ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMATICAS
DEPARTAMENTO DE MATEMATICAS

PROYECTO DE GRADUACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN SEGUROS Y RIESGOS FINANCIEROS

TEMA

MEDICIÓN DEL RIESGO ASOCIADO A LAS CARACTERÍSTICAS DE LOS VEHÍCULOS MEDIANTE LA APLICACIÓN DE TÉCNICAS DE ANÁLISIS MULTIVARIADO PARA EL ESTUDIO DE LA SINIESTRALIDAD EN EL RAMO DE SEGUROS DE VEHÍCULOS EN LA CIUDAD DE GUAYAQUIL.

AUTOR:
ING. GUILLERMO FERNANDO ORDÓÑEZ LOOR

Guayaquil-Ecuador
2016

DEDICATORIA

A mi Dios, porque todo lo que soy y tengo es gracias a sus bendiciones, su amor se transformó en lo que conozco y llamo Papá y Mamá, quienes hicieron un buen trabajo conmigo, no se dar un paso sin su voluntad.

Guillermo O.

AGRADECIMIENTO

A mi Dios, porque he sentido su mano desde pequeño, me ha guiado a través de mis padres y hermanos, sus bendiciones y misericordia son infinitas. A mi Padre Bolívar y a mi madre Clarita, porque simplemente hay cosas que no se pueden explicar, solo se lo que mi corazón siente y el ejemplo de lucha y esfuerzo constante lo aprendí de mis papas, a mi madre su amor, paciencia y control me ayudaron a sobrellevar mis problemas. A mis hermanos, que pasamos por muchas cosas tristes, pero aprendimos que juntos somos fuertes, a mi hermano mayor, sufrimos su perdida, pero lo llevo en mi corazón, él era un cuidador junto a mis padres. A mi tutora la Dra. Sandra García por brindarme su guía y valioso conocimiento en el desarrollo de este proyecto, a mis compañeros de trabajo por su apoyo y a mis buenos amigos que también me han dado ánimos a seguir en la lucha.

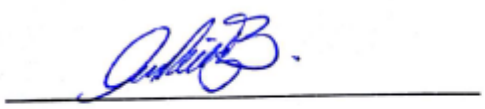
Guillermo O.

DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Graduación, me corresponde exclusivamente; el patrimonio intelectual del mismo, corresponde exclusivamente a la **Facultad de Ciencias Naturales y Matemáticas, Departamento de Matemáticas** de la Escuela Superior Politécnica del Litoral.

Ing. Guillermo Ordóñez Loor

TRIBUNAL DE GRADUACIÓN



Omar Ruiz Barzola, Ph.D.

PRESIDENTE DEL TIBUNAL

Sandra García Bustos, Ph.D.

DIRECTOR DE PROYECTO

M.Sc. Wehrli Pérez Caicer

VOCAL DEL TRIBUNAL

FIRMA DEL AUTOR DEL PROYECTO DE GRADUACIÓN

Ing. Guillermo Ordóñez Loor

TABLA DE CONTENIDO

CONTENIDO DE FIGURAS

CONTENIDO DE TABLAS

CAPÍTULO I

1. OBJETIVO Y GENERALIDADES

1.1. INTRODUCCIÓN

La aparición de los vehículos en la evolución del mundo ha sido un logro extraordinario para el progreso del hombre y acorde al avance tecnológico se ha vuelto indispensable para el desarrollo de las actividades cotidianas de la humanidad, pero su invento trae como consecuencia la aparición del riesgo producto de factores como los accidentes y robos, no puede desconocerse el problema severo que ellos ocasionan a la salud pública, debido a esto aparecen y se crean los seguros de vehículos en el mundo los que brindan cobertura a los siniestros que ocurren producto de los accidentes.

El exceso de velocidad, el alcoholismo, falta de uso de cinturón de seguridad están entre las principales causas que ocasionan daños y el fallecimiento por accidentes de tránsito de cerca de 1.25 millones de personas en el mundo, el 90% de las muertes ocurren en países de mediano y bajos ingresos [1]. Según el Banco Interamericano de Desarrollo (BID), en América Latina y El Caribe ocurren 17 muertes por cada 100.000 personas y este organismo estima que para el 2020 esta región podría tener la tasa más alta del mundo con 17 fallecidos por cada 100000 habitantes [2].

El Ecuador no es ajeno a estas cifras, según la Agencia Nacional de Tránsito de enero a diciembre del 2015 ocurrieron 35706 siniestros a nivel nacional, entre las causas más probables se tienen el no respetar las señales reglamentarias de tránsito con 13,71%, conducir de manera desatenta a las condiciones de tránsito con el 12,62%, rebasar los límites de velocidad con el 10,58%. En cuanto a las características, se tiene que los principales tipos de vehículos involucrados en siniestros de tránsito son automóviles con el 42%, motos con un 18% y camionetas con el 13% [3].

Se tiene además al mes de diciembre 2015 que los tipos de siniestros que tienen mayor porcentaje de incidencia en nuestro país, son los choques laterales con el 28,37%, atropellos con el 14,40% y estrellamientos con un 12,95%, los tres eventos cubren más del 50% de todos los tipos de daños reportados y que afectan tanto a los que conducen y sus acompañantes como a los peatones y ciudadanía en general, lo que preocupa y debe motivar a las autoridades gubernamentales a tomar medidas para mitigar el riesgo asociado a vehículos [3].

Las cifras presentadas evidencian una relación asociada a las características de los vehículos, además una falta de control y de precaución al momento de conducir, así como del desconocimiento de las señales de tránsito de la mayoría de los ecuatorianos y aunque existen cursos de conducción el número de accidentes de tránsito aún son grandes junto a la gravedad de los daños que estos causan, como se observa en la figura 1.

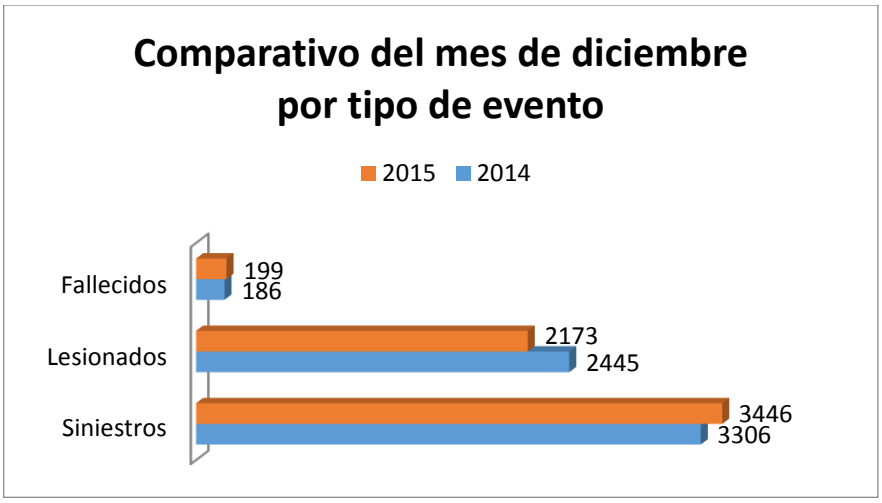


Figura 1: Tipos de Eventos

Fuente: Agencia Nacional de Tránsito (ANT) – Autor: Guillermo Ordóñez L.

Las cifras de los tipos de eventos ocurridos al mes de diciembre evidencian una reducción para el año 2015 respecto del año 2014. Para lesionados cuyo porcentaje de variación es del -11% [4]. Pero, existe un aumento para siniestros del 4% en variación [3]. Y por último, para

fallecidos 7% [5]. Esta alza en los índices de siniestralidad terminando el año 2015 son preocupantes ya que se mantienen altos.

1.2. PLANTEAMIENTO DEL PROBLEMA

Las estadísticas de accidentes de tránsito en el Ecuador son preocupantes por la cantidad de muertes y daños materiales que estos provocan, junto a la severidad con la que ocurren los siniestros, que además de causar las pérdidas de vidas humanas, también son notables las pérdidas económicas que se producen y en muchos casos incluso es irrecuperable el bien (vehículo) por el estado irreparable en el que queda. Las compañías de seguros han tomado parte de este problema, creando coberturas de seguros para vehículos ante la existencia de la necesidad de los usuarios de contar con un respaldo económico ante cualquier eventualidad que se pueda presentar, pero a medida que crece el número de personas que buscan asegurar sus vehículos, crece el riesgo de sufrir siniestros y aumenta la probabilidad del pago de liquidaciones por parte de las compañías de seguros, esto es un problema para las aseguradoras ya que no cuentan con un estudio técnico que les permita validar las características de los vehículos, muchas veces basan sus decisiones en reportes de los organismos oficiales y en estadísticas de la misma compañía sobre este tipo de seguro, pero no pueden decidir de manera óptima y más aún estimar si los vehículos asegurados son propensos a sufrir algún accidente.

Es necesario conocer si las características de los vehículos están asociadas directamente en los accidentes de tránsito, no existen análisis científicos de sus características como una posible y muy influyente causa en la ocurrencia de los siniestros en nuestro país.

Esta problemática se puede afrontar desde diferentes perspectivas, dependiendo de las características que se analicen acerca de los vehículos, es primordial contar con variables que se incluyen en

cada póliza, tales como marca, categoría, color, además variables como el registro de siniestros, tipo de vehículo asegurado, entre otras que se pueden obtener a través de una muestra de vehículos asegurados de una compañía aseguradora en la ciudad de Guayaquil, ya que por ser la ciudad más grande del Ecuador, cuenta con un importante número de vehículos a nivel nacional. También es de interés estimar la probabilidad de ocurrencia de siniestros de vehículos en la ciudad, además cuantificar la cantidad de siniestros acontecidos y la forma en que se relacionan.

El presente trabajo desarrolla un modelo para estimar la propensión al riesgo de los vehículos asegurados según sus características, a través de técnicas estadísticas multivariadas, tales como el análisis de regresión logística y el análisis de correspondencias múltiples, etc., que permitirán a las compañías de seguros estimar el riesgo asociado a las características de los vehículos y de esta manera clasificarlo o discriminarlo como proclive a sufrir o no un siniestro y decidir si se otorga cobertura. Se pueden citar algunas investigaciones realizadas en éste campo a nivel internacional, como los de: Chen H, Libo Caoa L, Logan D. (2012) [6]; Hemrit W, Ben Arab M, Raissi N. (2013) [7], que servirán de referencia para realizar el presente proyecto que será pionero en el país.

1.3. JUSTIFICACIÓN DEL PROBLEMA

El ramo de Seguros de Vehículos es el de mayor demanda en nuestro país, con una participación que supera el 25% del total de primas emitidas en todos los ramos de seguros entre los años 2011 y 2013, esto es debido a la frecuencia y severidad con la que ocurren accidentes de tránsito y más aún en las grandes ciudades [8]. Según La Agencia Nacional de Tránsito, las cifras en el mes de diciembre del 2015 fueron de 3446 siniestros, en la tabla 1, se presentan las cifras de accidentes desagregado por provincias [3].

Siniestros por Provincia a nivel Nacional (Trimestres 2015)						
Provincias	**Ene-Mar**	**Abr-Jun**	**Jul-Sep**	**Oct-Dic**	**Total 2015**	**%**
Azuay	312	325	356	380	1373	3,85%
Bolívar	45	45	43	50	183	0,51%
Cañar	69	71	93	75	308	0,86%
Carchi	46	46	45	36	173	0,48%
Chimborazo	135	156	142	177	610	1,71%
Cotopaxi	124	162	122	103	511	1,43%
El Oro	251	223	204	241	919	2,57%
Esmeraldas	81	89	115	136	421	1,18%
Galápagos	8	7	3	5	23	0,06%
Guayas	**1603**	**1679**	**1766**	**1751**	**6799**	**19,04%**
Imbabura	325	396	342	463	1526	4,27%
Loja	152	166	171	199	688	1,93%
Los Ríos	314	295	325	316	1250	3,50%
Manabí	348	260	290	319	1217	3,41%
Morona Santiago	36	41	36	43	156	0,44%
Napo	40	35	34	44	153	0,43%
Orellana	55	47	15	27	144	0,40%
Pastaza	38	36	25	20	119	0,33%
Pichincha	3675	4082	3813	4184	15754	44,12%
Santa Elena	118	113	85	95	411	1,15%
Santo Domingo De Los Tsáchilas	239	268	235	257	999	2,80%
Sucumbíos	35	36	23	35	129	0,36%
Tungurahua	396	411	423	505	1735	4,86%
Zamora Chinchipe	27	31	22	25	105	0,29%
Total	**8472**	**9020**	**8728**	**9486**	**35706**	**100%**
%	23,73%	25,26%	24,44%	26,57%	100,00%	

Tabla 1: Siniestros por Provincias

Fuente: Agencia Nacional de Tránsito (ANT) – Autor: Guillermo Ordóñez L.

Existe una notoria diferencia de enero a diciembre del 2015 en las provincias de Pichincha con el 44,15% y Guayas con 19,19%, ambas contabilizan más del 60% de los siniestros ocurridos en el país [3]. La tabla 2 muestra los siniestros por cantones de la provincia del Guayas.

Cantón	# Siniestros	Participación
Balao	4	0,66%
Balzar	5	0,82%
Colimes	2	0,33%
Daule	26	4,28%
Durán	56	9,23%
El Empalme	9	1,48%
El Triunfo	10	1,65%
Guayaquil	**377**	**62,11%**
Juján-A Baquerizo Moreno	2	0,33%
Lomas de Sargentillo	5	0,82%
Marcelino Maridueña	2	0,33%
Milagro	32	5,27%
Naranjal	22	3,62%
Naranjito	4	0,66%
Nobol	3	0,49%
Palestina	3	0,49%
Pedro Carbo	3	0,49%
Playas	2	0,33%
Salitre	2	0,33%
Samborondón	20	3,29%
Santa Lucía	2	0,33%
Simón Bolívar	2	0,33%
Yaguachi	14	2,31%
Total	**607**	**100,00%**

Tabla 2: Siniestros por cantones - Prov. Guayas – Diciembre 2015

Fuente: Agencia Nacional de Tránsito (ANT) – Autor: Guillermo Ordóñez L.

Las cifras muestran que el cantón Guayaquil tiene el 62,11% del total de accidentes de tránsito ocurridos en la provincia del Guayas, lo que la convierte en la de mayor índice de siniestralidad y ocupa los primeros puestos en el país.

Dadas las estadísticas de siniestros y la alta cobertura en el ramo de vehículos, se selecciona la ciudad de Guayaquil para el desarrollo del presente proyecto, ya que cuenta con un importante parque automotor y

además por ser la ciudad económicamente más importante y de mayor población en el país.

1.4. OBJETIVO GENERAL

Determinar un modelo para medir el riesgo asociado a las características de los vehículos en la ciudad de Guayaquil, mediante técnicas estadísticas multivariadas.

1.5. OBJETIVOS ESPECÍFICOS

1. Realizar un análisis descriptivo de la situación actual de siniestros en el ramo de seguros de vehículos en la ciudad de Guayaquil.
2. Identificar posibles relaciones entre diversas variables con el evento de ocurrencia de un siniestro.
3. Utilizar técnicas estadísticas multivariadas para medir el riesgo en la ocurrencia de un siniestro en el ramo de seguros de vehículos.
4. Comparar y analizar los modelos desarrollados.

El presente proyecto de investigación se compone de cuatro capítulos.

El Capítulo 1 consiste en objetivos y generalidades, como son la introducción, planteamiento del problema, justificación del problema, objetivo general y objetivos específicos de la investigación.

En el Capítulo 2, se elabora el marco teórico, como son las definiciones básicas de Estadística Descriptiva, Tablas de Contingencia, y de las técnicas de análisis multivariado como Regresión Logística y Análisis de Correspondencias Múltiples, sobre los que se basa esta investigación.

El Capítulo 3, consiste en el desarrollo y el análisis descriptivo de las variables de estudio, análisis de Tablas de Contingencia y el análisis e interpretación de los resultados obtenidos.

El Capítulo 4, se desarrollará el análisis multivariado de los modelos propuestos: Análisis de Correspondencias Múltiples y el Análisis de Regresión Logística, además la interpretación de los resultados obtenidos.

Y una parte final, que está formado por las Conclusiones y Recomendaciones de la investigación.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. INTRODUCCIÓN

Este capítulo abarcará la teoría que será usada como base para el desarrollo de los capítulos 3 y 4, estará compuesta en tres partes, una primera parte que comprenderá la estadística descriptiva, una segunda parte compuesta por la teoría sobre las tablas de contingencia y una tercera parte comprenderá la teoría sobre los modelos multivariados seleccionados para esta investigación, como son el análisis de correspondencias múltiples y el análisis de regresión logística.

La primera parte comprenderá las definiciones de la estadística descriptiva como son: media, mediana, moda, varianza, desviación estándar, coeficiente de variación, coeficiente de asimetría y de curtosis, percentiles, cuartiles, diagrama de caja, distribución de frecuencias, histograma de frecuencias, graficas de barras y de pastel.

La segunda parte estará comprendida por la teoría acerca de las tablas de contingencia, como son su definición, como se construye, su estructura gráfica, y prueba de hipótesis respectiva.

Por último, la tercera parte contiene el desarrollo de la teoría de los dos modelos multivariados que son la base del presente estudio, como son el análisis de correspondencias y el análisis de regresión logística.

2.2. ESTADÍSTICA DESCRIPTIVA

2.2.1. MEDIA MUESTRAL

Dado un conjunto de observaciones $X_1, X_2, X_3, ..., X_n$, una muestra de tamaño n tomada de una población de interés. La media aritmética de la

muestra denotada como $\overline{x}$ se define matemáticamente como: (Walpole et al., 2012) [9].

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

2.2.2. MEDIANA MUESTRAL

Dado un conjunto de observaciones $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ de una muestra ordenada de manera creciente, tomada de una población. La mediana de la muestra denotada como $\tilde{x}$ se define matemáticamente como: (Walpole et al., 2012) [9].

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , \text{si n es impar} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right], & \text{si n es par} \end{cases}$$

2.2.3. MODA MUESTRAL

Dado un conjunto de observaciones $x_1, x_2, x_3, ..., x_n$ de una muestra tomada de una población. La moda de la muestra se define como el dato x_i que se repite con mayor frecuencia. Si dentro del conjunto, existen dos valores que se repiten con mayor y la misma frecuencia, entonces se dice que dicho conjunto es bimodal. Si aparecen más de dos valores con la misma y mayor frecuencia, entonces se dice que dicho conjunto de datos es multimodal. En caso de no existir ningún valor que se repita dentro del conjunto de datos, se dice que no existe moda. (Triola, 2009) [10].

2.2.4. VARIANZA MUESTRAL

La varianza de una muestra de n observaciones $x_1, x_2, x_3, ..., x_n$, se define como la suma de los cuadrados de las desviaciones entre la media aritmética de la muestra $\bar{x}$ y los valores observados, dividida para $n-1$. La varianza de una muestra se denota como s^2 y se define matemáticamente como: (Mendenhall et al., 2010) [11].

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

2.2.5. DESVIACIÓN ESTÁNDAR MUESTRAL

La desviación estándar de una muestra de n observaciones $x_1, x_2, x_3, ..., x_n$, es igual a la raíz cuadrada positiva de la varianza de la muestra s^2. Se denota como S y se define matemáticamente como: (Mendenhall et al., 2010) [11].

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

2.2.6. COEFICIENTE DE VARIACIÓN MUESTRAL

El coeficiente de variación de n observaciones $x_1, x_2, x_3, ..., x_n$, de una muestra, describe la relación entre la desviación estándar y la media de la muestra. Se denota como CV y se expresa en porcentajes, matemáticamente se define como: (Triola, 2009) [10].

$$CV = \left(\frac{s}{\bar{x}} \right).100\%$$

2.2.7. COEFICIENTE DE ASIMETRÍA MUESTRAL DE FISHER

El coeficiente de asimetría de Fisher para un conjunto de n observaciones $x_1, x_2, x_3, ..., x_n$, pertenecientes a una muestra obtenida de una población, mide la proximidad de las observaciones a su media aritmética $\overline{x}$. Se denota como s_k y se define matemáticamente como:

$$s_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)S^3}$$

Donde S es la desviación estándar de la muestra.

Los valores que toma este coeficiente tienen las siguientes características:

Si $s_k < 0$, la distribución tiene una asimetría negativa, es decir la cola de la distribución se alarga para valores inferiores a la media de la muestra.

Si $s_k = 0$, la distribución es simétrica.

Si $s_k > 0$, la distribución tiene una asimetría positiva, es decir la cola de la distribución se alarga para valores mayores a la media de la muestra (Véliz, 2011) [12].

2.2.8. COEFICIENTE DE CURTOSIS MUESTRAL DE FISHER

El coeficiente de curtosis de Fisher para un conjunto de n observaciones $x_1, x_2, x_3, ..., x_n$, pertenecientes a una muestra obtenida de una población, es una medida del apuntamiento o achatamiento de la curva o distribución de la muestra, este coeficiente indica la cantidad de observaciones cercanas a la media de la muestra, es decir a mayor grado de curtosis más escarpada o apuntada será la curva. El coeficiente de curtosis se denota como k y se define matemáticamente como:

$$k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)S^4} - 3$$

Donde S es la desviación estándar de la muestra.

Los valores de estos coeficientes tienen las siguientes características:

Si k = 0, la distribución es Mesocúrtica, un ejemplo claro son las distribuciones normales.

Si k > 0, la distribución es Leptocúrtica, es decir la distribución muestral es más apuntada que las Mesocúrticas.

Si k < 0, la distribución es Platicúrtica, es decir la distribución muestral es más achatada que las Mesocúrticas (Véliz, 2011) [12].

2.2.9. PERCENTILES

Sea el conjunto $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ una muestra de n mediciones obtenida de una población con característica X, ordenadas de menor a mayor. Los percentiles dividen a la muestra en 100 partes iguales. El k-ésimo percentil de una muestra se denota como P_k donde k=1, 2, ⋯, 99, es un valor tal que al menos k% de las mediciones se encuentran en o por debajo de este valor y al menos el (100-k)% de ellas se encuentran en o por encima de P_k. (Anderson et al., 2008) [13].

Como calcular un percentil k:

Primero. Se deben ordenar los datos de menor a mayor.

Segundo. Se calcula el índice i, que es la posición que ocupa una observación dentro del conjunto ordenado. El índice se calcula como $i = \frac{kn}{100}$ donde k es el percentil deseado y n es el número de mediciones.

Por último, si:

(a) Sí $i \notin \mathbb{Z}$, se redondea este valor hacia arriba. El primer entero mayor que i denota la posición del percentil k.

(b) Si $i \in \mathbb{Z}$, el percentil k es el promedio de los valores en las posiciones i e i + 1. Es decir, $P_k = \frac{1}{2}\left[x_{(i)} + x_{(i+1)}\right]$

Por ejemplo, P_{50} es el percentil cincuenta, coincide además con el segundo cuartil que es la mediana de la muestra, el 50% de las mediciones tienen valores que son menores o iguales que P_{50} y el 50% restante, corresponde a las mediciones cuyos valores son mayores o iguales que P_{50}.

2.2.10. CUARTILES

Sea el conjunto $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$ una muestra de n mediciones obtenida de una población con característica X , ordenadas de menor a mayor. Los cuartiles se denotan como Q_1, siendo Q_1, Q_2 y Q_3, los mismos que dividen a la muestra en 4 partes iguales, cada parte contiene un 25% de las mediciones.

El primer cuartil, Q1, es un valor que es mayor o igual al 25% de las mediciones y menor al 75% de las mediciones restantes. El segundo cuartil, Q2, es un valor que es mayor o igual al 50% de las mediciones y menor al 50% de las mediciones restantes, cabe resaltar que Q2 coincide con la mediana de la muestra. El tercer cuartil, Q3, es un valor que es mayor o igual al 75% de las mediciones y menor al 25% de las mediciones restantes.

Calculo de los Cuartiles

Primero. Ordene las n mediciones de menor a mayor.

Segundo. Calcule $n\left(\frac{k}{4}\right)$, donde k=1,2,3. Si el valor resultante no es un entero, se debe redondear al entero siguiente y determine el valor ordenado que corresponde. Si el valor es un entero q, se calcula la media de la q-ésima y la (q+1)-ésimas mediciones ordenadas (Johnson, 2012) [14].

2.2.11. DIAGRAMA DE CAJA

Un diagrama de caja es una representación gráfica de las n observaciones de una muestra, cuya forma de caja se compone por el rango intercuartil (RIC), cuyos extremos son los cuartiles Q1 y Q3, y además contiene a la mediana de los datos. Los valores alejados en la muestra se representan mediante los denominados “bigotes” que se prolongan desde la caja. Si la muestra es grande, el diagrama presenta otras características importantes como el centro de localización, la variabilidad y el grado de asimetría.

Pasos para elaborar un diagrama de caja:

1. Se construye una caja acotada por dos extremos localizados sobre el primer y tercer cuartiles. La caja contiene el 50% de los datos centrales.
2. Se traza una línea vertical sobre el valor donde se localiza la mediana.
3. Se calcula el rango intercuartil (RIC), RIC = Q3 – Q1, con el cual se localizan los límites de un intervalo:

Límite inferior: Q1 – 1.5 RIC

Límite superior: Q3 + 1.5 RIC

Estos límites se ubican a 1.5 veces el rango intercuartil (RIC) por debajo del Q1 y 1.5 veces el RIC por encima del Q3, las mediciones que se encuentren fuera de estos límites son considerados como datos atípicos. También, pueden existir mediciones muy distantes del resto de los datos, menores que Q1 – 3 RIC o que sean mayores que Q3 + 3 RIC, a las cuales se las considera observaciones atípicas extremas.

4. Se grafican los bigotes a través de dos líneas rectas que se prolongan desde los extremos de la caja hasta los valores límites obtenidos en el paso anterior.
5. De existir, se encierran con un círculo las mediciones atípicas y mediante un asterisco a las mediciones atípicas extremas (Anderson et al., 2008) [13].

2.2.12. DISTRIBUCIÓN DE FRECUENCIAS

La distribución de frecuencias o también llamada tabla de frecuencias, permiten visualizar un conjunto de datos de manera resumida, lo que ayuda a la comprensión de los mismos, ya que describen la manera en que se distribuyen los valores del conjunto organizados en categorías en el caso que la naturaleza de los datos sea cualitativa o mediante clases o intervalos para el caso cuantitativo, es decir describe la variabilidad de los datos. Cada clase de la tabla contiene un número determinado de datos denominados frecuencias y la participación de cada una de ellas en el total de datos. (Véliz, 2011) [12]

Para el caso de las variables cualitativas, los conteos se realizan por cada una de las categorías definidas en cada variable. Por ejemplo, para la variable color del vehículo, donde se definen cuatro categorías: blanco, rojo, azul y negro, se cuenta el número de veces o la frecuencia de cada color o categoría en la distribución de los datos. En cambio, para variables cuantitativas, se cuenta las veces que ocurren o frecuencias de los valores distribuidos en cada clase o intervalo. (Alvarado y Obagi, 2008) [15]

Para un conjunto de n observaciones, las cuales se quieren presentar de manera resumida mediante una tabla de frecuencias, es importante tener en cuenta las siguientes definiciones para su construcción.

2.2.12.1. INTERVALOS DE CLASE

Son rangos de datos que dividen a las mediciones de la variable de estudio, acotados por un límite inferior y un límite superior, y se denotan por $[b_{i-1}, b_i)$.

2.2.12.2. MARCA DE CLASE

Es el punto medio del intervalo de clase y se obtiene promediando los límites superior e inferior. Se denota y calcula como: $m_i = \frac{b_{i-1}+b_i}{2}$.

2.2.12.3. FRECUENCIA ABSOLUTA

Es el número de datos o mediciones que contiene el intervalo de clase i. Se denota como f_i.

2.2.12.4. FRECUENCIA RELATIVA

Es el cociente entre el número de datos o mediciones que contiene el intervalo de clase i y el total de mediciones n . Se calcula como $\frac{f_i}{n}$.

2.2.12.5. FRECUENCIA ABSOLUTA ACUMULADA

Es el número de mediciones acumuladas hasta el intervalo i. Se denota y se calcula como:

$$\boldsymbol{F_i = \sum_{j=1}^{i} f_i}$$

2.2.12.6. FRECUENCIA RELATIVA ACUMULADA

Es la proporción de mediciones acumuladas hasta el intervalo i. Se denota y se calcula como:

$$\boldsymbol{\frac{F_i}{n} = \sum_{j=1}^{i} \frac{f_j}{n}}$$

2.2.12.7. PASOS PARA CONSTRUIR UNA TABLA DE FRECUENCIAS

Se siguen los siguientes pasos: (Rodríguez, 2007) [16].

1. Calcule el rango R de los datos, donde:

$$R = valor\ máximo - valor\ mínimo$$

2. Determinar el número de clases o intervalos i, para agrupar los datos. El número de clases debe estar entre 5 ≤ i ≤ 20. Existen varias reglas o criterios para calcular el número de clases o intervalos i, una de ellas es mediante: $2^i \geq n$, donde i es el número de clases y n es el total de observaciones. El número de clases a utilizar es el menor valor de i, tal que la desigualdad anterior se satisfaga. Otros criterios que se han definido para calcular el número de clases necesarios con las que se puede trabajar son: la

Regla Empírica tradicional, cuya expresión matemática es igual a la raíz cuadrada del número de observaciones, es decir, $\sqrt{n}$; y la Regla de Sturges, cuya fórmula para el cálculo del número de clases es $1 + log_2(n)$, donde n es el número de observaciones. Para estas dos reglas, si el número de clases obtenidas no es entero, se redondea al entero más próximo, este número redondeado sería la cantidad de clases a utilizar en la tabla de frecuencias (Alvarado y Obagi, 2008) [15].

3. Calcular la amplitud de las clases mediante:

$$Amplitud = \frac{R}{i} = \frac{Rango}{Número\ de\ clases}$$

4. Determinar los límites o extremos de cada clase.
5. Determinar la marca de cada clase.
6. Realizar el conteo de los datos para obtener la frecuencia absoluta correspondiente a cada clase.
7. Con la frecuencia absoluta, realizar el cálculo de la frecuencia relativa y la frecuencia relativa acumulada.

Se presenta a continuación la estructura de una tabla de frecuencias:

Clase $\boldsymbol{i}$	**Intervalo de clase**	$\boldsymbol{m_i}$	$\boldsymbol{f_i}$	$\boldsymbol{\frac{f_i}{n}}$	$\boldsymbol{F_i}$	$\boldsymbol{\frac{F_i}{n}}$
1	$[b_1, b_2)$	m_1	f_1	$\frac{f_1}{n}$	f_1	$\frac{f_1}{n}$
2	$[b_2, b_3)$	m_2	f_2	$\frac{f_2}{n}$	$f_1 + f_2$	$\frac{f_1 + f_2}{n}$
3	$[b_3, b_4)$	m_3	f_3	$\frac{f_3}{n}$	$f_1 + f_2 + f_3$	$\frac{f_1 + f_2 + f_3}{n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	$[b_i, b_{i+1})$	m_i	f_i	$\frac{f_i}{n}$	$f_1 + f_2 + f_3 + \cdots + f_i$	$\frac{f_1 + f_2 + f_3 + \cdots + f_i}{n}$

Tabla 3: Tabla de frecuencias

Fuente: Alvarado, J. A. y Obagi, J. J. (2008). Fundamentos de Inferencia Estadística (p. 20), Primera edición. Bogotá, Colombia: Pontificia Universidad Javeriana – Autor: Guillermo Ordóñez.

2.2.13. HISTOGRAMA DE FRECUENCIAS

El histograma de frecuencias es la representación gráfica mediante barras rectangulares verticales juntas de un conjunto de datos cuantitativos. Se construye a partir de los datos que han sido resumidos mediante una tabla de distribución de frecuencias. En el eje horizontal se colocan las escalas de la variable y en el eje vertical las escalas de las frecuencias absolutas o relativas. Se grafican las barras rectangulares, uno a continuación del otro, donde la base de cada rectángulo es proporcional a la amplitud del intervalo, y la altura es la respectiva frecuencia absoluta o relativa. Los histogramas son de mucha utilidad, debido a que proporcionan información relacionada con la simetría y dispersión de las observaciones (Anderson et al., 2008) [13].

2.2.14. GRÁFICAS DE BARRA

Una gráfica de barras o diagrama de barras, es una forma de representar un conjunto de datos asociados a las categorías de una variable cualitativa. Este tipo de gráfico se construye por lo general, ubicando sobre el eje horizontal, las etiquetas que representan a las categorías de la variable de estudio y en el eje vertical, se ponen los valores correspondientes a las frecuencias de cada categoría. Se grafican barras o rectángulos, las bases de los rectángulos, cuyo ancho es arbitrario, se localizan sobre las etiquetas (categorías) de la variable, y las alturas son proporcionales a cada etiqueta, es decir el valor de frecuencia de cada categoría. Las barras se disponen de manera separada para reflejar a cada categoría de la variable (Anderson et al., 2008) [13].

2.2.15. GRÁFICAS DE PASTEL

Una gráfica de pastel o gráfica circular, es un recurso estadístico que se utiliza para representar de manera gráfica las frecuencias asociadas a las categorías de una variable cualitativa. Una gráfica de pastel se elabora en primer lugar, graficando un círculo que identifique a todos los datos.

A continuación, se divide al círculo en partes, cada parte corresponde a la frecuencia relativa de cada categoría de la variable de estudio. El área de cada una de las partes es proporcional al número de datos en esa categoría (Anderson et al., 2008) [13].

2.3. TABLAS DE CONTINGENCIA

Son tablas de doble entrada, que sirven para analizar la existencia de una relación de dependencia o independencia entre dos variables cualitativas o factores de estudio, es decir un análisis bivariable. En las tablas de contingencia, una de las variables se usa para representar las filas y la otra para representar las columnas (Otero y Medina, 2016) [17].

Sean X e Y, las variables de las filas y de las columnas en ese orden, las cuales representan a dos variables cualitativas con r y c categorías respectivamente, entonces a un individuo cualquiera, se lo puede clasificar en una de las posibles r × c categorías existentes.

Los datos se distribuyen en las casillas de la tabla de contingencia, de acuerdo a los valores que toman según la fila y columna que le corresponden.

Dado que los datos se organizan en una tabla de contingencia, a esta tabla también se la denomina con el nombre de tabla de clasificación cruzada, con r filas y c columnas, o tabla r × c (Marín, 2016) [18].

El objetivo que se busca en una tabla de contingencia es determinar si existe o no una relación de dependencia entre la variable de las filas y la variable de las columnas, contrastando si las dos variables son independientes, mediante un test de significación para las hipótesis:

H_0: La variable de las columnas es independiente de la variable de las filas.

H_a: La variable de las columnas no es independiente de la variable de las filas.

Se presenta a continuación la estructura de una tabla de contingencia:

Categorías de la Variable X	Categorías de la Variable Y						
	1	2	⋯	j	⋯	c	Total Marginal Fila
1	f_{11}	f_{12}	⋯	f_{1j}	⋯	f_{1c}	$f_{1\bullet}$
2	f_{21}	f_{22}	⋯	f_{2j}	⋯	f_{2c}	$f_{2\bullet}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
i	f_{i1}	f_{i2}	⋯	f_{ij}	⋯	f_{ic}	$f_{i\bullet}$
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
r	f_{r1}	f_{r2}	⋯	f_{rj}	⋯	f_{rc}	$f_{r\bullet}$
Total Marginal Columna	$f_{\bullet 1}$	$f_{\bullet 2}$	⋯	$f_{\bullet \mathrm{j}}$	⋯	$f_{\bullet \mathrm{c}}$	n

Tabla 4: Tabla de contingencia

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez.

Dónde: (Freund, 2000) [19].

f_{ij}: Es la frecuencia observada en la categoría de la fila i y columna j.

$f_{i\bullet}$: Es el total de la frecuencia observada para la fila i.

$f_{\bullet \mathrm{j}}$: Es el total de la frecuencia observada para la columna j.

Estas frecuencias verifican que:

$$f_{i\bullet} = \sum_{j=1}^{c} f_{ij}$$

$$f_{\bullet \mathrm{j}} = \sum_{i=1}^{r} f_{ij}$$

Con esta información se pueden definir las probabilidades:

p_{ij}: Es la probabilidad de que un elemento caerá en la celda correspondiente a la fila i y columna j.

$p_{i\bullet}$: Es la probabilidad de que un elemento caerá en la fila i.

$p_{\bullet \mathrm{j}}$: Es la probabilidad de que un elemento caerá en la columna j.

Estas probabilidades deben cumplir que:

$$\sum_{i=1}^{r} p_{i\bullet} = 1 \ y \ \sum_{j=1}^{c} p_{\bullet \mathrm{j}} = 1$$

Y se pueden estimar por:

$$\widehat{p_{i\bullet}} = \frac{f_{i\bullet}}{n} \ , i = 1, 2, \cdots, r$$

$$\widehat{p_{\bullet \mathrm{j}}} = \frac{f_{\bullet \mathrm{j}}}{n} \ , j = 1, 2, \cdots, c$$

Bajo la hipótesis de independencia entre dos variables, se tiene que la frecuencia esperada en la categoría de la fila i y columna j es:

$$e_{ij} = n \, \widehat{p_{i\bullet}} \, \widehat{p_{\bullet \mathrm{j}}} = \frac{f_{i\bullet} \, f_{\bullet \mathrm{j}}}{n}$$

Por lo que, otra forma de definir este contraste de hipótesis es:

H_0: $p_{ij} = p_{i\bullet} \, p_{\bullet \mathrm{j}} \quad \forall i = 1, 2, \cdots, r \quad \forall j = 1, 2, \cdots, c$

H_a: $p_{ij} \neq p_{i\bullet} \, p_{\bullet \mathrm{j}}$ Para al menos un par de valores de i y j

El estadístico de prueba para la hipótesis de independencia es:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Que sigue aproximadamente una distribución χ^2 con $(r-1)(c-1)$ grados de libertad.

Existen dos métodos que permiten realizar el contraste de hipótesis de esta prueba, los cuales generan la siguiente regla de rechazo:

Método del valor-*p*: Rechazar H_0 si el valor-*p* ≤ *α*

Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2{}_{\alpha,(r-1)(c-1)}$

Donde *α* es el nivel de significancia, y las r filas y c columnas dan los $(r-1)(c-1)$ grados de libertad.

2.4. ANÁLISIS FACTORIAL DE CORRESPONDENCIAS

El análisis de correspondencias es una técnica multivariada que pretende reducir la dimensión de una tabla de datos formada por variables cualitativas con el objetivo de obtener un número pequeño de factores, lo que facilitará la interpretación del problema investigado y permitirá que este sea más sencillo de estudiar (Pérez, 2004) [20].

Al trabajar con variables cualitativas, o con variables cuantitativas categorizadas, el análisis de correspondencias posee dos características importantes. Por una parte, trabaja con frecuencias que son el resultado del cruce de dos o más variables. Por otro, cuando se cruzan dos variables, el análisis de correspondencias utiliza como individuos y variables las diferentes categorías existentes. A este método se lo conoce como análisis de correspondencias simples (ACS). Cuando el número de categorías corresponden a más de dos variables, el método se generaliza dando lugar al análisis de correspondencias múltiples (ACM) (Luque, 2000) [21].

En el análisis de correspondencias simples, los datos categóricos de las dos variables se pueden representar mediante una tabla de contingencia. En cambio, en el análisis de correspondencias múltiples, la tabla de contingencia se transforma en una hipertabla que contiene tres o más dimensiones, la cual no es sencilla de representar y que suele simplificarse en la denominada tabla de Burt (Pérez, 2004) [20].

El análisis de correspondencias tiene como finalidad crear relaciones entre variables cualitativas potenciando la información que proporcionan las tablas de contingencia, que sólo permiten comprobar si las variables de estudio están correlacionadas y el grado de intensidad de dicha correlación, con el propósito de encontrar algún modelo causal, o determinar la existencia de algún tipo de interrelación, usando test como el ji-cuadrado χ^2 para la correlación y el test V de Cramer para la intensidad de la correlación. El inconveniente en este tipo de análisis, es que no permite conocer cuáles son las categorías que causan la relación, tampoco cuales aportan poco a dicha relación. Precisamente por esto,

una de las ventajas del análisis de correspondencias, es que permite obtener conclusiones que no son posibles hacerlo con las tablas de contingencia, ya que, por medio de la extracción de las correlaciones entre categorías, se pueden definir similaridades o disimilaridades entre ellas, que en caso de encontrarse que se corresponden entre sí, se permitirá su agrupamiento (Luque, 2000) [21].

A través del análisis de correspondencias lo anterior es representado en un espacio dimensional de pocas variables sintéticas o factores, que pueden ser nombradas e interpretadas, de tal manera que resuman la mayor cantidad de información, y que con el aporte conjunto por medio de representaciones gráficas o mapas de correspondencias se puedan visualizar las relaciones obtenidas. Dado que el análisis de correspondencias se basa en el análisis factorial, las categorías al ser representadas en el espacio generado por las dimensiones, estas se obtienen como factores cuantitativos, debido a esto, se considera al análisis de correspondencias como un método que extrae variables cuantitativas ficticias a partir de variables cualitativas originales, ya que estas definen las relaciones que existen entre sus categorías (Pérez, 2004) [20].

El análisis de correspondencias, puede establecerse como un paso previo para la aplicación de otras técnicas multivariadas como el análisis de clúster, el análisis de regresión o el análisis discriminante, dado que hace posible la aplicación a un conjunto de datos cualitativos, porque consigue coordenadas métricas en el espacio dimensional definido por los factores (Luque, 2000) [21].

2.4.1. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

En la sección anterior se describió la aplicación del análisis factorial de correspondencias o análisis de correspondencias simple sobre dos variables de tipo cualitativa, descomponiéndose a la vez cada una de ellas en varias categorías.

Ahora se demostrará cómo se generaliza dicho método cuando se tienen más de dos variables cualitativas, mediante el análisis de correspondencias múltiples.

Se conoce que el análisis factorial de correspondencias simple es adecuado para el caso de tablas de contingencia, donde los datos son dispuestos en las combinaciones de las distintas modalidades o categorías de dos variables (caracteres) de tipo cualitativo. Si se cruza en una tabla de contingencia la variable fila I con categorías desde i=1,2,…,n; con la variable columna J con categorías desde j=1,2,…,p; se define kij para caracterizar la cantidad de observaciones que corresponden simultáneamente a la categoría i de la variable I y a la categoría j de la variable J. En el caso de tener más de dos caracteres o variables categóricas J1,J2,…,Jq; ya no es posible trabajar con tablas de contingencia y la representación de los datos se torna difícil. Sin embargo, es posible estudiar todas las relaciones entre las distintas categorías de todas las variables cualitativas existentes, el presente trabajo describe la metodología del análisis de correspondencias múltiples como generalización del análisis factorial de correspondencias simples (Pérez, 2004) [20].

2.4.2. FORMULACIÓN DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

En el análisis de correspondencias múltiples los datos se disponen en la llamada tabla disyuntiva completa o tabla Z, la cual está formada por un conjunto I de n individuos o filas, un conjunto J de Q variables o caracteres cualitativos cada una con un conjunto $1,\cdots,m_k$ de categorías o modalidades excluyentes, donde el número de modalidades o categorías es igual a $J=\sum_{k=1}^{Q} m_k$. Esta tabla disyuntiva completa Z, cuya dimensión es I × J tiene la siguiente forma: (Pérez, 2004) [20].

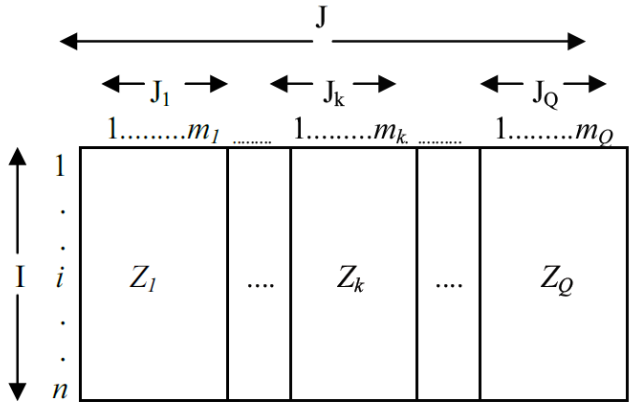


Figura 2: Tabla disyuntiva completa Z

Fuente: Pérez, C. (2004). *Técnicas de análisis multivariante de datos*. (p. 247), Madrid: Pearson Prentice Hall. - Autor: Guillermo Ordóñez.

Donde $Z = Z_1 \ldots . Z_k \ldots . Z_Q$.

En la tabla disyuntiva completa Z, el elemento Z_{ij} puede tomar el valor 0 o 1, dependiendo si el individuo i ha elegido la modalidad j o no, dicho de otra manera si está o no influenciado por la categoría j. Debido a esto, cada rectángulo de la tabla Z puede considerarse, a pesar de no serlo, como una tabla de contingencia cuyos valores son 0 o 1. De otra manera, la tabla Z consiste en Q subtablas adosadas, con el propósito que se puedan representar simultáneamente todas las modalidades o categorías (columnas) de todos los individuos (filas). Como las modalidades son excluyentes, cada subtabla consta de un solo 1 en cada fila (Pérez, 2004) [20].

Para mantener la notación que se usa en el análisis de correspondencias simple, a los elementos Z_{ij} se los representará como k_{ij}, de esta manera el lector estará familiarizado con la notación habituada hasta ahora. Por lo tanto, se tiene que:

$$Z_{ij} = k_{ij} = 0 \text{ ó } 1$$

$k_{i\bullet} = \sum_j K_{ij} = Q$ = número de modalidades (cada subtabla tiene un único 1 en cada fila.

$k_{\bullet j} = \sum_i K_{ij} =$ número de individuos que poseen modalidad j.

$\frac{f_{ij}}{f_{i\bullet}} = \frac{k_{ij}}{k_{i\bullet}} = \frac{1}{Q} =$ inverso del número de modalidades (0 si el individuo no elige j).

2.4.3. OBTENCIÓN DE LOS FACTORES: TABLA DE BURT

Sea la matriz $V = \frac{1}{Q} D^{-1} B$, V es una matriz de varianzas y covarianzas la cual se debe diagonalizar para obtener los factores, donde B es la tabla de Burt. Se tiene que $B = Z'Z$ (es el producto entre la matriz de datos transpuesta por sí misma) es una matriz simétrica que se compone de Q^2 bloques o agrupaciones, de tal manera que en la diagonal $Z'_k Z_k$ sus bloques son tablas diagonales que cruzan una variable con ella misma, donde los efectivos de cada categoría $k_{\bullet j}$ son los elementos de la diagonal. Aquellas agrupaciones que se encuentran fuera de la diagonal son tablas de contingencia que se obtienen a partir del cruce de dos en dos de las características $Z'_k Z_k$ y sus elementos son las frecuencias de asociación de las dos modalidades respectivas. D^{-1} , es una matriz diagonal cuyos elementos diagonales son los de la matriz de Burt, y sus elementos restantes son iguales a cero. La tabla de Burt tiene la siguiente forma: (Pérez, 2004) [20].

	J_1	J_2	$\cdots$	J_Q
J_1	$0 \cdot .0$	C_{12}	$\cdots$	C_{1Q}
J_2	C_{21}	$0 \cdot .0$	$\cdots$	C_{2Q}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
J_Q	C_{Q1}	C_{Q2}	$\cdots$	$0 \cdot .0$

Figura 3: Tabla de Burt

Fuente: Pérez, C. (2004). *Técnicas de análisis multivariante de datos*. (p. 248), Madrid: Pearson Prentice Hall. - Autor: Guillermo Ordóñez.

Las fórmulas de transición vienen dadas por las siguientes expresiones:

$$f_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i.}} \right) G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{Q} \sum_{j=1}^{p} k_{ij} G_\alpha(j)$$

$$G_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{.j}} \right) f_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{k_{.j}} \sum_{i=1}^{n} k_{ij} F_\alpha(i)$$

Estas fórmulas hacen posible representar los puntos línea y los puntos columna sobre los mismos gráficos de manera simultánea, relacionando así los resultados en los dos subespacios.

Si partimos que kij = 1 cuando el individuo i toma la modalidad j y cero en caso contrario, la proyección de un punto individuo i sobre el eje α, Fα(i), es el baricentro, excepto un coeficiente de dilatación 1/$\sqrt{\lambda\alpha}$, de las proyecciones de los puntos modalidades sobre el eje Gα(j). A todas las modalidades le corresponde el mismo peso 1/Q. De manera similar, la proyección de un punto modalidad j sobre el eje α, Gα(j), es el baricentro, excepto un coeficiente de dilatación 1/$\sqrt{\lambda\alpha}$, de las proyecciones de los puntos individuos que poseen esa modalidad sobre el eje Fα(i), todos ellos afectados del mismo peso k.j.

En el análisis factorial de correspondencias múltiples (ACM), el centro de gravedad de la nube de puntos de cada variable N(j) es √fi., el cual puede asemejarse a una distribución uniforme 1/√n, es decir:

$$k_{i\bullet} = \sum_j k_{ij} = Q \rightarrow \sum_i k_{i\bullet} = nQ \rightarrow f_{i\bullet} = \frac{1}{n}.$$

Dado que mediante la distribución marginal del centro de gravedad de la nube de modalidades N(J) se obtiene el centro de gravedad de la subtabla IxJk, el centro de gravedad de las modalidades de cada variable, cada una ponderada por su peso, es el mismo que el del N(J), es decir, 1/√n. Ya que sólo acoge una variable, la suma de cada fila es 1 y el total de la tabla es n, por lo tanto fi.=1/n.

Las modalidades de cada variable están centradas alrededor del origen, sin que todas tengan el mismo signo, esto es porque el análisis factorial de correspondencia es centrado y el centro de gravedad de las modalidades de una variable coincide con el centro de gravedad del conjunto J, y con el origen.

Una ayuda para la interpretación de cada fila y columna en el análisis factorial de correspondencias, es a través del cálculo de la contribución de una variable Jk al factor α como:

$$CTA_\alpha(J_k) = \sum_{j \in J_k} CTA_\alpha(j)$$

Que es la suma de las contribuciones de las modalidades de la variable Jk.

Sea G el centro de gravedad, la inercia debida a la modalidad j es igual a:

$$I(j) = f_{.j}.d^2(G,j) = f_{.j}\sum_{i=1}^{n}\left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \sqrt{f_{i.}}\right)^2 = \frac{k_{.j}}{nQ}\sum_{i=1}^{n}\left(\frac{k_{ij}/nQ}{k_{.j}1/n} - 1/\sqrt{n}\right)^2 = \frac{1}{Q}\left(1 - \frac{k_{.j}}{n}\right)$$

Donde la fracción de inercia correspondiente a una modalidad j es mayor siempre que el efectivo de esa modalidad sea menor, por consiguiente, es recomendable suprimir las modalidades que muy pocas veces son

seleccionadas, en este caso se construye otra modalidad al unirla con la más próxima.

La inercia de una variable es la suma de las inercias de sus modalidades o categorías, es decir:

$$I(J_k) = \sum_{j \in J_k}^{n} I(j) = \sum_{j \in J_k}^{n} \frac{1}{Q}\left(1 - \frac{k_{.j}}{n}\right) = \frac{1}{Q}(m_k - 1)$$

De la cual se tiene que la fracción de inercia debida a una variable es función creciente de la cantidad de modalidades de respuesta que posee. Si una variable tiene una excesiva cantidad de categorías o modalidades y dado el caso que su efectivo sea muy pequeño, pueden ocurrir influencias extremas, para evitarlo se aconseja que las modalidades se reagrupen en una cantidad que sea razonable y se mantenga el sentido.

La inercia total es la suma de las inercias de todas las modalidades:

$$I = \sum_{j \in J_k} I(J_k) = \sum_{k} \frac{1}{Q}(m_k - 1) = \frac{J}{Q} - 1$$

Donde J/Q es el promedio de modalidades por variable cualitativa o carácter. En efecto, la inercia total depende únicamente del número de modalidades y del de preguntas. Si se da el caso de tener dos variables, y cada una con dos modalidades, los resultados se pueden analizar tanto por Análisis de Correspondencia Múltiple (ACM) como por Análisis Factorial de Correspondencias (AFC).

Si se analiza la información mediante el análisis de correspondencias múltiples, se conseguirá siempre la misma inercia (J/Q-1=1), aunque se obtendrá dos ejes. De existir mucha relación de dependencia entre las dos variables, el primer eje acumulará gran parte de la inercia (casi 1) y el segundo muy poca, mientras que, si las dos variables son totalmente independientes, ambos factores acumularán la mitad (½) de la inercia cada uno.

En el caso del análisis factorial de correspondencias, se obtiene solo un factor que acumula el 100% de la inercia total, esta inercia dependerá del grado de relación que exista entre las categorías o modalidades. De esta manera, si las modalidades están demasiado relacionadas, la inercia tenderá a un valor alto, y si están poco relacionadas, la inercia se acercará a cero (Pérez, 2004) [20].

2.5. REGRESIÓN LOGÍSTICA

Muchas veces en el campo de la ciencia, suelen presentarse situaciones en las cuales el investigador desea predecir la ocurrencia o no de cierto evento en función de un grupo de variables explicativas de interés, donde debido a la naturaleza que presentan estas variables no es posible poder representarlas de manera cuantitativa.

Este problema conlleva a encontrar un modelo que determine la relación entre un conjunto de variables independientes o explicativas de tipo tanto cualitativo como cuantitativo y una variable dependiente binaria o dicotómica que tan solo toma dos posibles valores que determinan características opuestas o mutuamente excluyentes, como, por ejemplo: la propensión de un vehículo a sufrir o no un accidente de tránsito. Un posible modelo seria el análisis discriminante, no obstante, las características descritas de las variables independientes sobre la coexistencia de variables cualitativas y cuantitativas, incumple la asunción de normalidad multivariada, por lo que no podría usarse. Otra opción que queda descartada es el modelo de regresión lineal múltiple, en este caso es debido a las limitaciones que presenta en relación a la condición dicotómica de la variable dependiente o respuesta. En el presente capítulo se desarrollará una técnica de análisis multivariante llamada regresión logística múltiple, la misma que solventa los inconvenientes y limitaciones de los modelos antes mencionados, ya que se ajusta a la naturaleza de las variables referidas (Luque, 2000) [21].

La regresión logística múltiple es una técnica de análisis multivariante que se aplica cuando se quiere analizar la relación entre una o más variables independientes de tipo cualitativo o cuantitativo y una variable dependiente o variable respuesta, dicotómica o binaria. Además, para las variables independientes de tipo cualitativo en el modelo, cada una con k categorías, se deberá generar k – 1 variables de tipo DUMMY por cada una de ellas. La variable DUMMY toma solo dos valores 0 o 1, si toma el valor de 1 significa que un individuo corresponde a una determinada categoría de la variable cualitativa y 0 en caso contrario, de tal forma que todas las categorías de la variable independiente cualitativa queden representadas adecuadamente en el modelo (Álvarez, 2008) [22].

Los objetivos que se buscan conseguir mediante el modelo de regresión logística múltiple no solamente es determinar la existencia o no de relación entre las variables independientes y la variable dependiente, sino también medir el grado de intensidad de la relación y estimar o predecir la probabilidad de que el evento definido por la variable dependiente ocurra o no en función de los valores que tomen las variables independientes (Luque, 2000) [21].

2.6. MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

Sea y una variable dependiente que puede tomar dos únicos valores 0 o 1, el modelo de regresión logística múltiple trata de estimar la probabilidad de que ocurra cierto evento, es decir la probabilidad de que ocurra una respuesta binaria (0 o 1), con lo cual, la distribución de probabilidad es: $y=1$ con probabilidad p, y 0 con probabilidad $1-p$, esto en función de los valores que tome el grupo de variables independientes (explicativas), $x_{1i}, x_{2i}, \cdots, x_{ni}$, que pueden ser de tipo cualitativas o cuantitativas (Luque, 2000) [21].

El modelo de regresión logística múltiple se define mediante la siguiente expresión matemática:

$$E(y_i) = P(y_i = 1 | x_{1i}, x_{2i}, \cdots, x_{ni}) = p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}$$

o simplemente expresado como:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}}$$

De manera similar, este modelo se puede representar por medio de las siguientes expresiones:

$$\frac{p_i}{1 - p_i} = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}$$

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}$$

Para hacer la estimación de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$, y el ajuste de este modelo se usará el método de estimación por máxima verosimilitud (EMV).

En este método, primero se construye la función de verosimilitud denotada como L, donde L es una función de parámetros desconocidos que se usa para expresar la probabilidad de los datos observados. Los estimadores denominados maximoverosímiles de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$, que maximizan la función L son los valores de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$. Es decir, los valores obtenidos $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ mediante este método son los estimadores de los parámetros desconocidos $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$, los mismos que maximizan la probabilidad de que con ellos se puedan obtener los valores observados (Luque, 2000) [21].

Una vez que se tenga ajustado el modelo y obtenidos los coeficientes maximoverosímiles $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$, se estima $\hat{p}$, es decir la probabilidad de que un sujeto escoja la opción binaria 1 dado determinado valor de las variables independientes $x_{1i}, x_{2i}, \cdots, x_{ni}$, mediante la siguiente expresión:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}}$$

Una expresión similar es la llamada “odds” o “ventaja” y se define como:

$$\frac{\hat{p}_i}{1 - \hat{p}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}$$

Esta expresión es el cociente entre la probabilidad de que ocurra el suceso de estudio, que en el caso de la regresión logística es siempre $y_i = 1$, y la probabilidad de que no ocurra tal suceso. De otra manera, estima la ventaja o preferencia de un sujeto hacia la opción 1 de la variable dependiente frente la opción 0, para cada valor que tomen las variables independientes (Luque, 2000) [21].

Para que el modelo del odds sea lineal, se debe aplicar el logaritmo natural a la ventaja o preferencia por la respuesta 1 frente a la respuesta 0, este procedimiento es denominado transformación logística, mientras que la probabilidad p_i se formula mediante un modelo no lineal o modelo logístico. Al tomar el logaritmo a las ventajas o también llamado logit, se hace más sencillo la interpretación de los coeficientes del modelo en cuanto a la relación entre las variables (Pérez, 2004) [20].

El modelo logit, se expresa de la siguiente manera:

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}$$

2.6.1. PRUEBA DE SIGNIFICANCIA

El siguiente paso, es verificar la validez del modelo de regresión logística múltiple mediante la prueba de significancia, que al igual que en la regresión lineal múltiple, se plantean y realizan contrastes de hipótesis sobre los coeficientes de regresión obtenidos tanto de manera grupal, es decir, se formula una prueba para comprobar la significancia del modelo en general, como individual, que significa hacer contrastes de hipótesis para cada coeficiente estimado en el modelo. Estas pruebas permiten constatar si la relación entre el grupo de variables independientes que integran el modelo ajustado y la variable dependiente o respuesta es significativa (Luque, 2000) [21].

Las hipótesis para comprobar la significancia general del modelo, se plantea de la siguiente forma:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

$$H_1 : \text{Al menos algún } \beta_i \neq 0, \forall i = 1,2,\cdots,n$$

El estadístico de prueba para este contraste de hipótesis, es el test G, que se define como:

$$G = -2\ln\left[\frac{\text{Verosimilitud del modelo sólo con la constante } (L_0)]}{\text{Verosimilitud del modelo seleccionado } (L_p)}\right]$$

También llamado prueba de la razón de verosimilitud, cuya distribución es una ji-cuadrada χ^2 con p-1 grados de libertad, donde p es el número de parámetros del modelo en estudio. El estadístico G, se basa en la función de verosimilitud de cada modelo, finalmente, compara la probabilidad de que las cantidades estimadas por cada modelo sean representativas de los valores observados de la variable dependiente (Luque, 2000) [21].

La zona de rechazo que se determina para esta prueba, está dada por el método del valor crítico que establece la siguiente regla:

Método del valor crítico: Rechazar H_0 si $G \geq \chi^2_{\alpha,(p-1)}$

Donde α es el nivel de significancia y p-1 son los grados de libertad.

Si se rechaza la hipótesis nula, se deduce que el modelo es significativo.

Si al realizar la prueba estadística G, se evidencia que el modelo en general es significativo, compete hacer una prueba de significancia para cada coeficiente de regresión. Esta prueba individual se basa en el estadístico W de Wald, y sirve para evaluar si la contribución que hace cada variable independiente al modelo es significativa (Luque, 2000) [21].
El planteamiento de las hipótesis para probar la significancia individual es:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0, \forall i = 1,2,\cdots,n$$

Y el valor del estadístico de prueba de Wald se define como:

$$\mathrm{W} = \left(\frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)} \right)^2$$

Donde W es igual al cuadrado del cociente entre el estimador del coeficiente de la variable independiente y el estimador de su error estándar. Este estadístico sigue una distribución ji-cuadrada χ^2, con 1 grado de libertad en el caso de que la variable independiente sea cuantitativa, mientras que, si la variable independiente es cualitativa, el número de grados de libertad es igual al número de categorías menos uno (Luque, 2000) [21].
El criterio de decisión para esta prueba se realiza mediante el método del valor crítico que se define como:

Método del valor crítico: Rechazar H_0 si $\mathrm{W} \geq \chi^2_{\alpha,1}$ o

Rechazar H_0 si $\mathrm{W} \geq \chi^2_{\alpha,(p-1)}$

Donde α es el nivel de significancia, p es el número de categorías de la variable cualitativa y $(p-1)$ son los grados de libertad.
Si se rechaza la hipótesis nula, se concluye que la variable independiente es estadísticamente significativa.

2.6.2. MEDIDAS DE LA BONDAD DE AJUSTE

Las medidas de la bondad del ajuste, son pruebas o métodos que sirven para evaluar el grado de efectividad absoluta del modelo analizado, en relación a la explicación de la variable dependiente, es decir, determinan la aproximación de los valores estimados $\hat{y}_i$ a los valores reales observados y_i (Luque, 2000) [21].

Las formas de medir la bondad del ajuste de un modelo son: las basadas en pruebas de hipótesis y las que son equivalentes al coeficiente de determinación R2 utilizado en la regresión lineal múltiple.

2.6.2.1. BONDAD DEL AJUSTE USANDO CONTRASTE DE HIPÓTESIS

Esta clase de medidas se fundamentan en el siguiente contraste de hipótesis:

H_0: El modelo seleccionado ajusta bien los datos

H_1: $\neg H_0$

A través de un estadístico que sigue una distribución conocida. Los tipos de medidas de bondad del ajuste son:

2.6.2.1.1. DESVIANZA (DEVIANCE)

Sea D, un estadístico de prueba llamado desvianza, el cual se define como:

$$D = -2 \ln\left[\frac{\text{Verosimilitud del modelo seleccionado o ajustado}}{\text{Verosimilitud del modelo saturado o completo}}\right]$$

Donde un modelo saturado es el que tiene tanto parámetros como datos y cuya predicción de los valores observados es correcta.

La distribución del estadístico D es una ji-cuadrada χ^2 con N - p grados de libertad, donde N es el número de observaciones y p la

cantidad de parámetros incluidos en el modelo. En ciertos programas estadísticos se acostumbra usar la expresión -2 In likelihood o -2 Logaritmo de la verosimilitud para referirse a la desvianza de un modelo concreto.

Para realizar el contraste de hipótesis de esta prueba, se emplea el método del valor crítico que define el siguiente criterio de rechazo:

Método del valor crítico: Rechazar H_0 si D $\geq \chi^2_{\alpha,N-p}$

Donde α es el nivel de significancia, y $N-p$ son los grados de libertad.

En el caso que no se rechace la hipótesis nula, la conclusión es que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.2. PRUEBA DE LA JI-CUADRADO

Este tipo de test, se basa en el cálculo de un estadístico ji-cuadrado χ^2, el cual mide el grado de desajuste que podría existir si se compara la cantidad observada de respuestas afirmativas con la estimación de la probabilidad que realiza el modelo, estas comparaciones se efectúan para cada uno de los prototipos de predictores presentes. El prototipo o patrón de predictores, son cada una de las distintas combinaciones de valores que pueden tomar las variables independientes incorporadas en el modelo (Luque, 2000) [21].

Como ejemplo, si se tienen las variables Sexo (1 =hombre; 0= mujer) y Etnia (1 = mestizo; 0= afro), las combinaciones de estas determinan cuatro patrones de covariables, dado que cada individuo que integra la muestra se puede clasificar en uno de los siguientes grupos o patrones: hombre-mestizo; hombre-afro; mujer-mestiza; y mujer-afro.

Si el número de patrones de predictores es menor que el número de observaciones, es decir M<N, el estadístico χ^2 será:

$$\chi^2 = \sum_{i=1}^{M} \frac{m_i\left(y_i - \hat{p}_i\right)^2}{\hat{p}_i\left(1-\hat{p}_i\right)}$$

Donde:

m_i : es el número de casos incluidos en cada patrón de predictores.

y_i : es la respuesta binaria de la variable dependiente.

$\hat{p}_i$: es la probabilidad estimada por el modelo para el patrón de covariables i.

En el caso de que las muestras sean grandes, la distribución del estadístico χ^2 será una ji-cuadrada con M - p grados de libertad.

Si se tienen variables cuantitativas continuas, es muy probable que $M \approx N$, por tanto, el estadístico χ^2 se expresaría como:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - \hat{p}_i\right)^2}{\hat{p}_i\left(1-\hat{p}_i\right)}$$

El problema que se presenta cuando $M \approx N$, es la obtención de valores p erróneos, sin embargo, en aquellos casos donde el modelo ajustado es el correcto, se sugiere utilizar la prueba χ^2 con N-p grados de libertad, de esta forma se obtendrá resultados adecuados.

Para realizar el contraste de hipótesis de esta prueba, se aplica el siguiente criterio de decisión:

Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_{\alpha,M-p}$ o

Rechazar H_0 si $\chi^2 \geq \chi^2_{\alpha,N-p}$

Donde α es el nivel de significancia, M es la cantidad de patrones de predictores, N es el número de observaciones, p el número de parámetros contenidos en el modelo, $M-p$ y $N-p$ son los grados de libertad.

En el caso que no se rechace la hipótesis nula, la conclusión es que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.3. PRUEBA DE HOSMER-LEMESHOW

Esta prueba es adecuada para aplicarse en modelos que se componen de una o varias variables independientes que sean continuas y que, el número de patrones de predictores sea aproximadamente igual al número de casos observados, es decir $M \approx N$.

El procedimiento para esta prueba es como sigue, primero se estiman las probabilidades $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_N$, luego estas probabilidades se deben ordenar de menor a mayor (una para cada caso observado) y a continuación juntarlas en diez grupos, de tal forma, que el primer grupo contenga los $n_1 = N/10$ individuos que tengan las probabilidades estimadas más bajas, y así hasta tener en el último grupo los $n_{10} = N/10$ individuos con las probabilidades estimadas más altas. A estos grupos se los conoce como deciles de riesgo.

El estadístico de prueba de Hosmer-Lemeshow se denota como $\hat{C}$, y se calcula mediante una tabla de 2 x 10, que incluya a las frecuencias observadas y estimadas para cada uno de los diez grupos (Luque, 2000) [21].

La distribución del estadístico $\hat{C}$, es una ji-cuadrada χ^2 con 10-2=8 grados de libertad y se define como:

$$\hat{C} = \sum_{k=1}^{10} \frac{\left(o_k - n_k \bar{p}_k\right)^2}{n_k \bar{p}_k \left(1 - \bar{p}_k\right)}$$

Donde:

n_k es el número de patrones de predictores del grupo k-ésimo.

$o_k = \sum_{i=1}^{n_k} y_i$ es decir, el número de respuestas afirmativas registradas para la variable respuesta (y = 1) para los n_k patrones de predictores.

$\bar{p}_k = \sum_{i=1}^{n_k} \frac{m_i \hat{p}_i}{n_k}$ es la media de la probabilidad estimada.

Para realizar el contraste de hipótesis de esta prueba, se aplica el siguiente criterio de decisión:

Método del valor crítico: Rechazar H_0 si $\hat{C} \geq \chi^2_{\alpha,8}$

Donde *α* es el nivel de significancia, y 8 son los grados de libertad. En el caso que no se rechace la hipótesis nula, la conclusión es que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.4. BONDAD DEL AJUSTE: EFICACIA PREDICTIVA

Otra forma de estimar la bondad del ajuste del modelo elegido, radica en la comparación de las predicciones del modelo con los datos observados, este procedimiento se realiza a través de una tabla de clasificación. Esta es una tabla de doble entrada, en la cual se organizan los casos que integran la muestra conforme a los valores observados de la variable dependiente (respuesta dicotómica) y a los valores pronosticados por el modelo estimado, de tal manera, que para todos los casos donde la probabilidad estimada sea mayor o igual a 0.5 (valor de corte), se los ordenaran dentro del grupo que contenga la característica representada por la variable dependiente; y en caso contrario, las observaciones cuya probabilidad estimada es menor que 0.5, se las ordenará dentro del

grupo que indique inexistencia de la característica representada por la variable dependiente (Luque, 2000) [21].

La tabla de clasificación tiene la siguiente estructura:

Observado		**Pronosticado**		
		Variable dependiente		**Porcentaje correcto**
		0	**1**	
Variable dependiente	**0**	a	b	$\frac{a}{a+b}$
	1	c	d	$\frac{d}{c+d}$
Tasa global de aciertos				$\frac{a+d}{a+b+c+d}$

Tabla 5: Tabla de clasificación

Fuente: Datos de la investigación - Autor: Guillermo Ordoñez.

Donde:

a y d son los casos que el modelo clasificó de manera correcta.

b y c los casos que el modelo clasificó de manera incorrecta.

De tal forma se pueden definir los siguientes índices:

Tasa global de aciertos= $\frac{a+d}{a+b+c+d}$

Especificidad= $\frac{a}{a+b}$

Sensibilidad= $\frac{d}{c+d}$

Para poder verificar que el modelo tiene una elevada eficacia predictiva, se debe constatar la significación estadística de la tasa global de aciertos, por lo cual, se usara el test de Huberty, en esta prueba, se calcula en primer lugar el número esperado de casos correctamente clasificados debidos al azar:

$$e = \frac{(a+b)^2 + (c+d)^2}{a+b+c+d}$$

Luego, se contrastan las hipótesis:

H_0 : Número de casos correctamente clasificados por el modelo no difiere de la clasificación esperada sólo por efecto del azar.

H1: $\neg H_0$

El estadístico de prueba para realizar el contraste de hipótesis es:

$$Z^* = \frac{(a+d-e)\sqrt{a+b+c+d}}{\sqrt{e(a+b+c+d-e)}}$$

El cual sigue una distribución normal estándar.

Para realizar el contraste de hipótesis de esta prueba, se aplica el siguiente criterio de decisión:

Método del valor crítico: Rechazar H_0 si $|Z^*| \geq Z_{\alpha/2}$

Donde α es el nivel de significancia.

En el caso que se rechace la hipótesis nula, la conclusión es que la tasa de aciertos del modelo es significativamente mayor que la que se obtendría debido al azar.

2.6.2.2. BONDAD DEL AJUSTE: MEDIDAS SIMILARES A R^2

Otras medidas que permiten evaluar la bondad de ajuste de un modelo de regresión logística semejantes al coeficiente R^2 utilizado en la regresión lineal son:

1. La R cuadrado de Cox y Snell
2. La R cuadrado de Nagelkerke

2.6.2.2.1. R CUADRADO DE COX Y SNELL

Este coeficiente de determinación es generalizado y se utiliza para estimar la porción de varianza que las variables independientes explican de la variable respuesta, cuyos valores oscilan entre 0 y 1.

Se basa en comparar la razón entre el logaritmo de la verosimilitud para el modelo completo y el logaritmo de la verosimilitud para el modelo reducido que comprende solo del término independiente (Bernal, 2014) [23].

$$R^2 = 1 - \left(\frac{\hat{L}_C}{\hat{L}_0} \right)^{\frac{2}{N}}, \quad 0 \leq R^2 < 1$$

Donde:

$\hat{L}_C$: Es el logaritmo de la verosimilitud para el modelo completo.

$\hat{L}_0$: Es el logaritmo de la verosimilitud para el modelo reducido.

N: Es el número de observaciones.

Si el valor del R cuadrado de Cox y Snell es próximo a 1, es un indicativo que el ajuste es bueno.

2.6.2.2.2. R CUADRADO DE NAGELKERKE

Este coeficiente es una corrección del R cuadrado de Cox y Snell, toma un valor máximo que es menor a 1. Además, corrige la escala del estadístico para cubrir el rango completo de 0 a 1 (Bernal, 2014) [23].

$$\overline{R}^2 = \frac{R^2}{R^2_{Max}} \quad \text{Donde: } R^2_{Max} = 1 - \left(\hat{L}_0 \right)^{\frac{2}{N}}$$

Si el valor del R cuadrado de Nagelkerke es próximo a 1, es un indicativo que el ajuste es bueno.

CAPÍTULO III

3. ANÁLISIS ESTADÍSTICO UNIVARIADO Y TABLAS DE CONTINGENCIA

3.1. INTRODUCCIÓN

Este capítulo abarcará una primera parte compuesta por la descripción y codificación de las variables consideradas para el estudio, luego una segunda parte que, en base a estas variables se realizará el análisis estadístico univariado y el análisis de tablas de contingencia. El análisis se efectuará sobre una muestra de 92,780 vehículos, proporcionada por una aseguradora de la ciudad de Guayaquil.

La estructura de la muestra proporcionada por la aseguradora, está comprendida por variables relacionadas a las características de los vehículos, sobre las cuales se basa el presente estudio y que servirán para realizar los respectivos análisis estadísticos.

La mayoría de las variables consideradas son de naturaleza cualitativa, cuyo análisis se realizará por medio de tablas de frecuencias, diagramas de barras y de pastel. Para el caso de las variables cuantitativas, el análisis se realizará a través del cálculo de medidas de tendencia central, medidas de dispersión, medidas de asimetría y medidas de curtosis, percentiles, tablas de frecuencias, histogramas y diagramas de cajas.

3.2. DESCRIPCIÓN Y CODIFICACIÓN DE LAS VARIABLES DE LA INVESTIGACIÓN

En primer lugar, se procederá a realizar la descripción y codificación de las variables independientes de tipo cualitativa.

Variable: Categoría

Descripción:

Variable cualitativa dicotómica que identifica las categorías del vehículo bajo estudio.

Esta variable toma únicamente dos valores o categorías que se codifican como:

Codificación:

Codificación	Categoría
0	LIVIANO
1	PESADO

Tabla 6: Variable: Categoría

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Marca

Descripción:

Variable cualitativa de clase nominal, que identifica la marca del vehículo bajo estudio.

Los valores o categorías que puede tomar esta variable junto a su codificación son los siguientes:

Codificación:

Codificación	Marca
0	CHEVROLET
1	SUZUKI
2	HYUNDAI
3	NISSAN
4	KIA
5	TOYOTA
6	FORD
7	TUKO
8	MAZDA
9	HINO
10	MITSUBISHI
11	RENAULT

12	VOLKSWAGEN
13	SKODA
14	HONDA
15	GREAT WALL
16	OTRA

Tabla 7: Variable: Marca

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Tipo

Descripción:

Variable cualitativa de clase nominal, que identifica el tipo del vehículo bajo estudio.

Los valores o categorías que puede tomar esta variable junto a su codificación son los siguientes:

Codificación:

Codificación	Tipo
0	SEDÁN
1	TODOTERRENO
2	CAMIONETA
3	MOTO
4	CAMIÓN
5	OTRO

Tabla 8: Variable: Tipo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Color

Descripción:

Variable cualitativa de clase nominal, que identifica el color del vehículo bajo estudio.

Los valores o categorías que puede tomar esta variable junto a su codificación son los siguientes:

Codificación:

Codificación	Color
0	BLANCO
1	PLATEADO
2	NEGRO
3	PLOMO
4	AZUL
5	ROJO
6	BEIGE
7	ACERO
8	GRIS
9	VERDE
10	DORADO
11	CONCHO DE VINO
12	CELESTE
13	AMARILLO
14	OTRO

Tabla 9: Variable: Color

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Siniestro

Descripción:

Variable dependiente o respuesta de naturaleza cualitativa dicotómica, a través de la cual se determina si el vehículo bajo estudio sufrió o no un siniestro.

Esta variable toma únicamente dos valores que se codifican como:

Codificación:

Codificación	Siniestro
0	No hubo Siniestro
1	Hubo Siniestro

Tabla 10: Variable: Siniestro

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Por último, se describirá y codificará la única variable cuantitativa.

Variable Año

Descripción:

Variable cuantitativa discreta, que identifica el año de fabricación del vehículo.

Codificación:

Puesto que se necesita realizar el análisis de tablas de contingencia, esta variable se la codificará para transformarla en cualitativa. Cabe resaltar que para el análisis univariado, la variable año se trabaja sin codificación alguna.

Para la codificación se dividirá el conjunto de observaciones en 4 partes iguales, cada una contendrá un 25% de los datos, para esto se necesita calcular los estadísticos: mínimo, percentil 25, percentil 50, percentil 75 y máximo. Este procedimiento permitirá formar intervalos, dentro de los cuales se clasificarán a los datos. Se presenta a continuación la codificación de esta variable.

Codificación	Año
1	[mínimo, Percentil 25)
2	[Percentil 25, Percentil 50)
3	[Percentil 50, Percentil 75)
4	[Percentil 75, máximo]

Tabla 11: Variable: Año

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

3.3. ANÁLISIS ESTADÍSTICO UNIVARIADO DE LAS VARIABLES DE LA INVESTIGACIÓN

Variable: Categoría

En la Tabla 12, se observa que los vehículos livianos tienen una mayor participación con un 90.53% de un total de 92,780 vehículos asegurados en la ciudad de Guayaquil, y en menor proporción con un 9.47% correspondiente a los vehículos pesados. La representación gráfica de la distribución se presenta en la Figura 4.

Categoría	Frecuencia Absoluta	Frecuencia Relativa
LIVIANO	83991	90.53%
PESADO	8789	9.47%
Total	**92780**	**100%**

Tabla 12: Categoría del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

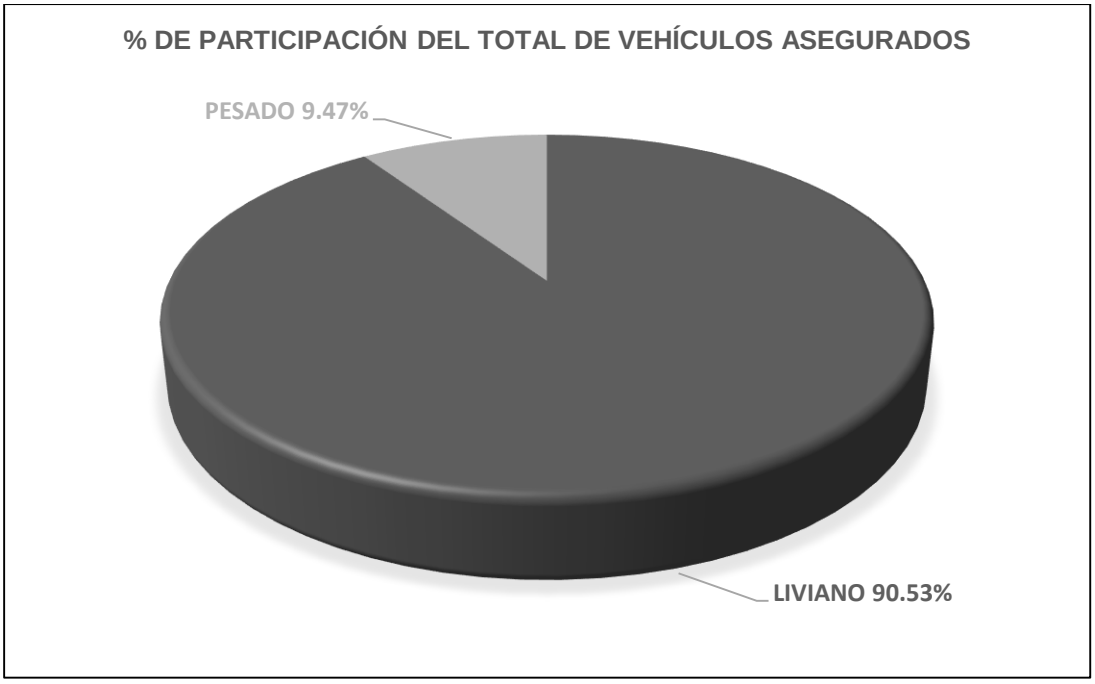


Figura 4: Categoría del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Marca

En la Tabla 13, se puede notar que, de un total de 92,780 vehículos asegurados en la ciudad de Guayaquil, la mayor participación la tienen los vehículos de marca Chevrolet con un 28.78%, seguido por las marcas Suzuki con 12.13%, Hyundai con 7.97% y Nissan con el 5.81%. Estas cuatro marcas abarcan más del 50% en el mercado asegurado. Los vehículos de marca TUKO que corresponden básicamente a las motos tienen un 2.85% de participación. Cabe resaltar que en la categoría OTRA, se encuentran agrupadas todas las marcas que individualmente no son representativas en el conglomerado total. La representación gráfica de la distribución se presenta en la Figura 5.

Marca	Frecuencia Absoluta	Frecuencia Relativa
CHEVROLET	26700	28.78%
SUZUKI	11252	12.13%
HYUNDAI	7396	7.97%
NISSAN	5389	5.81%
KIA	5298	5.71%
TOYOTA	4742	5.11%
FORD	4078	4.40%
TUKO	2644	2.85%
MAZDA	2563	2.76%
HINO	2490	2.68%
MITSUBISHI	2197	2.37%
RENAULT	2130	2.30%
VOLKSWAGEN	1773	1.91%
SKODA	1625	1.75%
HONDA	1185	1.28%
GREAT WALL	576	0.62%
OTRA	10742	11.58%
Total	**92780**	**100.00%**

Tabla 13: Marca del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

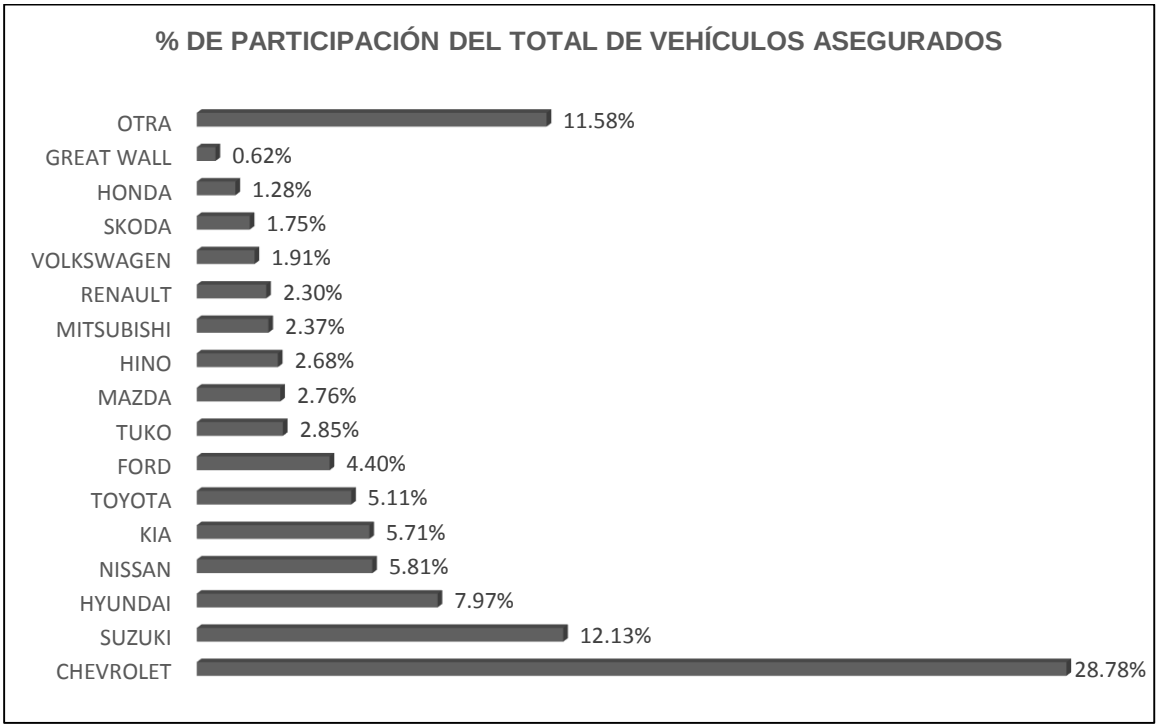


Figura 5: Marca del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Tipo

En la Tabla 14, se puede observar que, de un total de 92,780 vehículos asegurados en la ciudad de Guayaquil, la mayor participación la tienen los vehículos de tipo Sedán con un 30.62%, seguido por los vehículos de tipo Todoterreno con 24.20%, y Camioneta con un 17.58%. Estos tres tipos de vehículos abarcan más del 70% en el mercado asegurador. Las Motos tienen una considerable participación con un 14.80%. Cabe resaltar que en la categoría **OTRA**, se encuentran agrupados todos los tipos de vehículos que individualmente no son representativos en el conglomerado total. La representación gráfica de la distribución se presenta en la Figura 6.

Tipo	Frecuencia Absoluta	Frecuencia Relativa
SEDÁN	28412	30.62%
TODOTERRENO	22451	24.20%
CAMIONETA	16309	17.58%
MOTO	13733	14.80%
CAMIÓN	6504	7.01%
OTRO	5371	5.79%
Total	**92780**	**100.00%**

Tabla 14: Tipo de Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

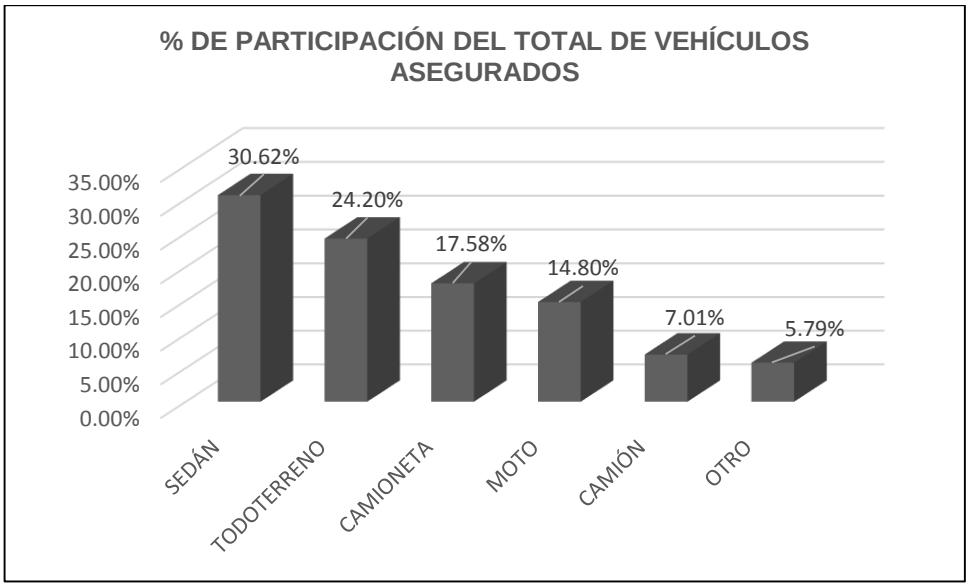


Figura 6: Tipo de Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Color

En la Tabla 15, se puede observar que el color de vehículo que predomina es el Blanco con un 20.91% de participación de un total de 92,780 vehículos asegurados en la ciudad de Guayaquil, seguido por los vehículos de color Plateado con 16.65%, y Negro con 13.80% de participación. Entre estos tres colores de vehículos abarcan más de la mitad con 51.36% del mercado asegurado.

Los colores de vehículos que menos participación tienen son el Celeste y el Amarillo, con 0.77% y 0.62% de participación respectivamente. En la categoría **OTRO**, se encuentran agrupados aquellos colores de vehículos y sus combinaciones que individualmente no son representativos en el conglomerado total. La representación gráfica de la distribución se presenta en la Figura 7.

Color	Frecuencia Absoluta	Frecuencia Relativa
BLANCO	19403	20.91%
PLATEADO	15452	16.65%
NEGRO	12801	13.80%
PLOMO	9842	10.61%
AZUL	8509	9.17%
ROJO	8099	8.73%
BEIGE	3061	3.30%
ACERO	2942	3.17%
GRIS	2651	2.86%
VERDE	2359	2.54%
DORADO	2330	2.51%
CONCHO DE VINO	2100	2.26%
OTRO	1945	2.10%
CELESTE	715	0.77%
AMARILLO	571	0.62%
Total	**92780**	**100.00%**

Tabla 15: Color del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

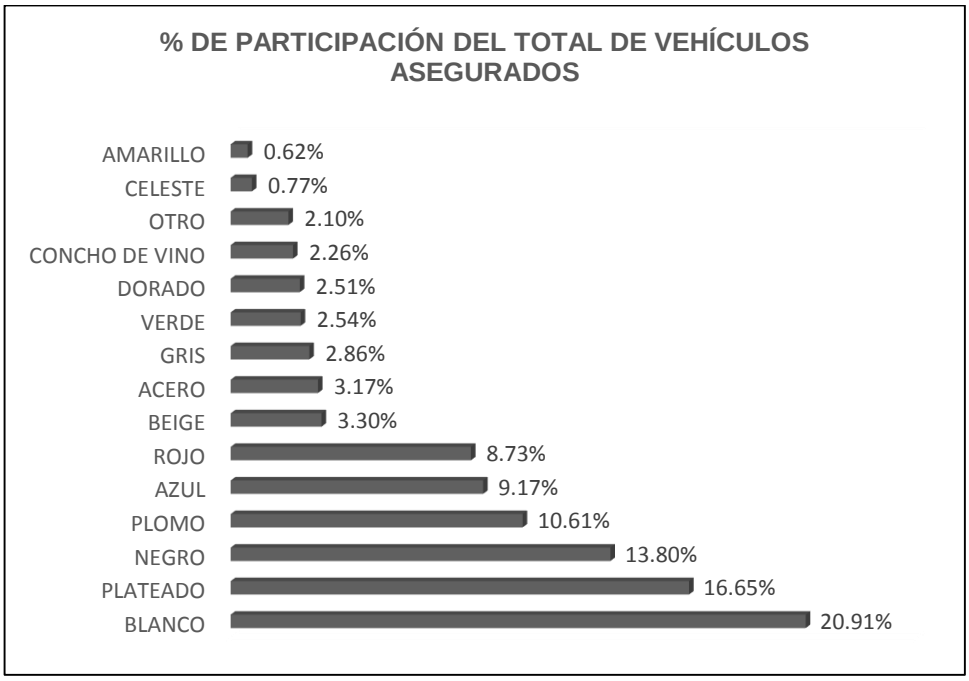


Figura 7: Color del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable: Siniestro

En la Tabla 16, se observa que el 89.12% de un total de 92,780 vehículos asegurados en la ciudad de Guayaquil no sufrieron siniestro, mientras que el 10.88% de vehículos sufrieron siniestro. La representación gráfica de la distribución se presenta en la Figura 8.

Siniestro	Frecuencia Absoluta	Frecuencia Relativa
No hubo Siniestro	82685	89.12%
Hubo Siniestro	10095	10.88%
Total	**92780**	**100%**

Tabla 16: Siniestro del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

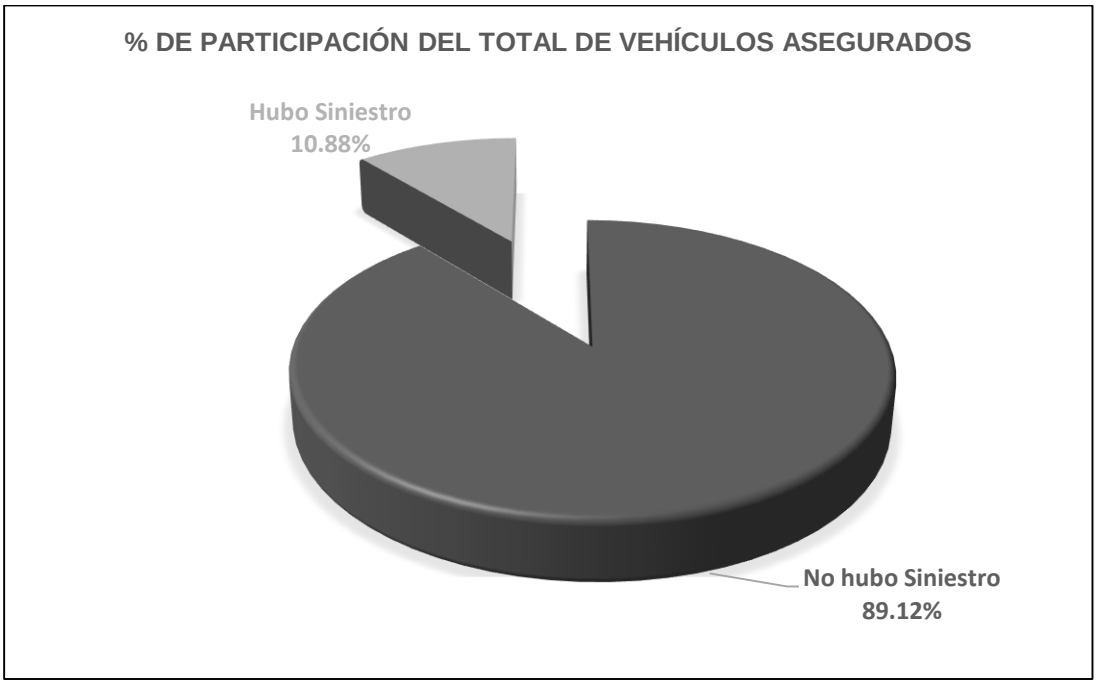


Figura 8: Siniestro del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

Variable Año

En la Tabla 17, podemos notar que los 92,780 vehículos de la muestra asegurados en la ciudad de Guayaquil, tienen al 2008.96 como año promedio de fabricación de estos vehículos. El año que tiene la mayor frecuencia de fabricación es el 2014, la desviación estándar es 5.32 años, la distribución es asimétrica negativa, tiene cola a la izquierda dado que su coeficiente de asimetría es -2.05, lo cual se puede observar en el histograma de la Figura 9. El coeficiente de curtosis es 7.48, lo que significa que la distribución es Leptocúrtica, es decir, tiene una forma más puntiaguda que la distribución normal.

n	92780
Media	2008.964
Mediana	2010
Moda	2014
Desviación estándar	5.3282
Varianza	28.39

Asimetría		-2.057
Curtosis		7.483
Rango		71
Mínimo		1944
Máximo		2015
Percentiles	**25**	2007
	50	2010
	75	2013

Tabla 17: Año de fabricación del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

En la Tabla 17, se presenta además los percentiles 25, 50 y 75, los cuales indican que un 25% de los vehículos han sido fabricados antes del 2007, el 50% de los vehículos fueron fabricados entre los años 2007 y 2010 y un 25% de los vehículos fueron fabricados después del año 2013.

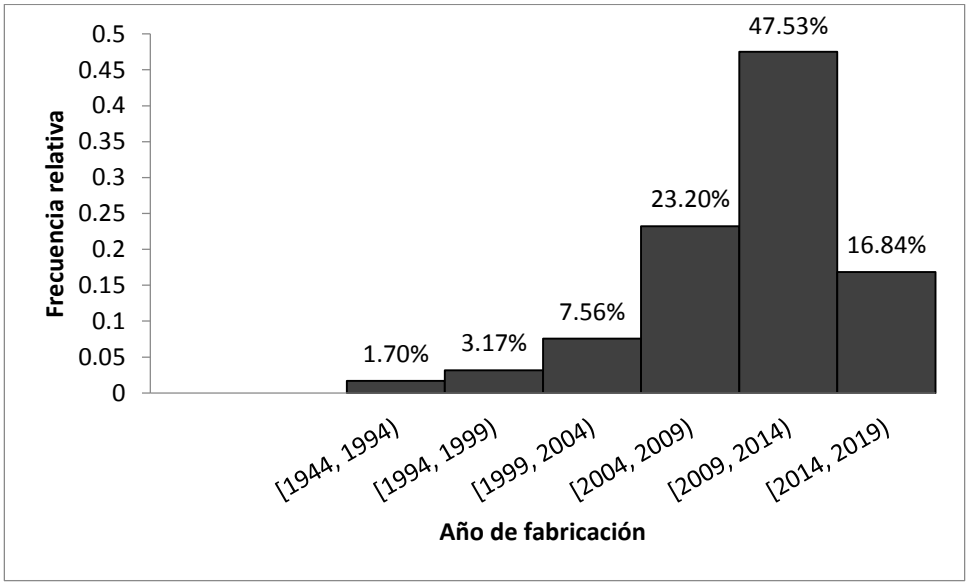


Figura 9: Histograma - Año de fabricación del Vehículo

Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

En la Figura 10 se representan los percentiles en un Diagrama de Cajas, en el cual se puede observar que el 25% de los vehículos que fueron fabricados antes del 2007 están más dispersos que el 25% de los vehículos que han sido fabricados después del 2013.

También, se puede notar que existe una alta concentración de vehículos fabricados entre los años 2007 y 2013.

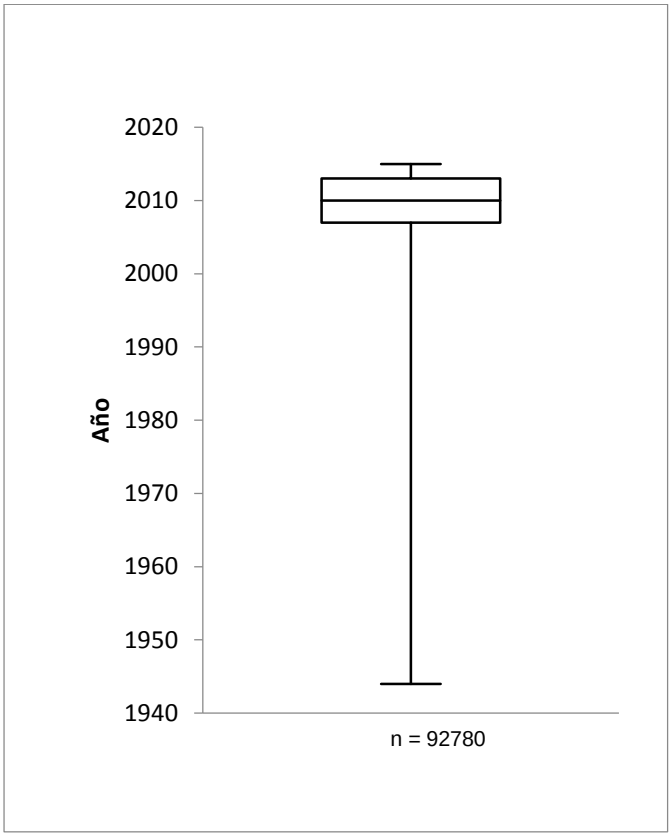


Figura 10: Diagrama de Cajas - Año de fabricación del Vehículo
Fuente: Datos de la investigación - Autor: Guillermo Ordóñez

3.4. ANÁLISIS DE TABLAS DE CONTINGENCIA

El análisis de las tablas de contingencia sirve para determinar si existe algún tipo de relación entre las variables objeto de estudio, en esta investigación se analizará de existir, la relación entre la variable respuesta Siniestro con las variables independientes descritas anteriormente.

Siniestro vs Categoría

Para realizar el análisis se plantea el siguiente contraste de hipótesis:

Ho: El siniestro del vehículo es independiente de la categoría del vehículo

vs.

H1: No es verdad Ho

En la Tabla 18 se presentan los resultados del cruce y del contraste entre estas dos variables, donde se observa que el valor del estadístico de prueba es 382.59, y dado que el valor-p de la prueba es menor a 0.05, lo que significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, el siniestro del vehículo es dependiente de su categoría.

VEHÍCULO		**SINIESTRO**		**Total**
		No hubo Siniestro	**Hubo Siniestro**	
CATEGORIA	**LIVIANO**	74309	9682	83991
	PESADO	8376	413	8789
Total		82685	10095	92780
Resultado Prueba Chi-Cuadrado				
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**	
Chi-cuadrado de Pearson	382.59	1	0	

Tabla 18: Tabla de Contingencia y Prueba Chi-cuadrado – Siniestro vs Categoría

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Siniestro vs Marca

Para realizar el análisis se plantea el siguiente contraste de hipótesis:

Ho: El siniestro del vehículo es independiente de la marca del vehículo

vs.

H1: No es verdad Ho

En la Tabla 19 se presentan los resultados del cruce y del contraste entre estas dos variables, donde se observa que el valor del estadístico de prueba es 1,945.56, y dado que el valor-p de la prueba es menor a 0.05,

lo que significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, el siniestro del vehículo es dependiente de su marca.

VEHÍCULO		SINIESTRO		Total
		No hubo Siniestro	Hubo Siniestro	
MARCA	CHEVROLET	23576	3124	26700
	FORD	3564	514	4078
	GREAT WALL	496	80	576
	HINO	2404	86	2490
	HONDA	1081	104	1185
	HYUNDAI	6425	971	7396
	KIA	4419	879	5298
	MAZDA	2187	376	2563
	MITSUBISHI	1992	205	2197
	NISSAN	4525	864	5389
	RENAULT	1728	402	2130
	SKODA	1297	328	1625
	SUZUKI	10720	532	11252
	TOYOTA	4046	696	4742
	TUKO	2639	5	2644
	VOLKSWAGEN	1524	249	1773
	OTRA	10062	680	10742
Total		82685	10095	92780

Resultado Prueba Chi-Cuadrado			
Estadístico de Prueba	Valor	Grados de Libertad	Valor-p
Chi-cuadrado de Pearson	1,945.56	16	0

Tabla 19: Tabla de Contingencia y Prueba Chi-cuadrado – Siniestro vs Marca

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Siniestro vs Tipo

Para realizar el análisis se plantea el siguiente contraste de hipótesis:

Ho: El siniestro del vehículo es independiente del tipo de vehículo

vs.

H1: No es verdad Ho

En la Tabla 20 se presentan los resultados del cruce y del contraste entre estas dos variables, donde se observa que el valor del estadístico de prueba es 3,024.14, y dado que el valor-p de la prueba es menor a 0.05, lo que significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, el siniestro del vehículo es dependiente de su tipo.

VEHÍCULO		**SINIESTRO**		**Total**
		No hubo Siniestro	**Hubo Siniestro**	
TIPO	**CAMIÓN**	6171	333	6504
	CAMIONETA	14589	1720	16309
	MOTO	13630	103	13733
	SEDÁN	23659	4753	28412
	TODOTERRENO	19505	2946	22451
	OTRO	5131	240	5371
Total		82685	10095	92780
Resultado Prueba Chi-Cuadrado				
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**	
Chi-cuadrado de Pearson	3,024.14	5	0	

Tabla 20: Tabla de Contingencia y Prueba Chi-cuadrado – Siniestro vs Tipo

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Siniestro vs Color

Para realizar el análisis se plantea el siguiente contraste de hipótesis:

Ho: El siniestro del vehículo es independiente del color del vehículo

vs.

H1: No es verdad Ho

En la Tabla 21 se presentan los resultados del cruce y del contraste entre estas dos variables, donde se observa que el valor del estadístico de prueba es 878.98, y dado que el valor-p de la prueba es menor a 0.05, lo que significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, el siniestro del vehículo es dependiente de su color.

VEHÍCULO		SINIESTRO		Total
		No hubo Siniestro	Hubo Siniestro	
COLOR	ACERO	2922	20	2942
	AMARILLO	528	43	571
	AZUL	7717	792	8509
	BEIGE	2526	535	3061
	BLANCO	17626	1777	19403
	CELESTE	598	117	715
	CONCHO DE VINO	1803	297	2100
	DORADO	1957	373	2330
	GRIS	2262	389	2651
	NEGRO	11507	1294	12801
	PLATEADO	13511	1941	15452
	PLOMO	8493	1349	9842
	ROJO	7411	688	8099
	VERDE	2066	293	2359
	OTRO	1758	187	1945
Total		82685	10095	92780
Resultado Prueba Chi-Cuadrado				

Estadístico de Prueba	Valor	Grados de Libertad	Valor-p
Chi-cuadrado de Pearson	878.98	14	0

Tabla 21: Tabla de Contingencia y Prueba Chi-cuadrado – Siniestro vs Color

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Siniestro vs Año

Para realizar el análisis se plantea el siguiente contraste de hipótesis:

Ho: El siniestro del vehículo es independiente del año de fabricación del vehículo

vs.

H1: No es verdad Ho

En la Tabla 22 se presentan los resultados del cruce y del contraste entre estas dos variables, donde se observa que el valor del estadístico de prueba es 584.29, y dado que el valor-p de la prueba es menor a 0.05, lo que significa que existe evidencia estadística para rechazar la hipótesis nula.

Por lo tanto, el siniestro del vehículo es dependiente del año de fabricación.

VEHÍCULO		SINIESTRO		Total
		No hubo Siniestro	Hubo Siniestro	
Años agrupados por cuartil	1	19634	1789	21423
	2	23249	3306	26555
	3	17052	2853	19905
	4	22750	2147	24897
Total		82685	10095	92780

Resultado Prueba Chi-Cuadrado			
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**
Chi-cuadrado de Pearson	584.29	3	0

Tabla 22: Tabla de Contingencia y Prueba Chi-cuadrado – Siniestro vs Año

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

CAPÍTULO IV

4. ANÁLISIS ESTADÍSTICO MULTIVARIADO

4.1. INTRODUCCIÓN

Este capítulo abarca el desarrollo de la última parte del análisis estadístico, que se comprende de las técnicas estadísticas multivariadas como lo son el Análisis de Correspondencias múltiples y el Análisis de Regresión Logística.

En este estudio, se encontró que las variables Marca, Color y Tipo tienen muchas categorías, debido a esto y con la finalidad de favorecer la interpretación de los resultados, estas variables serán redefinidas, creando nuevas categorías que agrupen a las categorías originales de acuerdo a su peso, es decir, a la frecuencia de siniestralidad que presentaron estas categorías.

En el caso de las variables Categoría y Siniestro como sólo tienen dos categorías se las analizará sin realizar cambio alguno.

Por último, para la variable Año, como se mencionó en el capítulo 3 será recodificada, agrupando los años de fabricación de acuerdo al percentil correspondiente, ya que para los análisis multivariados que aplicaremos se requiere que esta variable sea cualitativa, por lo cual se la renombrará a Año según percentil.
Los análisis que se desarrollarán se procesarán usando el software estadístico IBM SPSS versión 22.

A continuación, se presentan en las tablas 23, 24 y 25, la recategorización y recodificación de las variables descritas anteriormente.

MARCA					
Codificación	**Siniestralidad**	**Intervalo**	**Marca**	**Frecuencia**	**Frecuencia relativa**
GM1	ALTA	f >=1000	CHEVROLET	3124	30.95%
GM2	MEDIA	600<= f <1000	HYUNDAI	971	9.62%
			KIA	879	8.71%
			NISSAN	864	8.56%
			TOYOTA	696	6.89%
			OTRA	680	6.74%
GM3	BAJA	300<= f <600	SUZUKI	532	5.27%
			FORD	514	5.09%
			RENAULT	402	3.98%
			MAZDA	376	3.72%
			SKODA	328	3.25%
GM4	MUY BAJA	f<300	VOLKSWAGEN	249	2.47%
			MITSUBISHI	205	2.03%
			HONDA	104	1.03%
			HINO	86	0.85%
			GREAT WALL	80	0.79%
			TUKO	5	0.05%

Tabla 23: Recategorización de la variable Marca

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

COLOR					
Codificación	**Siniestralidad**	**Intervalo**	**Marca**	**Frecuencia**	**Frecuencia relativa**
GC1	ALTA	f >=1000	PLATEADO	1941	19.23%
			BLANCO	1777	17.60%
			PLOMO	1349	13.36%
			NEGRO	1294	12.82%
GC2	MEDIA	600<= f <1000	AZUL	792	7.85%
			ROJO	688	6.82%
GC3	BAJA	200<= f <600	BEIGE	535	5.30%
			GRIS	389	3.85%
			DORADO	373	3.69%
			CONCHO DE VINO	297	2.94%
			VERDE	293	2.90%

GC4	MUY BAJA	f<200	OTRO	187	1.85%
			CELESTE	117	1.16%
			AMARILLO	43	0.43%
			ACERO	20	0.20%

Tabla 24: Recategorización de la variable Color

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

TIPO					
Codificación	**Siniestralidad**	**Intervalo**	**TIPO**	**Frecuencia**	**Frecuencia relativa**
GT1	ALTA	f>=3000	SEDÁN	4753	47.08%
GT2	MEDIA	1000<=f<3000	TODOTERRENO	2946	29.18%
			CAMIONETA	1720	17.04%
GT3	BAJA	f<1000	CAMIÓN	333	3.30%
			OTRO	240	2.38%
			MOTO	103	1.02%

Tabla 25: Recategorización de la variable Tipo

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

4.2. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

El análisis que se describe en el presente capitulo, tiene como objetivo explicar el problema investigado en base a los resultados obtenidos, a través de la reducción del conjunto de variables a sólo dos dimensiones.

Dimensión	**Total (autovalor)**	**Inercia**	**% de varianza**
1	1.991	0.332	33.188
2	1.518	0.253	25.308
Total	3.510	0.585	
Media	1.755	0.292	29.248

Tabla 26: Resumen del modelo

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

La Tabla 26 presenta el resumen obtenido del modelo, donde se puede identificar el autovalor, inercia y porcentaje de varianza para cada dimensión.

De los resultados obtenidos en esta tabla, se destaca la inercia, que indica la proporción o porcentaje de la varianza de los datos explicada por cada una de las dimensiones. Se puede observar en la columna de inercia que el 58.5% de la variabilidad total de los datos es explicada por las dimensiones incluidas en el modelo, donde la primera dimensión explica el 33.2% de la variabilidad de los datos y el 25.3% de la variabilidad restante es explicada por la segunda dimensión.

La Tabla 27 presenta las medidas discriminantes por variable, donde se puede identificar las puntuaciones que tiene cada variable en la dimensión respectiva, las cuales determinan en cuanto discrimina cada variable en cada una de las dimensiones, es decir, permiten distinguir la importancia que tiene cada variable de acuerdo a la dimensión obtenida.

Variables	**Dimensión**		**Media**
	1	**2**	
CATEGORIA	0.369	0.300	0.335
MARCA	0.348	**0.531**	0.439
TIPO	**0.763**	0.000	0.382
COLOR	0.239	**0.332**	0.286
SINIESTRO	0.077	0.000	0.039
AÑO SEGÚN PERCENTIL	0.194	**0.354**	0.274
Total activo	1,991	1,518	1,755
% de varianza	33,188	25,308	29,248

Tabla 27: Medidas discriminantes por variable

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

De los resultados obtenidos en esta tabla, se puede observar que las discriminaciones se dan de la siguiente manera: la variable Tipo de vehículo están mayormente relacionadas con la dimensión 1, mientras que la Marca, el Color y el año según percentil del vehículo tienen más relación con la segunda dimensión.

En la Figura 11, se representa mediante un gráfico de dos dimensiones las medidas discriminantes que se han descrito, donde podemos notar

que se forman dos grupos de variables de acuerdo a la proximidad entre ellas. En el primer grupo se visualiza solamente a la variable Tipo de vehículo, dado que marca una lejanía con respecto al resto, mientras que en el segundo grupo se aglomeran las variables Categoría, Año según percentil, Color y Marca del vehículo, las cuales se distinguen por la relación de cercanía que tienen entre ellas en el gráfico.

Por último, se tiene a la variable Siniestro, que presenta puntuaciones muy bajas en las dos dimensiones, y como se puede observar en el grafico está muy cercana al origen, debido a esto, esta variable no es muy explicativa dentro del conjunto de variables bajo estudio, lo que significa que no se encuentra relacionada con ninguno de los dos grupos de variables.

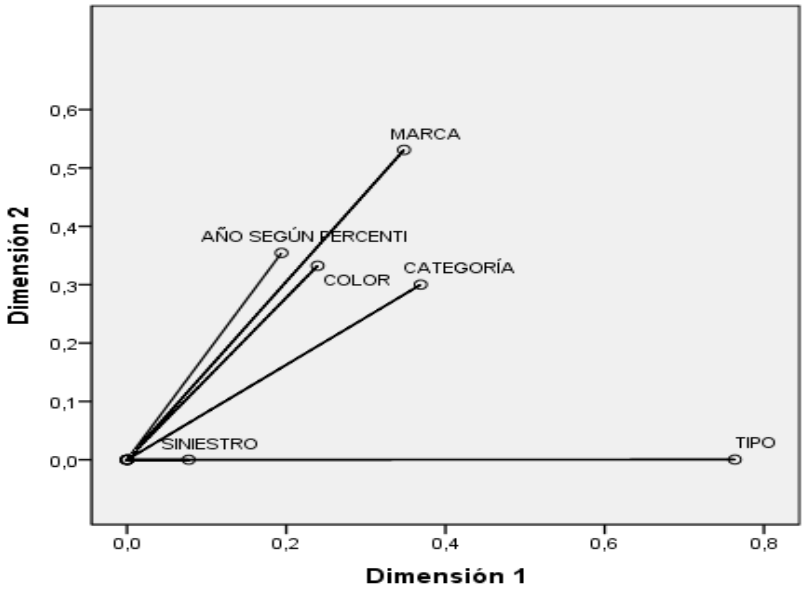


Figura 11: Medidas discriminantes por variable

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

La Figura 12 presenta el diagrama conjunto de puntos de categorías de las variables de estudio, donde podemos identificar correspondencias entre las distintas categorías existentes. Es decir, permite distinguir patrones de acuerdo a la proximidad y las relaciones que se dan entre las categorías.

En este gráfico, se pueden observar las asociaciones que presentan las diferentes categorías del conjunto de variables de estudio con las categorías de la variable siniestro, cuyo análisis es el siguiente:

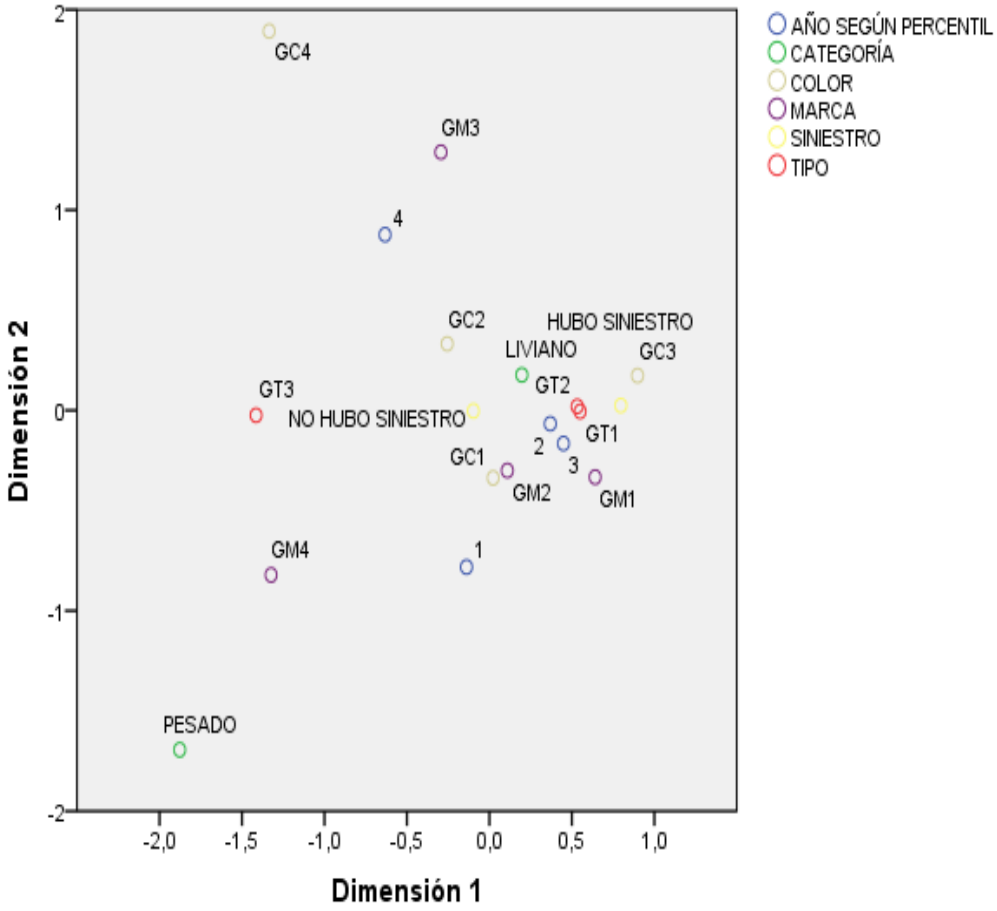


Figura 12: Diagrama conjunto de puntos de categorías

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Para las variables AÑO SEGÚN PERCENTIL y SINIESTRO, notamos la siguiente relación entre las categorías: los vehículos fabricados a partir del año 2013, PERCENTIL 3, están más relacionados con la existencia de siniestralidad, categoría HUBO SINIESTRO, en tanto que aquellos vehículos fabricados entre los años 2007 y 2010, PERCENTIL 2, tienden a no sufrir siniestros, categoría NO HUBO SINIESTRO.

Para el caso de la variable CATEGORÍA, se tiene que, los vehículos LIVIANOS están más asociados con la categoría NO HUBO SINIESTRO, aunque también se observa una relación con la categoría HUBO SINIESTRO debido a su proximidad, mientras que los vehículos PESADOS no tienen relación alguna con las categorías de la variable SINIESTRO.

Para la variable COLOR, se observa que los vehículos de color, en la categoría GC3, BEIGE, GRIS, DORADO, CONCHO DE VINO y VERDE, están más relacionados con la siniestralidad, mientras que los vehículos de color, categoría GC1 y GC2, PLATEADO, BLANCO, PLOMO, NEGRO, AZUL y ROJO, tienden a no sufrir siniestros.

Para la variable MARCA, podemos notar que los vehículos de marca (categoría GM1), CHEVROLET, están más propensos a sufrir siniestros, y los vehículos de marca (categoría GM2) HYUNDAI, KIA, NISSAN, TOYOTA y OTRA, están mas asociados a no sufrir siniestros.

En cuanto al TIPO, se puede observar que los vehículos de tipo (categoría GT1 y GT2) SEDÁN, TODOTERRENO y CAMIONETA, están más propensos a sufrir siniestros, aunque también se nota cierta relación con la categoría NO HUBO SINIESTRO debido a su cercanía en el gráfico.

En términos generales de acuerdo a las relaciones encontradas podemos decir que: los vehículos LIVIANOS, de tipo SEDÁN, TODOTERRENO y CAMIONETA, de marca CHEVROLET, con los colores BEIGE, GRIS, DORADO, CONCHO DE VINO y VERDE, fabricados recientemente, a partir del año 2013, tienen más propensión a sufrir siniestros; mientras que los vehículos LIVIANOS, de marca HYUNDAI, KIA, NISSAN, TOYOTA y OTRA, con los colores PLATEADO, BLANCO, PLOMO, NEGRO, AZUL y ROJO, fabricados entre los años 2007 y 2010, son los menos propensos a sufrir un siniestro.

4.3. ANÁLISIS DE REGRESIÓN LOGÍSTICA

En este apartado, se pretende hallar un modelo de regresión logística que describa la relación entre el conjunto de variables independientes y la variable dependiente que son motivo de estudio en esta investigación, cuyo análisis tiene como objetivo identificar cuáles son los factores que influyen significativamente en la probabilidad de que un vehículo sufra un siniestro.

Para encontrar un modelo de regresión logística óptimo, se debe proceder a realizar varias pruebas utilizando el proceso de introducción manual de variables a SPSS, en primer lugar se introdujo todas las variables al modelo y luego se fue eliminando de manera sucesiva aquellas variables que no resultaron significativas, además se probó con interacciones entre las variables, este procedimiento se repitió hasta que se obtuvo un conjunto de variables significativas, de esta forma se logró definir el modelo que mejor se ajusta a los datos observados. La descripción y análisis de este modelo se realizará en base a los resultados obtenidos por la introducción de las variables que finalmente lo integran como son Color, Marca y Tipo.

Para poder emplear el análisis de Regresión Logística, se procedió a realizar la recodificación de las tres variables debido a que son categóricas, creando variables DUMMY cuyo número es igual al número de categorías de la variable original pero disminuida en uno. Por ejemplo, para la variable MARCA como tiene cuatro categorías GM1, GM2, GM3 y GM4, se crearon tres variables DUMMY, MARCA(1), MARCA(2) y MARCA(3), con la finalidad de tener una categoría de referencia, que ayude en la interpretación de los resultados. En la Tabla 28 se presentan las codificaciones de estas variables categóricas.

Variables originales y sus categorías		Codificación de variables DUMMY		
		COLOR(1)	COLOR(2)	COLOR(3)
COLOR	GC1	1	0	0
	GC2	0	1	0
	GC3	0	0	1
	GC4	0	0	0
		MARCA(1)	MARCA(2)	MARCA(3)
MARCA	GM1	1	0	0
	GM2	0	1	0
	GM3	0	0	1
	GM4	0	0	0
		TIPO(1)	TIPO(2)	
TIPO	GT1	1	0	
	GT2	0	1	
	GT3	0	0	

Tabla 28: Codificaciones de variables categóricas

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Se puede notar en la Tabla 28, que la última categoría de cada variable es la categoría de referencia, es decir, no tiene representación en variable DUMMY, por ejemplo, para la variable MARCA que tiene cuatro categorías, GM1, GM2, GM3 y GM4, si un vehículo pertenece al grupo GM1, entonces la variable DUMMY MARCA(1) tomará el valor de 1, caso contrario tomará el valor de 0; si un vehículo pertenece al grupo GM2, entonces la variable DUMMY MARCA(2) tomará el valor de 1, caso contrario tomará el valor de 0; y si un vehículo pertenece al grupo GM3, entonces la variable DUMMY MARCA(3) tomará el valor de 1, caso contrario tomará el valor de 0, quedando el grupo GM4 como la categoría de referencia.

Luego de haber realizado la codificación de las variables categóricas, se procederá a determinar la significación global del modelo con estas variables, la cual se basa en el valor del estadístico de prueba G, con el propósito de contrastar la significancia de los coeficientes del modelo de regresión logística.

Las hipótesis para probar la significación global son las siguientes:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$

H_1 : Al menos algún $\beta_i \neq 0, \forall i = 1,2,\cdots,n$

Que expresado en palabras equivale a:

H_0: Todos los coeficientes de las variables independientes son iguales a cero.

H_1: Al menos uno de los coeficientes de las variables independientes es diferente de cero.

Para realizar este contraste de hipótesis se usa la prueba de Omnibus que provee SPSS, del cual se obtuvieron los siguientes resultados: El valor del estadístico de prueba G es 3547.17, con 8 grados de libertad y el valor-p de la prueba es igual a 0 (menor a 0.05), lo que significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, al menos uno de los coeficientes de las variables independientes es distinto de cero.

Dado los resultados de la prueba de Omnibus, se sabe que al menos uno de los coeficientes de las variables independientes presentes en el modelo, es distinto de cero, por consiguiente, se debe identificar cuáles son las variables que tienen coeficientes que son estadísticamente significativos a un nivel de significancia de 0.05. Para esto, se utiliza una prueba de significancia individual para cada uno de los coeficientes de las variables independientes, esta prueba se basa en el estadístico W de Wald, por lo que se plantea el siguiente contraste de hipótesis.

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0, \forall i = 1,2,\cdots,n$

Que expresado en palabras equivale a:

H_0 : El coeficiente de la variable independiente es igual a cero.

H_1 : El coeficiente de la variable independiente es diferente de cero.

Los resultados obtenidos al efectuar la prueba de hipótesis a cada uno de los coeficientes de las variables independientes, se pueden visualizar en la Tabla 29, donde se puede afirmar que existe evidencia estadística para no rechazar la hipótesis nula por medio del estadístico de Wald, se puede observar para los coeficientes de las variables DUMMY, MARCA(2), MARCA(3) y COLOR(2), valores p asociados (p>0.05), lo que significa que estos coeficientes no son estadísticamente significativos.

En cambio, los coeficientes de las variables DUMMY, MARCA(1), TIPO(1), TIPO(2), COLOR(1) y COLOR(3), de acuerdo a sus estadísticos de prueba y valores-p respectivos (p<0.05), se rechaza la hipótesis nula del contraste de hipótesis planteado, lo cual significa que estos coeficientes son estadísticamente significativos.

Factores	Coeficientes de las variables (β)	Error estándar	Estadístico de Wald (W)	Grados de Libertad	Valor p	e^{β} (Odds Ratio)	Intervalo de Confianza (95%) para el Odds Ratio (e^{β})	
							Límite Inferior	Límite Superior
Constante	-3.763	0.071	2841.408	1	0	0.023		
MARCA(1)	-0.121	0.045	7.217	1	0.007	0.886	0.811	0.968
MARCA(2)	0.082	0.044	3.551	1	0.059	1.086	0.997	1.183
MARCA(3)	0.083	0.046	3.187	1	0.074	1.086	0.992	1.189
TIPO(1)	1.99	0.044	2037.355	1	0	7.318	6.712	7.978
TIPO(2)	1.607	0.044	1329.643	1	0	4.986	4.573	5.436
COLOR(1)	0.168	0.058	8.411	1	0.004	1.183	1.056	1.325
COLOR(2)	0.025	0.062	0.16	1	0.689	1.025	0.907	1.158
COLOR(3)	0.273	0.062	19.423	1	0	1.314	1.164	1.483

Tabla 29: Variables en el Modelo de Regresión Logística

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

Dado que un modelo de regresión logística debe estar formado sólo con variables que sean estadísticamente significativas, de acuerdo a los resultados descritos de la Tabla 29, se deben eliminar las variables DUMMY MARCA(2), MARCA(3) y COLOR(2) ya que sus coeficientes no son estadísticamente significativos y por lo tanto no aportan globalmente a la solución del modelo. En base a esto, en la siguiente Tabla 4.8, se presentan los resultados obtenidos.

Factores	**Coeficientes de las variables (β)**	**Error estándar**	**Estadístico de Wald (W)**	**Grados de Libertad**	**Valor p**	**e^{β} (Odds Ratio)**	**Intervalo de Confianza (95%) para el Odds Ratio (e^{β})**	
							Límite Inferior	**Límite Superior**
Constante	-3.763	0.071	2841.408	1	0	0.023		
MARCA(1)	-0.121	0.045	7.217	1	0.007	0.886	0.811	0.968
TIPO(1)	1.99	0.044	2037.355	1	0	7.318	6.712	7.978
TIPO(2)	1.607	0.044	1329.643	1	0	4.986	4.573	5.436
COLOR(1)	0.168	0.058	8.411	1	0.004	1.183	1.056	1.325
COLOR(3)	0.273	0.062	19.423	1	0	1.314	1.164	1.483

Tabla 30: Variables que integran finalmente el Modelo de Regresión Logística

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

En la Tabla 30, se puede notar que los coeficientes de las variables DUMMY, MARCA(1), TIPO(1), TIPO(2), COLOR(1) y COLOR(3), son significativos ya que sus respectivos valores-p son menores a 0.05 ($p<0.05$), por lo tanto, estas variables son las que finalmente componen el modelo.

Para evaluar la bondad del ajuste del modelo obtenido, se realiza la prueba de Hosmer-Lemeshow, la cual plantea el siguiente contraste de hipótesis:

H_0: El modelo seleccionado ajusta bien a los datos

H1: $\neg H_0$

El valor del estadístico de prueba chi-cuadrado es 12.16, con 8 grados de libertad y el valor-p de la prueba es igual 0.144 (mayor a 0.05), lo que significa, que existe evidencia estadística para no rechazar la hipótesis nula. Por lo tanto, el modelo seleccionado ajusta bien a los datos observados.

Otra forma que se usa para evaluar la bondad de ajuste del modelo, fue mediante la tasa global de aciertos que se obtiene de la tabla de clasificación de resultados que permite examinar la capacidad predictiva del modelo seleccionado, la cual plantea el siguiente contraste de hipótesis:

H_0 :Número de casos correctamente clasificados por el modelo no difiere de la clasificación esperada sólo por efecto del azar.

H1: $\neg H_0$

Los resultados de la clasificación se pueden visualizar en la Tabla 31, donde el porcentaje global de aciertos fue de 77.8%, además se presenta el test de Huberty que permite verificar la significación estadística de esta tasa y de esta manera contrastar las hipótesis anteriores. Este test permite obtener el número esperado de casos correctamente clasificados debidos al azar (e) que es igual a 74786.79, a partir del cual se calcula el valor del estadístico Z^* que se distribuye normalmente, cuyo valor es igual a 21.53. Si se compara este valor con el valor de 1.96 obtenido de la tabla de una distribución normal para un nivel de significancia de 0.05, se tiene que el valor del estadístico Z^* = 21.53 > 1.96, por lo que, se puede afirmar que existe evidencia estadística para rechazar la hipótesis nula, y se concluye que la tasa global de aciertos del modelo es significativamente mayor que la que se hubiera obtenido debido al azar. En otras palabras, el modelo seleccionado clasifica adecuadamente a los datos observados.

Observado		Pronosticado		
		SINIESTRO		Corrección de porcentaje
		NO	SÍ	
SINIESTRO	NO	69333	13352	83.9
	SÍ	7234	2861	28.3
Porcentaje global				77.8
$e = 74786.79$			$Z^* = 21.53$	

Tabla 31: Tabla de clasificación de resultados

Fuente: Datos de la investigación. Autor: Guillermo Ordóñez

En base a lo analizado de los resultados obtenidos, se determina que el modelo de regresión logística que permite estimar la probabilidad de que un vehículo sufra un siniestro para valores dados de las variables independientes, es:

$$p = \frac{e^{-3.763 - 0.121x_1 + 1.99x_2 + 1.607x_3 + 0.168x_4 + 0.273x_5}}{1 + e^{-3.763 - 0.121x_1 + 1.99x_2 + 1.607x_3 + 0.168x_4 + 0.273x_5}}$$

Que en forma resumida se puede expresar como:

$$p = \frac{1}{1 + e^{-3.763 - 0.121x_1 + 1.99x_2 + 1.607x_3 + 0.168x_4 + 0.273x_5}}$$

Donde:

X1: MARCA(1)

X2: TIPO(1)

X3: TIPO(2)

X4: COLOR(1)

X5: COLOR(3)

Una vez que se ha obtenido un modelo de regresión logística que ajusta a los datos observados, se realizará su interpretación a partir de las categorías cuyas variables DUMMY tienen el valor de 1. En esta investigación, las categorías cuyas variables DUMMY tienen el valor de 1 son las siguientes: GM1, GT1, GT2, GC1 y GC3, que corresponden a las variables que resultaron significativas en el modelo.

Se debe recordar que cada una de estas categorías agrupa a diferentes marcas, tipos de vehículo y colores respectivamente, en función de esto se puede interpretar lo siguiente:

La marca CHEVROLET (GM1), en un vehículo, constituye un factor de riesgo para la ocurrencia de un siniestro.

El que un vehículo pertenezca al tipo SEDÁN, TODOTERRENO o CAMIONETA constituye un factor de riesgo para la ocurrencia de un siniestro.
Un vehículo que sea de color PLATEADO, BLANCO, PLOMO, NEGRO, BEIGE, GRIS, DORADO, CONCHO DE VINO o VERDE, constituye un factor de riesgo para la ocurrencia de un siniestro.

En la tabla 30 se presentan los valores del Odds ratio, tal que, si un valor Odd ratio es mayor a 1, significa que es más probable que ocurra el evento de interés, HUBO SINIESTRO, en relación a la no ocurrencia de dicho evento; en caso contrario, si este valor es menor a 1, significa que es menos probable que ocurra el evento de interés, HUBO SINIESTRO. En base a esto se puede interpretar que:

Si un vehículo es de marca CHEVROLET (GM1), entonces es menos probable que sufra un siniestro a que si es de las marcas VOLKSWAGEN, MITSUBISHI, HONDA, HINO, GREAT WALL, TUKO (GM4, que es la categoría de referencia).

Si un vehículo es del tipo SEDÁN (GT1), entonces es más probable que sufra un siniestro a que sea del tipo CAMIÓN, MOTO u OTRO (GT3, que es la categoría de referencia).

Si un vehículo es del tipo TODOTERRENO o CAMIONETA (GT2), entonces es más probable que sufra un siniestro a que sea del tipo CAMIÓN, MOTO u OTRO (GT3, que es la categoría de referencia).

Si un vehículo es de color PLATEADO, BLANCO, PLOMO o NEGRO (GC1), entonces es más probable que sufra un siniestro a que sea de color CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

Si un vehículo es de color BEIGE, GRIS, DORADO, CONCHO DE VINO o VERDE (GC3), entonces es más probable que sufra un siniestro a que sea de color CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. De acuerdo a los resultados alcanzados en este proyecto sobre la siniestralidad en el ramo de seguros de vehículos en la ciudad de Guayaquil, se visualiza la necesidad que las compañías aseguradoras realicen estudios sobre este y los demás ramos que ofrecen, dado la carencia de los mismos, ya que en las pólizas recogen información valiosa tanto del cliente como del vehículo, pero no cuentan con una herramienta que les ayude a discriminar el riesgo de siniestralidad al brindar una cobertura, por lo que esta investigación se centró en proporcionar conocimiento a través del análisis de modelos estadísticos multivariados que permiten medir el riesgo asociado a las características de los vehículos, así como también las relaciones que existen entre las variables seleccionadas, lo cual genera un enorme beneficio para las aseguradoras ya que pueden disponer de esta valiosa información para entre otros aspectos, por ejemplo, decidir el valor a cobrar por la prima.
2. La ciudad, sería otro factor importante a analizar, ya que las estadísticas de accidentes de tránsito se concentran mayormente en las grandes ciudades como Guayaquil, por lo tanto, incorporar esta variable beneficiaría más a una investigación.
3. De los resultados obtenidos en el modelo de Regresión Logística, se deduce que los vehículos de marca CHEVROLET, del tipo SEDÁN, TODOTERRENO o CAMIONETA, y de color PLATEADO, BLANCO, PLOMO, NEGRO, BEIGE, GRIS, DORADO, CONCHO DE VINO o VERDE, constituyen un factor de riesgo para la ocurrencia de un siniestro. Sin embargo, aunque los vehículos de marca CHEVROLET, que es la más comercial, es menos probable que sufran siniestros a que sean de las marcas VOLKSWAGEN, MITSUBISHI, HONDA, GREAT WALL o TUKO.
4. De acuerdo al análisis de Correspondencias Múltiples, los vehículos LIVIANOS, de tipo SEDÁN, TODOTERRENO y CAMIONETA, de marca CHEVROLET, con los colores BEIGE, GRIS, DORADO, CONCHO DE VINO

y VERDE, fabricados recientemente, a partir del año 2013, tienen más propensión a sufrir siniestros.

5. De acuerdo al análisis de Correspondencias Múltiples, los vehículos LIVIANOS, de marca HYUNDAI, KIA, NISSAN, TOYOTA y OTRA, con los colores PLATEADO, BLANCO, PLOMO, NEGRO, AZUL y ROJO, fabricados entre los años 2007 y 2010, son los menos propensos a sufrir un siniestro.
6. Los vehículos del tipo SEDÁN, tienen más probabilidad de sufrir un siniestro a que sean del tipo CAMIÓN, MOTO u OTRO.
7. Los vehículos del tipo TODOTERRENO o CAMIONETA, tienen más probabilidad de sufrir un siniestro a que sean del tipo CAMIÓN, MOTO u OTRO.
8. Los vehículos de color PLATEADO, BLANCO, PLOMO, NEGRO, BEIGE, GRIS, DORADO, CONCHO DE VINO o VERDE, tienen más probabilidad de sufrir un siniestro a que sean de color CELESTE, AMARILLO, ACERO u OTRO.
9. Comparando los resultados obtenidos en los dos modelos de análisis multivariado utilizados en nuestro estudio, se encontró que ambas técnicas identificaron de manera similar los mismos factores en relación a la presencia de siniestralidad, sin embargo, en el análisis de regresión logística, el modelo obtenido identificó otros colores de vehículos que no fueron asociados en el análisis de correspondencias múltiples.

RECOMENDACIONES

1. Se recomienda a las compañías de seguros que proporcionen toda la información requerida por los investigadores, dado que de esta forma se podrían identificar otros posibles factores que puedan influir en la siniestralidad de un vehículo.
2. Seria de interés realizar un estudio que permita incorporar los datos relacionados con el perfil del cliente que asegura su vehículo, como edad, sexo, nivel de ingresos, nivel de instrucción, tipo de licencia, etc., de tal manera que permita identificar si estas características influyen en la siniestralidad.
3. En base a las estadísticas de accidentes de tránsito en el Ecuador, un factor importante de analizar seria la ciudad de residencia del asegurado, para verificar la influencia que tendría esta variable en la siniestralidad de vehículos, analizando las posibles relaciones conjuntas que existan entre el perfil del cliente, las características de los vehículos asegurados y la ciudad de origen, de tal manera que las aseguradoras puedan tener un mejor criterio para brindar cobertura a sus clientes.

BIBLIOGRAFÍA

[1] *Lesiones causadas por el tránsito*. (2016). *Organización Mundial de la Salud*. Recuperado el 10 de enero de 2016, de http://www.who.int/mediacentre/factsheets/fs358/es/

[2] *Acerca de la Estrategia - Banco Interamericano de Desarrollo.* (2016). *Banco Interamericano de Desarrollo*. Recuperado el 15 de enero de 2016, de http://www.iadb.org/es/temas/transporte/acerca-de-la-estrategia,6726.html

[3] *Descargables - Siniestros Diciembre 2015 - Agencia Nacional de Tránsito del Ecuador - ANT. (2016). Ant.gob.ec. Recuperado el 15 de enero de 2016, de* http://www.ant.gob.ec/index.php/descargable/file/3368-siniestros-diciembre-2015

[4] *Descargables - Lesionados Diciembre 2015 - Agencia Nacional de Tránsito del Ecuador - ANT. (2016). Ant.gob.ec. Recuperado el 15 de enero de 2016, de* http://www.ant.gob.ec/index.php/descargable/file/3369-lesionados-diciembre-2015

[5] *Descargables - Fallecidos Diciembre 2015 - Agencia Nacional de Tránsito del Ecuador - ANT. (2016). Ant.gob.ec. Recuperado el 15 de enero de 2016, de* http://www.ant.gob.ec/index.php/descargable/file/3367-fallecidos-diciembre-2015

[6] Chen H, Libo Caoa L, Logan D. (2012). *Analysis of Risk Factors Affecting the Severity of Intersection Crashes by Logistic Regression*. Traffic Injury Prevention. Volume 13, Issue 3, pages 300-307.

[7] Hemrit W, Ben Arab M, Raissi N. (2013). *The correspondence analysis between the key indicators and events of operational risk: a case study of the*

insurance sector in Tunisia. International Journal of Risk Assessment and Management. Volume 17.

[8] Quishpe, I. (2015). *Factores de riesgos de siniestralidad y cálculo de primas de los vehículos asegurados en el Ecuador mediante modelos lineales generalizados (proyecto de graduación de pregrado).* Escuela Politécnica Nacional, Quito, Ecuador.

[9] WALPOLE, R.E., MYERS, R.H., MYERS, S.L. Y YE, K. (2012). *Probabilidad y estadística para ingeniería y ciencias (Novena edición).* Ciudad de México, México: Pearson Educación de México, S.A. de C.V.

[10] TRIOLA, M. F. (2009). *Estadística. Décima edición*. Ciudad de México, México: Pearson Educación de México, S.A. de C.V.

[11] MENDENHALL, W., BEAVER, R.J., Y BEAVER, B.M. (2010*). Introducción a la probabilidad y estadística. Décimo tercera edición*. Ciudad de México, México: Cengage Learning Editores, S.A. de C.V.

[12] Véliz, C. (2011). *Estadística para la administración y los negocios. Primera Edición.* Ciudad de México, México: Pearson Educación de México, S.A. de C.V.

[13] Anderson, D. R., Sweeney, D. J. y Williams, T. A. (2008). *Estadística para administración y economía, Décima edición.* Ciudad de México, México: Cengage Learning Editores, S.A. de C.V.

[14] Johnson, R. A. (2012). *Probabilidad y estadística para ingenieros, Octava edición.* Ciudad de México, México: Pearson Educación, S.A. de C.V.

[15] Alvarado, J. A. y Obagi, J. J. (2008). *Fundamentos de Inferencia Estadística, Primera edición.* Bogotá, Colombia: Pontificia Universidad Javeriana.

[16] Rodríguez, L.E. (2007*). Probabilidad y Estadística Básica para Ingenieros*. Guayaquil, Ecuador: Escuela Superior Politécnica del Litoral. Instituto de Ciencias Matemáticas.

[17] Otero, J. and Medina, E. (2016). [online] Disponible en: https://www.uam.es/personal_pdi/economicas/eva/pdf/tab_conting.pdf [Recuperado el 5 de julio de 2016].

[18] Marín, J. (2016). Recuperado el 5 de julio de 2016, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema2Cat.pdf

[19] Freund, J., Miller, I., & Miller, M. (2000). *Estadística matemática con aplicaciones*. México: Pearson Educación.

[20] Pérez, C. (2004). *Técnicas de análisis multivariante de datos*. Madrid: Pearson Prentice Hall.

[21] Luque, T. (2000). *Técnicas de análisis de datos en investigación de mercados*. [Madrid]: Ediciones Pirámide.

[22] Álvarez, R. (2008). *Estadística multivariante y no paramétrica con SPSS*. Madrid: Ediciones Díaz de Santos.

[23] Bernal, E. (2014). *Bioestadistica básica para investigadores con SPSS*. [s. l.]: Bubok Publishing.

ANEXOS

ANEXO 1

SINIESTRALIDAD POR MARCA DE VEHÍCULO Y AÑO

MARCA	AÑO					Total general
	2011	2012	2013	2014	2015	
CHEVROLET	437	415	453	139	38	**1482**
FORD	62	101	91	21	0	**275**
GREAT WALL	11	16	26	22	5	**80**
HINO	21	3	8	2	0	**34**
HONDA	6	7	3	1	0	**17**
HYUNDAI	119	123	106	43	14	**405**
KIA	169	164	298	63	10	**704**
MAZDA	46	27	29	19	4	**125**
MITSUBISHI	17	31	3	0	3	**54**
NISSAN	115	175	137	50	7	**484**
OTRA	46	88	100	36	4	**274**
RENAULT	99	106	37	14	6	**262**
SKODA	43	37	79	16	0	**175**
SUZUKI	123	32	21	51	7	**234**
TOYOTA	45	53	132	23	0	**253**
TUKO	0	0	0	5	0	**5**
VOLKSWAGEN	69	47	17	4	0	**137**
Total general	**1428**	**1425**	**1540**	**509**	**98**	**5000**

ANEXO 2

SINIESTRALIDAD POR TIPO DE VEHÍCULO Y AÑO

TIPO DE VEHICULO	AÑO					Total general
	2011	2012	2013	2014	2015	
CAMIÓN	29	55	54	11	3	**152**
CAMIONETA	275	274	259	81	17	**906**
MOTO	8	14	7	43	1	**73**
OTRO	49	58	38	24	1	**170**
SEDÁN	638	666	749	200	51	**2304**
TODOTERRENO	429	358	433	150	25	**1395**
Total general	**1428**	**1425**	**1540**	**509**	**98**	**5000**

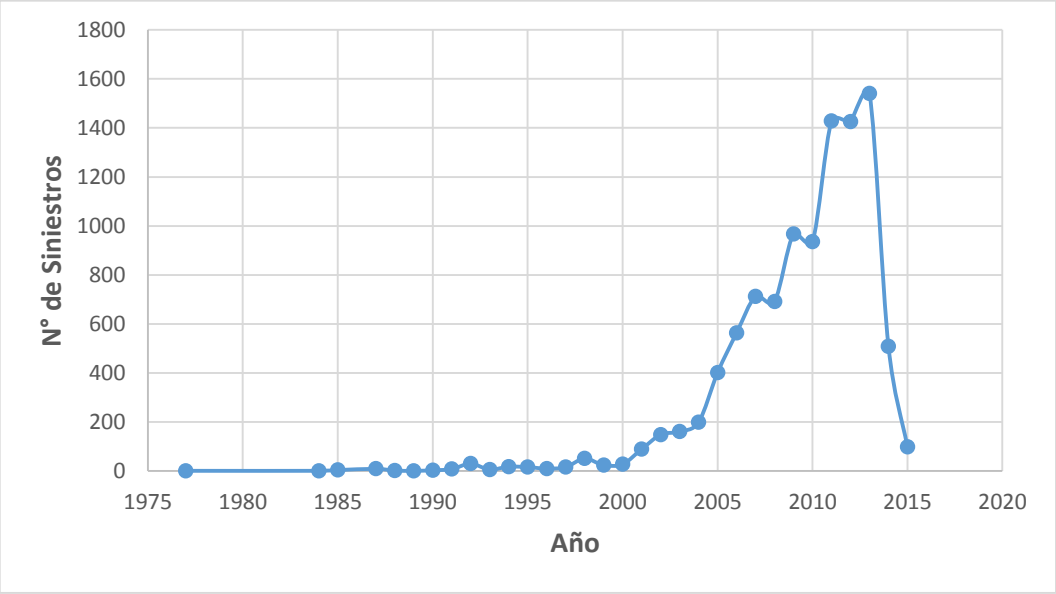
ANEXO 3

SINIESTRALIDAD POR COLOR DE VEHÍCULO Y AÑO

COLOR	AÑO					Total general
	2011	2012	2013	2014	2015	
ACERO	0	0	0	20	0	**20**
AMARILLO	2	2	31	2	0	**37**
AZUL	107	70	58	46	6	**287**
BEIGE	71	60	50	12	2	**195**
BLANCO	257	321	274	79	16	**947**
CELESTE	29	28	10	3	2	**72**
CONCHO DE VINO	49	47	67	28	9	**200**
DORADO	120	71	98	24	9	**322**
NEGRO	224	193	220	69	15	**721**
OTRO	20	36	32	9	0	**97**
PLATEADO	221	264	334	87	17	**923**
PLOMO	165	224	297	102	17	**805**
ROJO	83	97	64	25	5	**274**
VERDE	80	12	5	3	0	**100**
Total general	**1428**	**1425**	**1540**	**509**	**98**	**5000**

ANEXO 4

SINIESTRALIDAD HISTÓRICA DE VEHÍCULOS



ANEXO 5

COMPARATIVO DE SINIESTRALIDAD POR TIPO, MARCA Y AÑO

TIPO DE VEHICULO	MARCA	AÑO		Total general
		2014	2015	
CAMIONETA	CHEVROLET	30	9	**39**
	FORD	7	0	**7**
	GREAT WALL	7	2	**9**
	HYUNDAI	4	1	**5**
	KIA	7	0	**7**
	MAZDA	14	4	**18**
	NISSAN	3	0	**3**
	OTRA	2	1	**3**
	TOYOTA	7	0	**7**
SEDÁN	CHEVROLET	71	22	**93**
	FORD	2	0	**2**
	GREAT WALL	5	0	**5**
	HYUNDAI	20	11	**31**
	KIA	40	7	**47**
	NISSAN	29	5	**34**
	OTRA	6	2	**8**
	RENAULT	8	4	**12**
	SKODA	12	0	**12**
	TOYOTA	3	0	**3**
	VOLKSWAGEN	4	0	**4**
TODOTERRENO	CHEVROLET	34	7	**41**
	FORD	11	0	**11**
	GREAT WALL	10	3	**13**
	HYUNDAI	13	1	**14**
	KIA	14	3	**17**
	MAZDA	5	0	**5**
	NISSAN	18	2	**20**
	OTRA	8	1	**9**
	RENAULT	6	2	**8**
	SUZUKI	18	6	**24**
	TOYOTA	13	0	**13**
Total general		**431**	**93**	**524**