaño de la muestra de una proporción se encuentra la muestra a la cual se le aplica la encuesta diseñada.

4. Resultados

Como resultados se tiene primero que todo, la adaptación de los dos cuestionarios que se emplearon para formar el cuestionario final. Recordemos que estos cuestionarios ya se encuentran validados; el realizado por Gómera (2012) mostró una consistencia interna $>$0.7 y mostró la validez a través de las correlaciones con una serie de variables consideradas como criterio de validez teórica (valores ecológicos de la Escala de

Schwartz y valoración de la responsabilidad en el deterioro medioambiental) y el de Lafuente y Jiménez (2006) mediante el análisis de componentes principales categórico y en el cual realizaron test de fiabilidad para cada dimensión dando resultados positivos. En esta adaptación se realizaron algunos cambios del lenguaje y de la misma manera se implementaron algunas preguntas referentes específicamente al contexto de la UPTC, y otras referentes al uso de la tecnología y sus implicaciones en el medio ambiente, logrando así tener un cuestionario que midiera las cuatro dimensiones de la CA. A continuación se presenta la descripción de los 62 ítems iniciales y la manera como fueron codificadas las respuestas para cada dimensión:

RESUMEN DEL CUESTIONARIO			
Dimension	Pregunta	Item	Codificación
Afectiva	Los seres humanos están abusando seriamente del medio ambiente	1	Valoración escalar: 1. No se preocupa nada 2. Se preocupa poco 3. Se preocupa algo 4. Se preocupa bastante 5. Se preocupa mucho
Afectiva	Situación del medio ambiente en el mundo	2	Valoración escalar: 1. Es muy buena 2. Es buena 3. Es regular 4. Es mala 5. Es muy mala
Afectiva	Situación del medio ambiente en la propia localidad	3	Valoración escalar: 1. Es muy buena 2. Es buena 3. Es regular 4. Es mala 5. Es muy mala
Afectiva	Sentimiento ecologista	4	Valoración escalar: 1. Nada ecologista 2. Poco ecologista 3. Regular ecologista 4. Bastante ecologista 5. Muy ecologista

Afectiva	Inconvenientes con el auto privado	5	Valoración escalar: 1. Sufrimiento causado por loa accidentes de tránsito 2. Precio del combustible 3. Gastos de compra y mantenimiento 4. Efectos que provoca en la contaminación del aire
Afectiva	Introducir alguna mejora de carácter medio ambiental en la UPTC	7	Valoración escalar: 1. No considera mejoras 2. Considera una mejora 3. Considera dos mejoras 4. Considera mas de dos mejoras
Afectiva	Consideración del medio ambiente como uno de los problemas actuales que más le preocupan	9	Valoración escalar: 1. Si no recibe puntos 2. Si recibe un punto 3. Si recibe dos puntos 4. Si recibe tres puntos
Afectiva	Grado de Percepción de los problemas ambientales en la UPTC	10	Valoración escalar: 1. Muy mala percepción 2. Baja percerpción 3. Aceptable percepción 4. Buena percepción 5. Muy buena percepción
Afectiva	Preocupación y sentimiento de protección por el medio ambiente	15.1, 15.2, 15.4, 15.8, 15.9 y 15.10	Valoración escalar: 1. Totalmente de acuerdo 2. De acuerdo 3. Ni de acuerdo ni en desacuerdo 4. Desacuerdo 5. Totalmente en desacuerdo

Afectiva	Preocupación y sentimiento de protección por el medio ambiente	15.3, 15.5, 15.6, 15.7, 15.11, 15.12, 15.13	Valoración escalar: 1. Totalmente en Desacuerdo 2. En Desacuerdo 3. Ni de acuerdo ni en desacuerdo 4. De acuerdo 5. Totalmente en deacuerdo
Cognitiva	Grado de información sobre asuntos relacionados con el medio ambiente en la UPTC	11	Valoración escalar: 1. Muy poco informado 2. poco informado 3. Regular informado 4. Bastante informado 5. Muy informado
Cognitiva	Conocimiento del organismo dedicado a la protección del medio ambiente en la UPTC	12	Valoración escalar: 1. No lo conoce 2. Dice que lo conoce pero la respuesta no es correcta 3. Algo ha oido pero no sabe concretar 4. lo conoce y lo cita correctamente
Cognitiva	Conocimiento sobre situaciones relacionadas con el medio ambiente	16.1, 16.2, 16.3, 16.8, 16.10, 16.11, 16.13, 16.14	Valoración escalar: 1. Totalmente verdadera 2. Probablemente verdadera 3. No sabe, no contesta 4. Probablemente falsa 5. Totalmente falsa

Cognitiva	Conocimiento sobre situaciones relacionadas con el medio ambiente	16.4, 16.5, 16.6, 16.7, 16.9, 16.12, 16.15	Valoración escalar: 1. Totalmente falsa 2. Probablemente falsa 3. No sabe, no contesta 4. Probablemente verdadera 5. Totalmente verdadera
Conativa	Grado en que considera que la propia actividad cotidiana afecta al medio ambiente	6	Valoración escalar: 1. No, nada 2. Si, un poco 3. Si, regular 4. Si, bastante 5. Si, mucho
Conativa	Disposición a recibir formación / información ambiental	8	Valoración escalar: 1. Ninguna modalidad 2. Una modalidad 3. Dos modalidades 4. Tres modalidades 5. Cuatro o más modalidades
Conativa	Disposición a apoyar medidas destinadas a mejorar la protección del medio ambiente.	17.1, 17.2, 17.3, 17.7	Valoración escalar: 1. Totalmente en contra 2. Más bien en contra 3. Ni a favor ni en contra 4. Más bien a favor 5. Totalmente a favor
Conativa	Disposición a apoyar medidas destinadas a mejorar la protección del medio ambiente.	17.4, 17.5, 17.6	Valoración escalar: 1. Totalmente a favor 2. Más bien a favor 3. Ni a favor ni en contra 4. Más bien en contra 5. Totalmente en contra

Activa	Medio de transporte empleado para ir a clase.	13	Valoración escalar: 1. Automóvil particular 2. Transporte público 3. Moto 4. Bicicleta 5. Caminando
Activa	Principal razón para emplear ese medio de transporte	14	Valoración escalar: 1. Comodidad 2. Económica 3. Salud 4. Distancia / tiempo 5. Respeto por el medio ambiente
Activa	Comportamientos relacionados con el reciclaje de basuras y otros residuos sólidos urbanos	18.1-18.5 y 19.1-19.8	Valoración escalar: 1. No lo ha hecho ni lo haría 2. No lo ha hecho, pero está dispuesto a hacerlo 3. Lo ha hecho alguna vez 4. Lo hace siempre

Tabla 1: Codificación de los ítems para cada dimensión de la CA

Para validar el cuestionario se realizó una prueba piloto a 30 estudiantes de pregrado de la UPTC sede Tunja en la cual se observó que:

El cuestionario está un poco largo dado como sugerencia por algunos estudiantes a los que se les aplicó la encuesta, por tanto, cuando se le aplique a la muestra real, se debe analizar si hay algunas muy parecidas para seleccionar las que tienen más relevancia. Por ejemplo, la pregunta 15.7: "El equilibrio de la naturaleza es muy delicado y fácilmente alterable" y la pregunta 15.10 "El equilibrio de la naturaleza es lo bastante fuerte para hacer frente al impacto que los países industrializados le causan". En estas dos preguntas se observa que las dos se refieren al equilibrio de la naturaleza luego pueden ser muy parecidas, pero también se tiene que analizar que la primera está en sentido positivo y la segunda está en sentido negativo lo que serviría para indicar si el estudiante está respondiendo de manera consiente.

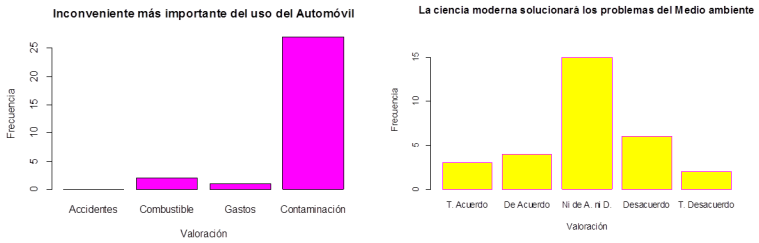
La otra observación es en las preguntas número 9 y 10, ya que en éstas, deben puntuar solo las tres problemáticas que considere más importantes y 5 de los 30 encuestados dieron una puntuación a todos los problemas que se presentaban en dichos puntos, por tanto, se aconseja dar una mejor redacción para que sea entendible que sólo tienen que marcar las tres más relevantes. En cuanto a la pregunta número 7 que es una pregunta semi cerrada, se observa que es muy difícil para codificarla, pero a la vez da más información ya que especifica la mejora de carácter medio ambiental a realizar en la UPTC, por tanto habría a considerar si es mejor dejarla, eliminarla o dejarla totalmente cerrada para mayor facilidad en el análisis.

En esta prueba piloto se realizó una base de datos y se hizo un análisis descriptivo donde se determinó con respecto al sentimiento de preocupación, es decir la dimensión afectiva, que los estudiantes tienen un

alto nivel; en cuanto al nivel cognitivo los estudiantes tienen un buen nivel, pero baja un poco con respecto a la dimensión afectiva y finalmente lo que tiene que ver con las actitudes y conductas, es decir la dimensión conativa y activa ya tienen un nivel bajo. Pues casi no realizan acciones en pro del medio ambiente aunque están en su mayoría dispuestos a realizarlas.

Para analizar la fiabilidad del cuestionario se nota que una encuesta no fue desarrollada de manera consciente, razón por la cual se trabaja solo con las 29 restantes. Así, se analiza la consistencia interna del cuestionario utilizando el alfa de Cronbach por medio del software R para cada una de las dimensiones de la CA.

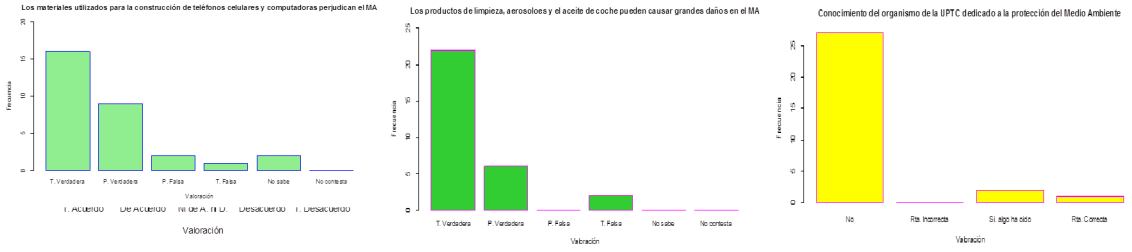
- Con respecto a la dimensión afectiva inicialmente eran 21 afirmaciones como se muestra en la TABLA 1, al realizar el análisis se excluyeron cinco de ellas, quedando solamente 16. Los ítems eliminados fueron: las preguntas 7,9 y 10 por los problemas obtenidos en la prueba piloto; la pregunta 5 por unanimidad de criterio, ya que la mayoría de los estudiantes encuestados escogieron la opción "Los efectos que provoca en la contaminación del aire", lo cual logra que disminuya la varianza y por tanto, el alfa de Cronbach. Finalmente, la pregunta 15.1 se excluyó debido a que la mayoría de los encuestados optaron por la opción intermedia: "Ni de acuerdo ni en Desacuerdo, lo cual no proporciona mucha información en la investigación.



((A)) Pregunta 5 ((B)) Pregunta 151

FIGURA 1: Preguntas 5 y 15.1

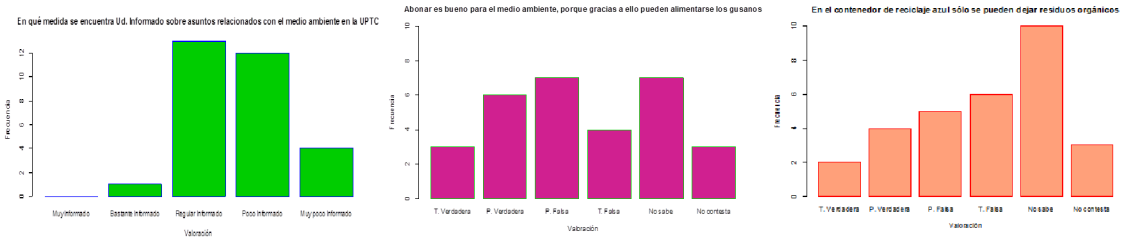
- Ahora, para la dimensión cognitiva, inicialmente eran 17 ítems quedando solamente 11. Las preguntas excluídas fueron: 16.6,16.12 y la pregunta 12 por unanimidad de criterio



((A)) P16.6 ((B)) P16.12 ((C)) P12

FIGURA 2: Preguntas 16.6, 16.12 y 12

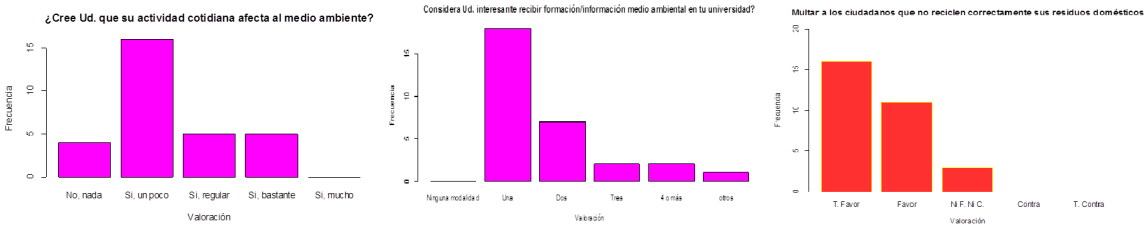
De la misma manera, se elimió la pregunta número 11, debido a que es una pregunta muy específica de la UPTC, aunque se puede observa la falta de información respecto a la entidad encargada de proteger al medio ambiente en esta institución. Igualmente, la pregunta 16.10 ya que permite cualquier tipo de respuesta válida y la pregunta 16.14 debido a que en sumayoría optaron por la opción intermedia "No sabe, no contesta".



((A)) P11 ((B)) P16.10 ((C)) P16.14

FIGURA 3: Preguntas 11, 16.10 y 14.14

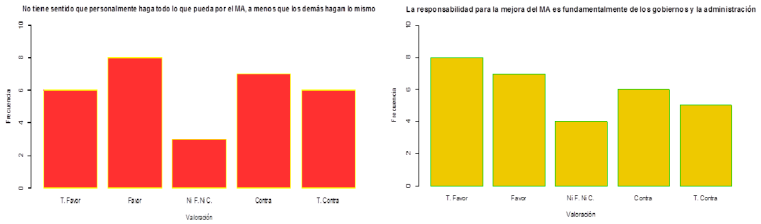
- Con respecto a la dimensión conativa, inicialmente eran 9 preguntas, quedando solamente 3. Se eliminaron las preguntas 6, 8 y 17.3 por unanimidad de criterio y como se dijo anteriormente, esto conlleva a disminuir el alfa de Cronbach.



((A)) P6 ((B)) P.8 ((C)) P17.3

FIGURA 4: Preguntas 6, 8 y 17.3

Igualmente, se elimiaron las preguntas 17.4 y 17.6 debido a que tenían aproximadamente el mismo porcentaje de puntuaciones bajas que altas y la pregunta 17.5 ya que muestra poca correlación y su item-test, no es significativo.



((A)) P17.4 ((B)) P.17.6

FIGURA 5: Preguntas 17.4 y 17.6

- Finalmente, para la dimensión activa se eliminó solo la pregunta 14, debido a que la respuesta es de las necesidades particulares del individuo, no tanto de la acción que realiza para mejorar el medio ambiente.

De esta manera, el cuestionario quedó reducido a 44 ítems, y a continuación se muestra el cálculo del alfa de Cronbach para cada dimensión ya teniendo sólo los 44 ítems.

Dimensión de la CA	Alfa de Cronbach
Afectiva	0.80
Cognitiva	0.59
Conativa	0.66
Activa	0.82

Tabla 2: Alfa de Cronbach

Lo cual, indica que el cuestionario tiene una fiabilidad buena, aunque un poco baja para la dimensión cognitiva, aunque para algunos autores desde 0.5 es bueno en la primera etapa de la investigación.

Para terminar, en la última etapa se deseaba conocer el tamaño de la muestra a la cual se le debe aplicar el cuestionario elaborado. Para esto, se tomó toda la población objeto que en este caso son 25521 estudiantes de pregrado de la UPTC sede Tunja, y aplicando la fórmula del tamaño de muestra para una proporción con una precisión esperada del 5 %, una proporción esperada del 50 % y un nivel de confianza del 95 %, obtenemos:

$$n = \frac{Z_{\alpha/2} pqN}{e^2(N-1) + Z^2_{\alpha/2} pq}$$

reemplazando:

$$n = \frac{(1.96)^2 * 0.5 * 0.5 * 25521}{(0.05)^2 * 25520 + 1.96 * 0.5 * 0.5} = 378$$

Finalmente, se considera que el estudiante tiene alto grado de CA si obtiene más de 163 puntos en el cuestionario elaborado

5. Conclusiones

Se presenta una herramienta que es una adaptación de dos cuestionarios validados y que son ampliaciones de la Escala NEP en cuanto al aspecto de percepción, la cual permite medir el grado de CA a partir de las cuatro dimensiones que trabaja esta definición mediante la suma de las respuestas dadas que son calculadas a partir de la codificación dada mediante la Escala tipo Likert.

Se validó el cuestionario por medio de la prueba piloto en el cual se detectaron algunas falencias y se da el respectivo análisis para mejorar el cuestionario en el momento que vaya a ser aplicado.

Se concluye que el cuestionario es confiable por medio del alfa de Conbrach que en su totalidad es de 0,8 aunque está un poco bajo en la dimensión cognitiva, pero en general es un instrumento confiable y logra medir las cuatro dimensiones de la CA.

Finalmente, se presenta la muestra respectiva a la cual se debe implementar el cuestionario por medio de un Muestreo Aleatorio simple y teniendo en cuenta la fórmula para estimar el tamaño de la muestra para un proporción.

Referencias Bibliográficas

Berenger, J., Corraliza, J., Moreno, M. y Rodríguez, L. (2002), 'La medida de las actitudes ambientales: propuesta de una escala de conciencia ambiental (Ecobarómetro)', *Intervención Psicosocial* **11**(3), 349–358.

Castanedo, C. (1995), 'Escala para la evaluación de las actitudes pro-ambientales (EAPA) de alumnos universitarios', *Revista Complutense de Educación* **6**(2), 254–277.

Cochran, W. (1996), *Técnicas de muestreo*, Continenta, S.A., México.

Dunlap, E., Van Liere, K., Mertig, A. y Jones, R. (2000), 'Measuring endorsement of the New Ecological Paradigm: A revised NEP scale', *Journal of Social Issues* **56**(3), 425–442.

Gómera, A., Villamandos, F. y Vaquero, M. (2012), 'Medición y categorización de la conciencia ambiental del alumnado universitario: Contribución de la universidad a su fortalecimiento.', *Profesorado. Revista de Curriculum y Formación de Profesorado* **16**(2), 193–212.

Irurtia, A. (2012), Conciencia ambiental en la educación secundaria: hacia una nueva percepción., Master's thesis, Universidad Pública de Navarra.

Jiménez, M. y Jiménez, M. a Lafuente, R. (2006), *La operacionalización del concepto de conciencia ambiental en las encuestas*, R. de Castro (Coord.). Persona, Sociedad y Medio ambiente.

Maldonado, S. (2007), 'Manual práctico para el diseño de la Escala Likert', *Xihmai* **2**(4), 14.

Martínez, C. (2012), *Estadística y Muestreo*, Bogotá, Colombia: Ecoe ediciones.

Sarndal, C., Swensson, B. y Wietman, J. (1992), *Model Assisted Survey Sampling*, Springer.

DESCRIPCIÓN Y CRUCE DE VARIABLES RELACIONADAS CON LA ACCIDENTALIDAD EN LA CIUDAD DE TUNJA
La información corresponde al periodo 2014 y 2015, las variables que se emplean están relacionadas con la víctima.

Especialización en Estadística

Eliana Ibeth Moyano Alba[1,a], Sandra Patricia Cárdenas Ojeda[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

Este artículo presenta la descripción y análisis de las diferentes variables que intervienen en un accidente de tránsito en Tunja. Esta información se encuentra recopilada en una base de datos disponible en la DIJIN de la Policía Nacional, la cual cuenta con información especifica que permite caracterizar a las victimas de dichos accidentes y también al accidente como tal. En el trabajo se hace un cruce de variables para determinar dependencia y nivel de asociación entre algunas de las variables de estudio. Finalmente, se muestran conclusiones del trabajo y bibliografía con la que se realizó el estudio.

***Palabras clave*:** accidentalidad, descripción univariada, tabla de contingencia, prueba de independencia, medidas de asociación.

Abstract

This article presents the description and the analysis of different variables that take part in the accidents of traffic in Tunja. This information is gather in a database in the DIJIN of the Policía Nacional, it has specific information that allows characterize to the victims of the accidents and the situation presented. In this work of application is a cross of variables to decide dependency and the association level among some study variables. Finally, it shows conclusions and bibliography of the word.

***Key words*:** accident, Univariate description, contingency table, test of independence, measures of association..

1. Introducción

La accidentalidad es considerada hoy en día como un problema de salud pública. De acuerdo con (Organización Mundial de la Salud 2004) se calcula que 25 % de todas las muertes debidas a lesiones, son resultado de las lesiones causadas por accidentes de tránsito. Las principales causas de estas muertes incluyen: conducir bajo la influencia del alcohol, manejar a alta velocidad y no usar el cinturón de seguridad.

La ciudad de Tunja, capital del departamento de Boyacá no es la excepción de este problema, y aunque ya se realizó un estudio hace 3 años acerca de la movilidad en esta ciudad, donde se describieron algunos aspectos de la accidentalidad, no hay reportes de estudios actuales que muestren cuales son los posibles

[a]Estudiante de Especialización en Estadística. E-mail: elim1220@hotmail.com
[b]Profesor asistente. E-mail: sandra.cardenas@uptc.edu.co

factores, los sitios más frecuentes de ocurrencia, los afectados por un accidente, entre otras variables que conducen a caracterizar estas situaciones en la ciudad. Además, es fundamental establecer las relaciones que se pueden dar entre las distintas variables de estudio.

Teniendo en cuenta la descripción anterior, la pregunta que pretende responderse con esta investigación es: ¿Qué características tiene la accidentalidad en la ciudad de Tunja Boyacá y cómo se relacionan unas variables con otras?

Los accidentes de tránsito son considerados como un gran problema mundial, ya que es muy común escuchar a diario noticas sobre este tema donde se ven involucrados tanto peatones como conductores, y donde las consecuencias de estos incidentes van desde daños materiales, hasta personas heridas y en el peor de los casos, el fallecimiento de algunas de ellas. Por ende, se hace necesario estudiar acerca de los factores que son causales de dicha accidentalidad, pues es de esta manera como se puede tener mejor conciencia para tomar medidas preventivas que eviten que esta "epidemia" siga siendo la autora de tantas tragedias.

Por medio de la realización de este proyecto, se pretende abordar una temática de gran interés, ya que la accidentalidad es un aspecto que afecta directamente a los Tunjanos y que con base en su estudio y divulgación, pueden formularse soluciones a dicho problema y posiblemente crear programas que permitan la prevención de accidentes en las vías de Tunja, así como la efectiva concientización vial de las personas al volante acerca del cumplimiento de las normas de tránsito.

Por otro lado, el estudio se realiza a través de información disponible en una base de datos facilitada por la Policía Nacional, más específicamente, el departamento de la DIJIN, la cual será sometida a un análisis estadístico que permita jerarquizar, determinar y analizar cómo ha sido la accidentalidad en la ciudad los últimos tres años. Por medio del análisis de las diferentes variables, se pretender establecer ciertas características que se ven involucradas en la accidentalidad, así como observar las relaciones que se manifiestan entre algunas de ellas.

Para la realización de estudio, se toma como referencia diferentes estudios realizados a nivel local, nacional e internacional acerca de la accidentalidad y los diferentes factores que se ven implicados. En el año 2012, la alcaldía de Tunja a través del convenio interadministrativo 010 con el grupo de investigación GIDPOT de la facultad de Ingeniería de la Universidad Pedagógica y Tecnológica de Colombia, realizó una caracterización de la movilidad en la ciudad donde se describieron aspectos urbanísticos, infraestructuras y sistemas de regulación y control, los sistemas de transporte presentes, el tránsito en la ciudad, transporte de mercancía, y lo finalmente, lo que más interesa en esta investigación, la seguridad vial y la accidentalidad. En esta investigación se logró caracterizar algunos aspectos relevantes del fenómeno de la accidentalidad en la ciudad, haciendo primero una comparación de ésta con los índices observados en otras ciudades, así como la identificación del tipo de vehículo que más se ve implicado en un accidente y los lugares de Tunja donde más se concentran estos casos. También se hace un análisis de las consecuencias de dicha accidentalidad: registros de morbilidad y mortalidad. Para llevar a cabo este estudio, los investigadores se valen de información disponible en las bases de datos de Medicina Legal de la ciudad.

En Colombia, la Policía Nacional realizó un estudio acerca de la incidencia del factor humano en la accidentalidad vial en Colombia por medio de un diseño descriptivo-correlacional. La información para el estudio fue recolectada a través de un cuestionario y encuesta tipo Likert. Entre los resultados que se obtuvieron esta que los conductores con nivel educativo superior tienen menor participación en accidentes y que los peatones con menor nivel educativo inciden en más conductas riesgosas (Norza, Granados, Useche, Romero y Moreno 2014).

Según estadísticas del Ministerio de Salud (Perdomo 2000) afirma que el accidente de tránsito a nivel nacional es la novena causa de mortalidad general y la segunda causa de muerte violenta en Colombia. En el año 2008 el Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLyCF) registró 5.290 defunciones y 40.377 personas lesionadas por causas relacionadas con accidentes de tránsito, y estimó que en promedio se presentaron quince muertos y ciento veinticinco lesiones cada día, y una muerte y cinco lesiones cada hora (Martínez 2008)

Un accidente de tránsito, según (Congreso, Colombia 2002) es un evento generalmente involuntario, generado al menos por un vehículo en movimiento, que causa daños a personas y bienes involucrados en él e igualmente afecta la normal circulación de los vehículos que se movilizan por la vía o vías comprendidas en el lugar o dentro de la zona de influencia del hecho.

Para hacer referencia a un accidente de tránsito, es necesario tener en cuenta los actores viales que se pueden ver involucrados en este. Los conductores de un vehículo: automóvil, motociclista o ciclista, las personas que transitan por las vías ya sea caminando o transportándose en otro vehículo, son quienes se ven implicados cuando sucede un accidente. Tomando como referencia lo establecido por Congreso, Colombia (2002) en la ley 769 de 2002, se definen los diferentes actores viales así: un conductor es la persona habilitada y capacitada técnica y teóricamente para operar un vehículo; pasajero es la persona distinta del conductor que se transporta en un vehículo público; motociclista es una persona que se desplaza, bien sea como conductor o como pasajero, en una motocicleta; ciclista se entiende como el conductor de bicicleta i triciclo; peatón es la persona que transita a pie o por una vía. Estos últimos actores son los usuarios más vulnerables de las vías, porque carecen de protección ante un impacto; por ende, son proclives a padecer atropellos.

Por lo anterior, se hace necesario distinguir los diferentes tipos de accidentes viales, pues a partir de ellos se establece quienes son los actores viales que pueden estar implicados.

Dentro de las clases de accidentes de tránsito se encuentran las siguientes:

(a) Atropello, caracterizado por el encuentro de un vehículo con un peatón;

(b) Caída, caracterizada por el descenso o desprendimiento de un pasajero del vehículo en el que se transporta;

(c) Colisión, es embestirse dos o más vehículos en movimiento;

(d) Choque, es embestir un vehículo en movimiento contra otro detenido o contra obstáculos físicos;

(e) Volcamiento, es el giro de un vehículo en movimiento sobre su eje longitudinal o transversal respecto a su sentido de marcha, durante el cual apoya cualquier parte de su estructura después de abandonar la posición normal de rodaje, y

(f) Otros: cualquier accidente de tránsito no incluido dentro de la tipificación dada en Álvarez (2009).

Para la organización mundial de la salud citada por (Cabrera, Velásquez y Valladares 2009), uno de los factores de riesgo fundamentales implicados en la seguridad vial en el caso de los conductores y motociclistas es el exceso de velocidad, la conducción bajo los efectos del alcohol y la no utilización del cinturón de seguridad. Sin embargo, un accidente de tránsito no solo ocurre por las fallas humanas del conductor o del peatón, sino también por causas mecánicas del vehículo o factores climáticos que dificultan la visibilidad durante recorridos que se hacen por las diferentes vías.

Las causas de los accidentes de tránsito son múltiples y están relacionadas con factores propios del conductor, del vehículo y de las vías. Los factores humanos pueden constituir hasta el 90 % de la causalidad, entre estos factores se encuentran el exceso de velocidad, no respetar la señales de tránsito, la fatiga, factores propios de personalidad y el abuso de sustancias incluyendo el alcohol. La ingestión de bebidas alcohólicas es un factor de riesgo fuertemente asociado a las accidentes de tráfico (Martínez 2008).

2. Referente Conceptual

Para el análisis de la base de datos se requiere de la conceptualización acerca de los tipos de variables y escalas de medición de las mismas, la construcción de tablas de contingencia para asociar dos o más variables, a partir de las mismas, la utilización de las pruebas de independencia para determinar cuáles de ellas se relacionan y finalmente las medidas de asociación para variables de tipo nominal.

Esta conceptualización se basa en el libro análisis estadístico de datos categóricos de los profesores Díaz M. y Morales R. (2009).

La escala de medida de una variable categórica es un elemento importante para la selección del análisis estadístico apropiado. Una selección inadecuada de la escala de medida puede conducir a una estrategia estadística inapropiada que arrojaría conclusiones erróneas acerca de la realidad contenida en los datos.

Las variables dicotómicas son variables que tienen dos posibles respuestas, frecuentemente corresponden a la presencia o no de cierto atributo.

Las variables ordinales son aquellas que representan más de dos posibles resultados, y a veces estos resultados poseen un orden propio.

Si se dispone de variables con más de dos categorías, a las cuales no se les atribuye un orden, se tiene una variable de tipo nominal.

Una tabla de contingencia se asume como un arreglo bidimensional de f-filas por c-columnas (fc celdas). En general, se nota con n_{ij} a la frecuencia de la i-ésima modalidad de la variable fila y j- ésima de la variable columna.

2.1. Tablas de contingencia

Una tabla de contingencia de es un arreglo bidimensional, de una variable fila con $f-$categorías o modalidades y una variable columna con $c-$categorías, donde hay $f \times c$ celdas, las entradas de las celdas son las frecuencias o conteos del número de casos en cada una de las combinaciones de valores de ambas variables. En general, se nota con n_{ij} a la frecuencia de la i-ésima modalidad de la variable fila y j-ésima de la variable columna.

El total por fila o por columna está formado por las frecuencias marginales, y se notan por $n_{i.}$ (donde el punto señala que se suman columnas dentro de la fila i) y $n_{.j}$ (donde el punto señala que se suman filas dentro de la columna j), respectivamente.

La suma de las frecuencias por celda es igual a la suma de las frecuencias marginales e igual al número total de individuos seleccionados y clasificados; se nota por n.

De acuerdo con (Díaz M. y Morales R. 2009), la notación general, para una tabla de contingencia de $f-$filas y $c-$columnas, se muestra en la tabla 1.

Filas	Columnas						
	1	2	...	j	...	c	Total($n_{i.}$)
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
f	n_{f1}	n_{f2}	...	n_{fj}	...	$n_{f.}$	n
Total($n_{.j}$)	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..} = n$

Tabla 1: Tabla de contingencia

donde

- La frecuencia de la $i-$ésima modalidad de la variable fila y la modalidad $j-$ésima de la variable columna se escribe como n_{ij}.
- El total de observaciones en la $i-$ésima modalidad de la variable fila se nota por $n_{i.}$, es decir,

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij}$$

- El total de observaciones en la $j-$ésima modalidad de la variable columna se nota por $n_{.j}$; es decir,

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{fj} = \sum_{i=1}^{f} n_{ij}$$

- El número total de observaciones en la muestra se escribe con n, y es igual a la suma de los márgenes fila o columna, es decir,

$$n = \sum_{i=1}^{f} \sum_{j=1}^{c} n_{ij}$$

Por otra parte, las frecuencias pueden ser transformadas en proporciones o porcentajes. Un primer porcentaje se obtiene de dividir cada frecuencia n_{ij} por el número total de observaciones n; este porcentaje se escribe como f_{ij}, es decir,

$$f_{ij} = \frac{n_{ij}}{N} \times 100$$

la cantidad f_{ij} corresponde a la proporción o porcentaje de elementos que tienen los atributos i y j.

El segundo porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal fila $n_{i.}$, así:

$$f_{j|i} = \frac{n_{ij}}{ni.} \times 100$$

La cantidad $f_{j|i}$ es la proporción de elementos de cada celda, respecto al total de la fila i estas frecuencias corresponden al *perfil fila*.

El tercer porcentaje se obtiene al dividir cada frecuencia n_{ij} por la respectiva frecuencia marginal columna $n_{.j}$:

$$f_{ij} = \frac{n_{ij}}{n} \times 100$$

La cantidad $f_{i|j}$ es la proporción de elementos de cada celda, respecto al total de la columna j estas frecuencias corresponden al *perfil columna*.

2.2. Pruebas de independencia

Al disponer de la información en una tabla de contingencia, es posible indagar si las variables que constituyen dicha tabla son independientes o no.

la hipótesis nula de independencia está dada por:

Ho: La variable fila es independiente de la variable columna

La estadística de prueba que se empleada en el juzgamiento de esta hipótesis es:

$$\chi^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

que bajo la hipótesis nula de independencia, tiene distribución de probabilidad *ji-cuadrado* con $(f-1)\times(c-1)$ grados de libertad.

Se rechaza la hipótesis nula a un nivel α cuando se verifica que $\chi_0^2 > \chi^2_{(f-1)(c-1),\alpha}$

2.2. Medidas de asociación

A continuación algunas medidas relacionadas con la estadística ji-cuadrado, como se ver en (Díaz M. y Morales R. 2009, pág 38), algunas de estas son:

2.2.1. El coeficiente de contingencia

Es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando se tiene información clasificadora acerca de uno o ambos conjuntos de atributos. El grado de asociación entre dos conjuntos de atributos, sean ordinales o no, se puede describir mediante la siguiente fórmula:

$$C = \sqrt{\frac{\chi_0^2}{\chi_0^2 + n}}, \qquad \text{donde} \qquad \chi_0^2 = \sum_{i=1}^{f} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

La estadística C toma valores entre 0 y 1. Valores cercanos a cero muestran una baja asociación entre las variables, mientras que valores próximos a 1 indican una posible alta asociación.

2.2.2. El coeficiente (V) de Cramer

Este coeficiente tiene un valor máximo en tablas de contingencia de cualquier tamaño. Se define como

$$V = \sqrt{\frac{\chi_0^2}{nk}} \tag{3}$$

donde k =mín{f-1,c-1} es el menor número de modalidades fila (o columna) menos uno de la tabla de contingencia. Se trata de un coeficiente que toma el valor 1 cuando hay asociación perfecta entre los atributos, cualquiera que sea el tamaño de la tabla de contingencia.

3. Metodología

El desarrollo del trabajo de aplicación parte de una base de datos facilitada por la DIJIN de la Policía Nacional seccional Tunja, donde se describen los accidentes ocurridos en los años 2014 y 2015 en la ciudad a partir de diferentes variables. Por lo anterior, el enfoque que sobresale en este trabajo es cuantitativo con tipo de investigación descriptivo correlacional, pues se pretende hacer una descripción de las variables, y establecer relaciones entre las diferentes variables de estudio.

La base de datos sobre la accidentalidad cuenta con 286 filas y 16 columnas. Cada fila caracteriza una víctima de accidente y las columnas las variables de estudio:

Conducta (se refiere a si la persona murió o quedo lesionada), Zona (rural o urbana), Género de la víctima, Móvil agresor (tipo de trasporte en el que se movilizaba el victimario), Hipótesis(causa del accidente), Modalidad(tipo de accidente de tránsito), Edad(edad de la vćtima), Agrupación edad(menor, adolescente o adulto), Día de semana, Dirección (lugar exacto donde se presenta el accidente), Móvil víctima (el medio en que se trasportaba la víctima), Mes, Hora, Cargo persona (oficio u ocupación de la víctima), Hecho.id(código que se le asigna al accidente), Año(en que ocurre el accidente).

Con la variable edad de la víctima se hizo un análisis descriptivo univariado, haciendo uso de medidas de tendencia, de dispersión y localización, realización de gráficos (boxplot e histograma) a través del uso del software libre (R Core Team 2015) y Excel.

Luego de este proceso descriptivo, se construyeron tablas de contingencia, también usando (R Core Team 2015) para relacionar dos variables y determinar por medio de las prueba de independencia ji-cuadrado, la dependencia o no entre las diferentes parejas de variables que se han agrupado. Con las variables que presentaron alguna dependencia, se identificó el grado de asociación por medio de los coeficientes de contingencia y cramer.

Finalmente, se interpretaron y analizaron los resultados obtenidos para dar una caracterización de la accidentalidad en la ciudad de Tunja, y la consolidación de un informe dirigido a la entidad facilitadora de

la información.

La base de datos fue dividida en dos partes: una primera parte con variables que describían particularmente a las víctimas de los accidentes ocurridos en estos años; la otra parte, con las variables que específicamente hacían referencia a la descripción del accidente de tránsito.

4. Resultados

A continuación se presenta el análisis estadístico de cada una de las variables sobre la accidentalidad en Tunja.

Víctimas

Entre los años 2014 y 2015 se presentaron en la ciudad 243 accidentes de tránsito, los cuales generaron un total de 285 víctimas. Del total de accidentes, un 35 % ocurrieron en el año 2014 (99 accidentes) y el 65 % restante, en el año 2015 (186 accidentes).

Como se nombró con anterioridad, la base de datos hace un recuerdo de las víctimas describiendo aspectos como conducta (homicidio o lesión), género, edad, agrupación de edad, cargo u ocupación y móvil en el que la víctima de encontraba en el momento del accidente.

De las 285 víctimas de los accidentes en estos años, 15,09 % (43 personas) murieron a causa del accidente, mientras que 84,91 % (242 personas) resultaron lesionadas en este acto. La Figura 1 muestra la cantidad de muertes y lesiones que se presentaron en cada años, observándose una aumento considerable de lesionados del año 2014 al año 2015. De las 99 víctimas en 2014, 20 murieron y 79 resultaron lesionadas; para el año 2015, de las 186 víctimas, 23 murieron mientras que 163 quedaron lesionadas.

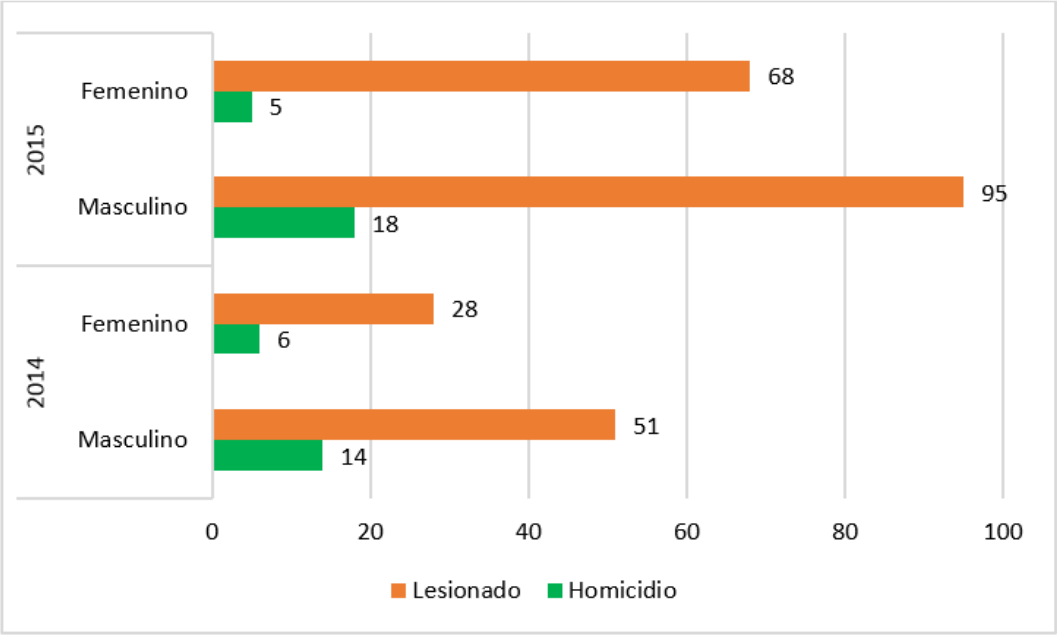


Figura 1: Cantidad de muertes y lesiones según el género en cada año. Fuente la Autora, 2016.

En la Figura 1 y la Tabla 2 se puede ver que en los accidentes de tránsito, el 62,46 % de las víctimas de los accidentes corresponde al género masculino y al género femenino un 37,54 %. Se observa en los dos años de estudio, que los hombres son quienes se ven más implicados en accidentes, aumentando la cifra de homicidios y lesiones de una año al otro.

AÑO	CONDUCTA						Total 2014-2015
	2014			2015			
GÉNERO	Homicidio	Lesionado	Subtotal	Homicidio	Lesionado	Subtotal	
Masculino	14	51	65	18	95	113	178
	14,14 %	51,52 %	65,66 %	9,68 %	51,08 %	60,75 %	62,46 %
Femenino	6	28	34	5	68	73	107
	6,06 %	28,28 %	34,34 %	2,69 %	36,56 %	39,25 %	37,54 %
Subtotal	20	79	99	23	163	186	285
% por año	20,20 %	79,80 %	100 %	12,37 %	87,63 %	100 %	100 %

Tabla 2: Conducta clasificada por género. Fuente la Autora, 2016

Del año 2014 a 2015, el porcentaje de hombres muertos disminuyó y de lesionados se mantuvo estable. Para las mujeres, el porcentaje de homicidios también disminuyó pero el porcentaje de lesionadas aumentó en un 8 % aproximadamente.

Por otro lado, el promedio de edad de las víctimas que se ven implicadas en un accidente es de 28 años, observado que la tendencia de victimas está más presente en edades jóvenes. El 50 % de las personas está por debajo de 25 años.

En cuanto a las víctimas según su agrupación por edad, se observa (Tabla 3) que los adultos son quienes más se ven involucrados en un accidente, con una participación en el 2014 de 90,91 % y en el 2015, reportando un porcentaje de 90,86 %. En los dos años de estudio, no se presentaron muertes por accidente en adolescentes.

AÑO	AGRUPACIÓN EDAD								Total 2014-2015
	2014				2015				
CONDUCTA	Menor	Adolescente	Adulto	Subtotal	Menor	Adolescente	Adulto	Subtotal	
Homicidio	1	0	19	20	2	0	21	23	43
	1,01 %	0 %	19,19 %	20,20 %	1,08 %	0,00 %	11,29 %	12,37 %	15,09 %
Lesionado	5	3	71	79	8	7	148	163	242
	5,05 %	3,03 %	71,72 %	79,80 %	4,30 %	3,76 %	79,57 %	87,63 %	84,91 %
Subtotal	6	3	90	99	10	7	169	186	285
% por año	6,06 %	3,03 %	90,91 %	100 %	5,38 %	3,76 %	90,86 %	100 %	100 %

Tabla 3: Conducta clasificada por gupo de edad. Fuente la Autora, 2016

Respecto a la ubicación geográfica del accidente, en los años de estudio de accidentalidad, ocurren más accidentes en la zona urbana que en la zona rural. Para el año 2014, el 24,24 % de los accidentes ocurrieron en zona rural frente a un 75.76 % de los accidentes ocurridos en la zona urbana. En el año 2015, en la zona urbana ocurrieron el 65.59 % de los accidentes registrados para este año, notándose así, el aumento porcentual de víctimas en la zona urbana de un año a otro.

AÑO	ZONA						Total 2014-2015
	2014			2015			
CONDUCTA	Rural	Urbana	Subtotal	Rural	Urbana	Subtotal	
Homicidio	10	10	20	14	9	23	43
	10,10 %	10,10 %	20,20 %	7,53 %	4,84 %	12,37 %	15,09 %
Lesionado	14	65	79	50	113	163	242
	14,14 %	65,66 %	79,80 %	26,88 %	60,75 %	87,63 %	84,91 %
Subtotal	24	75	99	64	122	186	285
% por año	24,24 %	75,76 %	100 %	34,41 %	65,59 %	100 %	100 %

Tabla 4: Conducta clasificada por zona. Fuente la Autora, 2016

En la base de datos, también se hace una descripción de la ocupación de la víctima en el momento del accidente. En general se observó que la mayoría de víctimas tenían una ocupación laboral, representadas por el 73,33 %. Llama la atención, que entre las ocupación de las víctimas, las amas de casa se hayan visto implicadas en accidentes un 11,39 %.

Finalmente, otro aspecto importante a describir acerca de las víctimas del accidente, es determinar por cuál medio o móvil se desplazaban éstas en el momento del accidente (Figura 2) se observa que el mayor porcentaje de víctimas de accidentes de tránsito se movilizaban a pie, representando el 40 %.

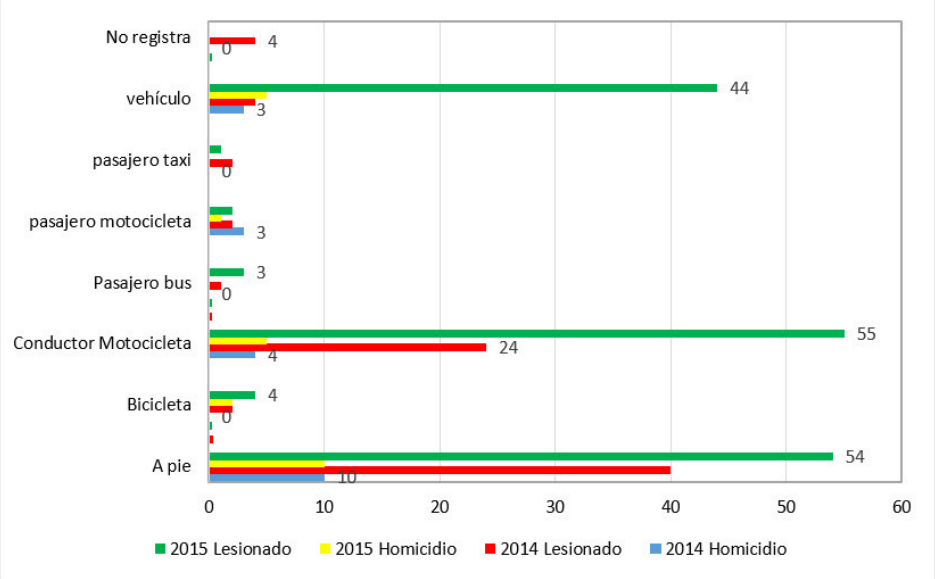


Figura 2: Conducta clasificada por móvil víctima

En el 2014, el 50,51 % de las victimas registradas se movilizaban a pie, del cual el 10,10 % murieron en los hechos y el 40,40 % sufrieron lesiones. Para el año 2015, está condición de movilidad de la víctima representó solo el 34,41 % del total de implicados, de los cuales, solo el 5,38 % culminaron en homicidio. La siguiente forma de movilización común entre las víctimas fue para conductores de motocicleta, con un 30,88 % (88 personas) del total.

Accidentes Como se mencionó, la base de datos sobre accidentalidad cuenta con 285 filas, cada una de las cuales describe a cada víctima del accidente, sin embargo, hay víctimas que se dieron en un mismo accidente, por esta razón fue necesario depurar una parte de la base de datos para solo obtener los registros sobre accidentes.

Para la descripción de los 243 accidentes que se presentaron entre 2014 y 2015, se tienen en cuenta las siguientes variables: móvil en que se trasportaba el agresor, zona del accidente y sector (dirección) en que se dio este, hipótesis del hecho y tipo de accidente, día, mes y hora en que se presentaron los hechos. En seguida se muestran los resultados obtenidos de este análisis.
El 36 % de los registros de accidentes son del 2014, mientras que el 64 % de estos ocurrieron en el año 2015.

Para los accidentes ocurridos en la zona urbana de la ciudad de Tunja, se tiene un registro de la dirección donde se dieron los accidentes. Para tabular e interpretar esta variable, fue necesario agrupar las direcciones según la sectorización de la ciudad descrita por la alcaldía. El mapa tenido en cuenta para esto se observa en la Figura 3.

Teniendo en cuenta el sector donde ocurrieron los accidentes, se puede ver que de los 88 accidentes dados en el 2014, el 22,73 % de estos ocurrieron en la zona rural de la ciudad mientras que el 77,27 % de los accidentes se presentaron en la zona urbana. Para el año 2015, 27,74 % de los accidentes fueron en la zona rural y el resto, un 72,26 % de los accidentes en la zona urbana de la ciudad de Tunja.

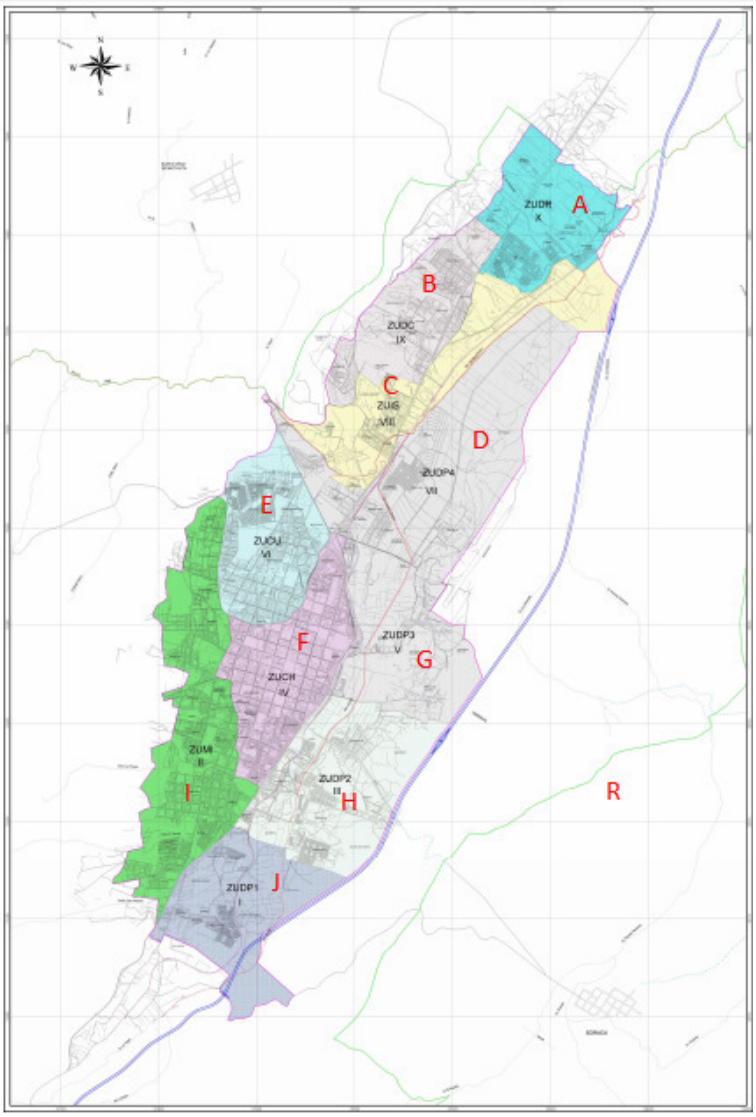


Figura 3: División Política por sectores de la ciudad de Tunja. Tomado de página web alcaldía de Tunja. Fuente la Autora, 2016.

En la Tabla 5 se resume la cantidad de accidentes que ocurrieron en cada sector de la zona urbana de la ciudad estos dos años de estudio. En general, el lugar de la ciudad donde más ocurrieron accidentes es en F, que hace referencia a la parte del centro de la ciudad. En la salida norte de la ciudad, hay una baja incidencia de accidentalidad, registrándose solo 2 accidentes entre estos años.

Para la variable móvil agresor se puede observar, que del total de accidentes registrados, el 53,03 % de los accidentes no registran el móvil en que se encontraba el victimario. Es más común que en un accidente de tránsito el agresor se movilice en un vehículo particular (25,93 %) seguido de los conductores de motocicleta (10,70 %).

Observando esta variable por años, en el 2014 los accidentes donde el victimario se encuentra en vehículo representan el 15,91 %. Para el año 2015, es donde se presenta un gran porcentaje de datos faltantes sobre esta variable, por lo cual es difícil interpretar cual es el móvil agresor más común de accidentes de este año.

SECTOR	2014		2015		Total 2014-2015	%
	No accidentes	%	No accidentes	%		
A	0	0	2	1,29	2	0,82
B	2	2,27	13	8,39	15	6,17
C	4	4,55	12	7,74	16	6,58
D	4	4,55	11	7,10	15	6,17
E	4	4,55	9	5,81	13	5,35
F	12	13,64	29	18,71	41	16,87
G	3	3,41	4	2,58	7	2,88
H	8	9,09	12	7,74	20	8,23
I	9	10,23	13	8,39	22	9,05
J	2	2,27	3	1,94	5	2,06
R	17	19,32	43	27,74	60	24,69
NO REGISTRA	23	26,14	4	2,58	27	11,11
TOTAL	88	100	155	100	243	100

Tabla 5: Accidentes clasificados por año y sector. Fuente la Autora, 2016

Por otro lado, es importante hablar acerca de la modalidad del accidente, información que se encuentra plasmada en la Tabla 6. El tipo de accidente que más se presentó estos años fue accidente de tránsito donde la víctima era el peatón, representando un 38,68 % del total de accidentes. El otro tipo de accidente que siguió en frecuencia a este, fue el accidente de tránsito donde la móvil de la víctima es una moto, con un 33,74 % del total de accidentes.

Observando cada año y cada zona de accidentalidad, se puede ver que en el 2014 en la zona rural, el accidente de tránsito donde la víctima se trasportaba a pie fue el de mayor ocurrencia, con un 9,09 % y respecto a la zona urbana, también ésta fue la modalidad de accidente con mayor ocurrencia, con un 34,09 % del total de accidentes ocurridos en este año. Para el año 2015, referente a la zona rural, el tipo de accidente que más se presentó fue en el que la víctima se movilizaba en vehículo, representando un 10,97 % de los accidentes, y respecto a la zona urbana, el accidente que más se presentó fue accidente de tránsito cuando la víctima estaba caminando. Solo el 2,46 % del total de accidentes fue por caída o volcamiento.

Luego de conocer la cantidad de accidentes y el tipo de accidente que se presentaron estos años, es necesario estudiar las posibles causas de estos. Para hablar de la hipótesis sobre la causa de ocurrencia de los accidentes, nuevamente hay que aclarar que hay valores faltantes para esta variable.

Estos datos NO REGISTRADOS representan el 13,99 % de los accidentes. En general, se observa que las causas más frecuentes de accidentes son impericia en el manejo (23,87 %), el cruzar sin observar (15.64 %) y desobedecer las señales de tránsito (11,52 %).

AÑO	ZONA						Total 2014-2015
	2014			2015			
MODALIDAD	Rural	Urbana	Subtotal	Rural	Urbana	Subtotal	
ACCIDENTE DE TRÁNSITO BICICLETA	0	4	4	3	3	6	10
	0,00 %	4,55 %	4,55 %	1,94 %	1,94 %	3,87 %	4,12 %
ACCIDENTE DE TRÁNSITO MOTO	7	22	29	10	43	53	82
	7,95 %	25,00 %	32,95 %	6,45 %	27,74 %	34,19 %	33,74 %
ACCIDENTE DE TRÁNSITO PEATÓN	8	30	38	11	45	56	94
	9,09 %	34,09 %	43,18 %	7,10 %	29,03 %	36,13 %	38,68 %
ACCIDENTE DE TRÁNSITO VEHÍCULO	4	11	15	17	19	36	51
	4,55 %	12,50 %	17,05 %	10,97 %	12,26 %	23,23 %	20,99 %
CAÍDA	0	1	1	1	1	2	3
	0,00 %	1,14 %	1,14 %	0,65 %	0,65 %	1,29 %	1,23 %
VOLCAMIENTO	1	0	1	1	1	2	3
	1,14 %	0,00 %	1,14 %	0,65 %	0,65 %	1,29 %	1,23 %
Subtotal	20	68	88	43	112	155	243
% por año	22,73 %	77,27 %	100,00 %	27,74 %	72,26 %	100,00 %	100,00 %

Tabla 6: Accidentes clasificados por modalidad y zona . Fuente la Autora, 2016

Particularmente en el año 2014, las dos causas más comunes de accidentalidad fueron cruzar sin observar y desobedecer señales de tránsito, mientras que para el año 2015, la más frecuente fue impericia en el manejo seguida de no mantener distancia de seguridad.

Finalmente, se hace una caracterización del tiempo en que ocurrieron los accidentes. El día en que más se presentaron accidentes fue el sábado con un 27,57 % del total de accidentes. El día con menos registro de accidentes fue el martes, con 14,81 %. Se presencian más accidentes de tránsito en la zona rural los fines de semana, es decir, sábado y domingo. Otra información sobre el día en que ocurrieron en los accidentes estos años, se presentan en la Tabla 7.

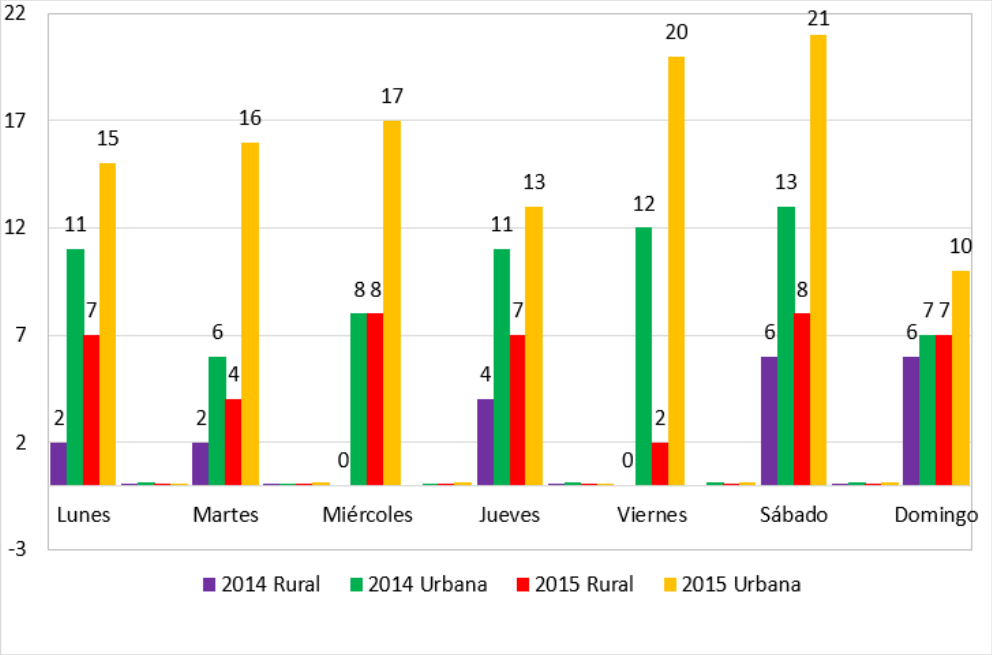


FIGURA 4: Día de la semana en que ocurren los accidentes. Fuente la Autora, 2016.

AÑO	ZONA						Total 2014-2015
	2014			2015			
DÍA	Rural	Urbana	Subtotal	Rural	Urbana	Subtotal	
Lunes	2	11	13	7	15	22	48
	2,27 %	12,50 %	14,77 %	4,52 %	9,68 %	14,19 %	19,75 %
Martes	2	6	8	4	16	20	36
	2,27 %	6,82 %	9,09 %	2,58 %	10,32 %	12,90 %	14,81 %
Miércoles	0	8	8	8	17	25	41
	0,00 %	9,09 %	9,09 %	5,16 %	10,97 %	16,13 %	16,87 %
Jueves	4	11	15	7	13	20	50
	4,55 %	12,50 %	17,05 %	4,52 %	8,39 %	12,90 %	20,58 %
Viernes	0	12	12	2	20	22	46
	0,00 %	13,64 %	13,64 %	1,29 %	12,90 %	14,19 %	18,93 %
Sábado	6	13	19	8	21	29	67
	6,82 %	14,77 %	21,59 %	5,16 %	13,55 %	18,71 %	27,57 %
Domingo	6	7	13	7	10	17	43
	6,82 %	7,95 %	14,77 %	4,52 %	6,45 %	10,97 %	17,70 %
Subtotal	20	68	88	43	112	155	243
% por año	22,73 %	77,27 %	100,00 %	27,74 %	72,26 %	100,00 %	100,00 %

TABLA 7: Accidentes clasificados por zona y día. Fuente la Autora, 2016

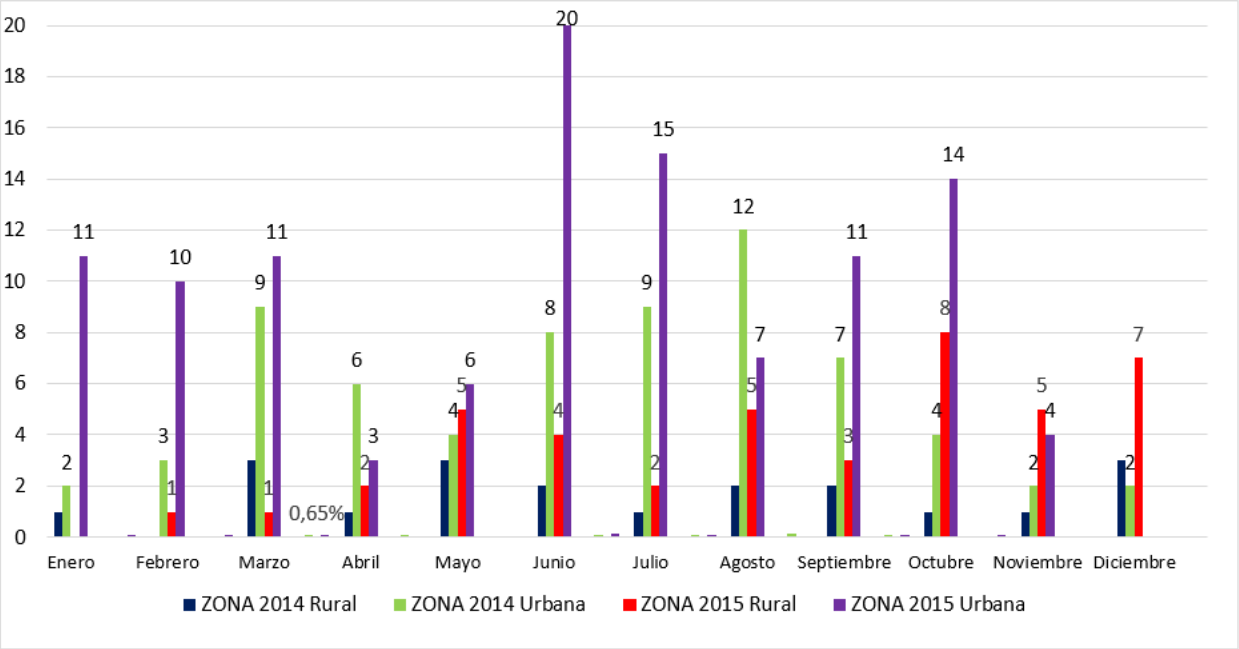


Figura 5: Mes en que ocurren los accidentes. Fuente la Autora, 2016.

También se identifica Junio el mes en el que más ocurrieron accidentes en estos dos años, con un porcentaje de 13,99 % que representan 34 accidentes de un total de 243. El mes con menos ocurrencia de estos fue abril, noviembre y diciembre, con un 4,94 % cada uno. En el año 2014, en particular, agosto fue en mes con más accidentes en Tunja, y en 2015, Junio vuelve a ser el mes con mayor número de accidentes.

AÑO	ZONA						Total 2014-2015
	2014			2015			
HORA	Rural	Urbana	Subtotal	Rural	Urbana	Subtotal	
Madrugada	4	10	14	3	6	9	23
	4,55 %	11,36 %	15,91 %	1,94 %	3,87 %	5,81 %	9,47 %
Mañana	4	19	23	14	39	53	76
	4,55 %	21,59 %	26,14 %	9,03 %	25,16 %	34,19 %	31,28 %
Tarde	7	15	14	13	17	30	44
	7,95 %	17,05 %	15,91 %	8,39 %	10,97 %	19,35 %	18,11 %
Noche	5	24	29	13	50	63	92
	5,68 %	27,27 %	32,95 %	8,39 %	32,26 %	40,65 %	37,86 %
Subtotal	20	68	88	43	112	155	243
% por año	22,73 %	77,27 %	100 %	27,74 %	72,26 %	100 %	100 %

Tabla 8: Accidentes clasificados por zona y hora. Fuente la Autora, 2016

Finalmente, la hora de mayor ocurrencia de accidentes en general en estos dos años fue en horas de la noche (18:00 a 23:59) con un porcentaje de 37,84 % del total de accidentes. En seguida se registran el 31,28 % de los accidentes en horas de la mañana (6:00 a 12:00). La hora de menor ocurrencia de accidentes en la ciudad es a la madrugada (12:00 a 6:00) con un 9,47 % del total de los accidentes. En el año 2014, la mayor

cantidad de accidentes se dieron en la zona urbana en horas de la noche con un 27,27 % de los 88 accidentes. Para el año 2015, de igual forma, se presentaron más accidentes en la zona urbana en horas de la noche con un 32,26 %.

Cruce de variables

La Tabla 9 muestra en resumen la aplicación de la prueba ji-cuadrado para determinar dependencia o no entre algunas de las variables de estudio. También, la determinación de coeficientes de asociación para las variables que nuestran dependencia.

CRUCE VARIABLES		CHI-CUADRADO	P-VALOR	COEFICIENTES	
				CRAMER	CONTINGENCIA
Conducta	Género	2,5189	0,1125	0,104	0,104
	Edad	1,97	0,3734	0,083	0,083
	Zona	13,41	0,0002502	0,228	0,228
	móvil víctima	12,367	0,08912	0,208	0,204
	móvil agresor	11,833	0,06581	0,204	0,2
	hipótesis	16,346	0,09016	0,239	0,233
	Modalidad	11,158	0,04834	0,198	0,194
	Día	9,1453	0,1656	0,179	0,176
	Mes	15,611	0,1562	0,234	0,228
	hora	2,5156	0,1127	0,104	0,104

Tabla 9: Estadísticas asociadas a las pruebas de independencia. Fuente la Autora, 2016

Según los resultados, se puedo observar una asociación entre la variable conducta y zona de ocurrencia del accidente, entre género y el móvil de la víctima, así como entre género y modalidad del accidente. También se muestra asociación entre móvil del agresor y la hipótesis del accidente. Modalidad e hipótesis también se encuentran altamente relacionadas. Se confirman los resultados anteriores con los valores de los coeficientes de Cramer y Contingencia.

CRUCE VARIABLES		CHI-CUADRADO	P-VALOR	COEFICIENTES	
				CRAMER	CONTINGENCIA
Género	edad	0,56408	0,7542	0,044	0,044
	móvil víctima	30,692	0,00007086	0,312	0,312
	hipótesis	7,282	0,8986	0,16	0,158
	zona	0,4518	0,5015	0,048	0,048
	modalidad	28,206	0,00003318	0,315	0,3
Móvil agresor	modalidad	41,183	0,08391	0,184	0,381
	sector	68,056	0,4071	0,216	0,468
	hipótesis	115,11	0,00002462	0,281	0,567
Modalidad	sector	50,683	0,6402	0,204	0,415
	hipótesis	132,63	2,068E-09	0,33	0,594
	hora	7,0884	0,9552	0,099	0,169
	día	7,0884	0,9552	0,2	0,408

Tabla 10: Continuación estadísticas asociadas a las pruebas de independencia. Fuente la Autora, 2016

5. Conclusiones

La caracterización de la accidentalidad en la ciudad de Tunja en los años 2014 y 2015, mediante la depuración y análisis de la base de datos facilitada por la Policía Nacional, permitió identificar con claridad algunas condiciones de las personas que se vieron implicadas en los diferentes accidentes, establecer aspectos demográficos y característicos de todos los accidentes ocurridos en este lapso y por medio del cruce de variables, determinar relaciones entre las diferentes variables de estudio. Desafortunadamente, no se obtuvieron datos solo el victimario o quien genera el accidentes, considerándose esta información relevante para hacer una caracterización más completa de la accidentalidad en la ciudad.

En el proceso de depuración de la base de datos se detectaron datos faltantes en variables como hipótesis y modalidad del accidente, datos que son de gran importancia a la hora de caracterizar la accidentalidad, sin embargo, a pesar de los datos faltantes se pudo llegar a conclusiones sobre estas variables.

En los años 2014 y 2015 se tuvieron en la ciudad 285 víctimas, 43 murieron y el resto sufrieron lesiones a causa del accidente. El promedio de la edad de las víctimas de los accidentes es de 28 años, siendo los adultos, con un 90 %, quienes más se ven involucrados en estos. En su mayoría, las víctimas son de género masculino y por lo general, ellas se encontraban en la zona urbana en el momento del incidente. La mayoría de las víctimas de accidentes tenían en el momento del suceso una estabilidad laboral. Las víctimas de los accidentes con más porcentaje del total se movilizaban a pie, seguidas de las que se movilizaban en moto. También se encontraron entre las víctimas, pasajeros de servicio público.

Referente a la caracterización de los accidentes, en estos dos años de estudio se reportaron 243 accidentes, 88 en 2014 y 155 en 2015. La mayor parte de los accidentes se ubicaron geográficamente en la zona urbana de la ciudad, siendo el sector centro donde más se presentaron accidentes y la salida norte de la ciudad donde menos registros se tuvieron. Por otro lado, en el momento del accidente, fue común observar que el agresor o victimario conducía un vehículo particular, y con menos frecuencia, el agresor se movilizaba en motocicleta. La posible causa del accidente que más se repite es impericia en el manejo. Finalmente, el día que más se registran accidentes es el sábado y el mes en junio, siendo la hora más común de ocurrencia en horas de la noche.

Se puede encontrar una dependencia entre las variables conducta-zona del accidentes, género y móvil de la víctima, género y modalidad del accidente, que está ligado con el móvil en que se desplaza la víctima, móvil agresor e hipótesis y modalidad del accidente e hipótesis.

Se recomienda a la Policía Nacional, institución encargada del registro de cada uno de los accidentes con las diferentes variables, que dentro de esta recolección se tenga en cuenta aspectos directamente relacionados con la persona que genera el accidente, pues esta información aporta aún más a la caracterización que se pretende hacer. En los vitimarios se pueden medir variables como: edad, género, experiencia de conducción, profesión, actividad laboral, entre otras.

Referencias Bibliográficas

Álvarez, A. (2009), 'Accidente de tránsito. Blog de enseñanza Medicina Forense', *Blog de enseñanza Medicina Forense* .
*http://alvarezunahvs.files.wordpress.com/2009/11/accidente-de-transito.pdf

Cabrera, G., Velásquez, N. y Valladares, M. (2009), 'Seguridad vial, un desafío de salud pública en la Colombia del siglo XXI', *Revista Fac. Nac. Salud Pública. Bogotá* **27**(2).

Congreso, Colombia (2002), *LEY 769. Código Nacional de Tránsito*, Colombia.

Díaz M., L. G. y Morales R., M. A. (2009), *Análisis estadístico de datos categóricos*, primera edn, Universidad Nacional de Colombia.

Martínez, L. (2008), 'Muertes y lesiones por accidente de tránsito en colombia: 2007', *Rev. Forensis* .

Norza, E., Granados, E., Useche, S., Romero, M. y Moreno, J. (2014), 'Componentes descriptivos y explicativos de la accidentalidad vial en Colombia: incidencia del factor humano', *Revista Criminalidad* .

Organización Mundial de la Salud (2004), 'The global burden of disease: a comprehensive assessment of mortality and disability from disease, injuries and risk factors in 1990 and projected to 2020', *Ginebra* .

Perdomo, M. (2000), 'Death in Traffic Accidents, Colombia, year 2000', *Instituto Nacional de Medicina Legal y Ciencias Forenses - INMLCF* .

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org/

CARACTERIZACIÓN DE LOS AUTOMOTORES QUE INGRESAN AL CENTRO DE DIAGNÓSTICO AUTOMOTRIZ SOGAMOSO LTDA. CEDAS

Especialización en Estadística

Alberto Zea Higuera[1,a], Dairo Sigifredo Gil Gil[1,b]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

[2]Escuela de Matemáticas y Estadística, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

En este artículo se presenta la descripción univariada y multivariada de los datos recolectados mediante una base datos que corresponde a todos los automotores que ingresan al Centro de Diagnóstico Automotriz Sogamoso Ltda -CEDAS con el fin de adquirir el certificado de revisión técnico-mecánica y de emisiones contaminantes. Esta información se indagó teniendo en cuenta el análisis descriptivo de los datos y la teoría de colas con el fin de llegar a mejorar el servicio y los tiempos de intervención de los automotores, brindando un consolidado con la información más relevante para dicha empresa.

***Palabras clave*:** caracterización, descripción de datos,tablas de contingencia.

Abstract

This article univariate and multivariate description of the data collected is presented through a database corresponding to all motor vehicles entering the Sogamoso Automotive Diagnostic Center Ltda -CEDAS in order to acquire the certificate of technical and mechanical overhaul and emissions contaminants. This information is investigated taking into account the descriptive data analysis and queuing theory in order to reach and improve service intervention times of automotive, providing a consolidated with the most relevant information for that company.

***Key words*:** characterization, description of data, contingency tables.

1. Introducción

El Centro de Diagnóstico Automotriz Sogamoso Ltda -CEDAS, es el responsable de la atención de más del 60 % de los automotores que necesitan de la revisión técnico-mecánica en la ciudad de Sogamoso. Bajo esta premisa, es necesario encaminar todos los esfuerzos para la obtención de estándares de calidad muy altos, mejorando la disponibilidad y confiabilidad de esta empresa, y así seguir obteniendo la acreditación del Organismo Nacional de Acreditación de Colombia - ONAC.

En el CEDAS cada moto que llega se le realiza una toma de fotos, luego la revisión visual de frenos, de luces, de gases, y sonometría. Para los carros, ya sean tipo liviano o pesado se les realiza una inspección visual, toma de fotos, diagnóstico de pista, un test line donde se revisa la suspensión, el alineador, los frenos, diagnóstico de luces y análisis de gases.

[a]Estudiante de Especialización en Estadística. E-mail: alzeahig10@gmail.com
[b]Profesor asistente. E-mail: dsigifre@gmail.com

Como el parque vehicular de la ciudad de Sogamoso ha venido creciendo, la empresa se preocupa por los tiempos que tienen que esperar los clientes que llevan su automotor al CEDAS, por lo tanto, el sistema objeto de este estudio es verificar si los horarios de los funcionarios han sido establecidos correctamente de acuerdo a la clasificación interna de la compañía. Actualmente se cuenta con una base de datos, de la cual se tiene acceso a todos ellos, excepto los que son confidenciales para la empresa.

A partir de febrero del año en curso la empresa de diagnóstico automotriz CEDAS ve la necesidad de tener un consolidado de las características de las variables que conforman una base de datos que se lleva dentro de la empresa a partir del 01 de agosto del 2013 hasta el 01 enero 2016.

Es decir, que desea conocer las características de los automotores que ingresan al sistema para identificar duración del servicio promedio para los tipos de carro que ingresan a la empresa, línea, modelo, servicio, cilindraje, marca, clase y tipo de falla, determinando además, los vehículos que no cumplen con las especificaciones, los parámetros establecidos en su primera revisión y la razón por la cual la reprobaron. Además identificar cuáles meses del año desde la fecha de inicio son pico y valle, las horas pico y valle y los días de mayor y menor cobertura.

En consecuencia, se plantea el siguiente interrogante:

¿Cuáles son los tiempos recomendables de asignación del servicio para realizar la revisión técnico-mecánica y emisión de gases y las características de los diferentes automotores que lo requieren?

Este artículo está inspirado en los planteamientos de (Lizarazo Mansilla, Galindo Romero et al. 2013), en cuanto al crecimiento automotriz propone realizar los servicios de revisión vehicular por medio de unidades móviles.

González Restrepo y Sepúlveda Abalo (2010), realizaron una simulación de tiempos estudiando cómo evitar que en las horas pico se formen las largas colas a la espera del cambio de un semáforo ubicado en una de las calles principales de Pereira.

Galván Zacarías, Melo Álvares y Alcantara de Vasconcellos (2014), (2014) desarrollaron una investigación de los centros de servicio automotriz de América Latina destacando los conceptos de la inspección ambiental y las principales tendencias de la evolución futura de la inspección técnico vehicular.

Diaz H (2005) Díaz. H, (2005), utiliza un modelo de reingeniería de procesos para el centro de auto-lavado en Santa Lucia, con estudio de colas. En este documento se observa los procedimientos estadísticos para mejorar eficiencia y fortalecer sus ventajas competitivas, con estudio de colas propone el modelo y reduce el tiempo en el sistema en un 50 % e incrementa en un 30 % el nivel de eficiencia.

por otra parte se busca que CEDAS se beneficie con este trabajo, con un conocimiento más amplio de los automotores que acceden al centro de diagnostico automotriz, ademas realizar una interpretación de la congestión que se presenta en las estaciones internas de la empresa en determinadas horas del día, esto con el fin de orientar el objetivo de la organización hacia un mejor servicio a la comunidad.

2. Referente Conceptual

Las técnicas de análisis exploratorio de datos permiten analizar la información exhaustivamente y detectar las posibles anomalías que puedan presentar las observaciones. J. W Tuckey fue uno de los primeros en la introducción de este tipo de análisis.

Los estadísticos descriptivos más habitualmente utilizados han sido la media y la desviación típica. Sin embargo, el uso automático de estos índices no es muy aconsejable.

La media y la desviación típica son índices convenientes sólo cuando la distribución de datos es aproximadamente normal o, al menos simétrica y unimodal. Pero las variables objeto de estudio no siempre cumplen estos requisitos. Por lo tanto, es necesario un examen a fondo de la estructura de los datos, se debe tener en cuenta que tan solo validamos para este procesamiento el supuesto de normalidad que son: diferencia de medias y la prueba de homogeneidad.

Para datos cuantitativos es aconsejable comenzar con el gráfico de tallo y hojas o histograma digital. El paso siguiente suele ser examinar la presencia de valores atípicos (outliers) en el conjunto de datos; en cuanto al tratamiento que se le dio a la recolección de los datos (metodología del censo) para este caso no se le da

un tratamiento especial a datos sospechosos como outliers ya que se le dio fiabilidad a la información que brindo el afectado en el momento del registro (Gil, 2012).

3. Metodología

No obstante que, como se describió en párrafos anteriores, CEDAS, dispone de una base de datos altamente confiable, la información anterior a 2014 adolece de algunas inconsistencias, que pueden alterar el análisis del servicio que presta la entidad. No se incluye 2016, sólo está actualizada hasta finales de enero de dicho año. La base está compuesta por 34.063 registros de los cuales al descartar los sospechosos de inconsistencias se reduce a 26.012 registros lo que constituye una gran muestra que corresponde al 76.4 % de registros de la base lo que garantiza una altísima cobertura.

En CEDAS, a medida que va llegando el cliente con su automotor se registran las características en una base de datos en Excel, altamente confiable.

Las variables más relevantes que se tomaron para el estudio son:

- **OT_FECHAREG:** es la fecha y hora a la que llega el cliente a adquirir la orden de trabajo.
- **OT_FECHATER:** es la fecha y hora de salida del automotor.
- **OT_NREV:** es el número de revisiones a las que se ha sometido el vehículo este valor es el número de devoluciones que ha tenido el vehículo.
- **OT_TIPOREVISION:** es una característica donde se especifica qué tipo de vehículo es decir si es liviano pesado o corresponde a una moto.
- **OT_RESULTADO:** esta variable indica si el automotor fue rechazado o aprobado para poderle entregar el certificado de la revisión técnico-mecánica y emisión de gases.
- **VH_LINEA:** indica la línea del vehículo de acuerdo a su marca.
- **VH_SERVICIO:** esta variable da información si se trata de un vehículo particular, de servicio público u oficial.
- **VH_TIPO:** corresponde al tipo de automotor si es pesado liviano o moto.
- **VH_MODELO:** corresponde al año de fabricación del automotor pero esta se transformó a número de meses desde la fabricación hasta la revisión.
- **VH_CLASE:** corresponde a la clase de vehículo. (Automóvil, motocicleta, microbús, etc.)
- **VH_CILINDRAJE:** esta variable muestra el cilindraje del vehículo medido en centímetros cúbicos.
- **VH_COMBUSTIBLE:** muestra que tipo de combustible utiliza el automotor.
- **VH_TIPOMOTOR:** la información de esta variable sirve para conocer de cuantos tiempos es la moto y para carro no aplica.

Con base en estos datos se recogieron las variables más relevantes para el estudio y poderle brindar posibles soluciones al problema de la empresa. Con esta información se definieron las variables de interés según el resultado de las pruebas de los automotores, y se hizo elaboración de la base de datos luego se depuro usando el paquete R para eliminar inconsistencias y por ultimo se proceso la información más relevante y consolidada por medio de tablas de contingencia, diagramas de caja y gráficos de barras.

4. Resultados

Para esta sección del análisis, se resaltarán los resultados más relevantes y los consolidados que más aporten al estudio, teniendo en cuenta aspectos como: total de certificados vendidos por año, tipo de vehículos que son rechazados en su primer prueba, pruebas de frenos, tipos de combustible su antigüedad entre otros.

Se presentarán tablas de consolidados, contingencia y diagramas de barras que faciliten la interpretación de los datos y la situación de cada caso.

4.1. Descripción de la demanda

CEDAS hace revisión técnico-mecánica (T-M) y expide el certificado correspondiente, básicamente a tres tipos de vehículos (Ver tabla 1).

	Livianos	Motos	Pesado	Total
N^o de Automotores	8291	10236	7485	26012
%	31.9	39.4	28.8	100

Tabla 1: Fuente: Zea Alberto, con autorización de CEDAS (2016)

Obsérvese que de los vehículos a los que presta este servicio la entidad, sobresalen las motos, correspondiendo a un 39,4 % del total. Además, los vehículos livianos y los pesados, conforman dos grupos equi-repartidos de aproximadamente 8.000 vehículos cada uno (ver tabla 2).

Tipo de vehículo y valor del certificado				
	Livianos	Motos	Pesado	Total
N^o CERTIFICADOS	6801	8649	6204	216554
%	31.4	39.9	28.7	100
Costo(en $)de la revisión T-M	136680	84785	221574	$\cdots$

Tabla 2: Precios de enero 16 de 2016 por tipo de vehículo
Fuente: Zea Alberto, con autorización de CEDAS (2016)

4.2. Tipos de combustible

Los automotores utilizan tres tipos de combustible encargado de propulsar motores; diésel correspondiente al 29.9 % de los automotores que han ingresado durante los últimos dos años, motores de gasolina que son la mayoría con un 68.7 % y, algunos motores de gasolina les adecuan gas que son el 1.3 % esta adecuación es para ahorrar combustible y dinero. En la gráfica se visualiza que los vehículos más antiguos son los que mas han sido transformados para que el motor funcione con gas y con gasolina, puesto que estos aun siendo de tipo liviano tenían un cilindraje alto y consumían demasiado combustible. Observando también que hay varios vehículos de motores diésel que se consideran antiguos.

	N^o de Automotores	%
Gasolina	17883	68.7
Gasolina-gas	341	1.3
Diésel	7748	29.9
Total	26008	100.0

Tabla 3: Fuente: Zea Alberto, con autorización de CEDAS (2016)

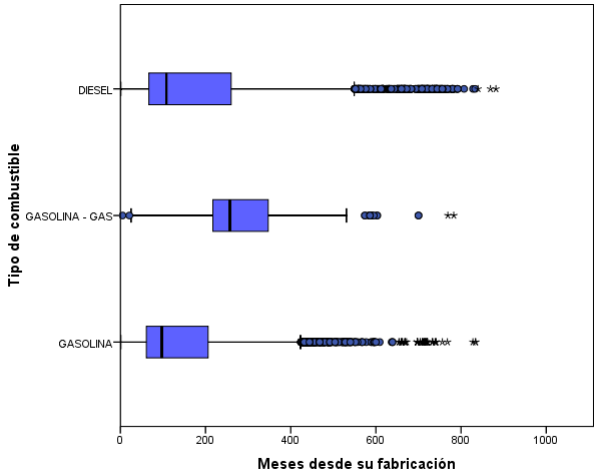


Figura 1: Fuente: Zea Alberto, con autorización de CEDAS (2016)

A los automotores se les realiza un análisis de emisión de gases, cuando el motor es a gasolina o gas-gasolina y opacidad cuando es diésel, y se hace tanto en ralentí como en crucero. Esto se realiza con equipos calibrados y aprobados por la CORPOBOYACÁ.

	N^o de Automotores	%
Rechazado	2299	12.6
Aprobado	15910	87.4
Total	18209	100.0

Tabla 4: Fuente: Zea Alberto, con autorización de CEDAS (2016)

	N^o de Automotores	%
Rechazado	79	1.1
Aprobado	6856	98.9
Total	6935	100.0

Tabla 5: Fuente: Zea Alberto, con autorización de CEDAS (2016)

Los resultados de las pruebas de gases y opacidad, se pueden observar en las tablas que nos muestran que el mayor porcentaje de rechazos lo presentan los automotores que trabajan con gasolina, y que para los automotores que trabajan con diésel es mínimo el porcentaje de rechazo. Cosa que favorece el medio ambiente, puesto que los vehículos de motor diésel consumen menos combustible que los de gasolina, pero causan cuatro veces más contaminación atmosférica que el resto, pues emiten niveles muy superiores de dióxido de nitrógeno (NO2) y partículas en suspensión, dos de los principales contaminantes del aire.

4.3. Resultado prueba de frenos de servicio

	N^o de Automotores	%
Rechazado	1396	5.4
Aprobado	24611	94.6
Total	26007	100.0

Tabla 6: Fuente: Zea Alberto, con autorización de CEDAS (2016)

Una de las pruebas más importantes por la cual son muy exigentes con los resultados es la de frenado, puesto que la falta de frenos es una de las mayores causas de alto riesgo de accidentalidad, y en CEDAS

LTDA es la comprobación más importante que se realiza en la revisión técnico mecánica, consiste en una prueba en la cual se miden las fuerzas de frenado de la motocicleta o vehículo. Sumando la fuerza de los ejes de la moto o vehículo respecto al peso total se determina la eficacia que tiene la moto o vehículo para frenar.

De los 26007 de los automotores que se sometieron a la prueba de frenos el 5.4 % han sido rechazados.

		AÑO DE REVISIÓN					
		2014			2015		
		Resultado pruebas de frenos de servicio		Total	Resultado pruebas de frenos de servicio		Total
		Rechazado	Aprobado		Rechazado	Aprobado	
Particular	N^o Automotores	321	10258	10579	346	8332	8678
	%	3.0	97.0	100.0	4.0	96.0	100.0
Público	N^o Automotores	280	2913	3193	435	2982	3417
	%	8.8	91.2	100.0	12.7	87.3	100.0
Oficial	N^o Automotores	7	67	74	7	46	53
	%	9.5	90.5	100.0	13.2	86.8	100.0
Especial	N^o Automotores	0	1	1	0	8	8
	%	0.0	100.0	100.0	0.0	100.0	100.0
Total	N^o Automotores	608	13847	13847	788	11368	12156
	%	4.4	100.0	100.0	6.5	93.5	100.0

Tabla 7: Fuente: Zea Alberto, con autorización de CEDAS (2016)

Cabe resaltar que los automotores que más han sido rechazados son los vehículos que prestan el servicio público, puesto que la prueba queda registrada con la fuerza de frenado de cada una de las llantas lo cual hace que repruebe en la estación de frenado.

4.3. Resultado de las horas pico durante los años 2014 y 2015

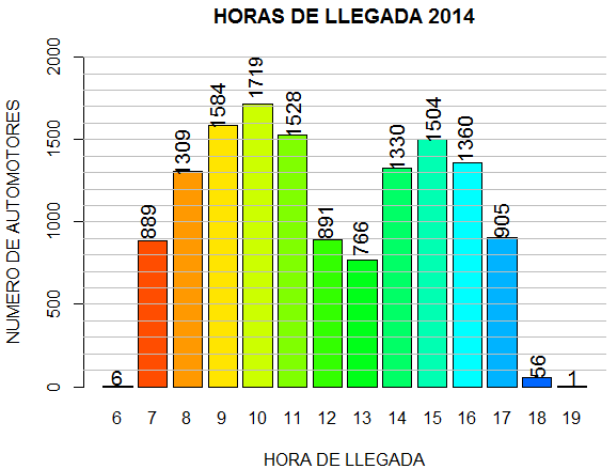


Figura 2: Fuente: Zea Alberto, con autorización de CEDAS (2016)

Se debe nombrar que los vehículos que son devueltos al reprobar en la estación de freno en su mayoría son los vehículos de tipo pesado. En el 2014 fueron devueltos el 59 %, y para el 2015 el 66.6 %, cifras demasiado altas puesto que es una de las estaciones más importantes que requieren de su aprobación inmediata para evitar posibles accidentes.

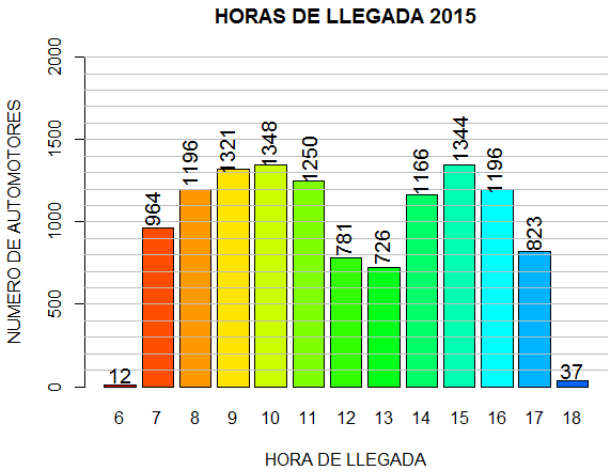


FIGURA 3: Fuente: Zea Alberto, con autorización de CEDAS (2016)

En cedas la asignación de turnos a sus técnicos, es decir, a los que hacen las pruebas es de 2 tipos. Uno es de 7-11 am y de 1-5 pm y el otro es de 8 am a 1 pm y de 3 -6 pm. En la gráfica se puede observar que para los dos años, las horas de mayor afluencia son de 9-11 am y de 3-4 pm, observando que en estas horas están todos los técnicos, es de esperar que en estos horarios no se dé embotellamientos.

5. Conclusiones

- Debido a que el porcentaje de automotores que son rechazados es alto, y tienen que volver una segunda y tercera vez, estos probablemente vuelvan en horarios pico, haciendo que se genere los cuellos de botella o embotellamientos, por lo tanto se le recomienda a la empresa, agendar los vehículos que han sido devueltos para algún horario valle.

- Se le recomienda a la empresa hacer un posterior estudio de reingeniería puesto que se vieron varias falencias que este estudio no puede solucionar.

- Para evitar posibles embotellamientos, se le recomienda a la empresa realizar un estudio en la ciudad de Sogamoso donde se aprecie la idea de trabajar con el centro de diagnóstico automotriz móvil, es decir, que se trabaje a domicilio, por lo menos para las motos.

- Queda la satisfacción de brindar a entes importantes de la ciudad de Sogamoso, una herramienta tan importante con la que a la fecha no se contaban, que es el consolidado con información de mucha relevancia con la cual estas dependencias pueden acompañar su puesta en acción.

Referencias Bibliográficas

Ausín, C. (2003), 'Análisis bayesiano de sistemas de colas', *Universidad Carlos III de Madrid* .

Diaz H, C. (2005), 'Modelo de reingeniería de procesos para el centro de', *SANTA LUCÍA, AUTO LAVADO and OAXACA, OAX* .

Díaz, L. G. (2009), 'Análisis de datos categóricos', *Universidad Nacional de Colombia Sede Medellín* .

Galván Zacarías, A., Melo Álvares, O. y Alcantara de Vasconcellos, E. (2014), 'Inspección técnica vehicular en américa latina'.

González Restrepo, M. y Sepúlveda Abalo, E. J. (2010), 'Aplicación de teoria de colas en los semaforos para mejorar la movilidad en la carrera 7 entre calles 15 y 20 de la ciudad de pereira'.

Lizarazo Mansilla, K. E., Galindo Romero, A. A. et al. (2013), 'Estudio de factibilidad para diseñar y construir unidades móviles para el abastecimiento de los centros de diagnóstico automotor cda.'.

Taha, H. A. (2004), *Investigación de operaciones*, Pearson Educación, México.

Visauta Vinacua, B. (1997), 'Análisis estadístico con spss para windows', *Editorial McGraw-Hill* .