ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS
DEPARTAMENTO DE MATEMÁTICAS

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 QUITO.

AUTOR:
ING. JOHNNY JOSÉ JIMÉNEZ CONTRERAS

Guayaquil-Ecuador
2016

DEDICATORIA

Quiero dedicar esta investigación a Dios, porque ha sido mi guía, ha permitido que pueda superar muchos obstáculos de diversa índole que he tenido que atravesar durante mi periodo de estudios. Sobre todo, me ayudó a superar la pérdida de madre, momento en el cual sentí que me derrumbaba y que todo perdía sentido para mí, justo en ese momento, me dio la fortaleza para seguir adelante, ya que sin su ayuda no hubiese podido concluir con éxito esta maestría, que hoy luego de varios años llega a su fin.

A mi mamá, Anita, quien me hubiese gustado que esté hoy conmigo, pero lamentablemente por decisión de Dios, no me pudo acompañar, pero estoy seguro que, desde el cielo ella hace me da su apoyo.

A mi papá, José, porque siempre estuvo ahí para apoyarme, guiarme y darme sus sabios consejos.

A mi esposa Lisbeth, porque ella ha sido, es y será un pilar fundamental en mi vida.

A mi hermana, Carol, y mis sobrinos Mathew y Anita Valentina, porque llenan de alegría mis días.

AGRADECIMIENTO

Quiero agradecer en primer lugar a Dios, porque él ha hecho lo que soy hoy.

A mi tutora Ph. D. Sandra García, por brindarme su apoyo incondicional en esta investigación.

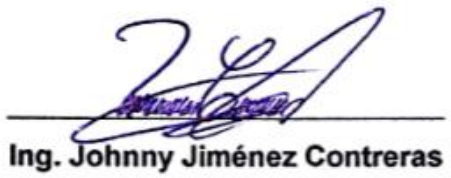
A mi jefa, la Ing. Ángela Flores, quien en aquellos momentos críticos me supo comprender y brindar su ayuda.

A la Universidad Politécnica Salesiana, porque confió en mí y me ha permitido desarrollarme como profesional.

A mis compañeros y profesores de la maestría por brindarme sus conocimientos y experiencias.

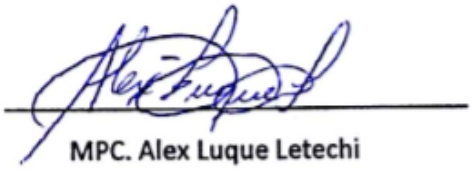
DECLARACIÓN EXPRESA

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.



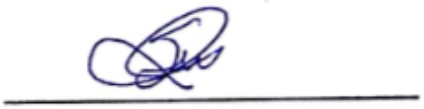
Ing. Johnny Jiménez Contreras

TRIBUNAL DE GRADUACIÓN



MPC. Alex Luque Letechi

PRESIDENTE DEL TIBUNAL



Sandra García Bustos, Ph.D.

DIRECTOR DE PROYECTO

Máster Marlon Manya Orellana

VOCAL DEL TRIBUNAL

FIRMA DEL AUTOR DEL PROYECTO DE GRADUACIÓN

Ing. Johnny Jiménez Contreras

TABLA DE CONTENIDO

CONTENIDO DE FIGURAS

CONTENIDO DE TABLAS

CAPÍTULO I

1. OBJETIVO Y GENERALIDADES

1.1. INTRODUCCIÓN

En nuestra era el transporte tiene múltiples usos y es imprescindible para llevar a cabo nuestras actividades diarias. Se usa para trasladarse al trabajo, centros de estudios, hospitales, salir de viaje, etc. y la forma en que lo usamos determina el nivel de riesgo a que nos encontramos expuestos a diario en las carreteras. El nivel de riesgo obedece a varias causas, como físicas, sociales y culturales, entre las que se puede mencionar, el país de residencia, tipo de usuario, edad, la ubicación geográfica: urbana o rural, la velocidad, tipo de vehículo, nivel de educación y el consumo de alcohol. Esta exposición puede ocasionar que exista riesgo de muerte o de heridas graves para todos los usuarios de transporte, y es transcendental que se tenga información relevante sobre las causas mencionadas, ya que se puede actuar preventivamente y evitar futuras muertes o lesiones en los individuos [1].

Los accidentes de tránsito constituyen una de las principales causas de muerte e invalidez a nivel mundial, generan en las familias de las víctimas mucho dolor por la pérdida de sus seres queridos e incertidumbre por el desconocimiento de una futura recuperación en los familiares que resultaron heridos producto de un evento de esta naturaleza. Algunos organismos internacionales han realizado algunos estudios relacionados con los accidentes de tránsito, entre ellos el Informe para el mejoramiento de la seguridad vial en el mundo, realizado por la Organización Mundial de la Salud en colaboración con las comisiones regionales y otros asociados del Grupo de Colaboración de las Naciones Unidas para la Seguridad Vial [2]. En este informe se menciona que producto de accidentes de tránsito, anualmente fallecen a nivel mundial aproximadamente 1.25 millones de personas y hasta 50 millones quedan gravemente heridas como consecuencia de las lesiones

producidas. El continente africano es el que presenta la mayor tasa de accidentes de tránsito, en cambio el continente europeo tiene la menor tasa. El 50% de los fallecimientos en accidentes de tránsito se presentan en los usuarios que tienen mayor vulnerabilidad en las vías como los transeúntes, atletas, ciclistas y motociclistas.

El Ecuador no es ajeno a la realidad existente a nivel mundial y también se ve afectado por un considerable número de accidentes de tránsito, estas cifras alarmantes son registradas por el organismo oficial a nivel nacional que es la Agencia Nacional de Tránsito (ANT). Según este organismo, las cifras del número de siniestros en diciembre de 2014 fueron 3306, en comparación con los siniestros acontecidos en diciembre de 2015 que fueron 3446, produciéndose un incremento del 4% [3]. El número de fallecidos en diciembre de 2014 fue de 186, mientras que en diciembre de 2015 fue de 199, produciéndose un incremento del 7% [4]. La cantidad de lesionados a diciembre de 2014 fue de 2445, por el contrario, en diciembre de 2015 fue de 2173, registrándose una reducción del 11% [5].

La Figura 1 muestra el comparativo de los meses de diciembre de 2014 y 2015 de los eventos acontecidos a nivel nacional producto de accidentes de tránsito.

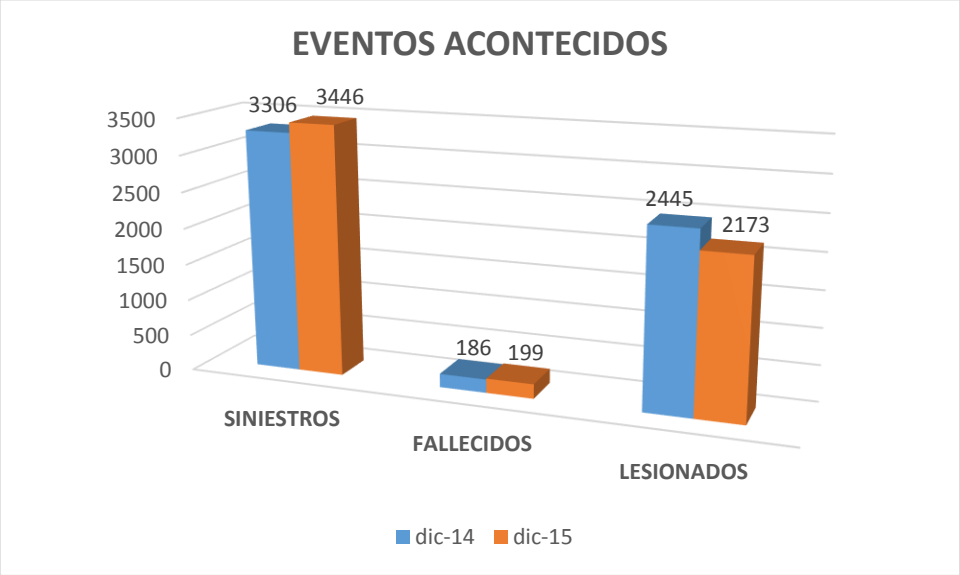


Figura 1: Eventos acontecidos

Fuente: Agencia Nacional de Tránsito - Autor: Johnny Jiménez

Esta problemática se presenta por diversos factores que constituyen las probables causas de los eventos acontecidos en diciembre de 2015. Entre las principales causas a nivel nacional de acuerdo al mayor número de siniestros producidos tenemos:

1. No respetar las señales reglamentarias de tránsito, tales como: disco pare, ceda el paso, luz roja del semáforo, etc. con un total de 439 siniestros.
2. Conducir desatento a las condiciones de tránsito, entre las que se encuentran: pantallas de video, comida, maquillaje o cualquier otro elemento distractor, con un total de 576 siniestros.
3. Conducir el vehículo superando los límites máximos de velocidad, con un total de 417 siniestros [3].

Estas causas generan que estos eventos acontecidos afecten no sólo a los conductores y sus acompañantes, sino también a los peatones en general, lo que constituye un serio problema que debe preocupar y motivar a las autoridades de gobierno, a tomar medidas que se ejecuten a través del ente regulador nacional de tránsito, para tratar de mitigar el riesgo de los vehículos a sufrir siniestros de tránsito.

Además de estas causas, los siniestros también se presentan de acuerdo al tipo de vehículo involucrado, de tal manera que, en diciembre de 2015, el porcentaje de vehículos que presentaron siniestros de tránsito a nivel nacional, son: Automóviles (42%), motos (18%), camionetas (13%), jeep (12%), bus (5%), camión (6%) y otros (4%). Dentro de la categoría otros se incluyen a: Tráileres, especiales (transportes de valores, grúas, etc.), volquetas, furgonetas y tanqueros. [3].

1.2. PLANTEAMIENTO DEL PROBLEMA

El número elevado de accidentes de tránsito motiva a que los propietarios de vehículos busquen opciones para estar protegidos ante un posible siniestro de tránsito. Los dueños de vehículos consideran necesario contar con una protección ante una eventualidad que se pudiera producir y de esta manera compensar económicamente los daños producidos a través del pago de una prima que paga el asegurado a la empresa aseguradora, de tal manera que las pérdidas que sufra el asegurado se reduzcan significativamente y no causen tanto impacto del que hubiera sufrido de no contar con un seguro.

Estos accidentes responden a varias causas, de las que se puede destacar como muy influyente a las características de los vehículos, esto constituye un problema para las aseguradoras, ya que no disponen de estudios técnicos que les permitan validar si un vehículo cualesquiera, en base a sus características, sea más o menos propenso al riesgo de tener un siniestro de tránsito, lo cual hace que la probabilidad de que una aseguradora se enfrente a un mayor número de reclamos por pago de indemnizaciones se incremente. Las aseguradoras al asumir el riesgo de asegurar un vehículo, suelen basar su decisión en las estadísticas que estas poseen sobre sus vehículos asegurados, lo cual conlleva a que la decisión tomada no sea la más adecuada. Este problema motiva a que las aseguradoras se vean en la necesidad, de encontrar una forma técnica para tomar decisiones óptimas, que les

permitan estimar la probabilidad de que un vehículo sea propenso al riesgo de sufrir un siniestro, ya que, en nuestro país al no existir este tipo de estudios técnicos, no se conoce si estas características influyen en las ocurrencias de siniestros de esta naturaleza.

Para afrontar esta problemática es primordial tener información relacionada con las características que definen a los vehículos. Por lo cual se necesitan las variables que se incluyen en cada póliza, tales como marca, categoría, color, tipo de vehículo asegurado, además variables relacionadas con el registro de siniestros, que serán obtenidas a través de una muestra de vehículos asegurados pertenecientes a una compañía aseguradora de la ciudad de Quito.

Esta problemática se puede afrontar desde diferentes perspectivas, dependiendo de las características que se analicen en los vehículos, es de interés observar las causas que producen dichas variables en la probabilidad de ocurrencia de siniestros de vehículos en la ciudad de Quito, además en la cantidad de siniestros acontecidos y la forma en que se relacionan, lo que a través de técnicas estadísticas multivariadas, tales como el análisis de correspondencias múltiples y el análisis de regresión logística, que permitirán determinar la propensión al riesgo de un vehículo según sus características, para poder clasificarlo o discriminarlo dentro del grupo de los proclives a sufrir o no un siniestro, en base a las variables anteriormente indicadas, y de esta manera decidir si otorgarle o no la cobertura.

1.3. JUSTIFICACIÓN DEL PROBLEMA

El Ecuador tiene un serio problema, la gran cantidad de accidentes de tránsito, por tal motivo, el producto de mayor demanda es el ramo de Seguros de Vehículos, debido a la alta frecuencia con que ocurren accidentes de tránsito, tanto así que, del total de primas emitidas en el año 2014, este ramo tuvo una participación del 25.8% respecto a todos los ramos comercializados por las aseguradoras [6].

A nivel nacional, el total acumulado de siniestros a diciembre de 2015 fue de 35706. Las provincias que tuvieron más siniestros fueron Pichincha con 15754 y Guayas con 6799, estos totales representan aproximadamente el 44.12% y 19.04% del total de siniestros [3].

La Tabla1 muestra los siniestros en forma trimestral por provincia a nivel nacional con valores acumulados a diciembre de 2015.

Provincia	Ene-Mar	Abr-Jun	Jul-Sep	Oct-Dic	Total a Dic. 2015	%
Azuay	312	325	356	380	1.373	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	1.603	1.679	1.766	1.751	6.799	19,04
Imbabura	325	396	342	463	1.526	4,27
Loja	152	166	171	199	688	1,93
Los Ríos	314	295	325	316	1.250	3,50
Manabí	348	260	290	319	1.217	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	**3.675**	**4.082**	**3.813**	**4.184**	**15.754**	**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	1.735	4,86

Zamora Chinchipe	27	31	22	25	105	0,29
Total	**8.472**	**9.020**	**8.728**	**9.486**	**35.706**	**100,00**
%	**23,73**	**25,26**	**24,44**	**26,57**	**100,00**	

Tabla 1: Siniestros por provincia a nivel nacional a diciembre de 2015

Fuente: Agencia Nacional de Tránsito - Autor: Johnny Jiménez

Los cinco tipos de siniestros, con valores acumulados a diciembre de 2015, que más se presentaron a nivel nacional fueron:

1. Los choques laterales con un total de 10129, lo que representa a un 28.37% del total de siniestros.
2. Los atropellos con un total de 5140, lo que corresponde a un 14.40% del total de siniestros.
3. Los estrellamientos con un total de 4624, lo que representa a un 12.95% del total de siniestros.
4. Los choques posteriores con un total de 4068, lo que corresponde a un 11.39% del total de siniestros.
5. Las pérdidas de pista con un total 3471, lo que representa a un 9.72% del total de siniestros [5].

La Tabla 2 muestra los siniestros en forma trimestral por tipo a nivel nacional con valores acumulados a diciembre de 2015.

Tipo	Ene-Mar	Abr-Jun	Jul-Sep	Oct-Dic	Total a Dic. 2015	%
Choque lateral	2.368	2.566	2.465	2.730	10.129	28,37
Atropello	1.284	1.318	1.180	1.358	5.140	14,40
Estrellamiento	1.034	1.162	1.169	1.259	4.624	12,95
Choque posterior	1.000	978	1.010	1.080	4.068	11,39
Pérdida de pista	671	884	897	1.019	3.471	9,72
Rozamiento	642	665	647	643	2.597	7,27
Choque frontal	451	446	433	449	1.779	4,98
Colisión	250	314	297	300	1.161	3,25
Volcamiento	281	259	234	280	1.054	2,95
Caída de pasajero	180	223	210	197	810	2,27
Otros	237	122	115	108	582	1,63
Arrollamiento	74	83	71	63	291	0,81
Total	**8.472**	**9.020**	**8.728**	**9.486**	**35.706**	**100,00**
%	**23,73**	**25,26**	**24,44**	**26,57**	**100,00**	

Tabla 2: Siniestros por tipo a nivel nacional a diciembre de 2015

Fuente: Agencia Nacional de Tránsito - Autor: Johnny Jiménez

Se puede notar que históricamente durante el año 2015, la Provincia de Pichincha fue la que registró la tasa más alta de siniestralidad. En el mes de diciembre de 2015, en esta provincia se presentaron en total 1511 siniestros, donde vale la pena destacar que sólo en la Ciudad de Quito se produjeron 1449 siniestros, lo cual representa un 95,90% a nivel provincial [3].

Por esta razón se escogió a la ciudad de Quito, que cuenta con un amplio parque automotor para realizar esta investigación; ya que, a pesar de no ser la ciudad más poblada del Ecuador, se presentan accidentes de tránsito con mayor frecuencia.

La Tabla 3 muestra los siniestros de tránsito en la provincia de Pichincha según cantones en diciembre de 2015.

Provincia	Cantón	N° Siniestros	%
Pichincha	Cayambe	5	0,33
	Mejía	36	2,38
	Pedro Moncayo	6	0,40
	Puerto Quito	3	0,20
	Quito	**1.449**	**95,90**
	Rumiñahui	10	0,66
	San miguel de los bancos	2	0,13
	Total	**1.511**	**100,00**

Tabla 3: Siniestros de tránsito según cantones en diciembre de 2015

Fuente: Agencia Nacional de Tránsito - Autor: Johnny Jiménez

Año tras año el parque automotor va en aumento, este incremento es directamente proporcional al riesgo de sufrir un siniestro de tránsito, lo cual conlleva a que más personas necesiten contratar un seguro de vehículos.

Dentro del ámbito de los seguros se han realizado algunas investigaciones a nivel internacional, para estudiar la siniestralidad, como los de (Das & Sun, 2016) [7]; (Paefgen, Staake, & Fleisch, 2014) [8]; (Hemrit, Arab, & Raissi, 2013) [9], pero a nivel nacional no existen estudios de similares características, lo que hace de esta investigación pionera dentro del mercado asegurador ecuatoriano y servirá como base para futuros trabajos en este campo.

1.4. OBJETIVO GENERAL

Determinar un modelo para medir el riesgo asociado a las características de los vehículos en la ciudad de Quito, 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 Quito.
2. Identificar posibles relaciones entre diversas variables con el evento de ocurrencia de un siniestro.
3. Utilizar técnicas estadísticas multivariadas para modelar la probabilidad de ocurrencia de un siniestro en el ramo de seguros de vehículos.
4. Comparar y analizar los modelos desarrollados.

La presente investigación consta de cuatro capítulos.

El Capítulo 1 consta de todo lo relacionado al planteamiento del problema, justificación del problema, objetivo general y objetivos específicos de la investigación.

El Capítulo 2, estará compuesto por el marco teórico, relacionado con las definiciones básicas de Estadística Descriptiva, Tablas de Contingencia, y además otras técnicas como Análisis de correspondencias múltiples y Regresión Logística, sobre los cuales se basará esta investigación.

El Capítulo 3, se realizará el análisis descriptivo de las variables, análisis de Tablas de Contingencia, así como la interpretación de los resultados obtenidos.

El Capítulo 4, se realizará el análisis multivariado, a través de los modelos propuestos: Análisis de correspondencias múltiples y Regresión Logística, así como la interpretación de los resultados obtenidos.

Y la parte final, está formada por las Conclusiones y Recomendaciones de la investigación.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. INTRODUCCIÓN

Si un investigador ha recolectado un conjunto de observaciones para realizar un estudio, no es una buena idea trabajar con los datos en bruto, ya que no le brindarían la información que necesita si previamente no los trata, y le resultaría muy difícil entenderlos, por lo cual debe tratar de organizarlos y resumirlos para que estos presenten algún tipo de patrón y sean fáciles de interpretar.

En esta investigación se definirán algunos conceptos que sustentan los diversos análisis que se usarán a lo largo de la misma, de tal manera que para la descripción univariada de datos, como el caso de datos cualitativos se definirán las distribuciones de frecuencia, las gráficas de barras y de pastel; en cambio para los datos de tipo cuantitativo, se tendrán histogramas, diagramas de caja, así como también, la definición de estadísticos, entre los cuales están la media, mediana, varianza, desviación estándar, coeficientes de asimetría y curtosis, cuartiles y percentiles.

Para analizar la posible relación entre dos variables se definirán las tablas de contingencia, finalizando con el análisis multivariado, donde se tratarán los conceptos relacionados con el análisis de correspondencias múltiples y el análisis de regresión logística.

2.2. ESTADÍSTICA DESCRIPTIVA

2.2.1. MEDIA MUESTRAL

Sean $x_1, x_2, \cdots, x_n$ un conjunto de n observaciones recolectadas a partir de una población. Este conjunto se considera una muestra de la población bajo estudio. La media muestral o aritmética de ese conjunto

de datos, la cual se denota por $\overline{x}$, se define mediante la fórmula matemática como: (Walpole et al., 2012) [10].

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

2.2.2. MEDIANA MUESTRAL

Sean $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ una muestra de n observaciones, recolectadas de una población de interés, ordenadas de menor a mayor. La mediana muestral de este conjunto de datos, la cual se denota por $\widetilde{x}$, se define mediante la fórmula matemática como: (Walpole et al., 2012) [10].

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

2.2.3. MODA MUESTRAL

Sean $x_1, x_2, \cdots, x_n$ un conjunto de n observaciones, las cuales representan una muestra tomada de una población. La moda muestral de este conjunto de datos es el valor que aparece con mayor frecuencia, es decir es el dato que más se repite. Sin embargo, si dentro de este conjunto aparecen dos valores con la misma frecuencia, y ésta es la mayor, ambos valores son modas; por lo tanto, se considera al conjunto de datos como bimodal. Si en el conjunto de datos existen más de dos valores que aparecen con la misma frecuencia, y ésta es la mayor, cada uno de estos valores son modas; por lo tanto, se considera al conjunto de datos como multimodal. Por el contrario, si ningún valor se repite dentro del conjunto de datos, entonces se dice que no existe moda (Triola, 2009) [11].

2.2.4. VARIANZA MUESTRAL

Sean $x_1, x_2, \cdots, x_n$ un conjunto de n observaciones muestrales. La varianza de este conjunto de datos es la suma de los cuadrados de las desviaciones entre cada dato y su media muestral $\bar{x}$, dividida entre (n-1). La varianza muestral se denota por S^2 y se calcula con la fórmula: (Mendenhall et al., 2010) [12].

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

2.2.5. DESVIACIÓN ESTÁNDAR MUESTRAL

Para un conjunto de n observaciones muestrales, $x_1, x_2, \cdots, x_n$, se define a la desviación estándar muestral, como la raíz cuadrada positiva de la varianza muestral, es decir, es igual a S.

La desviación estándar muestral matemáticamente es igual:

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

La desviación estándar muestral cumple con dos propiedades sustanciales:

1. El valor de la desviación estándar muestral S puede incrementarse considerablemente si se incluyen en la muestra, uno o más valores extremos, es decir, observaciones que se encuentren muy alejadas de las demás.
2. Las unidades de la desviación estándar muestral S, por ejemplo: gramos, segundos, metros, etcétera, son las mismas de los datos originales (Triola, 2009) [11].

2.2.6. COEFICIENTE DE VARIACIÓN MUESTRAL

El coeficiente de variación CV de un conjunto de n observaciones muestrales $x_1, x_2, \cdots, x_n$, describe la relación entre la desviación estándar muestral y la media muestral. El coeficiente de variación muestral, si se expresa como un porcentaje, viene dado por: (Triola, 2009) [11].

$$CV = \frac{S}{\bar{x}} * 100\%$$

2.2.7. COEFICIENTE DE ASIMETRÍA MUESTRAL DE FISHER

El coeficiente de asimetría de Fisher es una medida que permite determinar la simetría de la distribución de un conjunto de n observaciones muestrales, $x_1, x_2, \cdots, x_n$, respecto a su media muestral $\bar{x}$ (Véliz, 2011) [13].

Este coeficiente se puede calcular como:

$s_k = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)S^3}$, donde S es la desviación estándar muestral.

Dependiendo de los valores que se obtengan, el coeficiente de asimetría se puede interpretar como:

Si la distribución es asimétrica positiva, entonces $s_k > 0$. Esto significa que la cola de la distribución se va alargando para valores mayores que la media muestral.

Si la distribución es simétrica, entonces $s_k = 0$. Esto significa que la distribución se adapta a la forma de la campana de Gauss.

Si la distribución es asimétrica negativa, entonces $s_k < 0$. Esto significa que la cola de la distribución se va alargando para valores menores que la media muestral.

2.2.8. COEFICIENTE DE CURTOSIS MUESTRAL DE FISHER

El coeficiente de curtosis de Fisher es una medida que permite determinar el grado de apuntamiento o achatamiento de la distribución de un conjunto de n observaciones muestrales, $x_1, x_2, \cdots, x_n$.

A través de este coeficiente se puede tener una idea de cuántos datos se encuentran próximos a la media muestral $\overline{x}$, de tal manera que, a mayor grado de curtosis, la distribución tendrá más apuntamiento, es decir será más escarpada (Véliz, 2011) [13].

Este coeficiente se puede calcular como:

$k = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{(n-1)S^4} - 3$, donde S es la desviación estándar muestral.

Dependiendo de los valores que se obtengan, el coeficiente de curtosis se puede interpretar como:

Si $k > 0$, entonces la distribución es Leptocúrtica. Esto significa que existe una gran concentración de datos en la región central de la distribución. Si se compara esta distribución con la distribución normal, se observaría que es más puntiaguda que la normal.

Si $k = 0$, entonces la distribución es Mesocúrtica. Esto significa que existe una concentración de datos aproximadamente normal en la región central de la distribución. Si se compara esta distribución con la distribución normal, se tendría que es tan achatada como la normal y ambas distribuciones serían muy parecidas.

Si $k < 0$, entonces la distribución es Platicúrtica. Esto significa que existe una baja concentración de datos en la región central de la distribución. Si se compara esta distribución con la distribución normal, se observaría que es más achatada que la normal, en otras palabras, corresponde a una distribución con aspecto aplanado o en forma de meseta.

2.2.9. PERCENTILES

Sean $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ una muestra de n observaciones, recolectadas de una población de interés, ordenadas de menor a mayor. Los percentiles son valores que dividen a este conjunto de observaciones en 100 partes iguales, cada una contiene un 1% de las observaciones. Los percentiles se representan como P_k donde k=1, 2, ⋯, 99. El subíndice k indica el percentil deseado, y divide a los datos en dos partes. Por ejemplo, P_{10} es el percentil diez y divide al conjunto de datos de tal manera que el 10% de las observaciones tienen valores que son menores o iguales que P_{10} y el 90% restante, corresponde a las observaciones cuyos valores son mayores o iguales que P_{10}.

En términos generales, el percentil k es un valor tal que k% de las observaciones son menores o iguales que este valor y (100 – k) % de las observaciones restantes son mayores o iguales que este valor (Anderson et al., 2008) [14].

Pasos para el cálculo del percentil k:

1. Se ordenan los datos en orden ascendente, es decir de menor a mayor.
2. Calcular el índice i, que representa 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 observaciones.
3. (a) Si $i \notin \mathbb{Z}$, 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, $P_k = \frac{1}{2}\left[x_{(i)} + x_{(i+1)}\right]$

2.2.10. CUARTILES

Sean $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ una muestra de n observaciones, recolectadas de una población de interés, ordenadas de menor a mayor. Los cuartiles son valores que dividen a este conjunto de observaciones en 4 partes iguales, cada una contiene un 25% de las observaciones. Los cuartiles se representan como Q_1, Q_2 y Q_3 y dividen al conjunto de datos en dos partes.

El primer cuartil, o Q_1, o percentil 25, es un valor tal que el 25% de las observaciones son menores o iguales que este valor y el 75% restante de las observaciones son mayores o iguales que este valor.

El segundo cuartil, o Q_2, o percentil 50, es un valor tal que 50% de las observaciones son menores o iguales que este valor y 50% restante de las observaciones son mayores o iguales que este valor. Algo importante que se debe conocer, es que el segundo cuartil coincide con la mediana.

El tercer cuartil, denotado por Q_3, o percentil 75, es un valor tal que el 75% de las observaciones son menores o iguales que este valor y el 25% restante de las observaciones son mayores o iguales que este valor.

Los cuartiles se pueden obtener a través de los pasos para calcular percentiles (Anderson et al., 2008) [14].

2.2.11. DIAGRAMA DE CAJA

Un diagrama de caja es un gráfico de los datos basado en el resumen de cinco valores a saber: valor mínimo, primer cuartil Q_1, segundo cuartil Q_2 (mediana), tercer cuartil Q_3 y valor máximo. Un diagrama de caja es especialmente útil para examinar al mismo tiempo ciertas características importantes, como la dispersión y la simetría de los datos o si hay presencia de valores atípicos y cuando se comparan dos conjuntos de muchos datos.

Para elaborar un diagrama de caja se siguen los siguientes pasos: (Anderson et al., 2008) [14].

1. Se construye una caja cuyos extremos se encuentren localizados sobre el primer y tercer cuartiles. En esta caja se encuentran contenidos 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 intercuartílico (RIC), RIC = $Q_3 – Q_1$, el cual sirve para poder localizar los límites de un intervalo. Estos límites se localizan 1.5 veces el rango intercuartílico (RIC) por debajo del Q_1 y 1.5 veces el rango intercuartílico (RIC) por encima del Q_3. El intervalo que se forma es [Q_1 – 1.5*RIC, Q_3 + 1.5*RIC], de tal forma que cualquier observación que se encuentre fuera de este intervalo es considerada como una observación atípica.

Además, pueden existir observaciones que sean menores que Q_1 – 3*RIC o que sean mayores que Q_3 + 3*RIC, a las cuales se las considera observaciones atípicas extremas, es decir observaciones que se encuentran muy alejadas del resto de los datos.

4. Se grafican dos líneas rectas que salen desde los extremos de la caja hasta los valores menor y mayor de los límites calculados en el paso anterior, a dichas líneas rectas se les denomina bigotes.
5. Por último, en caso de que existan, mediante un círculo se localizan a las observaciones atípicas y mediante un asterisco a las observaciones atípicas extremas.

2.2.12. DISTRIBUCIÓN DE FRECUENCIAS

La distribución de frecuencias es útil para poder resumir, mediante una tabla, una gran cantidad de datos, ya que se logra comprender de mejor manera la naturaleza de los mismos. A la distribución de frecuencias también se le llama tabla de frecuencias, porque permite listar valores de los datos, tanto por categorías, en forma individual o por medio de intervalos, asociándolas a sus respectivas frecuencias o conteos. Las tablas de frecuencias son importantes ya que permiten resumir datos de naturaleza cualitativa o cuantitativa (Triola, 2009) [11].

En el caso de las variables cualitativas, los conteos se realizan en función de las categorías que se definan para la variable de estudio. Por ejemplo, para la variable estado civil, en la cual quedan definidas cinco categorías: soltero, casado, divorciado, viudo y unión libre, su distribución de frecuencias será simplemente el conteo de los valores asociados a cada categoría. Por el contrario, si se dispone de una variable cuantitativa, se distribuyen sus valores en intervalos, dentro de los cuales se registra el conteo de las respectivas ocurrencias (Alvarado y Obagi, 2008) [15].

Supongamos que disponemos de un conjunto de n observaciones que se quieren resumir a través de una tabla de frecuencias, es importante conocer algunas definiciones que serán de utilidad para su construcción (Alvarado y Obagi, 2008) [15].

2.2.12.1. INTERVALOS DE CLASE

Son aquellos grupos en los que se dividen a las observaciones de la variable de interés. Estos intervalos tienen un límite inferior y un límite superior, y se denotan por $[a_{i-1}, a_i)$.

2.2.12.2. MARCA DE CLASE

Es el valor promedio de los límites del intervalo de clase correspondiente. Se denota y calcula como: $m_i = \frac{a_{i-1}+a_i}{2}$

2.2.12.3. FRECUENCIA ABSOLUTA

Representa la cantidad de observaciones contenidas dentro del intervalo de clase i. Se denota como f_i.

2.2.12.4. FRECUENCIA RELATIVA

Es la relación entre el número de observaciones contenidas dentro del intervalo de clase i y el total de observaciones disponibles, n. Se calcula como $\frac{f_i}{n}$.

2.2.12.5. FRECUENCIA ABSOLUTA ACUMULADA

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

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

2.2.12.6. FRECUENCIA RELATIVA ACUMULADA

Es la proporción de observaciones acumuladas hasta el intervalo i. Se denota y se calcula como: $\frac{F_i}{n} = \sum_{j=1}^{i} \frac{f_j}{n}$

2.2.12.7. PASOS PARA CONSTRUIR UNA TABLA DE FRECUENCIAS

Se pueden seguir los siguientes pasos: (Rodríguez, 2007) [16]

1. Obtener el rango de los datos (R).

   $$R = valor\ máximo - valor\ mínimo$$

2. Seleccionar el número de clases o intervalos C, para agrupar los datos. El número de clases debe estar entre 5 y 20, es decir 5≤C≤20

   Una regla simple para aproximar el número de clases es: $2^C \geq n$. Donde, C es el número de clases a utilizar y n es el número total de observaciones.

   El número de clases a utilizar es el menor valor de C que garantiza que la desigualdad anterior se satisfaga.

   Existen otros criterios que se han definido para obtener la cantidad de clases con las que se puede trabajar, entre las cuales tenemos la Regla empírica tradicional, que indica que la cantidad de clases es igual a la raíz cuadrada del número de observaciones, es decir, $\sqrt{n}$; y la Regla empírica de Sturges, que indica que las clases a utilizar está dada por la fórmula $1 + log_2(n)$, donde n es el número de observaciones. En estas dos reglas, si el número de clases obtenidas no es entero, se redondea al entero más próximo, y se trabaja con una cantidad de clases igual al valor entero redondeado (Alvarado y Obagi, 2008) [15].

3. Obtener la amplitud de las clases:

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

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

Una tabla de frecuencias tiene la siguiente estructura:

Clase i	Intervalo de clase	Marca de clase m_i	Frecuencia absoluta f_i	Frecuencia relativa $\frac{f_i}{n}$	Frecuencia absoluta acumulada F_i	Frecuencia relativa acumulada $\frac{F_i}{n}$
1	$[a_1, a_2)$	m_1	f_1	$\frac{f_1}{n}$	f_1	$\frac{f_1}{n}$
2	$[a_2, a_3)$	m_2	f_2	$\frac{f_2}{n}$	$f_1 + f_2$	$\frac{f_1 + f_2}{n}$
3	$[a_3, a_4)$	m_3	f_3	$\frac{f_3}{n}$	$f_1 + f_2 + f_3$	$\frac{f_1 + f_2 + f_3}{n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
C	$[a_c, a_{c+1}]$	m_c	f_c	$\frac{f_c}{n}$	$f_1 + f_2 + f_3 + \cdots + f_c$	$\frac{f_1 + f_2 + f_3 + \cdots + f_c}{n}$

Tabla 4: 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: Johnny Jiménez.

2.2.13. HISTOGRAMA DE FRECUENCIAS

Para representar en forma gráfica un conjunto de datos cuantitativos sea usa el histograma de frecuencias. Se lo construye con base en datos que previamente hayan sido resumidos mediante una tabla de frecuencias. Se coloca sobre el eje horizontal la variable de interés y sobre el eje vertical la frecuencia absoluta o frecuencia relativa. Se grafican rectángulos, uno a continuación del otro, cuyas bases son los valores de la variable que se encuentran agrupados en los intervalos de clase, y las alturas de estos rectángulos son las correspondientes frecuencias absolutas o relativas. Los histogramas son utilizados porque proveen información relacionada con la simetría y dispersión de las observaciones (Anderson et al., 2008) [14].

2.2.14. GRÁFICAS DE BARRA

Una gráfica de barras o diagrama de barras, es un gráfico utilizado para representar las frecuencias asociadas a las categorías de una variable cualitativa. Para su construcción, por lo general, sobre el eje horizontal, se colocan las etiquetas para representar a las categorías de la variable de interés y sobre el eje vertical se colocan los valores de frecuencia de cada categoría. Se grafican rectángulos o barras de un ancho arbitrario, cuyas bases se colocan sobre las categorías de la variable y cuyas alturas se extienden hasta el valor de frecuencia de la categoría actual. En este tipo de gráfico las barras deben estar separadas para reflejar a cada categoría de la variable (Anderson et al., 2008) [14].

2.2.15. GRÁFICAS DE PASTEL

Las gráficas de pastel brindan otra forma de representación gráfica de las frecuencias asociadas a una variable cualitativa. Para elaborar una gráfica de pastel, primero se grafica un círculo que representa a todos los datos. A continuación, se usa la frecuencia relativa para subdividir el círculo en segmentos, o partes, que corresponden a la frecuencia relativa de cada categoría de la variable. El área de cada uno de los segmentos es proporcional al número de casos en esa categoría (Anderson et al., 2008) [14].

2.3. TABLAS DE CONTINGENCIA

Supongamos que disponemos de dos variables cualitativas o factores, y se requiere analizar si existe una relación de dependencia o independencia entre ellas, esta posible relación se la estudia a través de las denominadas tablas de contingencia. En las tablas de contingencia, una de las variables es usada para representar las filas y la otra para representar las columnas (Otero y Medina, 2016) [17].

Sean X la variable de las filas e Y la variable de las columnas, 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.

En las casillas de la tabla de contingencia se colocan las observaciones, de acuerdo a los valores que toman según la fila y columna que le corresponden.

Como las observaciones se disponen en una tabla de contingencia, comúnmente a esta tabla se la conoce con el nombre de tabla de clasificación cruzada, con r filas y c columnas, o tabla r × c (Marín, 2016) [18].

En general, es de interés determinar si existe o no una relación de dependencia entre la variable de las filas y la variable de las columnas, por lo cual se busca contrastar si las dos variables son independientes, es decir, se debe realizar 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.

Para poder efectuar esta prueba de hipótesis, se deben clasificar a los individuos en una tabla de contingencia que tiene la forma:

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 5: Tabla de contingencia
Fuente: Datos de la investigación - Autor: Johnny Jiménez.

Donde: (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 que pertenece 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.

Las probabilidades mencionadas 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 cual, otra manera de expresar 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 probar 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 para poder 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 factorial que busca reducir la dimensión de una tabla de datos formada por variables cualitativas, con la finalidad de obtener un número pequeño de factores, cuya interpretación a posteriori hará que el problema investigado sea más simple de estudiar (Pérez, 2004) [20].

El análisis de correspondencias al trabajar con variables cualitativas, o con variables cuantitativas categorizadas (por ejemplo, si se define la variable edad, pero categorizada en distintos rangos de edad) hace que posea dos características importantes. Primero, trabaja con frecuencias que son el resultado de cruzar dos o más variables. Segundo, cuando se cruzan dos variables, utiliza como individuos y variables las diferentes categorías existentes. Esto hace posible aplicar el llamado análisis de correspondencias simple (ACS), pero cuando las categorías pertenecen a más de dos variables, el método se generaliza y se obtiene el análisis de correspondencias múltiples (ACM) (Luque, 2000) [21].

Para el análisis de correspondencias simple, los datos de las dos variables se representan a través de una tabla de contingencia, pero en el análisis de correspondencias múltiples, la tabla de contingencia se convierte en una hipertabla que puede tener tres o más dimensiones, la

cual no es muy fácil de representar y suele resumirse en la llamada tabla de Burt (Pérez, 2004) [20].

A través del análisis de las tablas de contingencia sólo se puede comprobar si las variables estudiadas se encuentran asociadas entre sí, con la finalidad de encontrar algún modelo causal, o simplemente determinar si existe algún tipo de interrelación, usando diversos test como el ji-cuadrado χ^2. El inconveniente en este tipo de análisis es que no permite saber cuáles categorías son las que ocasionan esa relación ni cuáles son las que poco aportan a dicha asociación. Es por esto, que una de las ventajas del análisis de correspondencias, es que permite llegar a conclusiones que el análisis de las tablas de contingencia por sí sólo no puede encontrar, ya que, a través de la extracción de las relaciones entre categorías, permite definir similaridades o disimilaridades entre ellas, que en caso de detectarse que se corresponden entre sí, se permitirá su agrupamiento (Luque, 2000) [21].

Por medio del análisis de correspondencias todo esto queda representado en un espacio dimensional de pocas variables sintéticas o factores, los cuales pueden ser nombrados e interpretados, y deben ser tales que resuman la mayor cantidad de información, que con la ayuda de representaciones gráficas o mapas de correspondencias se pueden visualizar conjuntamente las relaciones obtenidas. Como el análisis de correspondencias se basa en el análisis factorial, las dimensiones que crean el espacio donde se representan las categorías, hace que se obtengan como factores cuantitativos, por tal motivo se considera al análisis de correspondencias como un método que extrae variables ficticias cuantitativas partiendo de variables cualitativas originales, ya que estas definen las relaciones que existen entre sus categorías (Pérez, 2004) [20].

El análisis de correspondencias, puede constituirse en un paso intermedio para la aplicación de otras pruebas multivariadas como el análisis clúster, el análisis de regresión o el análisis discriminante. De esta manera hace posible la aplicación a un conjunto de datos

cualitativos, ya que se obtienen coordenadas métricas en el espacio dimensional definido por los factores (Luque, 2000) [21].

2.4.1. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

El análisis de correspondencias múltiples, se considera una generalización del análisis de correspondencias simple, ya que el número de caracteres o variables cualitativas que se emplean es mayor que dos.

No se podría trabajar con una tabla de contingencia de 2x2, como en el caso del análisis de correspondencias simple, ya que la representación tabular de los datos ahora se complica, por lo cual es necesario realizar un análisis de correspondencias múltiples, porque permite estudiar las relaciones entre las modalidades de todas las variables cualitativas consideradas.

En este trabajo se abordará el análisis de correspondencias múltiples debido a la naturaleza propia de la investigación, y en la medida de lo posible se tratará de unificar la notación con la utilizada en el análisis de correspondencias simple, con la finalidad de lograr una mejor comprensión de los conceptos aquí expuestos.

2.4.2. FORMULACIÓN DEL ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES (ACM)

En el análisis de correspondencias múltiples se disponen los datos en una tabla Z que se denomina tabla disyuntiva completa, la cual está formada por un conjunto I de n individuos o filas, un conjunto J de Q variables o caracteres cualitativos y por un conjunto de modalidades o categorías excluyentes $1, \cdots, m_k$ para cada variable cualitativa. El número de modalidades o categorías es igual a $\mathrm{J} = \sum_{k=1}^{Q} m_k$ (Pérez, 2004) [20].

Esta tabla disyuntiva completa Z, la cual es de dimensión I × J puede ser representada de manera general como:

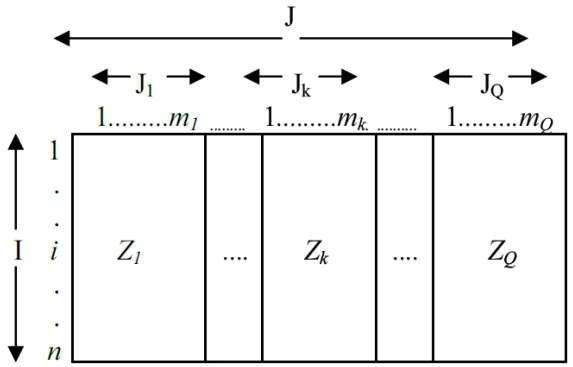


Figura 2: Tabla disyuntiva completa Z

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

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

En la tabla Z cada elemento Z_{ij} puede tomar el valor 0 o 1, dependiendo si el individuo i ha elegido o no la modalidad j. De acuerdo a esto, cada uno de los rectángulos en los que se divide la tabla Z, puede considerarse, a pesar de no serlo, como una tabla de contingencia que contiene a los valores 0 o 1. En otras palabras, la tabla Z está formada por Q subtablas adosadas, que permiten representar simultáneamente a todas las modalidades o categorías (columnas) de todos los individuos (filas). Como las modalidades o categorías son excluyentes, cada subtabla tendrá un solo 1 en cada fila.

A los elementos Z_{ij} se los representará como k_{ij}, para guardar analogía con la notación que se usa en el análisis de correspondencias simple, de esta manera el lector se sentirá familiarizado por la utilización de una notación semejante a la acostumbrada. Bajo estas consideraciones tenemos 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) (Pérez, 2004) [20].

2.4.3. OBTENCIÓN DE LOS FACTORES: TABLA DE BURT

Si se quieren obtener los factores se debe diagonalizar la matriz V, tal que $V = \frac{1}{Q} D^{-1} B$, donde D^{-1} es una matriz diagonal cuyos elementos diagonales son los de la matriz Burt, y el resto de sus elementos son iguales a cero y B es la matriz o tabla de Burt. Esta última se puede calcular multiplicando la matriz de datos transpuesta por sí misma, es decir, $B = Z'Z$.

La matriz de Burt es simétrica y está formada por Q^2 bloques, de tal manera que sus bloques de la diagonal $Z'_k Z_k$, cuyos elementos son tablas diagonales que cruzan una variable con ella misma, siendo los elementos de la diagonal los efectivos de cada modalidad $k_{\bullet j}$. Los bloques que no se encuentran en la diagonal representan tablas de contingencia que se obtienen cruzando las tablas de características de dos en dos $Z'_k Z_k$ cuyos elementos son las frecuencias de asociación de las dos modalidades correspondientes (Pérez, 2004) [20].

La tabla de Burt en forma general se puede representar como:

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

Figura 3: Tabla de Burt

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

Para representar al mismo tiempo los puntos línea y los puntos columna sobre los mismos gráficos, de tal manera que se relacionen los resultados en los dos subespacios, se utilizan las siguientes fórmulas de transición: (Pérez, 2004) [20].

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

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

Donde:

- $F_\alpha(i)$ (proyección de un punto individuo i sobre el eje α) representa el baricentro de las proyecciones de los puntos modalidades sobre el eje $G_\alpha(j)$. Dichas modalidades se encuentran influenciadas por un peso 1/Q.
- $G_\alpha(j)$ (proyección de un punto modalidad j sobre el eje α) representa el baricentro de las proyecciones de los puntos individuos que tienen esa modalidad sobre el eje $F_\alpha(i)$, los cuales se ven influenciados por el peso $k_{\bullet j}$.

La nube de puntos variables representada por N(j), tiene como centro de gravedad a $\sqrt{f_{i\bullet}}$, la que puede asemejarse a una distribución uniforme $\frac{1}{\sqrt{n}}$, porque $k_{i\bullet} = \sum_j k_{ij} = Q \rightarrow \sum_i k_{i\bullet} = nQ \rightarrow f_{i\bullet} = \frac{1}{n}$.

Las modalidades o categorías de cada variable, las cuales están ponderadas por su peso, tienen centro de gravedad igual a $\frac{1}{\sqrt{n}}$, que es el mismo que el de la nube de modalidades N(j), puesto que el centro de gravedad de la subtabla I × J_k se calcula a través de su distribución marginal. Debido a que se escoge sólo una variable, la suma en cada línea es igual a 1, y como el total en la tabla es n, entonces se tiene que $f_{i\bullet} = \frac{1}{n}$. (Pérez, 2004) [20].

Para ayudar a la interpretación de cada fila y columna, se calcula la contribución de una variable J_k asociada al factor α, la cual es definida como la suma de las contribuciones de las modalidades o categorías de la variable, la cual queda expresada como $CTA_{\alpha}(J_k) = \sum_{j \in J_k} CTA_{\alpha}(j)$.

La inercia debida a la modalidad o categoría j, dado el centro de gravedad G, se calcula por:

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

La inercia de una variable J_k se define como la suma de las inercias de sus correspondientes modalidades, y se calcula por:

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

La inercia total I se define como la suma de las inercias de las modalidades correspondientes, y se calcula por:

$$I=\sum_k I(J_k)=\sum_k \frac{1}{Q}(m_k-1)=\frac{J}{Q}-1$$

En esta expresión el cociente J/Q indica la media del número de modalidades o categorías por variable cualitativa, por lo cual la inercia total depende solamente de la cantidad de modalidades y de preguntas.

Si nos encontramos en presencia de dos variables, cada una con dos modalidades, se pueden analizar los resultados a través del Análisis Factorial de Correspondencias (AFC), o por Análisis de Correspondencias Múltiple (ACM).

Cuando se usa el primer análisis, se obtiene un solo factor que recolecta el 100% de la inercia total, de tal manera que, si las modalidades están muy relacionadas, la inercia debería tener un valor elevado; por el contrario, si este valor es muy cercano a cero, indicaría que las modalidades están poco relacionadas.

En cambio, si se utiliza el segundo análisis, se obtendrían dos ejes, pero la inercia obtenida será siempre la misma, es decir, $\frac{J}{Q}-1=1$. Cuando el primer eje recolecta gran parte de la inercia (un valor muy próximo a 1) y el segundo muy poca, es un indicador de que existe mucha relación entre las variables; por el contrario, si los dos factores recolectan un valor de inercia igual a 1/2 para cada uno, es un indicador de que existe total independencia entre las variables (Pérez, 2004) [20].

2.5. REGRESIÓN LOGÍSTICA

En muchas situaciones de la vida real, existen circunstancias en las cuales un investigador está interesado en predecir si un evento ocurre o no, en función de algunas variables explicativas. Se puede encontrar que, en algunas ocasiones, todas las variables de interés no se puedan representar de manera cuantitativa, lo cual no haría factible que se encuentre un modelo que estudie la relación de una variable dependiente con dos o más variables independientes, todas ellas cuantitativas, como en la regresión lineal múltiple (Luque, 2000) [21].

Supongamos que disponemos de un conjunto de variables, de las cuales una es dicotómica, que se considera como la variable dependiente o

variable respuesta, y además se tienen dos o más variables cualitativas o cuantitativas, que se consideran como variables independientes. Si se desea estudiar la relación entre esta variable dependiente y las demás variables independientes, se aplica una técnica de análisis multivariante conocida como regresión logística múltiple.

En la regresión logística múltiple, la variable dependiente, se dice que es dicotómica o binaria, porque esta sólo puede tomar dos valores 0 o 1, que definen características opuestas o mutuamente excluyentes, por ejemplo: ser o no propenso a sufrir un accidente de tránsito. Además, en el modelo para cada variable independiente cualitativa, se tendrán que generar variables DUMMY, por ejemplo, si se tiene una variable cualitativa con K categorías, se deberán generar K – 1 variables DUMMY, cada una de las cuales tomará el valor 0 o 1, es decir, que la variable DUMMY tomará el valor de 1 si un individuo pertenece a una determinada categoría de la variable cualitativa y tomará el valor de 0 en caso contrario, de tal manera que todas las categorías de la variable cualitativa puedan ser representadas en el modelo (Álvarez, 2008) [22].

En la regresión logística múltiple además de encontrar la relación entre la variable dependiente y las demás variables independientes, se busca determinar la medida de dicha relación y estimar la probabilidad de que ocurra o no el evento que define la variable dependiente en función de los valores que tomen las variables independientes (Luque, 2000) [21].

Una técnica multivariante que se podría utilizar en algunas situaciones parecidas a las descritas, es el análisis discriminante, pero este sólo admite variables cuantitativas. En nuestro caso, al poder combinarse tanto variables cuantitativas como cualitativas, se debe usar la regresión logística múltiple.

2.6. MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

Sean, y una variable dependiente, que toma valores que sólo pueden ser: 1 con probabilidad p o 0 con probabilidad $1-p$, y $x_1, x_2, \cdots, x_n$ las variables independientes, que pueden ser cuantitativas o cualitativas.

Un modelo de regresión logística múltiple busca estimar la probabilidad de que ocurra un determinado evento, es decir, la probabilidad de que un individuo i elija una respuesta binaria (1 o 0) en función de los valores que tomen las variables independientes, $x_{1i}, x_{2i}, \cdots, x_{ni}$ (Luque, 2000) [21].

La expresión matemática del modelo de regresión logística múltiple está dado por:

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

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 equivalente, 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 poder realizar el ajuste de este modelo y la estimación de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ se usa el método de estimación por máxima verosimilitud (EMV).

En primer lugar, construye una función L, que se denomina función de verosimilitud, la cual es usada para expresar la probabilidad de los datos observados como una función de parámetros desconocidos. Los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ que permiten maximizar la función L, son los llamados estimadores maximoverosímiles de los parámetros $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$. Es decir, que este método calcula los valores

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ que son estimadores de los parámetros desconocidos $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$, y son tales que maximizan la probabilidad de que con ellos se puedan obtener los valores observados (Luque, 2000) [21].

Cuando ya se haya ajustado el modelo y encontrado los estimadores maximoverosímiles $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$, la estimación de la probabilidad $\hat{p}$ viene dada por:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}}$$, que estima la probabilidad de que un individuo elija la respuesta binaria 1 dado un determinado valor de las variables independientes $x_{1i}, x_{2i}, \cdots, x_{ni}$.

Expresiones equivalentes a la anterior son:

$$\frac{\hat{p}_i}{1 - \hat{p}_i} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni}}$$, que se denomina “odds” o “ventaja”, y se define como el cociente entre la probabilidad de que ocurra el evento de interés, que en regresión logística es siempre $y_i = 1$, y la probabilidad de que dicho evento no ocurra, es decir, estima la ventaja o preferencia de un individuo por la respuesta 1 de la variable dependiente frente a la respuesta 0 para cada valor de las variables independientes (Luque, 2000) [21].

Si se toma el logaritmo natural al odds, se obtiene un modelo lineal que se denomina transformación logística o simplemente modelo logit, a diferencia de la probabilidad p_i, que constituye un modelo no lineal o modelo logístico. Este modelo en función del logaritmo del odds, es importante porque permite que los coeficientes del modelo se puedan interpretar sencillamente en términos de independencia o asociación entre variables (Pérez, 2004) [20].

El modelo logit, se puede expresar como:

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

2.6.1. PRUEBA DE SIGNIFICANCIA

La prueba de significancia en la regresión logística múltiple es parecida a la prueba de significancia en la regresión múltiple. En primer lugar, se realiza una prueba para probar la significancia global del modelo, por lo cual se deben plantear y contrastar hipótesis sobre los coeficientes de regresión en forma conjunta. Esto permite verificar si las variables independientes que están presentes en el modelo ajustado están relacionadas significativamente con la variable dependiente (Luque, 2000) [21].

Las hipótesis para probar la significancia global son las siguientes:

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

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

Esta prueba de significancia global se basa en el valor del estadístico de prueba G que se define como:

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

El estadístico G se distribuye como una ji-cuadrada χ^2 con p-1 grados de libertad, donde p es el número de parámetros en el modelo bajo estudio.

El estadístico G está basado en la función de verosimilitud de cada modelo y permite comparar la probabilidad de que los valores estimados por cada modelo representen a los valores observados de la variable dependiente (Luque, 2000) [21].

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

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 concluye que el modelo global es significativo.

Si después de haber realizado la prueba G, esta indicó que sí existe una significancia global, lo siguiente a realizar es determinar si la contribución que hace cada variable independiente al modelo es significativa, para lo cual se utiliza una prueba de significancia individual para cada coeficiente de regresión, que está basada en el estadístico W de Wald (Luque, 2000) [21].

Las hipótesis para probar la significancia individual son:

$$H_0 : \beta_i = 0$$

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

El valor del estadístico de prueba W se define como:

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

Este estadístico es igual al cuadrado de la razón entre el estimador del coeficiente de la variable independiente y el estimador de su error estándar, y sigue una distribución una ji-cuadrada χ^2 con 1 grado de libertad si la variable independiente es cuantitativa, pero si la variable independiente es cualitativa, el número de grados de libertad es igual al número de categorías menos uno (Luque, 2000) [21].

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

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

Cuando nos referimos a la bondad de ajuste de un modelo, tratamos de realizar aquellas pruebas que permiten evaluar el grado de efectividad absoluta del modelo en consideración, en relación con una descripción adecuada de la variable dependiente. Es decir, estas pruebas buscan determinar cuán cercanos se encuentran los valores estimados de los valores observados (Luque, 2000) [21].

Existen dos formas de medir la bondad de ajuste de un modelo, las cuales están basadas en contraste de hipótesis y en medidas similares al coeficiente de determinación R^2 usado en regresión lineal múltiple.

2.6.2.1. BONDAD DEL AJUSTE USANDO CONTRASTE DE HIPÓTESIS

En estas medidas se contrasta la hipótesis nula H_0, la cual plantea que el modelo que el investigador ha seleccionado se ajusta bien a los datos, a través de un estadístico que sigue una distribución conocida.

Es decir:

H_0: El modelo seleccionado ajusta bien a los datos

H_1: $\neg H_0$

Las medidas de bondad del ajuste son:

2.6.2.1.1. DESVIANZA (DEVIANCE)

La desvianza se calcula a través del estadístico de prueba D, que se define como:

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

El estadístico D sigue una distribución una ji-cuadrada χ^2 con N - p grados de libertad, donde N es el número de observaciones y p el número de parámetros presentes en el modelo. En los

paquetes estadísticos se suele utilizar la expresión -2 In likelihood o -2 Logaritmo de la verosimilitud para representar a la desvianza de un determinado modelo.

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla 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.

Si no se rechaza la hipótesis nula, se concluye que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.2. PRUEBA DE LA JI-CUADRADO

Se basa en el cálculo de un estadístico ji-cuadrado χ^2, el cual mide el grado de discordancia que podría existir si se comparan el número observado de respuestas afirmativas con la estimación de la probabilidad que hace el modelo. Estas comparaciones se realizan para cada uno de los patrones de predictores presentes, es decir, a cada una de las diferentes combinaciones de valores que pueden tomar las variables independientes incluidas en el modelo (Luque, 2000) [21].

Por ejemplo, sean las variables cualitativas: Sexo (1 =varón; 0= mujer) y Religión (1 = católico; 0= musulmán), estas determinan cuatro patrones de covariables, ya que cada uno de los individuos que pertenecen a la muestra se pueden clasificar en cualquiera de los siguientes grupos o patrones: varón-católico; varón - musulmán; mujer-católica; y mujer-musulmana.

Si el número M de patrones de predictores es menor que el número N de observaciones, el estadístico χ^2 se define como:

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

Donde:

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

y_i : es el valor de la respuesta binaria en la variable dependiente.

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

Si se tienen muestras grandes, el estadístico χ^2 sigue una distribución ji-cuadrada con M - p grados de libertad.

Cuando se tengan variables cuantitativas continuas, es muy probable que $M \approx N$, por lo cual, el estadístico χ^2 se definiría como:

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

Un inconveniente en este caso, es que se pueden obtener valores p erróneos, no obstante, en aquellos casos en los cuales el modelo ajustado sea el correcto, la prueba χ^2 se podría utilizar con N-p grados de libertad, obteniéndose de esta manera resultados adecuados.

Para poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

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 el número de patrones de predictores, N es el número de observaciones, p el número de

parámetros presentes en el modelo, $M-p$ y $N-p$ son los grados de libertad.

Si no se rechaza la hipótesis nula, se concluye que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.3. PRUEBA DE HOSMER-LEMESHOW

Es conveniente aplicarla en los modelos que contengan una o varias variables independientes que sean continuas y cuyo número de patrones de predictores sea aproximadamente igual al número de casos observados, es decir, $\mathrm{M} \approx \mathrm{N}$.

Primero se calculan las probabilidades estimadas $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_{\mathrm{N}}$ luego se ordenan de menor a mayor estas probabilidades (una para cada caso observado) y a continuación se agrupan en diez grupos, de tal manera, que en el primer grupo se encuentren los $\mathrm{n}_1 = \mathrm{N}/10$ individuos que tengan las probabilidades estimadas más bajas, y así sucesivamente hasta que en el último grupo los $\mathrm{n}_{10} = \mathrm{N}/10$ individuos tengan las probabilidades estimadas más altas. Los grupos mencionados se conocen como deciles de riesgo.

El estadístico utilizado para esta prueba es, $\hat{C}$, y se lo calcula a través de una tabla de 2 x 10, que contenga a las frecuencias observadas y estimadas para cada uno de los diez grupos (Luque, 2000) [21].

El estadístico $\hat{C}$, sigue una distribución 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 poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

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.

Si no se rechaza la hipótesis nula, se concluye que el modelo seleccionado ajusta bien a los datos.

2.6.2.1.4. BONDAD DEL AJUSTE: EFICACIA PREDICTIVA

Esta prueba se basa en comparar las predicciones del modelo seleccionado con los datos observados, a través de una tabla de clasificación. Esta tabla es de doble entrada y organiza los casos que forman la muestra de acuerdo a los valores observados de la variable dependiente y a los valores pronosticados por el modelo estimado, de tal forma que, a todos los casos para los cuales la probabilidad estimada sea mayor o igual a 0.5 (valor de corte), los clasificará dentro del grupo de los que posean la característica representada por la variable dependiente (1); y en los casos donde la probabilidad estimada es menor que 0.5, los clasificará

dentro del grupo de los que no poseen la característica representada por la variable dependiente (0) (Luque, 2000) [21].

La tabla de clasificación mencionada se puede representar como:

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 6:Tabla de clasificación

Fuente: Datos de la investigación - Autor: Johnny Jiménez.

Donde:

a y d son los casos que el modelo clasificó correctamente.

b y c los casos que el modelo clasificó erróneamente.

Con la tabla de clasificación 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 determinar que el modelo tiene una buena eficacia predictiva se debe comprobar la significación estadística de la tasa global de aciertos, para lo cual, por medio del test de Huberty, primero se calcula 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.

H_1: $\neg H_0$

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 poder realizar el contraste de hipótesis de esta prueba, se aplica el método del valor crítico que genera la siguiente regla de rechazo:

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

Donde α es el nivel de significancia.

Si se rechaza la hipótesis nula, se concluye 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

Existen otras medidas similares a R^2 usado en regresión lineal que permiten medir la bondad de ajuste de un modelo de regresión logística.

Estas medidas 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

Es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente

explicada por las variables independientes. Los valores fluctúan entre 0 y 1.

Se basa en la comparación del logaritmo de la verosimilitud para el modelo completo respecto al logaritmo de la verosimilitud para el modelo reducido que sólo incluye el 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 cercano a 1, es un indicativo de un buen ajuste.

2.6.2.2.2. R CUADRADO DE NAGELKERKE

Es una forma corregida de la R cuadrado de Cox y Snell. Posee un valor máximo que es menor a 1. Además, corrige la escala del estadístico para abarcar 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 cercano a 1, es un indicativo de un buen ajuste.

CAPÍTULO III

3. ANÁLISIS ESTADÍSTICO UNIVARIADO Y TABLAS DE CONTINGENCIA

3.1. INTRODUCCIÓN

En este capítulo se realizará la primera parte del análisis estadístico, que se centrará en el análisis estadístico univariado y análisis de tablas de contingencia. El análisis se aplicará a una muestra de 62474 vehículos, que fue proporcionada por una aseguradora de la ciudad de Quito.

La muestra proporcionada por la aseguradora, presenta ciertas características o variables asociadas a los vehículos, que serán estudiadas para realizar los análisis estadísticos descritos en el capítulo 1.

La mayoría de las variables que se utilizan son cualitativas. La descripción para las variables cuantitativas se realizará a través del cálculo de medidas de tendencia central, medidas de dispersión, medidas de asimetría y medidas de curtosis, tablas de frecuencias, histogramas y diagramas de caja; en cambio para las variables cualitativas se realizará por medio de tablas de frecuencias, diagramas de barras y diagramas de pastel.

Antes de realizar los análisis estadísticos es importante conocer información más específica sobre las variables, por lo cual, se realizará la descripción y codificación de las mismas, permitiendo viabilizar la aplicación de los análisis estadísticos propuestos anteriormente.

3.2. DESCRIPCIÓN Y CODIFICACIÓN DE LAS VARIABLES DE LA INVESTIGACIÓN

Primeramente, se realizará la descripción y codificación de las variables cualitativas.

Variable: Categoría

Descripción:

Variable cualitativa dicotómica que permite determinar la categoría del vehículo bajo estudio.

La variable puede tomar dos valores (categorías) posibles, los cuales se codifican como:

Codificación:

Codificación	Categoría
0	LIVIANO
1	PESADO

Tabla 7: Variable: Categoría

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Variable: Marca

Descripción:

Variable cualitativa, de carácter nominal que permite determinar la marca del vehículo bajo estudio.

La variable puede tomar los siguientes valores (categorías) posibles, los cuales se los codifica como:

Codificación:

Codificación	Marca
0	CHEVROLET
1	HYUNDAI
2	KIA
3	TOYOTA
4	NISSAN
5	SUZUKI

6	FORD
7	VOLKSWAGEN
8	MAZDA
9	RENAULT
10	GREAT WALL
11	MITSUBISHI
12	HONDA
13	HINO
14	SKODA
15	OTRA

Tabla 8: Variable: Marca

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Variable: Tipo

Descripción:

Variable cualitativa, de carácter nominal que permite determinar el tipo del vehículo bajo estudio.

La variable puede tomar los siguientes valores (categorías) posibles, los cuales se los codifica como:

Codificación:

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

Tabla 9: Variable: Tipo

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Variable: Color

Descripción:

Variable cualitativa, de carácter nominal que permite determinar el color del vehículo bajo estudio.

La variable puede tomar los siguientes valores (categorías) posibles, los cuales se los codifica como:

Codificación:

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

Tabla 10: Variable: Color del vehículo

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Variable: Siniestro

Descripción:

Variable dependiente o respuesta, de naturaleza cualitativa dicotómica, que permite determinar si el vehículo bajo estudio sufrió o no un siniestro.

La variable puede tomar dos valores posibles, los cuales se codifican como:

Codificación:

Codificación	Siniestro
0	No hubo Siniestro
1	Hubo Siniestro

Tabla 11: Variable: Siniestro

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Finalmente, se realizará la descripción y codificación de la única variable cuantitativa que está presente en la investigación.

Variable Año

Descripción:

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

Codificación:

Esta variable se codifica para convertirla en cualitativa, con la finalidad de poder realizar el análisis de las tablas de contingencia, ya que para el análisis univariado la variable se trabaja sin codificación alguna.
Para la codificación se dividirá el conjunto de observaciones en 4 partes iguales, cada uno conteniendo un 25% de los datos, por lo cual es necesario el cálculo de los estadísticos: mínimo, percentil 25, percentil 50, percentil 75 y máximo.
De esta manera se pueden formar intervalos, dentro de los cuales se clasifican a los datos.

La variable puede tomar cuatro valores posibles, los cuales se codifican como:

Codificación	Año
1	[mínimo, Percentil 25)
2	[Percentil 25, Percentil 50)

3	[Percentil 50, Percentil 75)
4	[Percentil 75, máximo]

Tabla 12: Variable: Año

Fuente: Datos de la investigación - Autor: Johnny Jiménez

3.3. ANÁLISIS ESTADÍSTICO UNIVARIADO DE LAS VARIABLES DE LA INVESTIGACIÓN

Variable: Categoría

De los 62474 vehículos, la Tabla 13 muestra que los vehículos livianos tienen la mayor participación con un 96,55% del total de vehículos asegurados en la ciudad de Quito y en un porcentaje muy bajo corresponde a los vehículos pesados, con una participación del 3,45%. La Figura 4 permite observar con mayor detalle las frecuencias relativas de la variable Categoría.

Categoría	Frecuencia Absoluta	Frecuencia Relativa
LIVIANO	60319	96,55%
PESADO	2155	3,45%
Total general	**62474**	**100%**

Tabla 13: Categoría del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

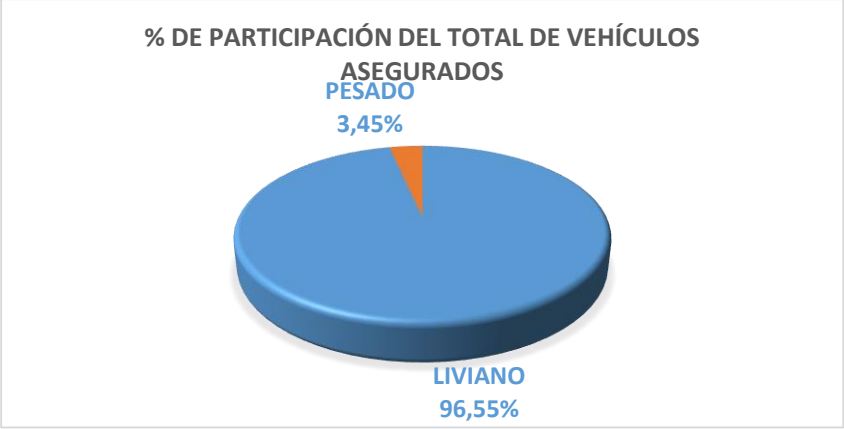


Figura 4: Categoría del Vehículo

Fuente: Datos de la investigación - Autor: Johnny Jiménez

Variable: Marca

De los 62474 vehículos, la Tabla 14 muestra que los vehículos de la marca Chevrolet tienen la mayor participación con un 33,26% del total de vehículos asegurados en la ciudad de Quito, le siguen los vehículos de la marca Hyundai con una participación del 12,31%, en tercer lugar, se encuentran los vehículos de marca Kia, con un 9,49%. Entre las tres marcas acumulan más del 50% de participación en el mercado. En la categoría **OTRA,** se encuentran aquellas marcas que individualmente tuvieron una participación muy baja, pero que en conjunto representan un 6,99% del total. La Figura 5 permite observar con mayor detalle las frecuencias relativas de la variable Marca.

Marca	Frecuencia Absoluta	Frecuencia Relativa
CHEVROLET	20776	33,26%
HYUNDAI	7689	12,31%
KIA	5928	9,49%
TOYOTA	4906	7,85%
NISSAN	3266	5,23%
SUZUKI	2955	4,73%
FORD	2568	4,11%
VOLKSWAGEN	2534	4,06%
MAZDA	2271	3,64%
RENAULT	2092	3,35%
GREAT WALL	888	1,42%
MITSUBISHI	862	1,38%
HONDA	569	0,91%
HINO	432	0,69%
SKODA	374	0,60%
OTRA	4364	6,99%
Total General	**62474**	**100%**

Tabla 14: Marca del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

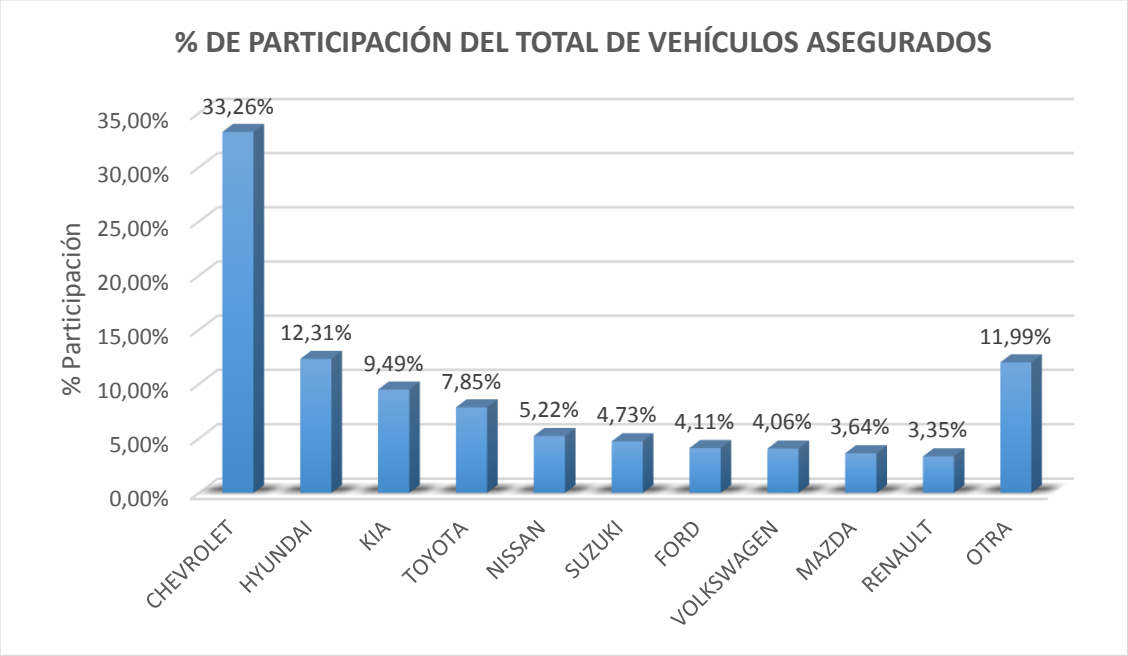


Figura 5: Marca del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Variable: Tipo

De los 62474 vehículos, la Tabla 15 muestra que los vehículos de tipo sedán tienen la mayor participación con un 43,13% del total de vehículos asegurados en la ciudad de Quito, le siguen los vehículos de tipo todoterreno con una participación del 35,44%, en tercer lugar, se encuentran los vehículos de tipo camioneta, con un 14,24%. Entre los tres tipos de vehículos acumulan más del 90% de participación en el mercado. En la categoría **OTRO,** se encuentran aquellos tipos de vehículos que no pertenecen a ninguna de las demás categorías de la variable de estudio, y que individualmente tuvieron una participación muy baja, pero que en conjunto representan un 2,02% del total. La Figura 6 permite observar con mayor detalle las frecuencias relativas de la variable Tipo.

Tipo	Frecuencia Absoluta	Frecuencia Relativa
SEDÁN	26947	43,13%
TODOTERRENO	22141	35,44%
CAMIONETA	8896	14,24%
CAMIÓN	1896	3,03%

MOTO	1335	2,14%
OTRO	1259	2,02%
Total General	**62474**	**100%**

Tabla 15: Tipo de Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

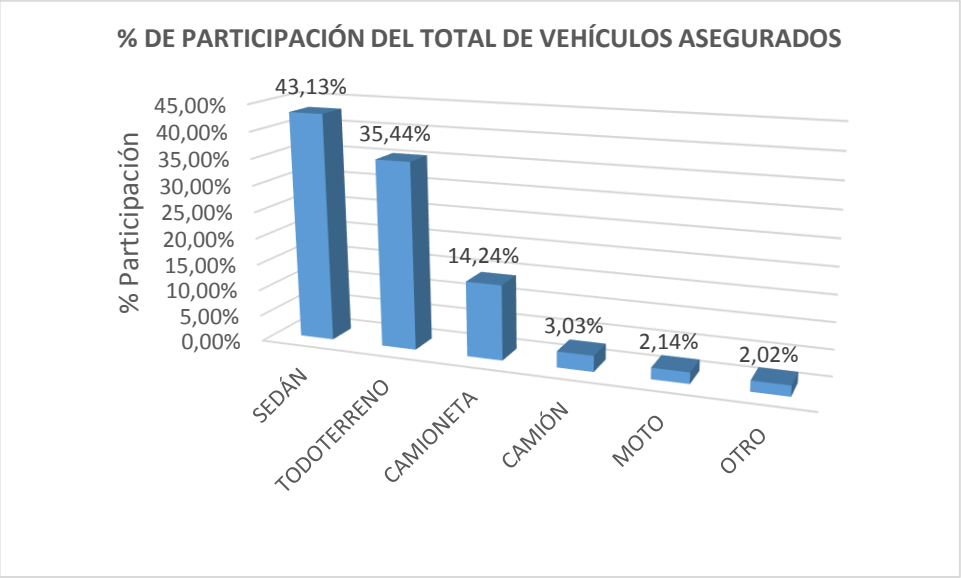


Figura 6: Tipo de Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Variable: Color

De los 62474 vehículos, la Tabla 16 muestra que los vehículos de color plateado tienen la mayor participación con un 19,93% del total de vehículos asegurados en la ciudad de Quito, le siguen los vehículos de color blanco con una participación del 19,24%, en tercer lugar, se encuentran los vehículos de color plomo, con un 15,30%. Estos tres colores usados por los vehículos acumulan más del 50% de participación en el mercado. En la categoría **OTRO,** se encuentran aquellos colores de vehículos y sus combinaciones que no pertenecen a ninguna de las demás categorías de la variable de estudio, que individualmente tuvieron una participación muy baja, pero que en conjunto representan un 1,53% del total. La Figura 7 permite observar con mayor detalle las frecuencias relativas de la variable Color.

Color	Frecuencia Absoluta	Frecuencia Relativa
PLATEADO	12452	19,93%
BLANCO	12023	19,24%
PLOMO	9560	15,30%
NEGRO	6523	10,44%
ROJO	6449	10,32%
AZUL	4335	6,94%
BEIGE	2058	3,29%
DORADO	2037	3,26%
VERDE	1879	3,01%
CONCHO DE VINO	1702	2,72%
GRIS	1532	2,45%
CELESTE	580	0,93%
AMARILLO	378	0,61%
ACERO	13	0,02%
OTRO	953	1,53%
Total General	**62474**	**100%**

Tabla 16: Color del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

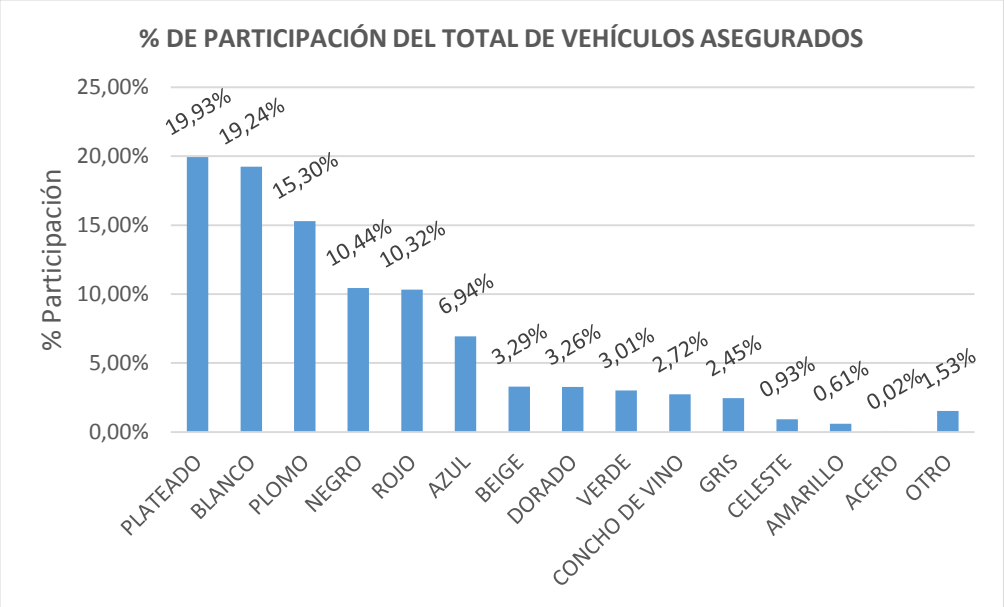


Figura 7: Color del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Variable: Siniestro

De los 62474 vehículos, la Tabla 17 muestra 87% de total de vehículos asegurados en la ciudad de Quito no sufrieron un siniestro, mientras que el 13% corresponde a los vehículos que sufrieron un siniestro. La Figura 8 permite observar con mayor detalle las frecuencias relativas de la variable Siniestro.

Siniestro	Frecuencia Absoluta	Frecuencia Relativa
No hubo Siniestro	54352	87%
Hubo Siniestro	8122	13%
Total general	62474	100%

Tabla 17: Siniestro del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

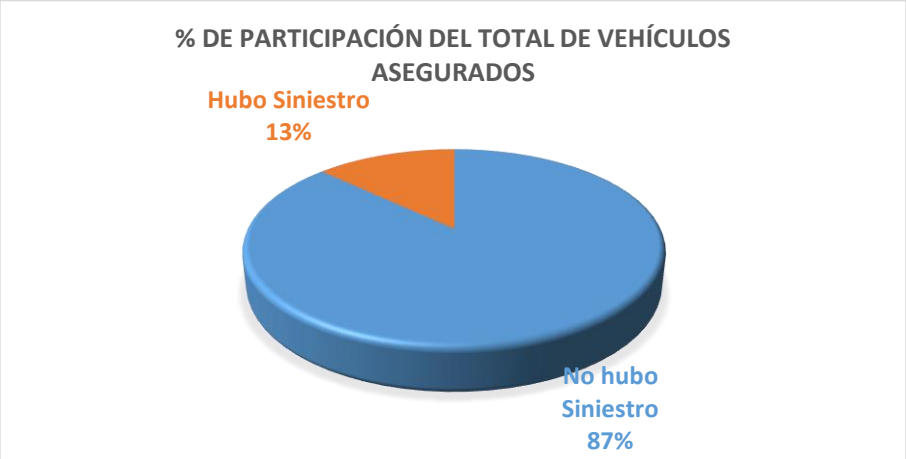


Figura 8: Siniestro del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Variable Año

De los 62474 vehículos, la Tabla 18 muestra que los vehículos asegurados en la ciudad de Quito tienen a 2009.53 como año promedio de fabricación, el año de fabricación que más se repite es el año 2013, la desviación estándar es 4.35, la distribución es asimétrica negativa, tiene cola a la izquierda, ya que su coeficiente de asimetría es -1.90, lo cual se puede apreciar en el histograma de la Figura 9. La distribución es Leptocúrtica, es

decir, es más puntiaguda que la distribución normal, porque su coeficiente de curtosis es 7.29.

n		62474
Media		2009.53
Mediana		2011.00
Moda		2013.00
Desviación estándar		4.35
Varianza		18.95
Asimetría		-1.90
Curtosis		7.29
Rango		66
Mínimo		1949
Máximo		2015
Percentiles	25	2007
	50	2011
	75	2013

Tabla 18: Año de fabricación del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Además, en la Tabla 18 se muestran los percentiles 25, 50 y 75, los cuales indican que un 25% de los vehículos tienen su año de fabricación menor a 2007, el 50% de los vehículos fueron fabricados entre 2007 y 2013 y un 25% de los vehículos fueron fabricados después del año 2013.

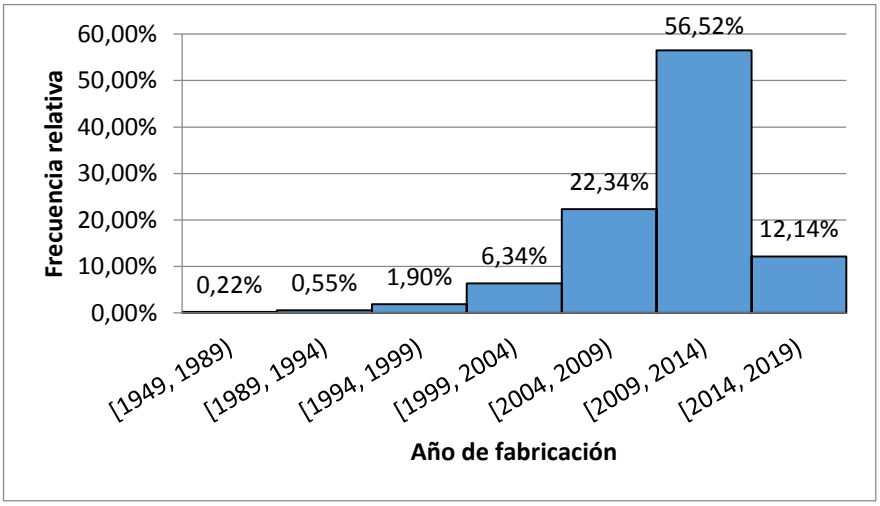


Figura 9: Histograma - Año de fabricación del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

En la Figura 10 se representan los percentiles mediante un Diagrama de Caja, en el cual se puede apreciar que el 25% de los vehículos que fueron fabricados antes del 2007 son más dispersos que el 25% de los vehículos que fueron fabricados después del 2013. Además, se puede apreciar una alta concentración de vehículos fabricados entre 2007 y 2013.

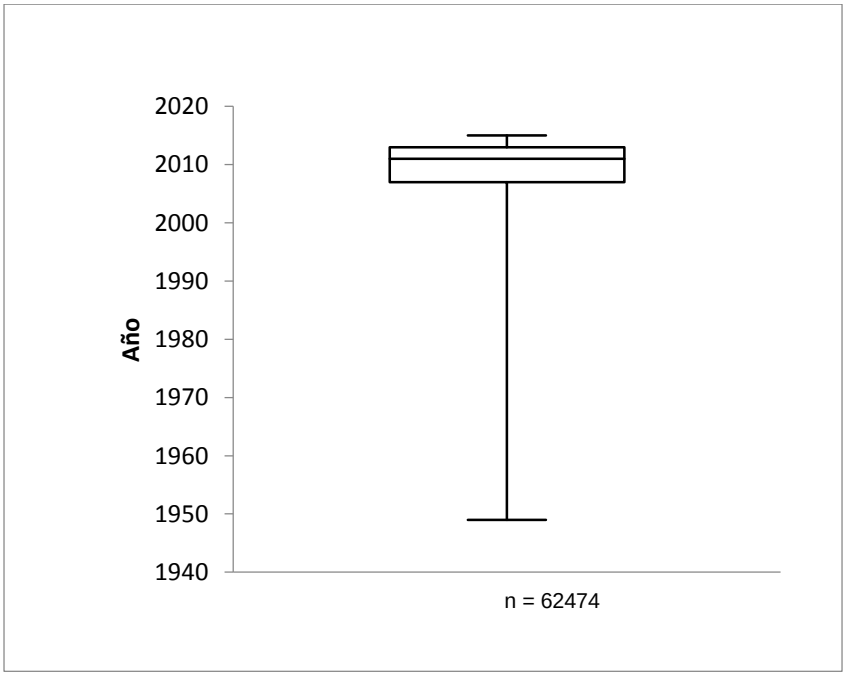


Figura 10: Diagrama de Cajas - Año de fabricación del Vehículo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

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. Se analizará la posible relación entre la variable Siniestro con las demás variables presentadas anteriormente.

Siniestro vs Categoría

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

H_0: El siniestro del vehículo es independiente de la categoría del vehículo

vs.

H_1: No es verdad H_0

En la Tabla 19 se presentan estos resultados, donde se observa que el valor del estadístico de prueba es 82.29, además el valor-p de la prueba es menor a 0.05, lo cual 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**	52338	7981	60319
	PESADO	2014	141	2155
Total		54352	8122	62474
Resultado Prueba Chi-cuadrado				
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**	
Chi-cuadrado de Pearson	82.29	1	0	

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

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Siniestro vs Marca

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

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

vs.

H_1: No es verdad H_o

En la Tabla 20 se presentan estos resultados, donde se observa que el valor del estadístico de prueba es 323.71, además el valor-p de la prueba es menor a 0.05, lo cual 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	18424	2352	20776
	FORD	2195	373	2568
	GREAT WALL	718	170	888
	HINO	397	35	432
	HONDA	492	77	569
	HYUNDAI	6593	1096	7689
	KIA	4943	985	5928
	MAZDA	1941	330	2271
	MITSUBISHI	773	89	862
	NISSAN	2773	493	3266
	RENAULT	1752	340	2092
	SKODA	333	41	374
	SUZUKI	2635	320	2955
	TOYOTA	4251	655	4906
	VOLKSWAGEN	2140	394	2534
	OTRA	3992	372	4364
Total		54352	8122	62474
Resultado Prueba Chi-cuadrado				
Estadístico de Prueba	Valor	Grados de Libertad	Valor-p	

Chi-cuadrado de Pearson	323.71	15	0

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

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Siniestro vs Tipo

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

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

vs.

H_1: No es verdad H_o

En la Tabla 21 se presentan estos resultados, donde se observa que el valor del estadístico de prueba es 232.02, además el valor-p de la prueba es menor a 0.05, lo cual 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	1749	147	1896
	CAMIONETA	7792	1104	8896
	MOTO	1285	50	1335
	SEDÁN	23067	3880	26947
	TODOTERRENO	19293	2848	22141
	OTRO	1166	93	1259
Total		54352	8122	62474
Resultado Prueba Chi-cuadrado				
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**	
Chi-cuadrado de Pearson	232.02	5	0	

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

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Siniestro vs Color

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

H_o: El siniestro del vehículo es independiente del color de vehículo

vs.

H_1: No es verdad H_o

En la Tabla 22 se presentan estos resultados, donde se observa que el valor del estadístico de prueba es 106.64, además el valor-p de la prueba es menor a 0.05, lo cual 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	9	4	13
	AMARILLO	346	32	378
	AZUL	3823	512	4335
	BEIGE	1766	292	2058
	BLANCO	10620	1403	12023
	CELESTE	505	75	580
	CONCHO DE VINO	1442	260	1702
	DORADO	1683	354	2037
	GRIS	1360	172	1532
	NEGRO	5597	926	6523
	PLATEADO	10816	1636	12452
	PLOMO	8234	1326	9560
	ROJO	5666	783	6449
	VERDE	1636	243	1879
	OTRO	849	104	953
Total		54352	8122	62474
Resultado Prueba Chi-cuadrado				
Estadístico de Prueba	Valor	Grados de Libertad	Valor-p	
Chi-cuadrado de Pearson	106.64	14	0	

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

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Siniestro vs Año

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

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

vs.

H_1: No es verdad H_o

En la Tabla 23 se presentan estos resultados, donde se observa que el valor del estadístico de prueba es 171.93, además el valor-p de la prueba es menor a 0.05, lo cual significa que existe evidencia estadística para rechazar la hipótesis nula, por lo tanto, el siniestro del vehículo es dependiente de su año de fabricación.

VEHÍCULO		**SINIESTRO**		**Total**
		No hubo Siniestro	**Hubo Siniestro**	
Años agrupados por cuartil	**1**	14233	1597	15830
	2	12241	1852	14093
	3	13555	2260	15815
	4	14323	2413	16736
Total		54352	8122	62474
Resultado Prueba Chi-cuadrado				
Estadístico de Prueba	**Valor**	**Grados de Libertad**	**Valor-p**	
Chi-cuadrado de Pearson	171.93	3	0	

Tabla 23: Tabla de Contingencia y Prueba Chi-cuadrado - Siniestro vs Año

Fuente: Datos de la investigación. Autor: Johnny Jiménez

CAPÍTULO IV

4. ANÁLISIS ESTADÍSTICO MULTIVARIADO

4.1. INTRODUCCIÓN

En este capítulo se realizará la parte final del análisis estadístico, que se centrará en el análisis estadístico multivariado que comprende el Análisis de Correspondencias múltiples y Análisis de Regresión Logística.

En esta investigación, debido a que las variables Marca, Color y Tipo tienen muchas categorías, y con la finalidad de favorecer la interpretación de los resultados, estas variables se redefinirán, reagrupando las categorías originales en nuevas categorías, las cuales se asocian de acuerdo a la frecuencia de siniestralidad que presentaron las categorías originales.

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

Por último, la variable Año, tal y como se mencionó en el capítulo 3 se la recodificará, agrupando los años de fabricación de acuerdo al percentil correspondiente, ya que para el análisis multivariado 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.

En las tablas 24, 25 y 26 se presentan con mayor detalle, la recategorización y recodificación de las variables anteriormente mencionadas.

MARCA					
Codificación	**Siniestralidad**	**Intervalo**	**Marca**	**Frecuencia**	**Frecuencia relativa**
GM1	ALTA	f ≥900	CHEVROLET	2352	28,96%
			HYUNDAI	1096	13,49%
			KIA	985	12,13%
GM2	MEDIA	500≤ f <900	TOYOTA	655	8,06%
			NISSAN	493	6,07%

GM3	BAJA	200≤ f <500	VOLKSWAGEN	394	4,85%
			FORD	373	4,59%
			OTRA	372	4,58%
			RENAULT	340	4,19%
			MAZDA	330	4,06%
			SUZUKI	320	3,94%
GM4	MUY BAJA	f<200	GREAT WALL	170	2,09%
			MITSUBISHI	89	1,10%
			HONDA	77	0,95%
			SKODA	41	0,50%
			HINO	35	0,43%

Tabla 24: Recategorización de la variable Marca

Fuente: Datos de la investigación. Autor: Johnny Jiménez

COLOR					
Codificación	**Siniestralidad**	**Intervalo**	**Color**	**Frecuencia**	**Frecuencia relativa**
GC1	ALTA	f ≥900	PLATEADO	1636	20,14%
			BLANCO	1403	17,27%
			PLOMO	1326	16,33%
			NEGRO	926	11,40%
GC2	MEDIA	500≤ f <900	ROJO	783	9,64%
			AZUL	512	6,30%
GC3	BAJA	200≤ f <500	DORADO	354	4,36%
			BEIGE	292	3,60%
			CONCHO DE VINO	260	3,20%
			VERDE	243	2,99%
GC4	MUY BAJA	f<200	GRIS	172	2,12%
			OTRO	104	1,28%
			CELESTE	75	0,92%
			AMARILLO	32	0,39%
			ACERO	4	0,05%

Tabla 25: Recategorización de la variable Color

Fuente: Datos de la investigación. Autor: Johnny Jiménez

TIPO					
Codificación	**Siniestralidad**	**Intervalo**	**TIPO**	**Frecuencia**	**Frecuencia relativa**
GT1	ALTA	f≥2000	SEDÁN	3880	47,77%
			TODOTERRENO	2848	35,07%
GT2	MEDIA	1000≤ f< 2000	CAMIONETA	1104	13,59%
GT3	BAJA	f<1000	CAMIÓN	147	1,81%
			OTRO	93	1,15%
			MOTO	50	0,62%

Tabla 26: Recategorización de la variable Tipo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

4.2. ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

El análisis que se describe en esta parte del proyecto tiene como finalidad explicar el problema investigado en base a los resultados obtenidos, mediante la reducción del conjunto de todas las variables objeto de estudio a sólo dos dimensiones.

En la Tabla 27 se muestra el resumen del modelo obtenido, donde se pueden identificar el autovalor, inercia y porcentaje de varianza para cada dimensión.

Dimensión	Total (autovalor)	Inercia	% de varianza
1	1.798	0.300	29.963
2	1.288	0.215	21.473
Total	3.086	0.515	
Media	1.543	0.257	25.718

Tabla 27: Resumen del modelo

Fuente: Datos de la investigación. Autor: Johnny Jiménez

De esta tabla se pueden interpretar los resultados obtenidos, donde se destaca la inercia, que indica la proporción o porcentaje de la varianza de los datos que es explicada por cada una de las dimensiones. En la columna Inercia se puede observar que en total existe un 51.5% de la variabilidad de

los datos que es explicada por las dimensiones incluidas en el modelo, donde la primera dimensión explica el 30% de la variabilidad de los datos y la segunda dimensión explica el 21.5% restante.

En la Tabla 28 se muestran las medidas discriminantes por variable, donde se pueden identificar las puntuaciones que tiene cada variable en la dimensión correspondiente, las cuales permiten ver cuánto discrimina cada variable en cada una de las dimensiones, en otras palabras, permiten identificar la importancia que tiene cada variable de acuerdo a la dimensión obtenida.

Variables	Dimensión		Media
	1	2	
CATEGORÍA	**0.755**	0.029	0.392
MARCA	0.179	**0.295**	0.237
TIPO	**0.777**	0.050	0.414
COLOR	0.046	**0.440**	0.243
SINIESTRO	0.014	0.012	0.013
AÑO SEGUN PERCENTIL	0.027	**0.462**	0.245
Total activo	1.798	1.288	1.543
% de varianza	29.963	21.473	25.718

Tabla 28: Medidas discriminantes por variable

Fuente: Datos de la investigación. Autor: Johnny Jiménez

De acuerdo a estos resultados se puede observar que, de acuerdo a las medidas de discriminación, las variables Categoría y Tipo están mayormente relacionadas con la dimensión 1, pero las variables Marca, Color y Año según percentil se encuentran relacionadas con la dimensión 2.

En la Figura 11 se representa mediante un gráfico de dos dimensiones las medidas discriminantes descritas anteriormente, donde se puede apreciar que se forman dos grupos de variables de acuerdo a la cercanía que estas presentan. En el primer grupo están las variables Categoría y Tipo, que se encuentran muy próximas entre sí, lo cual es un indicio de que existe

relación entre ellas, mientras que en el segundo grupo se distinguen las variables Año según percentil, Color y Marca, las cuales se encuentran relacionadas entre sí, dada su cercanía en el gráfico.

Por último, la variable Siniestro, debido a que presenta puntuaciones muy bajas en las dos dimensiones, hace que esta variable no sea muy explicativa dentro del conjunto de variables bajo estudio, lo cual indica que, no se encuentra relacionada con ninguno de los dos grupos de variables. Esto se puede apreciar como un punto muy cercano al origen.

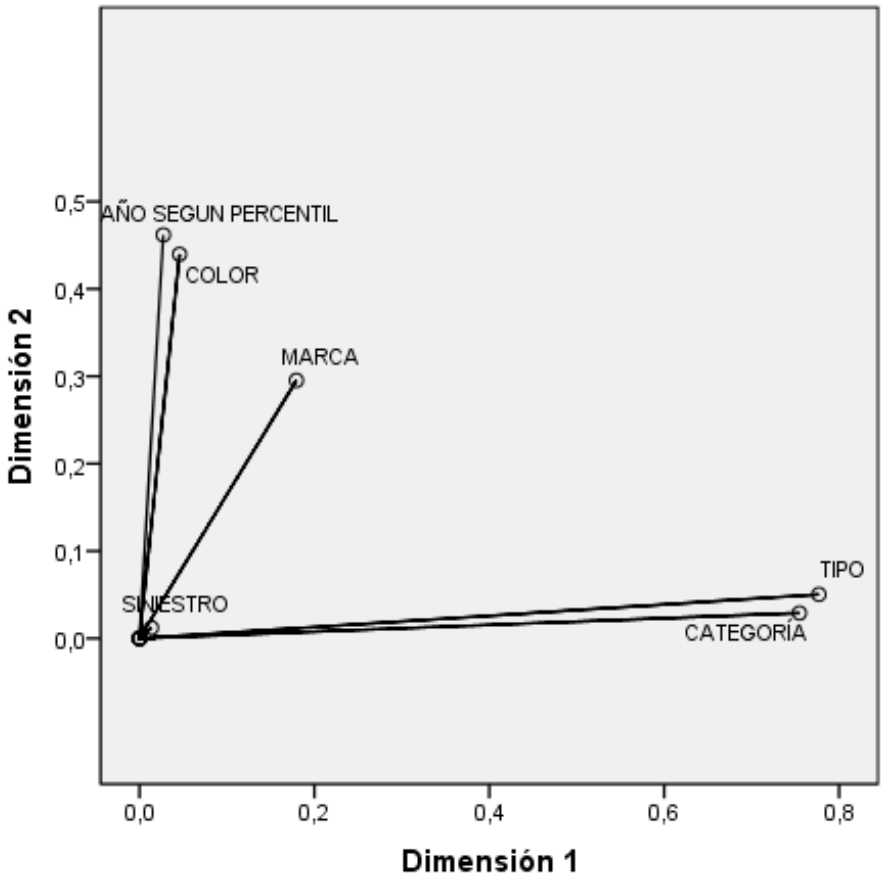


Figura 11: Medidas discriminantes por variable

Fuente: Datos de la investigación. Autor: Johnny Jiménez

En la Figura 12 se presenta el diagrama conjunto de puntos de categorías de las variables objeto de estudio, ya que permite representar las correspondencias propiamente dichas entre todas las variables y sus

respectivas categorías. Es decir, permite identificar patrones y establecer relaciones entre categorías de acuerdo a su proximidad.

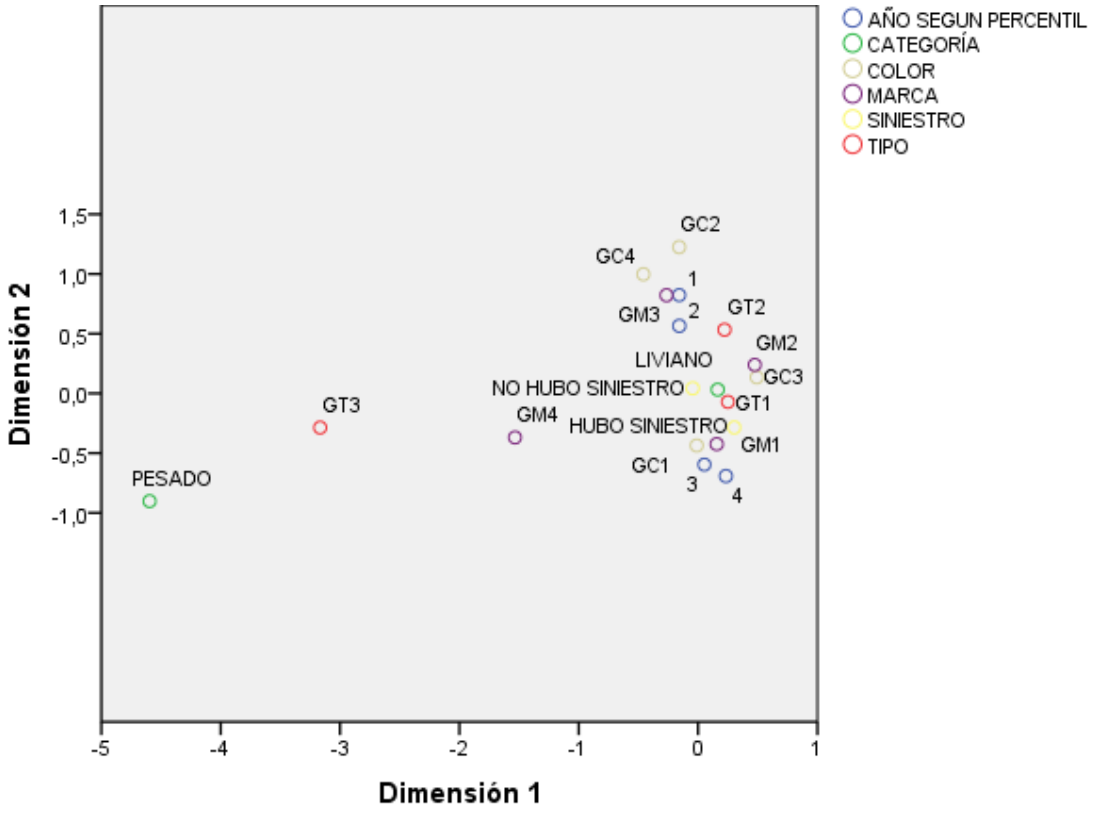


Figura 12: Diagrama conjunto de puntos de categorías

Fuente: Datos de la investigación. Autor: Johnny Jiménez

En este gráfico se pueden observar las relaciones entre las categorías de todas las variables y las categorías de la variable Siniestro, que al analizarlo detenidamente se obtienen los siguientes resultados:

La categoría GM2 que agrupa a las marcas TOYOTA y NISSAN se relaciona con la categoría NO HUBO SINIESTRO; mientras que la categoría GM1 que agrupa a las marcas CHEVROLET, HYUNDAI y KIA, se relaciona con la categoría HUBO SINIESTRO.

La categoría GC3 que agrupa a los colores DORADO, BEIGE, CONCHO DE VINO y VERDE se relaciona con la categoría NO HUBO SINIESTRO;

mientras que la categoría GC1 que agrupa a los colores PLATEADO, BLANCO, PLOMO y NEGRO, se relaciona con la categoría HUBO SINIESTRO.

La categoría GT2 que agrupa a los vehículos tipo CAMIONETA se relaciona con la categoría NO HUBO SINIESTRO; mientras que la categoría GT1 que agrupa a los vehículos tipo SEDÁN y TODOTERRENO, se relaciona con la categoría HUBO SINIESTRO.

La categoría LIVIANO se relaciona más con la categoría NO HUBO SINIESTRO, ya que están muy próximas, aunque también se observa cierta relación con la categoría HUBO SINIESTRO, ya que se puede observar cercanía entre ellas; mientras que la categoría PESADO no se relaciona con ninguna de las dos categorías HUBO SINIESTRO O NO HUBO SINIESTRO.

Las categorías 1 y 2 de la variable Año según percentil, que agrupan a los vehículos cuyo año de fabricación está dentro de los intervalos [1949, 2007) y [2007, 2011) respectivamente, se relaciona con la categoría NO HUBO SINIESTRO; mientras que las categorías 3 y 4 que agrupan a los vehículos cuyo año de fabricación está dentro de los intervalos [2011, 2013) y [2013, 2015], se relaciona con la categoría HUBO SINIESTRO.

Una vez vistas las diferentes asociaciones entre las categorías de las variables bajo estudio, se puede deducir que:

Los vehículos de marca CHEVROLET, HYUNDAI y KIA, que son de color PLATEADO, BLANCO, PLOMO o NEGRO, que son de tipo SEDÁN o TODOTERRENO, que son LIVIANOS y fueron fabricados entre los años 2011 y 2015, son las más propensos a sufrir un siniestro; mientras que los vehículos de marca TOYOTA y NISSAN, que son de color DORADO, BEIGE, CONCHO DE VINO o VERDE, que son de tipo CAMIONETA, que son LIVIANOS y fueron fabricados antes del 2011 son los menos propensos a sufrir un siniestro.

4.3. ANÁLISIS DE REGRESIÓN LOGÍSTICA

En esta parte del análisis se trata de encontrar un modelo de regresión logística que permita establecer la relación entre todas las variables independientes y la variable dependiente que son objeto de estudio en esta investigación, con la finalidad de explicar cuáles son los factores que influyen significativamente en la probabilidad de que un vehículo sufra un siniestro.

Se realizaron algunas pruebas tratando de encontrar el mejor modelo de regresión logística, para lo cual se utilizó el proceso de introducción manual de variables a SPSS, donde se probó primero con todas las variables y se fueron eliminando sucesivamente variables que no fueron significativas, además se probó con interacciones entre las variables, este proceso se fue repitiendo hasta que se encontró un conjunto de variables donde todas fueran significativas, de esa manera se logró determinar el modelo que mejor se ajusta a los datos observados. Esta parte se centrará solamente en la descripción y análisis de los resultados que fueron obtenidos en base a la introducción de las variables Color, Marca y Tipo, a través de las cuales se encuentra el modelo de regresión logística.

Para poder aplicar el análisis de Regresión Logística, dado que estas tres variables son categóricas, se procedió a la recodificación de las mismas, creando variables DUMMY cuyo número es igual al número de categorías de la variable original pero disminuida en uno. Por ejemplo, para la variable COLOR como tiene cuatro categorías GC1, GC2, GC3 y GC4, se crearon tres variables DUMMY, COLOR(1), COLOR(2) y COLOR(3), con la finalidad de tener una categoría de referencia, que ayude en la interpretación de los resultados. En la Tabla 29 se muestran las codificaciones de las variables categóricas.

Como se puede observar en la Tabla 29, la categoría de referencia es la última categoría de cada variable, la cual no está representada por ninguna variable DUMMY, por ejemplo, para la variable COLOR que tiene cuatro categorías (GC1, GC2, GC3 y GC4), si un vehículo pertenece al grupo GC1,

entonces la variable DUMMY COLOR(1) tomará el valor de 1, caso contrario tomará el valor de 0; si un vehículo pertenece al grupo GC2, entonces la variable DUMMY COLOR(2) tomará el valor de 1, caso contrario tomará el valor de 0; y si un vehículo pertenece al grupo GC3, entonces la variable DUMMY COLOR(3) tomará el valor de 1, caso contrario tomará el valor de 0, y GC4 es la categoría de referencia.

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 29: Codificaciones de variables categóricas

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Una vez que se ha realizado la codificación de las variables categóricas, se determinará la significancia global del modelo que contiene estas variables, la cual está basada en el valor del estadístico de prueba G, con la finalidad de contrastar la significancia de los coeficientes del modelo de regresión logística.

Las hipótesis para probar la significancia global son las siguientes:

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

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

Lo que en palabras se puede expresar en forma equivalente como:

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.

Este contraste de hipótesis se lo realizó a través de la prueba de Omnibus que provee SPSS, y se obtuvieron los siguientes resultados: El valor del estadístico de prueba G es 269.12, con 8 grados de libertad y el valor-p de la prueba es igual a 0 (menor a 0.05), por lo que existe evidencia estadística para rechazar la hipótesis nula. Por lo tanto, al menos uno de los coeficientes de las variables independientes es diferente de cero.

Como en la prueba de Omnibus se obtuvo que al menos uno de los coeficientes de las variables independientes que están presentes en el modelo, es diferente de cero, hay que determinar cuáles son las variables que tienen coeficientes que son estadísticamente significativos a un nivel de significancia de 0.05, por lo cual se utiliza una prueba de significancia individual para cada coeficiente de las variables independientes que se basa en el estadístico W de Wald, por lo tanto se va a realizar el siguiente contraste de hipótesis.

$$H_0 : \beta_i = 0$$

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

Lo que en palabras se puede expresar en forma equivalente como:

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

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

En la Tabla 30 se pueden observar los resultados de efectuar los contrastes de hipótesis a cada uno de los coeficientes de las variables independientes, donde se puede afirmar que existe evidencia estadística para no rechazar la hipótesis nula por medio del estadístico de Wald y con valores p asociados

(p>0.05), para los coeficientes de las variables DUMMY, MARCA(2), MARCA(3) y COLOR(2), lo cual significa que estos coeficientes no son estadísticamente significativos. Por el contrario, 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	-2,693	0,092	864,739	1	0	0,068		
MARCA(1)	-0,151	0,056	7,321	1	0,007	0,859	0,77	0,959
MARCA(2)	-0,067	0,063	1,132	1	**0,287**	0,936	0,828	1,058
MARCA(3)	-0,109	0,058	3,53	1	**0,06**	0,896	0,8	1,005
TIPO(1)	0,828	0,063	172,398	1	0	2,288	2,022	2,589
TIPO(2)	0,703	0,07	102,242	1	0	2,02	1,762	2,315
COLOR(1)	0,154	0,056	7,493	1	0,006	1,166	1,045	1,301
COLOR(2)	0,071	0,062	1,324	1	**0,25**	1,074	0,951	1,212
COLOR(3)	0,276	0,063	19,204	1	0	1,318	1,165	1,491

Tabla 30: Variables en el Modelo de Regresión Logística

Fuente: Datos de la investigación. Autor: Johnny Jiménez

Un modelo de regresión logística debe estar formado sólo con variables que sean estadísticamente significativas, por lo cual deben eliminarse las variables cuyos coeficientes no fueron estadísticamente significativos, es decir, se eliminan las variables DUMMY que globalmente no aportan en la solución del modelo, las cuales son: MARCA(2), MARCA(3) y COLOR(2). En base a esto se obtienen los siguientes resultados, los que aparecen en la Tabla 31.

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	-2,693	0,092	864,739	1	0	0,068		
MARCA(1)	-0,151	0,056	7,321	1	0,007	0,859	0,77	0,959
TIPO(1)	0,828	0,063	172,398	1	0	2,288	2,022	2,589
TIPO(2)	0,703	0,07	102,242	1	0	2,02	1,762	2,315
COLOR(1)	0,154	0,056	7,493	1	0,006	1,166	1,045	1,301
COLOR(3)	0,276	0,063	19,204	1	0	1,318	1,165	1,491

Tabla 31: Variables que quedan en el Modelo de Regresión Logística

Fuente: Datos de la investigación. Autor: Johnny Jiménez

En la Tabla 31 se puede observar 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 realizó la prueba de Hosmer-Lemeshow, la cual realiza el siguiente contraste de hipótesis:

H_0: El modelo seleccionado ajusta bien a los datos

H_1: $\neg H_0$

El valor del estadístico de prueba chi-cuadrado es 6.95, con 8 grados de libertad y el valor-p de la prueba es igual a 0.326 (mayor a 0.05), por lo cual, 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 utilizó para evaluar la bondad de ajuste del modelo fue a través de la tasa global de aciertos que se obtiene de la tabla de

clasificación de resultados que permite analizar la capacidad predictiva del modelo obtenido, 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.

H_1: $\neg H_0$

En la Tabla 32 se puede observar la clasificación de resultados, donde el porcentaje global de aciertos fue de 73%, además se presenta el test de Huberty que permite comprobar la significación estadística de esta tasa y de esta manera contrastar las hipótesis anteriores. De este test se obtiene el número esperado de casos correctamente clasificados debidos al azar (e) que es igual a 48341.82, a partir del cual se calcula el valor del estadístico Z^* que se distribuye normalmente, cuyo valor es igual a 26.17. 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^*= 26.17 > 1.96, por lo cual, 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**	43778	10574	80.5
	SÍ	6295	1827	22.5
Porcentaje global				**73.0**
$e = 48341.82$		$Z^* = 26.17$		

Tabla 32: Tabla de clasificación de resultados

Fuente: Datos de la investigación. Autor: Johnny Jiménez

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

$$p = \frac{e^{-2.693 - 0.151\ x_1 + 0.828\ x_2 + 0.703\ x_3 + 0.154\ x_4 + 0.276\ x_5}}{1 + e^{-2.693 - 0.151\ x_1 + 0.828\ x_2 + 0.703\ x_3 + 0.154\ x_4 + 0.276\ x_5}}$$

Que en forma resumida se puede expresar como:

$$p = \frac{1}{1 + e^{2.693 + 0.151 x_1 - 0.828 x_2 - 0.703 x_3 - 0.154 x_4 - 0.276 x_5}}$$

Donde:

X_1: MARCA(1)

X_2: TIPO(1)

X_3: TIPO(2)

X_4: COLOR(1)

X_5: COLOR(3)

Como a los datos observados ya se les ajustó un modelo de regresión logística, lo que ahora sigue es su interpretación. Para esto se considerarán las categorías que tienen el valor de 1 en las variables DUMMY, ya que a partir de estas categorías se realizará la interpretación. En esta investigación, las categorías cuyas variables DUMMY tienen el valor de 1 son las siguientes: GM1, GT1, GT2, GC1 y GC3. Estas categorías están asociadas 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, por lo cual, en función de esto se puede interpretar lo siguiente:

El que un vehículo sea de marca CHEVROLET, HYUNDAI o KIA (GM1) constituye un factor de riesgo para la ocurrencia de un siniestro.

El que un vehículo sea del tipo SEDÁN, TODOTERRENO o CAMIONETA constituye un factor de riesgo para la ocurrencia de un siniestro.

El que un vehículo sea de color PLATEADO, BLANCO, PLOMO, NEGRO, DORADO, BEIGE, CONCHO DE VINO o VERDE constituye un factor de riesgo para la ocurrencia de un siniestro.

En la tabla 31 se presentan los valores del Odds ratio, tal que, si el valor del odds ratio es mayor a 1, indica que es más probable que ocurra el evento de interés, HUBO SINIESTRO, en relación a la no ocurrencia de dicho evento; caso contrario, si este valor es menor a 1, indica 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 las marcas CHEVROLET, HYUNDAI o KIA (GM1), entonces es menos probable que sufra un siniestro a que sea de las marcas GREAT WALL, MITSUBISHI, HONDA, SKODA o HINO (GM4, que es la categoría de referencia).

Si un vehículo es del tipo SEDÁN o TODOTERRENO (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 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 GRIS, CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

Si un vehículo es de color DORADO, BEIGE, CONCHO DE VINO o VERDE (GC3), entonces es más probable que sufra un siniestro a que sea de color

GRIS, CELESTE, AMARILLO, ACERO u OTRO (GC4, que es la categoría de referencia).

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. De un total de 62474 vehículos asegurados en la ciudad de Quito, los vehículos livianos tuvieron una participación del 96.55% y los vehículos pesados tuvieron una participación del 3.45%.
2. Los vehículos de las marcas Chevrolet, Hyundai y Kia tuvieron la mayor participación, con 33,26%, 12.31% y 9.49% respectivamente, de un total de 62474 vehículos asegurados en la ciudad de Quito.
3. De los 62474 vehículos asegurados en la ciudad de Quito, los vehículos de tipo sedán tuvieron la mayor participación con un 43.13%, le siguen los vehículos de tipo todoterreno y camioneta con 35,44% y 14,24% respectivamente.
4. Los vehículos de color plateado, blanco y plomo tuvieron la mayor participación, con 19,93%, 19,24% y 15,30% respectivamente, de un total de 62474 vehículos asegurados en la ciudad de Quito.
5. De los 62474 vehículos asegurados en la ciudad de Quito, un 87% correspondió a vehículos que no sufrieron un siniestro, mientras que el 13% de vehículos sufrieron un siniestro.
6. El 25% de los vehículos tienen su año de fabricación menor a 2007, el 50% de los vehículos fueron fabricados entre 2007 y 2013 y un 25% de los vehículos fueron fabricados después del año 2013, considerando un total de 62474 vehículos asegurados en la ciudad de Quito.
7. De acuerdo a los resultados obtenidos del Análisis de Correspondencias Múltiples, los vehículos de marca CHEVROLET, HYUNDAI y KIA, que son de color PLATEADO, BLANCO, PLOMO o NEGRO, que son de tipo SEDÁN o TODOTERRENO, que son LIVIANOS y fueron fabricados entre los años 2011 y 2015, son las más propensos a sufrir un siniestro.
8. De acuerdo a los resultados obtenidos del Análisis de Correspondencias Múltiples, los vehículos de marca TOYOTA y NISSAN, que son de color DORADO, BEIGE, CONCHO DE VINO o VERDE, que son de tipo CAMIONETA, que son LIVIANOS y fueron fabricados antes del 2011 son los menos propensos a sufrir un siniestro.

9. De acuerdo a los resultados obtenidos del Análisis de Regresión Logística, los vehículos de marca CHEVROLET, HYUNDAI o KIA; del tipo SEDÁN, TODOTERRENO o CAMIONETA; y de color PLATEADO, BLANCO, PLOMO, NEGRO, ROJO, AZUL, DORADO, BEIGE, CONCHO DE VINO o VERDE, constituyen factores de riesgo para la ocurrencia de un siniestro.
10. Los vehículos de las marcas CHEVROLET, HYUNDAI o KIA, tienen menos probabilidad de sufrir un siniestro a que sean de las marcas GREAT WALL, MITSUBISHI, HONDA, SKODA o HINO.
11. Los vehículos del tipo SEDÁN o TODOTERRENO, tienen más probabilidad de sufrir un siniestro a que sean del tipo CAMIÓN, MOTO u OTRO.
12. Los vehículos de tipo CAMIONETA, tienen más probabilidad que sufran un siniestro a que sean del tipo CAMIÓN, MOTO u OTRO.
13. Los vehículos de color PLATEADO, BLANCO, PLOMO, NEGRO, DORADO, BEIGE, CONCHO DE VINO o VERDE, tienen más probabilidad de sufrir un siniestro a que sean de color GRIS, CELESTE, AMARILLO, ACERO u OTRO.
14. Comparando los resultados obtenidos en los dos modelos de análisis multivariado, 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 otras marcas y colores de vehículos que no fueron asociados en el análisis de correspondencias múltiples.

RECOMENDACIONES

1. Se recomienda a las empresas aseguradoras que provean toda la información requerida por los investigadores, ya que de esta manera se lograrían identificar otros posibles factores que pudieran influir en la siniestralidad de un vehículo.
2. Sería de interés realizar un estudio que permita incorporar 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 ocurrencia de un siniestro.
3. En base a las estadísticas de accidentes de tránsito a nivel nacional, un factor importante que las aseguradoras deberían analizar es la ciudad de residencia del asegurado, ya que se podría 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 puedan tener un mejor criterio para brindar cobertura a sus clientes.

BIBLIOGRAFÍA

[1] *OPS OMS | Informe sobre la situación de la Seguridad Vial en la Región de las Américas*. (2016). *Pan American Health Organization / World Health Organization*. Recuperado el 9 de enero de 2016, de http://www.paho.org/hq/index.php?option=com_content&view=article&id=10847&Itemid=41441&lang=es

[2] UNECE (2015). *Informe para el mejoramiento de la seguridad vial en el mundo*, OMS. (2015). Recuperado el 9 de enero de 2016, de https://www.unece.org/fileadmin/DAM/trans/doc/2015/wp1/UNSG_Report_70_386_Spanish.pdf

[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 - 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*

[5] *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*

[6] Zabala, Víctor. *Unidad de Investigación Económica y de Mercado*, *Equipo Editorial. Especial Seguros.* [en línea]. Julio 2015, [fecha de consulta: 16 de enero 2016]. Disponible en: http://www.ekosnegocios.com/revista/pdfTemas/1257.pdf.

[7] Das, S., Sun, X. (2016). *Association Knowledge For Fatal Run-Off-Road Crashes By Multiple Correspondence Analysis*. IATSS Research. Volume 39, Issue 2, 1 March 2016, Pages 146-155.

[8] Paefgen, J., Staake, T., Fleisch, E. (2014). *Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data.* Transportation Research Part A: Policy and Practice. Volume 61, March 2014, Pages 27-40.

[9] Hemrit, W., Arab, M.B., 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, Issue 2, 2013, Pages 107-147.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[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	324	408	382	188	40	**1342**
FORD	36	69	29	17	2	**153**
GREAT WALL	4	44	47	55	18	**168**
HINO	3	7	4	1	0	**15**
HONDA	5	0	1	0	0	**6**
HYUNDAI	133	156	303	129	32	**753**
KIA	110	134	360	136	29	**769**
MAZDA	40	19	29	42	3	**133**
MITSUBISHI	17	0	4	2	2	**25**
NISSAN	73	119	104	47	7	**350**
OTRA	11	24	37	12	2	**86**
RENAULT	73	66	37	41	9	**226**
SKODA	7	4	4	2	0	**17**
SUZUKI	72	56	34	13	8	**183**
TOYOTA	34	90	105	38	11	**278**
VOLKSWAGEN	73	49	26	21	0	**169**
Total general	**1015**	**1245**	**1506**	**744**	**163**	**4673**

ANEXO 2

SINIESTRALIDAD POR TIPO DE VEHÍCULO Y AÑO

TIPO DE VEHICULO	AÑO					Total general
	2011	2012	2013	2014	2015	
CAMIÓN	30	14	19	6	1	**70**
CAMIONETA	117	173	175	122	21	**608**
MOTO	9	7	12	6	1	**35**
OTRO	20	14	19	7	0	**60**
SEDÁN	480	621	794	385	86	**2366**
TODOTERRENO	359	416	487	218	54	**1534**
Total general	**1015**	**1245**	**1506**	**744**	**163**	**4673**

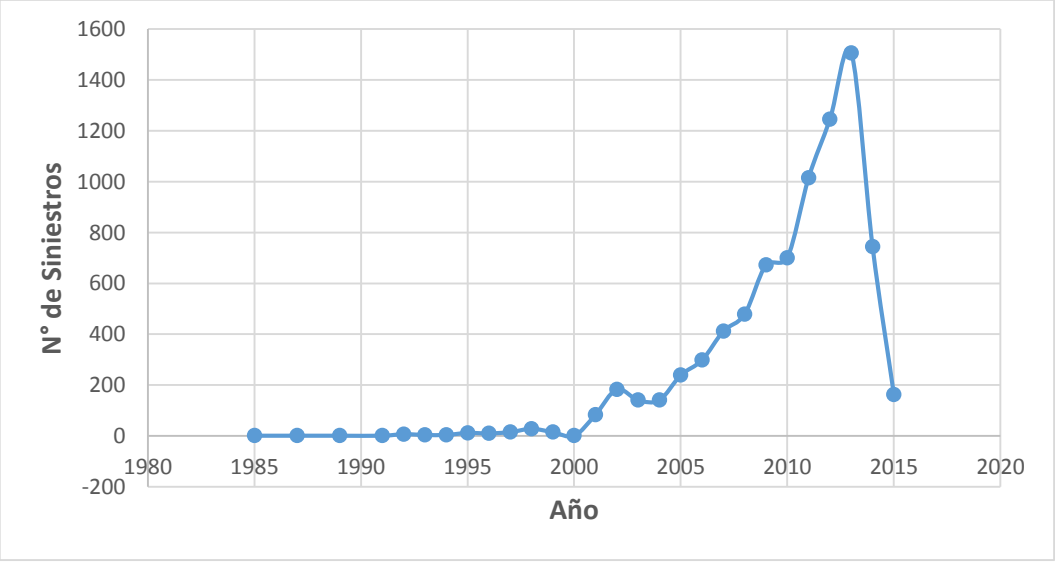
ANEXO 3

SINIESTRALIDAD POR COLOR DE VEHÍCULO Y AÑO

COLOR	AÑO					Total general
	2011	2012	2013	2014	2015	
AMARILLO	5	4	4	9	0	**22**
AZUL	44	76	55	30	7	**212**
BEIGE	29	35	36	15	3	**118**
BLANCO	154	218	239	112	20	**743**
CELESTE	21	8	20	1	1	**51**
CONCHO DE VINO	40	48	75	40	14	**217**
DORADO	73	90	93	62	14	**332**
GRIS	1	1	0	0	1	**3**
NEGRO	148	168	184	71	24	**595**
OTRO	7	20	25	14	0	**66**
PLATEADO	225	274	355	158	30	**1042**
PLOMO	161	205	314	176	36	**892**
ROJO	67	82	103	54	13	**319**
VERDE	40	16	3	2	0	**61**
Total general	**1015**	**1245**	**1506**	**744**	**163**	**4673**

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	7	**37**
	FORD	4	0	**4**
	GREAT WALL	19	3	**22**
	HYUNDAI	7	1	**8**
	KIA	3	0	**3**
	MAZDA	38	2	**40**
	MITSUBISHI	0	2	**2**
	NISSAN	4	0	**4**
	OTRA	1	1	**2**
	TOYOTA	15	5	**20**
	VOLKSWAGEN	1	0	**1**
SEDÁN	CHEVROLET	125	32	**157**
	GREAT WALL	7	1	**8**
	HYUNDAI	81	27	**108**
	KIA	85	15	**100**
	MAZDA	1	0	**1**
	NISSAN	33	4	**37**
	OTRA	4	0	**4**
	RENAULT	27	5	**32**
	SKODA	2	0	**2**
	TOYOTA	2	2	**4**
	VOLKSWAGEN	18	0	**18**
TODOTERRENO	CHEVROLET	29	1	**30**
	FORD	13	2	**15**
	GREAT WALL	29	14	**43**
	HYUNDAI	37	4	**41**
	KIA	47	13	**60**
	MAZDA	2	1	**3**
	MITSUBISHI	2	0	**2**
	NISSAN	10	3	**13**
	OTRA	3	0	**3**
	RENAULT	14	4	**18**
	SUZUKI	11	8	**19**
	TOYOTA	21	4	**25**
Total general		**725**	**161**	**886**