10

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES ACATLÁN

MÉTODO PARA CREACIÓN DE POBLACIONES SINTÉTICAS DE LA FLOTA VEHICULAR DE LA ZMCM MODELOS 1991-1998, CON BASE A LA EMISIÓN DE CONTAMINANTES DEL AIRE

T E S I S

Que para obtener el título de:
LICENCIADO EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN
p r e s e n t a :
NORMA GARCÍA SANTIBÁÑEZ SANTILLÁN

Asesores:

Ing. Elvira Beatriz Clavel Díaz (UNAM)
Dr. Adrian S. Barrera Roldán (IMP)

Septiembre 2000

DEDICATORIA

A ti Judith, que fuiste como una estrella fugaz, que aunque poco duro tu existencia fue lo suficiente para iluminar por donde pasaba.

.. y a ti pequeño Erick.

AGRADECIMIENTOS

Primero que nada gracias a Dios por todas la bendiciones y por las oportunidades que se me presentan día con día.

Gracias a mis padres por todo su apoyo y amor, gracias a mis hermanos que siempre me están apoyando, gracias a mis sobrinos que de una u otra forma alegran mi vida.

Gracias a ti Sergio porque siempre estas conmigo y por toda tu paciencia.

Gracias al Instituto Mexicano del Petróleo por brindarme la oportunidad de realizar este trabajo, en especial al grupo de Economía Ambiental coordinado por el Dr. Adrian Barrera y a sus colaboradores (Benjamin y Héctor).

Por último gracias a la Universidad Nacional Autónoma de México por la oportunidad que me dio de prepararme, a todos los profesores que contribuyeron a mi formación y especialmente gracias a la Profesora Beatriz Clavel por ayudarme tanto con este trabajo

ÍNDICE

PAG

III. CREACIÓN DE LA POBLACIÓN SINTÉTICA DE LA FLOTA VEHICULAR DE LA ZMCM

INTRODUCCIÓN

El término sintético o artificial se usa para algunas cosas de uso diario, por ejemplo, en zapatos de piel sintética, pasto sintético, inteligencia artificial, pero ¿cuál es el significado de sintético?, la definición de diccionario es: "Se dice de ciertos productos artificiales (hechos por la mano del hombre) que imitan o reproducen las características de otros naturales" [1].

Escuchar el término sintético o artificial puede causar desconfianza ya que no siempre la imitación es buena, de hecho en algunos productos lo sintético es sinónimo de mala calidad, lo que no siempre es cierto, por ejemplo, en medicina se habla de "corazón artificial", el cual ha ayudado a prolongar la vida a personas cuyo corazón ya no les funcionaba adecuadamente.

El término sintético se utiliza prácticamente en todas las áreas de nuestra vida diaria, incluso en la estadística, ya que no siempre es posible tener el total de la población necesario para el análisis y sólo se puede obtener una pequeña muestra de la población, por ejemplo si quisiéramos estudiar cuántos vehículos obtuvieron la calcomanía uno de verificación de los vehículos que circulan en el Distrito Federal, se tendría que contar con la información de todos los centros de verificación para poder obtener el resultado. Si tomamos en cuenta que esto implica invertir bastante tiempo en ir a cada uno de los centros de verificación (cuando se hace de manera personal) , además del costo involucrado . Algo que se podría hacer es crear el total de la población a partir de una muestra real. A esto se le conoce como poblaciones sintéticas.

Pero, ¿cómo crear una población sintética?. La respuesta es el objetivo de este trabajo, ya que su propósito es crear una población sintética para estudiar el comportamiento de la emisión de contaminantes de los automóviles particulares en la Zona Metropolitana de la Ciudad de México (ZMCM, de aquí en adelante así se le denominará), utilizando el método de creación de poblaciones sintéticas propuesto por Beckman y Baggerly (1996) en su artículo "*Creating synthetic baseline populations* ".

---

[1] http://www.logos.it/dictionary

Se eligió únicamente a los vehículos automotores porque son los que más contaminan el ambiente con sus emisiones, dos veces por año es necesario verificarlos para saber como están en cuanto a emisiones de contaminantes, además a la ZMCM por ser una de las más contaminadas del mundo, si no es que la más contaminada.

Las características de la ZMCM, de los combustibles utilizados y de los automóviles que circulan en ésta, además de las normas que regulan la emisión de contaminantes, se muestran en el capítulo I, ya que se necesita conocer las características de la población que se va a estudiar y así crear la población sintética lo más parecida a la realidad. Y para ello se necesita primero tener una muestra representativa de la población, es decir una muestra que cumpla con todas las características de la población.

En el capítulo II se ven las técnicas de muestreo para poder seleccionar una muestra, se explica el método para la creación de poblaciones sintéticas y algunos conceptos básicos del Método de simulación de Monte Carlo, que es el que se utilizó para la creación de la población.

Por último, en el capítulo III se ve la aplicación del método para la creación de poblaciones sintéticas en un caso práctico, en la emisión de contaminantes por los vehículos automotores. Los modelos de vehículos que se utilizan para este estudio van desde 1991 hasta 1998, y se tomó la muestra de las emisiones de los principales precursores del ozono en la ZMCM.

Para poder saber si la simulación fue adecuada en la creación de la población sintética, primero hay que delimitar el marco teórico, seleccionar la técnica de muestreo, recoger la información y agruparla, ya con la información y una vez que se conocen los estimadores se puede generar la población sintética. Todos estos pasos se irán explicando en el transcurso de este trabajo.

I

CONTAMINACIÓN Y LOS AUTOMÓVILES EN LA ZMCM.

1.1 ANTECEDENTES

La ciudad de México enfrenta el grave problema de contaminación atmosférica debido a las características fisiográficas y climáticas, pero principalmente a la explosión demográfica que trae consigo el que circulen más vehículos en la ZMCM. La principal fuente de contaminación del aire son los vehículos automotores de combustión interna (vehículos en lo sucesivo) con un 75%, y aunque en los últimos años se ha reducido la emisión de contaminantes, actualmente se expulsan a la atmósfera alrededor de 11 mil toneladas al día de contaminantes. Por lo que se han realizado importantes esfuerzos para reducirla como son la introducción de gasolinas con mayor índice de octano además de que se ha mejorado la tecnología de los vehículos, y se han implementado normas para poder controlar la emisión de contaminantes a la atmósfera además de programas como el "Hoy no circula" y diferentes planes de contingencias ambientales.

1.2 CARACTERÍSTICAS FISIOGRÁFICAS, CLIMÁTICAS Y DEMOGRÁFICAS DE LA ZMCM

La ZMCM posee una serie de características fisiográficas y climáticas únicas que contribuyen de manera importante al problema de contaminación entre las que se pueden mencionar:

Características fisiográficas

- La ZMCM se encuentra a una altura de 2 mil 240 metros, por lo que el contenido del aire es 23% menor que al nivel del mar. Esto hace que los procesos de combustión interna sean menos eficientes y produzcan por tanto una mayor cantidad de contaminantes.

- Está rodeada por las montañas de las Sierras del Ajusco, Chichinautzin, Nevada, Las Cruces, Guadalupe y Santa Catarina, las que constituyen una barrera física natural para la circulación del viento, impidiendo el desalojo del aire contaminado fuera del Valle.

- Se localiza dentro de la región central del país, por lo cual está sujeto también a la influencia de sistemas anticiclónicos, generados tanto en el Golfo de México como en el Océano Pacífico. Estos sistemas ocasionan una gran estabilidad atmosférica, inhibiendo el mezclado vertical del aire.

Características climáticas

- La ZMCM presenta con frecuencia inversiones térmicas que provocan el estancamiento de los contaminantes. Por las mañanas, la capa de aire que se encuentra en contacto con la superficie de los suelos adquiere una temperatura menor que las capas superiores, por lo que se vuelve más densa y pesada. Las capas de aire que se encuentran a mayor altura y que están relativamente más calientes actúan entonces como una cubierta que impide el movimiento ascendente del aire contaminado. Como se puede ver en la figura 1.

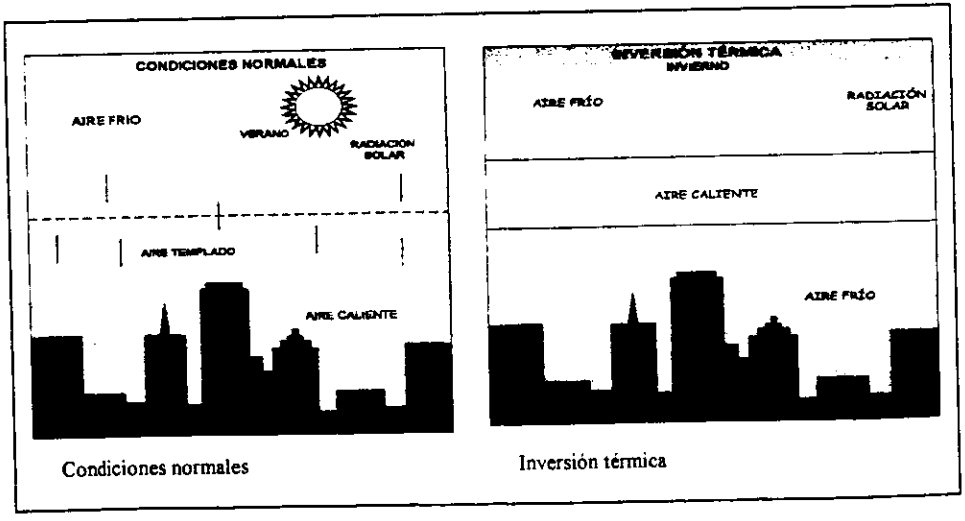

Figura 1. Condiciones atmosféricas que permiten la circulación del aire

- Recibe una abundante radiación solar debido a su latitud de 19° N, lo que hace que su atmósfera sea altamente ferroactiva. Esto es, en presencia de la luz solar los hidrocarburos y los óxidos de nitrógeno reaccionan fácilmente para formar ozono y otros oxidantes.

Características demográficas

Por otra parte, se ha comprobado que el aumento en la contaminación al medio ambiente está directamente relacionado con el número de personas que la habitan, así como por los procesos de industrialización ocurridos en el país. Entre las principales características demográficas que contribuyen a agudizar el problema de la contaminación encontramos los siguientes:

- La ZMCM está integrada de 16 delegaciones políticas del Distrito Federal y 28 municipios conurbanos del estado de México, por lo que se considera una de las metrópolis más grandes del mundo, sino es que la más grande. Aun cuando ha disminuido el crecimiento poblacional, en la actualidad alberga a 16.4 millones de habitantes, que representan el 18% de la población nacional.[1] Esto afecta de manera importante al problema de contaminación atmosférica porque a mayor número de personas mayor es la necesidad de transporte.

- Otro factor que contribuye al aumento de la contaminación en nuestra Ciudad y a que las tolvaneras sean más frecuentes, es la deforestación que ha sufrido la cuenca de México, ya que aproximadamente el 75% de la vegetación original ha sido devastada para satisfacer las demandas de industrialización y urbanización con una tasa promedio de crecimiento demográfico anual de hasta 5.6% .[2]

Debido a la gran cantidad de vehículos que circulan en la ZMCM, las políticas y programas contra la contaminación se han dirigido fundamentalmente a tratar de reducir el uso del vehículo privado, con un estricto control del estado mecánico de los vehículos y el uso de combustibles menos contaminantes.

[1] Comisión para el ahorro de energía
[2] Programa para Mejorar la Calidad del Aire en el Valle de México 1995-2000

1.2.1 PRINCIPALES CONTAMINANTES DE LOS COMBUSTIBLES EN LA ZMCM Y EFECTOS EN LA SALUD

Cuando hablamos de contaminación del aire, nos referimos a la alteración de su composición. El aire que respiramos generalmente está compuesto por los siguientes gases: un 78% de nitrógeno, 21% de oxígeno y 0.093% de argón. El resto son tan sólo pequeñas cantidades de helio, xenón, ozono y radón. En los primeros kilómetros de la atmósfera el aire contiene también una porción variable, de hasta un 4%, de vapor de agua.

El ozono es un contaminante secundario que se forma en la atmósfera mediante la acción de la luz solar sobre sus precursores: los hidrocarburos y óxidos de nitrógeno. La concentración de ozono cada día depende de las condiciones meteorológicas, las proporciones relativas entre hidrocarburos y óxidos de nitrógeno y la naturaleza de los hidrocarburos en la mezcla atmosférica. La relación entre el ozono y los vehículos, es que éstos emiten cantidades importantes de los precursores. Por lo tanto, los automóviles no emiten ozono directamente

Los vehículos que ocasionan mayores efectos nocivos al ambiente son los de uso particular constituido básicamente por coches y motocicletas, que emiten el 85% del total de contaminantes. Los vehículos de carga y pasajeros tanto foráneos como locales (aproximadamente 310 000 vehículos) conforman el 10%. El 5% restante lo ocupan los vehículos del transporte urbano y suburbano de pasajeros (microbuses, combis, taxis, autobuses y trolebuses).

Los vehículos contribuyen a la contaminación de la atmósfera en las siguientes cantidades: el monóxido de carbono constituye más del 90% del total de contaminantes en la atmósfera, las emisiones de hidrocarburos participan con el 30%, mientras que los óxidos de nitrógeno representan el 71%[3] . Estos son los tres contaminantes que se encuentran en mayor cantidad y se necesitan controlar su emisión. A continuación se describen cada uno de éstos, así como los principales efectos en la salud.

[3] Op cit.

➢ *Monóxido de Carbono (CO)*

El monóxido de carbono se forma debido a la combustión incompleta en los motores de los vehículos que utilizan gasolina. Las emisiones de monóxido de carbono dependen directamente de la afinación de los motores y de la eficacia en la combustión de los procesos industriales, de las condiciones y características del sistema vial, el tráfico y los diferentes medios de transporte. Es un gas incoloro y carece de olor. Debido al fuerte gradiente espacial que presenta este contaminante, las concentraciones encontradas en microambientes como en las banquetas de calles con intenso tránsito vehicular y en el interior de vehículos privados y públicos son mucho mayores que las concentraciones medidas simultáneamente en las estaciones fijas de análisis continuo. Esto significa que, a pesar de que no se exceda la norma en la estación, puede haber un número considerable de personas que se vean expuestas a niveles peligrosos de este contaminante. Por esto las concentraciones más altas de este gas se presentan en los periodos de mayor circulación vehicular. Es el contaminante en la ZMCM que encontramos en mayor cantidad en el aire y es muy difícil de eliminar.

Respecto a sus efectos en el ser humano el monóxido de carbono al ser inhalado entra al flujo sanguíneo y reduce el transporte de oxígeno a células y tejidos hasta causar daños al sistema nervioso central y cardiovascular (cuando se combina con la hemoglobina de la sangre se reduce automáticamente el transporte de oxígeno al cuerpo). También provoca una sobrecarga de trabajo al corazón. Esto quiere decir que este órgano deber realizar un mayor esfuerzo para bombear la sangre a todo el organismo. El monóxido de carbono frecuentemente se asocia con la disminución de la percepción visual, la capacidad de trabajo, la destreza manual y la habilidad de aprendizaje, también provoca dolor de cabeza, fatiga, somnolencia, fallos respiratorios y hasta la muerte.

➢ *Oxidos de Nitrógeno (NO y NO_2)*

Los óxidos de nitrógeno son contaminantes que por sí solos no representan un problema de salud pública, pero el óxido nítrico al oxidarse se convierte el bióxido de nitrógeno que sí representa un riesgo para la salud, especialmente si además están presentes los hidrocarburos. Dentro de estos compuestos están las cetonas, los aldehidos, los radicales alquino y los nitratos de peroxiacetilo, que provocan lagrimeo e irritación de garganta. El

óxido nítrico se deriva de los procesos de combustión; es un contaminante primario y juega un doble papel en materia ambiental, ya que se le reconocen efectos potencialmente dañinos de manera directa, al mismo tiempo que es uno de los precursores del ozono y otros oxidantes fotoquímicos, una vez que reacciona con la luz solar, produce compuesto tóxicos. Los Oxidos de Nitrógeno son gases de color café rojizo de olor picante.

La acumulación de bióxido de nitrógeno en el cuerpo humano constituye un riesgo para las vías respiratorias ya que se ha comprobado que puede alterar la capacidad de respuesta de las células en el proceso inflamatorio, como sucede con las células polimorfonucleares, macrófagos alveolares y los linfocitos, siendo más frecuente en casos de bronquitis crónica, además se dice que disminuye la visibilidad. En el caso de materiales provoca el desteñimiento de pinturas, en la vegetación ocasiona la caída prematura de las hojas e inhibición del crecimiento. Contribuye a la formación de la lluvia ácida (produce el ácido nítrico al combinarse con el agua y el oxígeno atmosférico) y es el principal generador de ozono.

- *Hidrocarburos (HC)*

Los hidrocarburos son compuestos orgánicos que contienen carbono e hidrógeno tales como benceno, tolueno, y formaldehído. Los hidrocarburos son un grupo de compuestos químicos que se forman durante la combustión incompleta de madera y combustible fósil. Las concentraciones de estos compuestos pueden ser bastante altas en las emisiones de los vehículos que usan diesel. Uno de los hidrocarburos más conocidos es el benzo-a.pireno.

Estos compuestos pueden ser absorbidos en el intestino y los pulmones. Existe bastante evidencia experimental que indica que los hidrocarburos son mutagénicos y carcinogénicos. Estudios específicos indican un riesgo mayor de desarrollar cáncer en personas ocupacionalmente expuestas a los hidrocarburos. Más específicamente, se ha encontrado que individuos que trabajan como conductores de camiones o mensajeros tienen un riesgo significativamente mayor de contraer cáncer de vejiga.

Los contaminantes como HC, CO y No_x pueden ocasionar problemas de consideración en la salud, dependiendo de la cantidad que se encuentre en la atmósfera, además son los

principales precursores del ozono; por esto la Secretaría de Salud estableció limites máximos permisibles de exposición en un determinado periodo de tiempo.

- *LÍMITES MÁXIMOS PERMISIBLES DE CONCENTRACIÓN DE CONTAMIANTES ATMOSFÉRICOS EN LA ZMCM*

Las normas que fijan los límites máximos permisibles de concentración de contaminantes, fueron publicadas por la Secretaría de Salud en el Diario Oficial de la Federación, los valores normados se muestran en la tabla 1.1.

Tabla 1.1. Valores normados para los contaminantes

Contaminante	Valores límite		
	Exposición agua		Exposición crónica
	Concentración y tiempo promedio	Frecuencia máxima aceptable	(Para protección de la salud de la población susceptible)
Ozono (O_3)	0.11 ppm (1 hora)	1 vez cada 3 años	-
Bióxido de nitrógeno	0.21 ppm (1 hora)	1 vez al año	-
Monóxido de carbono (CO)	11 ppm (8 horas)	1 vez al año	-

Fuente: Diario Oficial de la Federación del 3 de diciembre de 1994

La figura 2 nos muestra el número de días que se excedió la norma en la ZMCM en concentraciones de monóxido de carbono durante el periodo de 1988 a 1996. La norma indica que la concentración de monóxido de carbono no debe ser mayor de 11ppm en un periodo de 8 horas una vez al año y en la gráfica se puede ver por ejemplo que en 1991 se rebaso este límite por 77 días, aunque disminuyó para 1996 con 7 días que se excedió la norma pero aún es necesario disminuir la emisión de este contaminante a la atmósfera, ya que se sigue excediendo la norma.

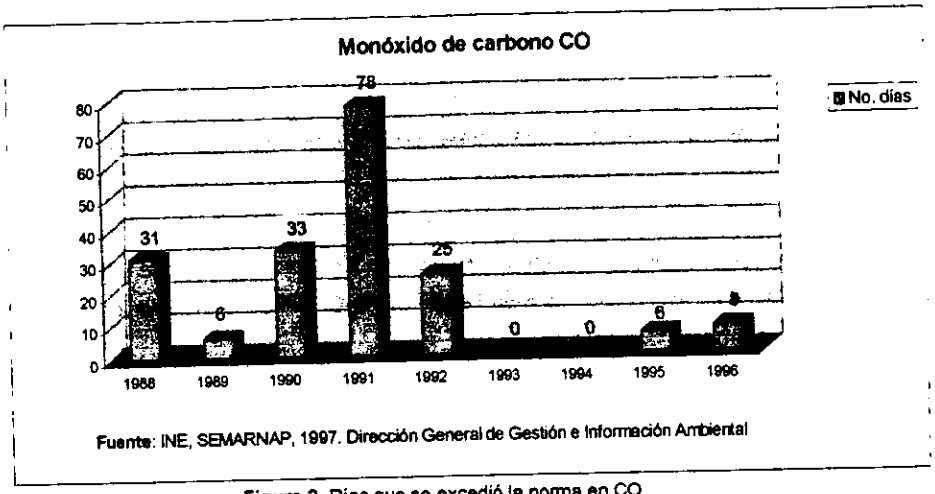

Figura 2. Días que se excedió la norma en CO

Para las emisiones de Bióxido de Nitrógeno se tiene para el período de 1988 a 1996 lo que se muestra en la figura 3.

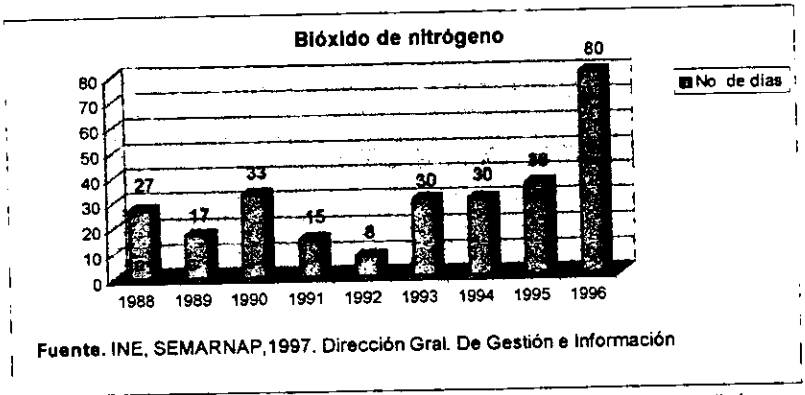

Figura 3. Número de días que se excedió la norma en emisiones de Bióxido de Nitrógeno

La norma establece no excederse más de una vez al año en emisiones de bióxido de nitrógeno. Los vehículos emiten monóxido de carbono como el contaminante más abundante (cerca dc dos millones de tonelada por año); siguen los hidrocarburos (por evaporación de combustión parcial, 300 mil toneladas por año); en menor cantidad están los óxidos de nitrógeno (90 mil toneladas por año); el de bióxido de azufre (12 mil toneladas al año); y relativamente pequeñas cantidades de compuestos como plomo (350 toneladas por año).

1.3 GASOLINA: FUENTE PRIMARIA DE CONTAMINACIÓN ATMOSFÉRICA EN LA ZMCM

La gasolina es una mezcla de varios cientos de compuestos llamados hidrocarburos que se obtienen del petróleo. Sirve para hacer funcionar a los motores de los vehículos de combustión interna, como son los automóviles, motocicletas, taxis, minibuses y algunos camiones de carga ligeros. A continuación se verá el proceso de combustión y cuáles son las principales fuentes de contaminación a la atmósfera.

1.3.1 COMBUSTIÓN Y FUENTES DE CONTAMINACIÓN

Durante la combustión, la energía contenida en los combustibles se transforma en calor y como resultado se generen gases contaminantes. Bajo condiciones ideales de combustión, la quema de estos energéticos generaría bióxido de carbono, vapor de agua y algunos gases inertes como el nitrógeno, los cuales no dañan la salud humana[4]. Sin embargo, ningún tipo de combustión es completo, por lo que se genera también un gran número de compuestos adicionales a los señalados, como se puede ver en la figura 4.

Las emisiones vehiculares dependen de la eficiencia de combustión y ésta, a su vez, de la relación aire/combustible adecuada para cada tipo de energético que alimenta al sistema. Cuando se tiene una mezcla con alto contenido de aire se generan grandes cantidades de monóxido de carbono y se reducen los hidrocarburos. Por el contrario, al tener una mezcla con alto contenido de combustible los hidrocarburos se incrementan y el monóxido de carbono disminuye.

En lugares situados por arriba del nivel del mar, la combustión es todavía menos eficiente debido a la baja presión atmosférica y a al consecuente deficiencia en la concentración de oxígeno en la atmósfera.

[4] El bióxido de carbono es un gas inerte, que se encuentra en forma natural en la atmósfera. Sin embargo. La quema de hidrocarburos genera un incremento excesivo en las emisiones de este gas, el cual es uno de los llamados "gases invernadero" que puede influir en el calentamiento global de la atmósfera. La única forma de reducir este fenómeno es disminuir el consumo de combustibles.

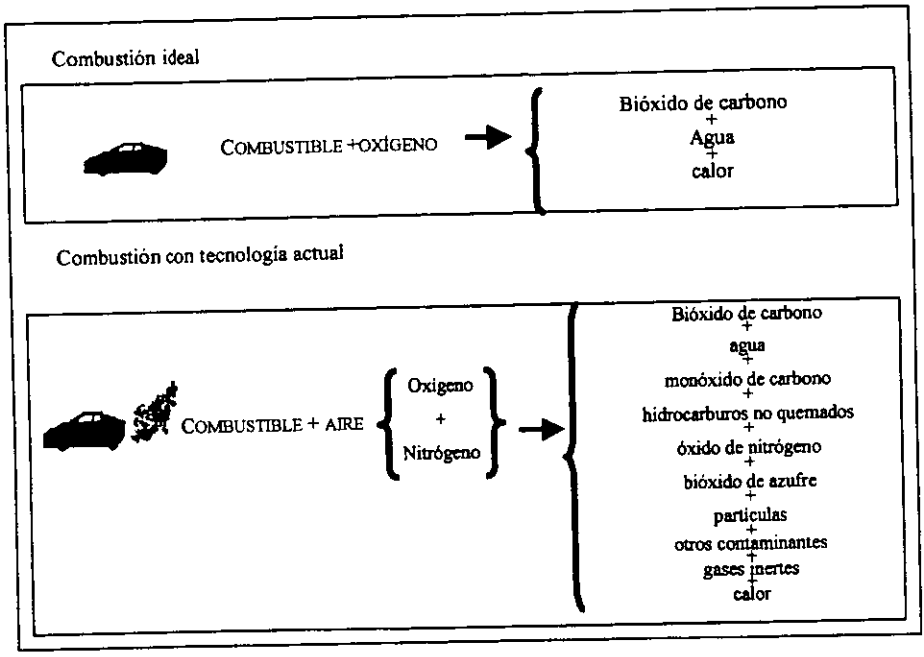

Figura 4. Combustión de los vehículos

Cualquier vehículo de combustión interna en circulación contamina la atmósfera de tres formas principalmente:

1. Mediante gases y humos derivados del sistema de combustión, ya sea por el tubo de escape, por el carburador, el tanque de gasolina o el cárter (que es un depósito que se encuentra en la parte inferior del motor y contiene el aceite para su lubricación).

2. Debido al ruido provocado por el motor y la carrocería.

3. Por partículas de caucho que se desprenden de la llantas. Estas partículas son peligrosas, ya que entran directamente al organismo por las vías respiratorias.

Los gases resultantes de la combustión incompleta pueden tener efectos sobre la salud humana como antes se señaló, la flora, la fauna y los materiales, dependiendo de su concentración y tiempos de exposición. Por ello, las emisiones de monóxido de carbono, óxidos de nitrógeno e hidrocarburos en el caso de vehículos a gasolina están normadas en nuestro país. En la tabla 1.2 se puede ver el inventario de emisiones que data de 1994 y se observa la manera en que contribuye cada uno de los sectores a la contaminación en la ZMCM.

Tabla 1.2. Inventario de emisiones 1994 (ton/año)

Inventario de emisiones 1994 (ton/año)							
Sector	Ton/año						
	PST	SO2	CO	NO_x	HC	Total	%
Industria	6,358	26,051	8,696	31,520	33,099	105,724	3
Servicios	1,077	7,217	948	5,339	398,433	413,014	10
Transporte	18,842	12,200	2,348,497	91,787	555,319	3,026,645	75
Vegetación y suelos	425,337	0	0	0	38,909	464,246	12
Total	451,614	45,468	2,358,141	128,646	1,025,760	4,009,629	100

Fuente: DDF, Gobierno del Estado de México, SEMARNAP, SS, 1996. Programa para mejorar la calidad del aire Valle de México 1995-200, pág. 74

Además de las industrias y los medios de transporte, existen otras fuentes de emisión de contaminantes: la combustión de gas en las estufas y hornos caseros; el combustóleo que utilizan las calderas de los hoteles y los baños públicos; los hornos de las panaderías; los generadores portátiles de energía que se utilizan en las construcciones, y el combustible que usan las máquinas pesadas, como revolvedoras de concreto y cemento, trascavos y grúas, entre otras. También contribuyen a aumentar la contaminación atmosférica: quemar basura y llantas al aire libre, defecar a la intemperie, fumar y quemar cohetes.

En la tabla 1.2 se ve que el transporte es de los que más contribuye a la contaminación en términos generales con un 75 % y como se vio anteriormente las emisiones de vehículos automotores son significativas en términos de precursores de ozono, como son hidrocarburos, óxidos de nitrógeno y desde luego monóxido de carbono. En la figura 3 se ve en qué proporción contribuyen los vehículos, minibuses, taxis etc.

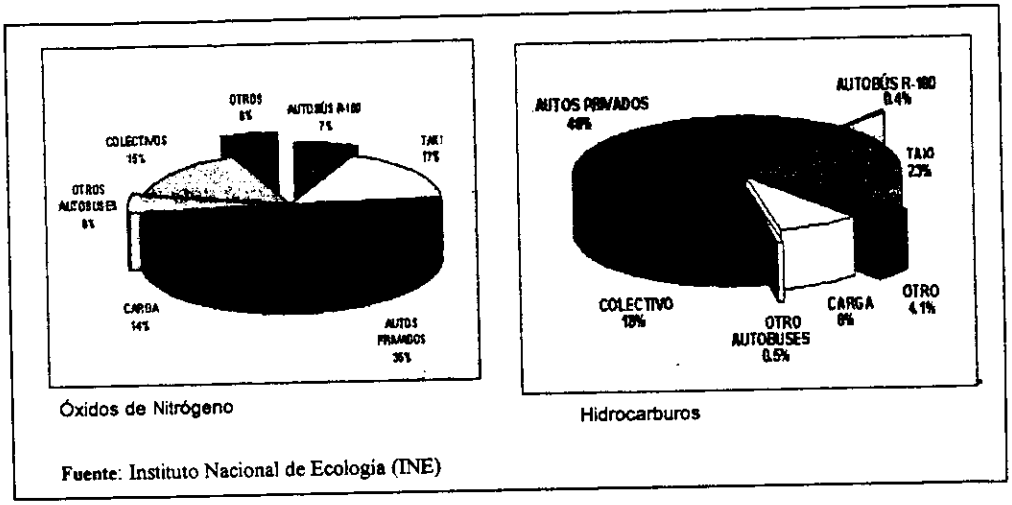

Figura 5. Inventario de emisiones de la ZMCM 1994. Contribución anula del sector transporte por tipo de vehículo

Se dice que los autos particulares son los que contaminan más el ambiente, como se puede ver en la figura 5, esto es su recorrido total (la suma de los recorridos de todos los autos particulares) es mayor al doble del recorrido de otros tipos de vehículos. Si bien los autos particulares pueden tener menores emisiones por unidad, son los más numerosos y en volumen contaminan más, aunado al hecho de que en ellos se transporta un número de personas menor al que utiliza transporte colectivo.

Los vehículos para que puedan funcionar necesitan de un combustible como la gasolina, a continuación se verá una breve historia del desarrollo de las gasolinas en México y de los principales componentes de estas.

1.3.2 GASOLINAS DE LA ZMCM

El uso masivo de las gasolinas empezó en la época de los años treinta con el desarrollo del automóvil. En esos tiempos los requerimientos de calidad de estos productos eran mínimos, situación derivada del incipiente desarrollo tecnológico de la industria de refinación automotriz, así como la inexistencia de regulaciones en materia de emisiones vehiculares.

En esa época las exigencias de calidad de la gasolina eran mínimas; básicamente se producía un combustible por destilación del crudo que resultaba de muy bajo octano[5] aproximadamente del orden de 57 (medido con base en el procedimiento del método de investigación, Research Octane Number –RON, que se explicará más adelante). En el caso de México, se le identificaba con el nombre de Gasolina.

En los años cuarenta la industria automotriz desarrolló motores de mayor relación de comprensión, los cuales demandaban de una gasolina con mayor octanaje, así en esas épocas nace en México el producto identificado con el nombre de Mexolina; éste presentaba un octanaje RON de alrededor de 70.

Entre los años 50's y 70's, las exigencias de calidad de las gasolinas cambiaron, como resultado de nuevos desarrollos tecnológicos de los motores para dar cumplimiento a los requerimientos se incorporaron nuevos procesos en la industria de refinación para producir naftas de alto octano. Para atender esta demanda, en nuestro país se comercializaban las siguientes gasolinas: Mexolina (70 octanos RON), Super Mexolina (80 octanos RON), Gasolmex (90 octanos RON) y la Pemex 100 (100 octanos RON).

En la década de los 70's, resultado del embargo petrolero en el Medio Oriente, se presenta a nivel mundial una crisis energética. Como respuesta a este evento, la industria automotriz diseña automóviles con menor peso, tamaño y una mayor economía de combustible (expresada como kilómetros recorridos por litro de gasolina consumida o millas por galón).

---

[5] El índice de octano, es la medida de la calidad y capacidad antidetonante de una gasolina y es indicativo del grado de eficiencia de la combustión, eliminando la presencia de explosiones múltiples dentro del motor, de forma tal que se produzca la máxima cantidad de energía útil.

Existen dos maneras de determinar el número de octano de una gasolina. La primera conocida por las siglas de RON (Research Octane Number), es una prueba que determina el desempeño de la gasolina en el motor bajo condiciones de operación moderadas y sin carga pesada (tal es el caso del comportamiento en la ciudad). La segunda, cuyas siglas de identificación son MON (Motor Octane Number), es una prueba que simula la operación de un motor en condiciones severas, altas velocidades y cargas elevadas (como es el caso del comportamiento en la carretera).

A fin de poder establecer el desempeño de la gasolina en los vehículos bajo cualquier condición de operación, en el ámbito internacional se emplea el parámetro que se denomina Índice de Octano (Antiknock Index, AKI). el cual se obtiene como la mitad de la suma de RON más el MON (su identificación internacional es (R+M)/2).

En respuesta, en nuestro país se eliminan los diferentes tipos de gasolinas comercializados, para dar paso a dos tipos de combustible identificados como Nova y Extra (la primera de 81 octanos y la segunda de 92 octanos, ambos valores expresados en RON).

Desde la época de los 30's hasta los finales de los 60's, la industria de refinación obtenía principalmente el incremento de octano a través de la incorporación de un aditivo antidetonante a partir de plomo (tretraetilo de plomo TEP); los valores típicos de concentración de metal fluctuaban entre 3 y 4 gramos por galón de gasolina (0.8 a 1.0 gramos por litro).

A mediados de los 70's, resultado de evaluaciones sobre el impacto a la salud del plomo y de la búsqueda de reducir la contribución de las emisiones vehiculares a la contaminación atmosférica, se inicia en los Estados Unidos el proceso de eliminación del plomo en las gasolinas. Esta acción demandó el desarrollo de nuevos procesos en la industria de la refinación que permitiesen sustituir el incremento del octano logrado con el TEP, por componentes obtenidos a través de la conversión de corrientes de bajo por alto octano. A partir de esta época se empieza el establecimiento de límites más estrictos a las emisiones generadas en los vehículos. La reducción y eliminación del plomo obedeció, como se citó anteriormente, en primer lugar, a los efectos nocivos para la salud que este metal tiene y, en segundo lugar, como una exigencia de la industria automotriz, quien incorporó la primera generación de dispositivos anticontaminantes (convertidores catalíticos) con el fin de satisfacer los requerimientos de las autoridades ambientales para obtener menos emisiones por distancia recorrida de los vehículos.

En lo que se refiere a las acciones que se han realizado para mejorar la calidad de los combustibles que consume el sector transporte de la ZMCM a partir de 1989, se puede ver en forma resumido en la Tabla 1.3.

Tabla 1.3. Mejoramiento en la calidad de combustibles

Mejoramiento en la calidad de los combustibles suministrados en la ZMCM
(1989 Reducción del 92% del contenido de plomo en gasolina Nova Plus (1989) Oxigenación de gasolinas, con le fin de favorecer una combustión más eficiente. (1990) Introducción de gasolina Magna Sin para cubrir la demanda de combustible de autos con convertidor catalitico. (1993) Introducción de Diesel Sin (0.05% de azufre en peso) (1996) Introducción de la gasolina Pemex-Magna con estrictas especificaciones de calidad. (1996) Introducción de la Gasolina Pemex Premium de 92 octanos.

Para satisfacer la demanda de los vehículos con convertidor catalítico a partir de 1989 se reduce considerablemente el uso del plomo en la gasolina por considerársele dañino para la salud, a partir de 1990 se introducen gasolinas con mayores índices de octano como es la Magna Sin (actualmente conocida como Pemex Magna) y en 1996 la comercialización de la gasolina Pemex Premium con mayor índice de octano. La gasolina Pemex Magna garantiza un octanaje de 87 R+M/2 libre de plomo y la gasolina Pemex Premium garantiza un octanaje de 92 R+M/2 es decir 5 puntos más que la gasolina Pemex Magna.

En la actualidad se están analizando combustibles alternos como es el gas natural o automóviles eléctricos, pero estas tecnologías son todavía muy caras, por lo que tal vez se tenga que esperar un poco para poder implementarlas.

Las características que sirven para especificar una gasolina son la presión de vapor que determina su volatilidad, el índice de octano que representa su facilidad de combustión y el contenido de compuestos específicos o de familias de ellos. En particular, se especifica la cantidad de compuestos considerados tóxicos o de los aditivos. En la tabla 1.4 se muestran estas características. Además puede verse como la gasolina Pemex-premium reduce los porcentajes de ciertos compuestos de las gasolinas.

Tabla 1.4. Características de las Gasolinas en la ZMCM

ESPECIFICACIÓN	UNIDADES	PEMEX-MAGNA	PEMEX-PREMIUM
Azufre	% en peso	0.050	0.17
Aromáticos	% en vol.	24.6	24.85
Olefinas	% en vol.	6.2	4.86
Presión vapor (PVR)	Psi (lb/pug^2)	6.8	7.05
Benzeno	% en vol.	0.86	0.50
Octano	(R+M)/2	88.6	92.2
Oxigeno	% en peso	1.27	0.89
Plomo	Kg/m^3	0.00039	0.0001

Fuente: Pemex, Refinación 1996.

A continuación se verá de que nos sirve reducir algunos de los parámetros o características de las gasolinas para disminuir la emisión de contaminantes.

Las principales causas por las que se emiten contaminantes por el uso de gasolinas son por la combustión, como ya se explicó, y por la evaporación, ya que gasolinas al evaporarse emiten hidrocarburos. Se han iniciado varios programas para contrarrestar la evaporación, uno de ellos es suministrar gasolinas con menor volatilidad, otro, la instalación de techo flotante en tanques de almacenamiento. Un tercero es la recuperación de vapores en las gasolineras durante el suministro y que los tanques de gasolina de vehículos tengan tapas con sello hermético, lo mismo que los recipientes usados para transportar o almacenar disolventes.

- *Parámetros de la volatilidad de la gasolina*

La volatilidad de una gasolina está determinada por tres parámetros: *la curva de destilación*, *la Presión de Vapor Reid* (conocida en la industria petrolera por las siglas PVR o en inglés como RVP) y la *relación vapor/líquido* (identificada como V/L).

Estos parámetros son indicadores que permiten controlar el comportamiento de la gasolina en los vehículos bajo cualquier condición climatológica, es decir, un arranque eficiente del motor tanto en climas fríos como en calientes. La volatilidad de una gasolina debe ser tal que permita que ésta se vaporice adecuadamente en la cámara de combustión, a fin de lograr un mezclado efectivo de la combinación aire-combustión, de tal forma que se obtenga el máximo aprovechamiento del combustible en el motor.

Si la gasolina es demasiado volátil se produce, en climas calientes, el fenómeno denominada sello de vapor (vapor lock), impidiendo el arranque del vehículo. Por otro lado, si el combustible es demasiado pesado y no tiene la volatilidad adecuada, el motor no encenderá en climas fríos, debido a que la gasolina se mantiene en forma líquida.

En suma, la volatilidad de la gasolina deberá estar bien balanceada, es decir, ni muy pesada ni muy volátil, para garantizar una operación eficiente de los motores bajo cualquier condición climatológica. Por tal motivo, este parámetro se ajusta de acuerdo a la estacionalidad de cada región.

El inmenso consumo de combustibles de los vehículos automotores constituye la principal fuente de emisiones contaminantes en el valle de México. El consumo total de combustibles fósiles en el Valle durante 1996 ascendió a cerca de 40 millones de litros diarios; más de 54% de estos combustibles correspondieron al transporte, que a su vez muestra un comportamiento ascendente y estrechamente vinculado con las fluctuaciones en la economía nacional. En gasolinas, el consumo promedio diario se incrementó de 1989 a 1996 en 13%, al pasar de 15.4 millones de litros diarios a 17.2 millones respectivamente.

Se ha visto como el uso de gasolinas contribuye a la emisión de contaminantes, pero sabemos que además de la calidad de las gasolinas existen otros factores que contribuyen a la emisión de contaminantes como los vehículos como se verá a continuación.

1.4 LOS AUTOMÓVILES EN LA ZMCM

En la Ciudad de México las primeras medidas que tomó el gobierno para reducir la emisión de contaminantes generadas por los vehículos automotores se iniciaron a principios de la década de los setenta. Los esfuerzos se centraron principalmente en la detención de unidades que emitían humos en forma ostensible, a través de su identificación visual, sin embargo los programas para concientizar a los automovilistas han evolucionado.

➢ *Evolución de las medidas para reducir la emisión de contaminantes*

En 1975 fueron incorporados los primeros analizadores manuales de gases para medir las concentraciones de monóxido de carbono e hidrocarburos, como apoyo a los operativos de detención de vehículos con emisiones ostensibles. Asimismo se inició la instalación de centros equipados para el diagnóstico de la afinación.

En años posteriores se buscó que los automovilistas acudieran a verificar las emisiones de sus unidades, en forma voluntaria y gratuita. Al mismo tiempo, el número de centros de verificación fue en aumento.

En 1985 por una propuesta de la SECOFI se prohibe la fabricación de motores de 8 cilindros. En 1989 se hizo obligatoria la verificación vehicular en la Zona Metropolitana de la Ciudad de México, tanto para los vehículos a gasolina como a diesel. A partir de entonces, en número de centros de verificación se incrementó rápidamente para atender la totalidad de la flota vehicular. Además se incluye la certificación de óxidos de nitrógenos (NO_x) para automóviles nuevos. En el mismo año de 1989 dadas las condiciones meteorológicas que dificultaban la dispersión de contaminantes, se inicia la aplicación obligatoria del programa "Hoy no circula" en la ZMCM, que sigue vigente hasta la fecha.

Para 1990 se incluye la verificación vehicular semestral en el D.F y Estado de México y la verificación semestral para camiones de carga que circulan por carreteras federales. En 1991 se introduce el convertidor catalítico[6] en vehículos ligeros de la Chrysler, General Motors, Ford, Nissan y Volkswagen. Y en 1994 la introducción de convertidores catalíticos de tres vías en todos los modelos nuevos de automóviles a gasolina.

El programa de verificación de emisiones, es una de las medidas tendentes a disminuir la contribución de vehículos en la contaminación ambiental. En 1996 se implementa el "Doble

[6] El convertidor catalítico es un dispositivo, antes de la salida de los gases de escape en forma de panal, que disminuye las emisiones de hidrocarburos, monóxido de carbono y óxidos de nitrógeno. Al reducir las emisiones de escape mediante este dispositivo, se evita la formación de ozono correspondiente.

El tiempo de vida medio del convertidor catalítico es 150,000Km, pero se modifica de acuerdo con el mantenimiento del vehículo y la gasolina utilizada. Por lo que los automóviles con convertidor catalítico deben utilizar gasolina sin plomo, ya que el plomo envenena al convertidor catalítico.

hoy no circula", con el cual a aquellos vehiculos que cuenten con calcomanía dos, en días de contingencia se les prohibe circular dos días a la semana, incentivando así la renovación del parque vehicular. Y por último a principios de 1999 se introdujo la calcomanía doble cero que exenta de la verificación por dos años a los vehículos nuevos siempre y cuando cuenten con tecnologia de punta.

➢ *Características tecnológicas de los automóviles que circulan en la ZMCM son:*

- Modelos anteriores a 1990 contienen carburadores, encendido electrónico, no contienen convertidor catalítico, relación de compresión hasta 9.1:1 y como se vio anteriormente a mayor relación de compresión mayor octanaje.

- Modelos de 1991 y 1996 contienen inyección electrónica (TBI, MPI, SMPI), encendido electrónico, control de circuito cerrado, convertidor catalítico de 3 vias, relación de compresión hasta 10:1.

- Modelos 1997 contienen inyección electrónica (MPI, SMPI), control de circuito cerrado con memoria adaptativa, sistema OBD, convertidor catalítico de tres vias y relación de compresión hasta 10.4:1.

Por lo que para los modelos de 1991 en adelante se necesitan gasolinas sin plomo y con un mayor índice de octano.

➢ *Factores que contribuyen a la contaminación de los vehículos*

Los autos mal afinados o mal mantenidos contaminan mucho más que los que reciben un mantenimiento adecuado, aunque también existen factores relacionados con la tecnología de vehículos (con carburador o inyección de combustible); su cilindrada; el modelo y la utilización de dispositivos anticontaminantes.

- *Tecnología:* La tecnología influye en que autos con carburador, producen mayores emisiones evaporativas que las que tiene sistema de inyección de combustible.

- *Cilindrada:* La cilindrada influye a la emisión de contaminantes a la atmósfera en general ya que entre más grande el motor, es decir a mayor cilindrada, se tiene mayores emisiones de escape.

- *Modelo:* La antigüedad, ya que usualmente los autos nuevos contaminan mucho menos que los viejos.

- *Dispositivos:* El utilizar dispositivos en general reduce la emisión de contaminantes, por ejemplo el cánister (dispositivo que evita las emisiones al aire de gasolina sin combustionar), reduce las emisiones evaporativas de hidrocarburos, mientras que el convertidor catalítico reduce en gran proporción las emisiones de escape de CO, HC y en menor proporción los óxidos de nitrógeno.

La persistencia de tecnología obsoleta y el mantenimiento inadecuado de los motores determinan que 20% de los vehículos que circulan en la ZMCM (los más antiguos y/o en condiciones mecánicas inapropiadas) emitan el 70% de las emisiones de hidrocarburos y el 55% de las de monóxido de carbono.

Además, otros factores que contribuyen a la emisión de contaminantes son: la antigüedad del parque vehicular, la tecnología de motores, la presencia de sistemas de control de emisiones, las condiciones y frecuencia del mantenimiento de los vehículos y además la cantidad de vehículos en circulación.

En 1996 el total de vehículos en la ZMCM se estimaba en 3 millones 157 mil 761 unidades. De esta cifra, los automotores de uso particular son los más numerosos, con 72% del total, como se muestra en tabla 1.5.

En la ZMCM existen 1.9 vehículos por cada diez habitantes. La flota vehicular en circulación ha venido aumentando en forma significativa durante los últimos años, con tasas de crecimiento anual de alrededor de 10% superior al ritmo de crecimiento de la población. Actualmente, la edad promedio de la flota vehicular es de 8.5 años, con una distribución ampliamente extendida en la que persiste una importante cantidad de vehículos viejos, aunque excepcionalmente resalta el hecho de que los taxis y microbuses son de modelos

relativamente recientes, como resultado de las políticas de renovación impuestas en el Distrito Federal.

Tabla 1.5. Parque vehicular en la ZMCM.

Tipo de vehículo	Unidades
Automóviles particulares	2,301,445
Taxis	91,652
Combis y microbuses	52,158
Autobuses urbanos	2,794
Autobuses suburbanos	1,284
Autobuses particulares	4,013
Transporte de carga	463,962
Otros	240,453
Total	3,157,761

(Programa de verificación vehicular, Programa de Placa Permanente, Registro vehicular del Edo. De México, etc.)

1.4.1 Normas ecológicas para emisión de contaminantes en vehículos nuevos y en circulación

A pesar de los problemas de calidad del aire en la ZMCM que se manifestaron desde fines de los años setenta, no fue sino hasta los últimos años de la década de los ochenta que se empezó a construir una infraestructura normativa y regulatoria para combatirla. En esos años se define las llamadas normas ecológicas de emisiones para vehículos nuevos y en circulación, las cuales desde su origen mostraron rezagos importantes con respecto a la normatividad establecida en Estados Unidos.

A partir de la creación de la Ley de Metrología y Normalización esta norma se convirtió en la Norma Oficial Mexicana (NOM), y posteriormente fue modificadas en función de las nuevas posibilidades tecnológicas de las empresas automotrices en México.

En la norma NOM-041-ECOL-1996, se establecen los límites máximos permisibles de emisión de gases contaminantes provenientes del escape de vehículos automotores en circulación que usan gasolina como combustible en la ZMCM. Se muestra la normas en la

tabla 1.6. Puede verse que todavía no se han normado para las emisiones de óxidos de nitrógeno, a pesar de ser uno de los principales precursores de la formación del Ozono. Además de que las mediciones de los contaminantes se realiza en partes por millón y porcentaje de volumen .

Tabla 1.6. NOM-041-ECOL-1996 que establece los límites máximos permisibles de emisión de gases contaminantes provenientes del escape de los vehículos automotores en circulación que usan gasolina como combustible

Año / Modelo	Hidrocarburos (HC)ppm	Monóxido de carbono (CO)% vol.	Oxígeno(O_2) máximo % vol.	Dilución ($CO+CO_2$)% vol. Mínima	máxima
1985 y anteriores	350	3.5	6	7	18
1986-1990	300	3.0	6	7	18
1991 y posteriores	200	2.0	6	7	18

*Los vehículos de cualquier año-modelo que cuenten con bomba de aire como equipo original, tienen un límite máximo en oxígeno de 15% de volumen

Fuente: Diario Oficial de la Federación 25 de febrero de 1997.

La NOM 042 establece los límites máximos permisibles de emisión de hidrocarburos, monóxido de carbono y óxidos de nitrógeno provenientes del escape de vehículos automotores nuevos en planta, con base a la cual se realizó este estudio, en ésta norma se manejan las mediciones en g/km. Esta norma oficial Mexicana es obligatoria para los fabricantes e importadores de vehículos automotores con peso de 400 a 3,857 kilogramos. en la tabla 1.7 se muestra esta norma.

Tabla 1.7. NOM-042

Año / Modelo	Niveles máximos permisibles de emisión (g/km)			
	Hidrocarburos	Monóxido de carbono.	Óxidos de Nitrógeno	g/prueba HC evaporativos
1994	0.25	2.11	0.62	n.a
1995 en adelante	0.25	2.11	0.62	2.0

n.a. : no aplica

Fuente: Diario Oficial de la Federación, 22 de octubre de 1993.

Otras alternativas de combustible para los vehículos de combustión interna son diesel, gas natural, el gas licuado, el etanol y el metanol, aunque estos dos alcoholes, por su alta volatilidad y desventajas en su manipulación y rendimiento, no se recomiendan en la ZMCM.

En un estudio realizado por el Instituto Mexicano del Petróleo se encontró que mediante determinaciones experimentales de las emisiones de vehículos en circulación, si se cambiaran los coches viejos, con más de 10 años en la ZMCM, se podría reducir la emisión de contaminantes, ya que, el 45% de los vehículos existentes en la ZMCM producen el 70% de las emisiones provenientes de las fuentes móviles y corresponden a vehículos que tienen más de 10 años; mientras que el 32% de vehículos que tiene menos de 5 años, contribuyen con menos del 5% de los contaminantes. El retirar estos coches viejos de la circulación, disminuirán apreciablemente las emisiones atribuidas a los vehículos; el cambio de coches viejos por nuevos también reducirá las emisiones en gran proporción.

Para reducir la formación de ozono, que es uno de los principales contaminantes en la ZMCM, es necesario controlar a sus principales precursores que son NO_x, HC y CO que en su mayoría son generados por los vehículos particulares automotores. Por lo tanto es de vital importancia controlar el problema de contaminación; las medidas que hasta ahora se han tomado no han sido suficientes.

Los principales problemas a los que se enfrenta uno para poder realizar investigaciones acerca de las emisiones es que la empresas que se dedican a hacer estos estudios tiene sus emisiones en gr/km y en la norma los encontramos en ppm (partes por millón) y es muy costoso hacer este tipo de estudios, además que para poder hacer un estudio que nos de cierto nivel de confianza se necesitaría una muestra lo suficientemente grande para poder hacer estimaciones, surgen las preguntas ¿si se genera una población sintética que tan parecida será a la población real?, y ¿mediante esta población sintética puedo hacer inferencias y que nivel de confianza tiene?.

En el siguiente capítulo se verá que se necesita para poder generar una población sintética además de como hacerle para escoger la muestra lo más representativa posible.

Si se comprueba que las poblaciones sintéticas son en realidad una representación confiable de la población podría ahorrarse mucho ya que no se necesitaría tomar muestras tan grandes para poder realizar los estudios.

Conceptos clave

Benzo_a.pireno: Hidrocarburo, sustancia tóxico que se encuentra en escapes de gases de motores de gasolina, hidrocarburos, crudo , asfalto, combustión de sustancias orgánicas, alquitrán de la madera y el carbón.

Carcinogénicos: Cualquier substancia o agente que produce cáncer, que acelera el desarrollo del cáncer o que actúa sobre una población para cambiar la frecuencia total del cáncer en términos del número de tumores o distribución, de acuerdo con el sitio y edad.

Combustibles fósiles: Son aquellos que provienen de descomposición de materia animal como el petróleo y el gas natural.

Mutágeno: Sustancia o agente, que causa mutación genética.

II

MARCO TEÓRICO CONCEPTUAL

2.1 INTRODUCCIÓN

Las técnicas estadísticas se utilizan en muchos aspectos de la vida diaria por ejemplo: en la actualidad el 82% de los navegantes de internet pertenecen al sexo masculino con 31 años de edad promedio[1] o que los vehículos automotores contribuyen en un 75% a la contaminación, etc. Para hacer este tipo de afirmaciones se necesita recolectar datos los cuales pueden ser una muestra representativa de la población, es decir, tiene que ser elegido de tal forma que refleje con precisión las características del fenómeno que se analiza, para poder mediante ésta muestra hacer inferencia estadística de la población (sacar conclusiones sobre una población a partir de una muestra). Otra manera para conocer las características de una población es el censo el cual resulta costoso, debido a que se necesita recabar la información de cada uno de los elementos de la población, además exige la movilización de muchos recursos humanos y su duración suele ser muy larga. Para el conocimiento de la población existen métodos alternativos que están constituidos por las muestras, cuya finalidad es construir modelos reducidos de la población total, con resultados extrapolables al universo del que se extraen, aunque habrá casos en los que por razones administrativas o porque así lo requiere la investigación se tendrá que hacer el recuento censal. En algunas ocasiones no es posible obtener el total de la muestra de la población real, así que se usará una población sintética, que se obtiene con base en las características de la población real y que pretende ser lo más parecido a ésta.

En este capítulo se verán algunas de las características que deben de tener las muestras, la manera en como se pueden agrupar los datos, de que tamaño se debe elegir la muestra para decir que es representativa y como se pueden validar los datos. Otra cuestión importante es en caso que no se puedan obtener todos los datos necesarios qué alternativas se tienen, como la creación de poblaciones sintéticas y qué tan válidas son, por

lo que se explicará el método para su creación. También se verá de una manera muy breve la técnica de simulación del método de Monte Carlo, con el cual se generó la población sintética.

2.2 MÉTODO ESTADÍSTICO Y LA INVESTIGACIÓN

La estadística a menudo ha sido clasificada como un método de investigación, asociado con o en contraposición a métodos tales como el estudio de casos, el análisis cronológico y la experimentación. La ciencia de la estadística tiene mucho que ofrecer al investigador en planeación. Estadística es una ciencia[2] y utiliza lo siguiente:

1. Colección y compendio de datos
2. Diseño de experimentos y reconocimientos
3. Medición de la variación, tanto en datos experimentales como de reconocimiento.
4. Estimación de parámetros de población y suministro de varias medidas de la exactitud y precisión de esas estimaciones.
5. Ensayo de hipótesis respecto a poblaciones.
6. Estudio de la relación entre dos o más variables

La estadística interviene en la investigación y/o el método científico, a través de la experimentación y la observación. Esto es, las observaciones experimentales y reconocimientos son partes integrantes del método científico, y esos métodos invariablemente conducen al empleo de técnicas de la estadística. Existen tres tipos de técnicas estadísticas, la descriptiva, inferencial y multivariada, aquí solo se describirán las dos primeras. Hay estadísticas que se recolectan mediante censos, otro tipo es el que se recolecta mediante encuestas por muestreo, donde sólo se emplea una fracción del universo para proveer información acerca del conjunto. Esto es, se examina sólo algunos de los objetos para poder extender la información al total.

[1] El universal, lunes 15 de marzo de 1999
[2] Freund, E., Smith R, (1989), pag. 19

2.2.1 ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

La *estadística descriptiva* nos permite describir y caracterizar la realidad cuantificada. Representa una primera fase del análisis estadístico que nos facilita información útil y valiosa para proseguir con el estudio inferencial o plantear investigaciones posteriores. Su contenido es cuantitativamente más reducido que el de la estadística *inferencial o multivariada*, más elemental y de fácil comprensión, pero no por ello menos importante, tanto por la información que nos proporciona como por constituir la base del estudio del contenido posterior de la estadística.

Un problema común para la *inferencia estadística* es determinar en términos de probabilidad si las diferencias observadas entre dos muestras significa que las poblaciones muestreadas son realmente diferentes. El procedimiento de *inferencia estadística* nos permite determinar si las diferencias observadas en las dos muestras están dentro del rango de la casualidad o no. Otro problema es determinar si una muestra de puntos es de alguna población especial. Y más aún es decidir si podemos legítimamente inferir que algunos grupos difieren entre si.

La estadística intenta obtener inferencia con respecto a la población basándose en la información contenida en la muestra. Las poblaciones se describen mediante medidas numéricas denominadas parámetros, la mayoría de las investigaciones estadísticas tratan de deducir la inferencia con respecto a uno o más parámetros de la población. La mayoría de los procedimientos de inferencia involucran *estimación* ó el *contraste de hipótesis*.

- *Estimación*

La estimación tiene muchas aplicaciones prácticas. Por ejemplo para la contaminación se podría estar interesado en estimar la proporción de *p* vehículos que se encuentran dentro de la norma en el primer semestre del año. Otros parámetros poblacionales importantes son la media μ, la varianza σ^2 y la desviación estándar σ de la población. Si se desea estimar el promedio de las emisiones de HC por vehículo en la ZMCM, se puede presentar la estimación de dos maneras distintas, se podría dar de un solo número por ejemplo 2.25 gr/km. La intención es que este número esté cerca de la media μ desconocida de la

población. Este tipo de estimación se denomina *estimación puntual*[3], ya que se da un solo valor o punto. También se podría decir que μ quedaría entre dos número, en este tipo de estimación se dan dos valores que se pueden utilizar para construir un intervalo que se espera que contenga el parámetro en estudio. Este segundo tipo de estimación se denomina *estimación por intervalo.*

➢ *Contraste de hipótesis*

El contraste de hipótesis o pruebas de hipótesis es en muchos aspectos similar al método científico. El científico o persona que está investigando, observa un fenómeno, establece una teoría y después prueba su teoría con respecto a la observación, ya que propone una teoría relativa a los valores específicos de uno o más parámetros, luego obtiene una muestra de la población y compara la observación con la teoría, si las observaciones se contraponen a la teoría, el investigador rechaza la hipótesis, en caso contrario puede concluir que la teoría es válida o bien que la muestra no detectó la diferencia entre los valores reales y los valores de los parámetros poblacionales. De manera similar ocurre con el contraste de hipótesis que tienen los siguientes elementos[4]:

1) *Hipótesis:* El primer paso para poder contrastar hipótesis es establecer las hipótesis, éstas son de dos tipos; *hipótesis nula* que se denota H_0 es una hipótesis de no efecto y normalmente se formula con el propósito de ser rechazada; y *la hipótesis alternativa* que se denota H_a o H_1 que es la propuesta operacional que el experimentador esta buscando, por ejemplo en una moneda se intentará saber si la moneda está cargada, por lo que H_0 será que la moneda no está cargada y H_a que la moneda está cargada. Las hipótesis H_a y H_0 son complementarias y mutuamente excluyentes.
2) *Elegir el estadístico de prueba.* Es el estimador propuesto para obtener información de la población a través de la muestra probabilística por ejemplo: la media, la varianza, etc.
3) *Nivel de significación.* Es una probabilidad muy pequeña que usualmente se denota como α y sus valores más comunes son de .05 y .01, la cual se compara con la probabilidad asociada con el valor del estadístico de prueba (esta tiene que ser menor o igual que el nivel de significación).

[3] Los estimadores puntuales pueden ser insesgados, consistentes, eficientes o sufiecientes dependiendo de sus propiedades (ver Anexo 1)
[4] Siegel S. & Castellan N., 1988, pags. 7-15.

4) *La distribución de muestreo.* La distribución de muestro es una distribución teórica, esto es, qué distribución tendríamos si se tuvieran todas las posibles muestras del mismo tamaño de la misma población, obtenidas aleatoriamente. Por ejemplo, el decir que la variable se distribuye normalmente con media=μ y desviación estándar=σ cuando *N* es grande. Aunque si el tamaño de la muestra es pequeña se utiliza la t-student.
5) *Región de rechazo.* Que se denota como RR especifica los valores del estadístico de la prueba para los cuales se rechaza la hipótesis nula, por ejemplo supóngase que se tiene la hipótesis nula H_0 : μ= 10 y se decide rechazar H_0 si se observa un valor de la media muestral $\bar{X}$ más grande que 12, por lo que el conjunto de valores mayores que 12 constituyen la región crítica o región de rechazo.
6) *Regla de decisión.* Si el valor del estadístico de prueba está en la región crítica, la decisión es rechazar H_0 ; si el valor del estadístico de prueba esta fuera de la región de rechazo (i.e. dentro de la región de aceptación) la decisión es que no se puede rechazar la H_0. Al tomar este tipo de decisiones se pueden cometer errores los cuales se pueden ver más claramente en la siguiente tabla.

		Hipótesis	
		Verdadera	Falsa
Decisión	Aceptarla	Correcta $1-\alpha$	Error Tipo II β
	Rechazarla	Error Tipo I α	Correcto $1-\beta$

β = P(Aceptar H_0 / H_0 es falsa)

α = P(Rechazar H_0 / H_0 es verdadera)

$1-\alpha$ = P(Aceptar H_0 / H_0 es verdadero), conocido como nivel de confianza o de significación.

$1-\beta$ = P(Rechazar H_0 / H_0 es falsa), se conoce como potencia de la prueba.

Cuando podemos suponer la naturaleza de la población de la cual las observaciones o datos fueron obtenidos se dice que es estadísticamente paramétrica. La distribución libre o no paramétrica resulta de técnicas en las cuales se requieren pocos requisitos. Si se quieren reducir los errores de muestreo se debe de incrementar el tamaño de la muestra.

Antes de poder elegir el modelo estadístico debe uno de fijarse en los supuesto que implica el utilizar cierto tipo de prueba, cuando éstos no se cumplen o no se está seguro del comportamiento que tiene la muestra se utiliza la estadística no paramétrica, ya que utilizar

un modelo estadístico establecido si la muestra no tiene las características los parámetros pueden caer en la región de rechazo.

2.3 ESCALAS DE MEDIDA

Medida es el proceso de mapeo o asignación de números a objetos u observaciones. El tipo de medida está en función de las reglas bajo las cuales éstos números son asignados a los objetos. Éstas son solo una forma para describir o agrupar los datos. Nos ayuda a establecer la relación entre los objetos que están siendo observados y el número asignado. Se discutirán cuatro de las escalas más generales: nominal, ordinal, intervalo y razón.

- *ESCALA NOMINAL O CATEGÓRICA*

  En medidas la escala nominal es el nivel más bajo que existe, cuando el número y otros símbolos son usados simplemente para clasificar un objeto, persona o característica. Esta escala también es conocida como una escala de clasificación. El tipo de relación que puede existir en este tipo de escala es la de equivalencia (=).

- *ESCALA ORDINAL O DE RANGO*

  Puede suceder que los objetos en una categoría de una escala sean no solo diferentes de los objetos de otra sino que también conservan algún tipo de relación entre ellos. Relaciones típicas entre las clases son; más alto que, mayor preferencia, más difícil, más maduro, etc. Tal relación puede ser denominada por el símbolo ">" el cual, en general significa mayor que, los significados dependen de la naturaleza de la relación que define la escala.

Dado un grupo de clases equivalentes (i.e. dada una escala nominal), si la relación mayor que, está entre algunos pero no todos los pares de clases, se tiene una escala parcialmente ordenada. Si la relación mayor que tiene todos los pares de clases así como un orden completo de rangos como es posible, se tiene una escala ordinal.

- *ESCALA DE INTERVALO*

Cuando una escala tiene todas las características de una escala ordinal, y además tiene la distancia o diferencia entre dos número cualesquiera, una medida considerablemente más fuerte que la ordinal se ha conseguido en el sentido de una escala de intervalo. Si en nuestro mapeo de varias clases de objetos, es preciso que se conozca qué tan grandes son los intervalos entre todos los objetos en la escala y el significado substancial, entonces se consigue medirlas por intervalos. Una escala por intervalos se caracteriza por una común y constante unidad de medida la cual designa un número a todos los pares de objetos en un orden. En este orden de medidas, la razón de dos intervalos cualesquiera es independiente de la unidad de medición o del punto cero. En una escala de intervalo, el cero y la unidad de medida son arbitrarias es decir el cero no indica ausencia del atributo que se mide.

- *ESCALA DE RAZÓN*

Cuando una escala o medida tiene todas las características de una escala de intervalo y además, tiene un punto cero como origen, es decir donde el cero indica ausencia de atributo entonces se tiene una escala de razón.

La escala nominal y de orden son las medidas más usadas en las ciencias de comportamiento y social. Los datos para cada escala nominal y ordinal deben ser analizados por métodos no paramétricos. Los datos medidos en escala intervalo o de razón pueden ser analizados por métodos paramétricos si el modelo estadístico es válido para los datos y cumple con los supuestos de cada modelo.

2.4 PRUEBAS ESTADÍSTICAS

Para poder elegir una prueba estadística apropiada se debe considerar la forma en la cual se obtuvo la muestra de datos o resultados, la naturaleza de la población, la hipótesis que se quiere probar, y el tipo de medida o escala la cual se empleara en la definición del manejo de las variables involucradas, todo esto determina cuál prueba estadística es la óptima o la más apropiada para analizar un grupo de datos o una muestra. Las dos grandes divisiones que tienen estas pruebas se basan en los diferentes niveles de medidas o clasificación (también conocidos como escalas) y son las pruebas paramétricas y no paramétricas y cada una de estas a su vez se pueden dividir por el número de variables independientes, número de grupos, si están correlacionadas y el tamaño de la muestra.

Una vez que se han analizado los diferentes tipo de escala, se verán las divisiones que éstas originan: pruebas paramétricas y no paramétricas. Una *prueba estadística paramétrica* especifica ciertas condiciones sobre la distribución de la respuesta en la población de la cual se obtuvo la muestra. Donde estas condiciones no son ordinariamente probadas se suponen válidas. El sentido de los resultados de una prueba paramétrica depende de la validación de estos supuestos. *Una prueba estadística no paramétrica* se basa en un modelo que especifica condiciones muy generales y ninguna determina la forma específica de la distribución de la cual la muestra se obtiene. Ciertos supuestos se asocian con la mayoría de las pruebas estadísticas no paramétricas, al saber que las observaciones son independientes, quizá el que las variables bajo estudio tienen subrayada continuidad, pero estos supuestos son menos y más claros que esas asociaciones con pruebas paramétricas. Además, como se verá, los procedimientos no paramétricas suelen probar diferentes hipótesis acerca de la población que los procedimientos paramétricos.

Finalmente, a diferencia de las pruebas paramétricas, las pruebas no paramétricas pueden ser aplicadas apropiadamente a datos medidos en una escala ordinal y otros datos en escala nominal o escala categórica.

Los criterios que se deben de considerar para elegir una prueba estadística y tomar la decisión en las hipótesis son: 1) la aplicabilidad o validación de las prueba (la cual incluye el nivel de medida y los demás supuestos de la prueba) y 2) el poder y eficiencia de la prueba.

Se pueden hacer las siguientes preguntas para saber que tan válido es usar determinada prueba.

1. De los métodos disponibles, paramétricos o no paramétricos, ¿cuál utiliza la información de la muestra apropiadamente?. Esto es, ¿cuál de las pruebas es válida?.
2. ¿Tengo los supuestos bajo los cuales un modelo estadístico o prueba se satisfacen?.
3. ¿Son las pruebas de hipótesis del modelo estadístico las apropiadas para la situación?.

VENTAJAS DE LAS PRUEBAS ESTADÍSTICAS NO PARAMÉTRICAS

- Si el tamaño de la muestra es muy pequeño, puede no ser una alternativa el usar las pruebas estadísticas no paramétricas a menos que se conozca exactamente la naturaleza de la población.
- Las pruebas estadísticas no paramétricas hacen pocos supuestos acerca de los datos y pueden ser más relevantes para una situación particular. Además, las pruebas de hipótesis para una prueba no paramétrica puede ser más apropiada para la investigación.
- Las pruebas estadísticas no paramétricas pueden analizar datos los cuales son inherentes en rangos, es decir, los métodos no paramétricos pueden categorizar o agrupar los datos como más o menos, mejor o peor y mayor o menor.
- Los métodos no paramétricos pueden tratar datos los cuales están simplemente clasificados o por categorías i.e. están medidos en una escala nominal.
- Son convenientes las pruebas no paramétricas cuando se trata de muestras hechas de observaciones de poblaciones muy diferentes probabilísticamente.
- Las pruebas estadísticas no paramétricas son típicamente más sencillas de asimilar o aplicar que las pruebas paramétricas. Además la interpretación suele ser más directa que la interpretación de las pruebas paramétricas.

LIMITANTES DE LAS PRUEBAS ESTADÍSTICAS NO PARAMÉTRICAS.

- Si todos los supuestos de un modelo estadístico paramétrico son conocidos por los datos y las hipótesis pueden ser probadas con una prueba paramétricas, entonces se

desecha la idea de utilizar pruebas estadísticas no paramétricas. Esto es expresado por el poder de eficiencia de la prueba no paramétrica.

- Las pruebas estadísticas no paramétricas no son sistemáticas, mientras que las pruebas estadísticas paramétricas han sido sistematizadas, y diferentes pruebas son simples variaciones en un tema central.
- Las pruebas no paramétricas tiene que ver con la conveniencia. Las tablas que se necesitan para las pruebas no paramétricas están muy esparcidas y aparecen en diferentes formatos.

Hasta el momento se ha visto la manera de dividir o agrupar los datos y de qué tipo pueden ser estas divisiones, estos datos a veces se muestran en tablas las cuales pueden ser en un sólo sentido o de dos o más sentidos, a continuación se verá de manera detallada las tablas de dos sentidos.

2.4.1 TÉCNICAS DE AGRUPACIÓN DE OBSERVACIONES

Los datos se agrupan en tablas de distribución de frecuencia las cuales se construyen al subdividir los datos de las mediciones u observaciones en intervalos de igual longitud. Los puntos que dividen a los datos deben elegirse de tal manera que resulte imposible que una medición sea ubicada en un punto de división, es decir, de tal manera que no exista ninguna ambigüedad con respecto a la clase a la que pertenece una observación en particular. El número de observaciones en una clase o intervalo recibe el nombre de frecuencia de clase. Al graficarse las frecuencias relativas de las clases contra sus respectivos intervalos en forma de rectángulos, se produce lo que comúnmente se conoce como histograma de frecuencia. Otra forma de agrupar los datos es en tablas de contingencia o tablas de clasificación cruzada, que se explicaran con más detalle por ser las que se utilizaran para la creación de poblaciones sintéticas.

2.4.2 TABLAS DE CONTINGENCIA

Las tablas de clasificación cruzada son conocidas como tablas de contingencia. Estas son utilizadas en problemas donde la información se obtiene contando no midiendo. Por lo tanto lo que se mide en estas tablas es la frecuencia observada de dos o más eventos quedando

una tabla de dimensiones rXc cuando la tabla es de dos dimensiones. Las frecuencias observadas se llaman frecuencias de celdas observadas. Suelen ser usadas para examinar la respuesta categórica de dos o mas variables cualitativas simultáneamente, es decir una respuesta conjunta. Por ejemplo, de una muestra de 500 vehículos si se desea saber el número de vehículos que cuentan con calcomanía cero, con calcomanía uno y con calcomanía dos, de acuerdo a la marca y modelo 1993 se generaría la tabla 2.1:

Tabla 2.1. Número de vehículos de acuerdo a su calcomanía

Calcomanía / Marca	0	1	2	Total
V.W.	20	30	50	100
Nissan	50	30	20	100
G.M.	30	40	30	100
Ford	40	40	20	100
Chrysler	35	35	30	100
Total	175	175	150	**500**

Nota: Estos datos fueron tomados únicamente como ejemplo.

Los datos de esta tabla pueden leerse de la siguiente manera: de la marca V.W., 20 cuentan con calcomanía cero, 30 con calcomanía uno y 50 con calcomanía dos; de los 500 vehículos que se tienen, 150 tienen calcomanía dos. Ya se vió como se pueden agrupar los datos, nos falta ver cuales son las diferentes técnicas para recolectar la información.

Para obtener la probabilidad por celda es con la siguiente fórmula:

$$p_{ij} = \frac{n_{ij}}{n} \quad i= 1,\ldots,r \; ; \; y \; j= 1,\ldots,c.$$

donde:

p_{ij} = probabilidad de cada celda.

n_{ij} = la observación en cada celda.

n = el total de observaciones en la tabla.

r = número de renglones en la tabla.

c = número de columnas en la tabla.

Existen varias pruebas estadísticas que se pueden usar cuando se tienen tablas de contingencia. Ya que los tipos de datos son enumerativos (o de conteo) existe una prueba para probar hipótesis con respecto a la distribución de probabilidad en una tabla de contingencia que es la prueba de bondad de ajuste X^2 . Que nos ayuda a saber si los datos que se tienen el la tabla de contingencia son parecidos a los esperado para cada celda. Se habla de esta prueba porque es la que se utilizó en el ejemplo, y la que ayudará a determinar si los datos siguen la distribución de frecuencia propuesta.

2.4.3 PRUEBA DE BONDAD DE AJUSTE CHI-CUADRADA X^2

Una medida o prueba que nos ayuda a ver la discrepancia entre la frecuencia observada y la frecuencia estimada es la prueba estadística X^2 (chi cuadrada)[5] La prueba chi-cuadrada fue propuesta por Karl Person en 1903, y posteriormente Sir Ronald Fisher la desarrollo más ampliamente y publicó las tablas de los valores críticos que aún son usadas. Muchas veces surge la necesidad de determinar si existe alguna relación entre dos rasgos diferentes en los que una población ha sido clasificada y en donde cada rasgo se encuentra subdividido en cierto número de categorías.

La fórmula que se utiliza para esta prueba es la siguiente:

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{[o_{ij} - e_{ij})]^2}{e_{ij}}$$

o bien:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{o_{ij}^2}{e_{ij}} - N$$

donde:

o_{ij} = el valor observado en cada celda

e_{ij} = el valor esperado de cada celda

N = El número total de observaciones en la tabla.

r = renglón y c = columna

[5] Existen métodos para obtener la frecuencia esperada de la celda como el método IPF (Iterative Proportional Fitting), multiplicación de marginales que se verán más adelante

Además el valor esperado de la celda se puede obtener por multiplicación de marginales:

$$e_{ij} = \hat{E}(n_{ij}) = n(\hat{p}_i \cdot \hat{p}_A) = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right) = \frac{r_i \cdot c_j}{n}$$

La estimación del valor esperado de la frecuencia de la celda observada n_{ij} para una tabla de contingencia, es igual al producto de sus respectivos totales de renglón y de columna, dividido entre la frecuencia total. Es decir:

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$$

También se necesita determinar el número apropiado de grados de libertad asociados con el estadístico de la prueba. Los grados de libertad asociados con una tabla de contingencia que tiene *r* renglones y *c* columnas siempre son iguales a (r-1)(c-1).

Freund J. (1989, p. 430) recomienda que cuando haya frecuencias de celda menor que 5 se combinen o colapsen algunas celdas y Siegel (1988, p. 49) dice que es necesario que más del 20 por ciento de las celdas tengan una frecuencia mayor que 5. Si estos requisitos no son reunidos por los datos en la forma en la cual fueron colectados originalmente, la investigación puede combinar categorías adyacentes para incrementar las frecuencias esperadas en varias celdas, esto se conoce también como colapsar.

Esta prueba se rechazará es decir se dirá que la tabla de contingencia no sigue esa distribución de frecuencia si el valor que se obtiene de X^2 excede al valor crítico que se obtiene de tablas del estadístico X^2_α con (r-1)(c-1) grados de libertad.

Es conveniente tener precaución con respecto al uso del estadístico X^2 como un método para analizar datos de tipo enumerativo. La determinación correcta de grados de libertad asociados con el estadístico X^2 es muy importante para localizar la región de rechazo. Si se especifica incorrectamente el número, se podría obtener conclusiones erróneas. Nótese también que no rechazar la hipótesis nula no implica que habría que aceptarla. Para muchas

aplicaciones prácticas tendríamos dificultades en establecer una hipótesis alterna que tenga sentido, y por lo tanto se desconocería la probabilidad de cometer el error tipo II[6].

2.5 DISEÑO DE MUESTRAS

Se sabe que el principal objetivo de la estadística es el de hacer inferencias acerca de una población con base en la información contenida en una muestra. Existen diferentes procedimientos estadísticos que se usan para analizar el conjunto de datos de una muestra y hacer inferencias acerca de algunas características de la población. Los diferentes métodos para seleccionar la muestra se llaman diseños de muestreo, que se usan para generar el conjunto de datos muestrales. El objetivo principal de las técnicas de muestreo es proporcionar esquemas para la extracción de una muestra que sea representativa de la población bajo estudio, es decir, que la muestra contenga las características de población, proporcionando así una cantidad específica de información a un costo mínimo.

Para seleccionar una muestra representativa se tendría que seleccionar aleatoriamente, esto con el fin de que todos los elementos de la población tengan la misma probabilidad de ser seleccionados. La precisión al hacerse las estimaciones, básicamente depende de dos factores: a) tamaño de la muestra y b) la variabilidad o heterogeneidad de la población.

a) *Tamaño de la muestra*: Mientras más grande sea la muestra, representará más fielmente a la población, tal que se pueden mejorar las estimaciones aumentando el tamaño de la muestra. Un ejemplo clásico es el de una moneda si se lanza la moneda 10 veces y de ésas 8 son águila y 2 sol, se podría pensar que la moneda está cargada, pero si se incrementa el número de lanzamientos y caen 52 veces águila y 48 sol, se ve que se aproxima a la probabilidad estimada de un 50 por ciento de probabilidad que caiga águila.

[6] Otra prueba que se puede utilizar es la de Kolmogorov-Smirnov. Como la distribución chi-cuadrada la prueba de Kolmogorov-Smirnov puede ser usada para probar la discrepancia entre una distribución empírica acumulada y alguna distribución teórica acumulada. El proceso se basa en clases en las cuales la distribución teórica y distribución observada tienen la más grande desviación absoluta. Esta desviación se compara con los valores en tablas. En general; para muestras pequeñas es conveniente utilizar la prueba de Kolmogorov. La prueba chi-cuadrada se dice que es muy poderosa cuando el tamaño de la muestra es n≥100, aunque algunos autores manejan que se pueden obtener buenos resultados con n≥30 Y por último para pruebas de tamaño menor que 10 la prueba de Cramer-Von Mises se dice que es la más apropiada.

b) La variabilidad o heterogeneidad: Para aumentar la precisión puede dividirse el marco muestral (si es que se dispone de los medios necesarios) en clases homogéneas llamadas estratos y seleccionar separadamente en cada estrato una muestra, garantizando en esta forma cualquier representación deseada de todos los estratos de la población.

Existen diferentes tipos de muestreo dependiendo si éste es probabilístico o no, además de otras características dependiendo de la población que se este estudiando.

2.5.1 Muestreo probabilístico y no probabilístico

Existe el muestreo no probabilístico y el probabilístico[7]. La característica esencial del muestro de probabilidad es que puede especificarse para cada elemento de la población la probabilidad que irá incluida en la muestra. En el caso más sencillo, cada uno de los elementos tiene la misma probabilidad de ser incluido, pero ésta no es una condición necesaria. Lo que sí es necesario es que para cada elemento debe haber alguna probabilidad especifica de ser incluida. El muestreo no probabilístico suele usarse cuando se tiene un amplio conocimiento del fenómeno que se investiga y cuando existen estudios previos al respecto, tal que el estadístico tiene antecedentes y el costo para la investigación es reducido. Este tipo de muestreo se recomienda también cuando no se desea un análisis profundo y preciso sobre las características del universo que se estudia. Este método resulta en ocasiones bueno, ya que capta con relativa facilidad las características de la población en estudio. No permite, por si mismo, llegar a estimaciones precisas pues no todos los sujetos tienen la misma probabilidad de ser elegidos y no se tiene la certeza de que la muestra sea representativa, resultando difícil realizar inferencias e imposible cuantificar el error en la estimación.

Las principales ventajas del muestreo de no probabilidad son la comodidad y la economía, ventajas que pueden superar a los riesgos supuestos en la no utilización del muestro de probabilidad. Entre estos tipos de muestreo se pueden mencionar los siguientes: *Muestreo accidental,* en éste muestreo se toman los casos que vienen a la mano, continuando con el proceso hasta que la muestra adquiere el tamaño precisado. "Si se utiliza una muestra accidental, solamente puede desearse que la equivocación no sea demasiado grande"[8];

[7] Selltiz C., Jahoda y Cook, 1965, pags. 566-600.
[8] Ibidem. p 568.

muestreo opinático o intencional, se caracteriza por un esfuerzo deliberado de obtener muestras "representativas" mediante la inclusión en la muestra de grupos supuestamente típicos, se utiliza frecuentemente en sondeos preelectorales de zonas que en ocasiones anteriores han marcado tendencias de voto; *muestreo causal o incidental*, se trata de un proceso en el que el investigador selecciona directamente o intencionalmente los individuos de la población, por ejemplo el utilizar un muestra a la que se tiene fácil acceso, como el caso de los voluntarios; *el muestreo de bola de nieve* que se realiza con algunos individuos los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra, este tipo de muestreo se utiliza cuando se hacen estudios en poblaciones llamadas marginales, grupos religiosos, determinado tipo de enfermos, etc.; y por último *muestreo por cuotas*, se fijan unas cuotas que consisten en un número de individuos que reúnen unas determinadas condiciones. Una vez determinada la cuota se eligen los primeros que se encuentran que cumplan con las características, este método se usa mucho en estudios de opinión y de mercado, este tipo de muestreo se conoce como una variante del muestreo estratificado ya que se divide a la población y se intenta recabar información de la muestra con respecto a las proporción que existe de cada estrato en la población total, por ejemplo si en la flota vehicular existe una proporción del 50 % de vehículos con convertidor catalítico en la muestra se tomará también el 50 % de vehículos con convertidor.

Con el muestreo probabilístico, se puede especificar para cada elemento de la población la probabilidad de ser incluido en la muestra. En el caso más sencillo, cada uno de los elementos tiene la misma probabilidad de ser incluido, o al menos para cada uno de los elementos debe de tener una probabilidad específica de ser incluido. Y dependerá del tipo de muestreo que se utilice.

Al seleccionar una muestra de n mediciones de una población finita de N mediciones, si el muestro se lleva a cabo de forma que todas las muestras posibles de tamaño n tengan la misma probabilidad de ser seleccionadas. Sólo estos métodos de muestreo probabilístico nos aseguran la representatividad de la muestra extraída por lo que son los más recomendables.

Dentro de los métodos de muestreo probabilístico encontramos los siguientes: *Muestreo aleatorio simple* donde se asigna un número a cada individuo de la población y mediante un medio mecánico (bolas dentro de una bolsa, tablas de números aleatorios, números

aleatorios generados con una calculadora, computadora, etc.) se eligen tantos sujetos u objetos como sea necesario para completar el tamaño de la muestra requerida, aunque es muy difícil de utilizar cuando se tiene una muestra muy grande; *muestreo aleatorio sistemático*, este procedimiento también requiere que se numere toda la población, pero solo se necesita un número aleatorio, se parte de ese número aleatorio y se toman individuos de *k* en *k* , siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra, se le conoce también como coeficiente de elevación[9] *k=N/n*, con este procedimiento se simplifica considerablemente la selección, pero existe el riesgo de introducir sesgo en la muestra al elegir los elementos de forma periódica, esto ocurre cuando el universo está ordenado en función de determinados criterios; *muestreo aleatorio estratificado*, consiste en considerar categorías típicas diferentes entre sí (estratos), cada estrato funciona independientemente, de este tipo de muestreo se hablará con mayor detalle más adelante; y *muestreo aleatorio por conglomerados*, en este muestreo la unidad muestral es un grupo de elementos de la población que forman una unidad, al que se llama conglomerado, y consiste en seleccionar aleatoriamente un cierto número de conglomerados e investigar todos los elementos pertenecientes a los conglomerados elegidos.

Hay que notar, que la aplicación de un determinado tipo de muestreo no es indiferente, depende de la información sobre el marco muestral, pero además condiciona el proceso y repercute en los errores muestrales, otro factor importante a considerar en el diseño muestral es el tamaño de la muestra que depende de diferentes factores ya sea los objetivos de la investigación o la estructura del marco muestral, además de las limitaciones económicas.
Finalmente hay que pasar a la estimación, donde se estudia que variables se van a estimar, con qué estimadores y con qué parámetros. En la práctica hay veces que se tiene que utilizar no los métodos más precisos sino los más viables, por ejemplo, cuando existe carencia de información o costos elevados de algunas aplicaciones.

2.5.2 Marco muestral

Para que se pueda realizar el diseño de la muestra y su posterior desarrollo, no basta con conocer las técnicas de muestreo sino que se necesita, además, acotar el universo y conocer las unidades que lo componen. Acotar el universo significa concretar perfectamente la población que va a ser objeto de estudio. Las unidades del universo acotado constituyen el

[9] Rodriguez Osuna J. Cuadernos metodológicos, 1991, pág. 26

marco del que se va a sacar la muestra. En la medida en que se conozca mejor el marco, se reducirá en primer lugar, los sesgos que se podrían introducir por su desconocimiento. El desconocimiento del universo no sólo afecta a la cobertura de la muestra, también es necesario conocer su distribución sobre el espacio - dónde se sitúa - y cuáles son sus características. Esto nos sirve para poder determinar el tamaño de la muestra total y sus dominios[10], para realizar la asignación, para calcular los coeficientes de ponderación y/o elevadores y para hacer la estratificación y el proceso de la selección.

En algunos casos los estudios van referidos a la población general o a un segmento de la misma, en estos casos la información sobre el universo se deduce de los censo y padrones generales de la población. Un problema de los censos es que esta información va envejeciendo con el tiempo y no se actualiza hasta que se realiza un nuevo censo o padrón.

El uso de muestras en lugar del censo o del total de la población se debe a en la mayor parte de los casos, a los alcances de las muestras, por ejemplo, reduce los costos puesto que no es lo mismo tomar solo una parte de la población que hacer un censo de esta, nos permite medir el error de muestro, si la muestra se elige de manera aleatoria nos permitirá conocer la probabilidad de ciertas observaciones.

Una limitación es que si no se eligió correctamente el marco muestral y no se analizó bien la distribución de la población se podría incrementar el error de la muestra o dar resultados que no son los que corresponden con el comportamiento de la población. Por lo que una muestra deberá estar diseñada de tal manera que los estados sean lo bastante parecidos a los estados que se tienen en el total de la población.

2.5.3 Tamaño de la muestra

Una vez que se ha acotado el universo y se tiene el marco muestral es necesario obtener el tamaño de la muestra, que este hace referencia al número de elementos del universo que se seleccionan para extraer de ellos la información que posteriormente se va a generalizar. Tamaño de muestra y precisión en las estimaciones son conceptos inseparables, ya que si aumenta el tamaño de la muestra lo hace el nivel de precisión de las estimaciones y si se quiere conseguir una mayor precisión, es necesario modificar el tamaño de la muestra.

[10] Dominios equivale a subpoblaciones, parte o fracción de la población original.

A medida que aumenta la precisión del estimador el intervalo de confianza se hace menor y, en ausencia de los sesgos, las posibles diferencias entre los valores de los parámetros poblacionales y los estimadores en el muestreo se hacen menores. Este se mide por el error de muestreo, que es la desviación típica del estimador, que refleja la dispersión de su distribución. En consecuencia, los errores de muestreo son concreción numérica del nivel de precisión de la estimación. Por eso hay que incluirlos en el estudio del tamaño de la muestra, y estos variaran dependiendo del tema o del tipo de estudio que se trata, ya que no es de la misma trascendencia estimar la precisión en la fabricación de vehículos que la precisión para estimar el consumo de una bebida refrescante. El nivel de precisión o error de muestro es el elemento más importante en la elección del tamaño de la muestra, puesto que condiciona, y en gran parte determina la dimensión de la muestra.[11]

Otro aspecto importante para determinar el tamaño de la muestra es el de la homogeneidad o heterogeneidad del universo, cuando la homogeneidad aumenta, es decir la varianza es menor, aumenta también el grado de precisión de la estimación para una determinada muestra.

El tipo de muestreo utilizado tiene también su incidencia en el tamaño de la muestra. En general, tomando como referencia el muestreo aleatorio simple, suele suceder que el muestreo por conglomerados es menos preciso y el estratificado más. En este último debido a que con la estratificación se divide a la población en subpoblaciones más homogéneas en su composición y heterogéneas entre si, con lo que se consigue disminuir la varianza total.

Se necesita establecer un nivel de confianza que se refiere a la probabilidad de acertar, es decir, a la probabilidad de que una estimación, en ausencia del sesgo, se ajuste a la realidad. Se debe fijar el nivel de confianza, de acuerdo con los objetivos de la investigación.

La distribución de la muestra entre los diferentes subconjuntos en que se puede dividir el universo bajo estudio se llama, en la terminología estadística, afijación[12] de la muestra o asignación de la muestra, dicha distribución puede realizarse de distintas maneras:

[11] Rodríguez Osuna J. Métodos de muestreo. 1991, págs 47-48.
[12] El término de afijación es usado por autores como Cochran, Abad entre otros en este trabajo se utilizará asignación.

1) *Simple* en donde se le asigna a cada estrato o subconjunto el mismo número de elementos.
2) *Proporcional*, donde la asignación a cada subconjunto depende de la proporción que haya en el universo de la población por ejemplo si en el universo en estudio existe un 50 % de vehículos con convertidor catalítico en la muestra se seleccionará un 50% de vehículos con convertidor.

3) *Óptima* donde se necesita sacar la desviación típica da cada estrato y se multiplica por el porcentaje que éste representa sobre el universo.

El cálculo del tamaño de la muestra se estudia primero, desde la formulación habitual en que los errores de muestro se expresan como errores absolutos, posteriormente se analizarán a partir del coeficiente de variación, es decir, de los errores relativos. Primero se verá el tamaño de muestra para universos pequeños.

UNIVERSOS PEQUEÑOS

a) *Tamaño de muestra para estimar la media*

Fórmula

$$n = \frac{NK^2\sigma^2}{Ne^2 + K^2\sigma^2}$$ [13]

donde:

n = Tamaño de la muestra.

N = Tamaño del universo.

K = Nivel de confianza.

σ^2 = Varianza de la población

e = Error de muestreo.

Como la varianza es desconocida se estima la varianza muestral con la fórmula:

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

[13] Rodríguez Osuna J. Ibid. Pág. 64.

b) *Tamaño de la muestra para estimar el total*

Fórmula

$$n = \frac{N^2 K^2 \sigma^2}{e^2 + NK^2\sigma^2}$$

c) *Tamaño de la muestra para estimar proporciones*

Fórmula

$$n = \frac{NK^2 P(1-P)}{(N-1)e^2 + K^2 P(1-P)}$$

donde:

P = Proporción de una categoría de la variable

$P(1-P)$ = Varianza. (es la varianza de una variable que se distribuye binomialmente)

UNIVERSOS GRANDES

Cuando se trata de universos grandes, la introducción del tamaño del universo en la muestra, lo que suele llamar <<corrección por poblaciones finitas>>, no modifica los cálculos y, sin embargo los complica. De ahí que, en estos casos si la fracción de muestro no supera el 5%[14], se puede prescindir de esta corrección con lo que la fórmula para calcular el tamaño de la muestra se simplifica considerablemente.

a) Tamaño de la muestra para estimar proporciones

Fórmula

$$n = \frac{K^2 P(1-P)}{e^2}$$

donde

n = Tamaño de la muestra.

P = Proporción de una categoría de la variable.

$P(1-P)$ = Varianza

[14] Cochran W. Técnicas de muestro, 1984, p. 49.

K = Nivel de confianza.

e = Error de muestreo.

El coeficiente de variación en términos de estadística inferencial presenta los errores de muestreo en porcentajes de la estimación, lo que facilita la interpretación de los resultados en comparación de las distintas estimaciones. La expresión matemática del coeficiente de variación en $cv = \frac{\sigma}{x}$, y referida en el supuesto de estimación de proporciones, es $cv = \frac{e_p}{P}$ donde e_p es el error de la estimación de la proporción y P la proporción. En las fórmulas de tamaño de la muestra si se quiere calcular el error relativo y no absoluto se trabaja con el coeficiente de variación, y la fórmula es la siguiente:

$$n = \frac{K^2(1-P)}{cv^2 P}$$

que permite el cálculo directo del tamaño de la muestra.

Hasta este momento se han visto las diferentes técnicas de muestreo, además de cómo definir el marco muestral y como poder determinar el tamaño de la muestra, ahora se verá de manera más detallada el muestreo estratificado puesto que es el que se utilizará para este trabajo por la manera en que se agruparon los datos.

2.6 MUESTREO ESTRATIFICADO

Una muestra aleatoria estratificada es una muestra aleatoria que se obtiene separando los elementos de la población en grupos disjuntos, llamados estratos, y seleccionando una muestra aleatoria simple dentro de cada estrato. La distribución de la muestra en función de los diferentes estratos se denomina asignación como se vio anteriormente, y puede ser de diferentes tipos; *simple*, donde a cada estrato le corresponde igual número de elementos muestrales; *proporcional* en el que la distribución se hace de acuerdo al peso (tamaño) de la población en cada estrato; y *optima*, se tiene en cuenta lo previsible de los resultados, de modo que se considera la proporción y la desviación típica, no es muy usado ya que no siempre se conoce la desviación.

➢ *Asignación proporcional*

Para calcular el tamaño de la muestra en este tipo de muestro se pueden utilizar las fórmulas que se vieron en el apartado anterior y una vez definido este, uno de los procedimientos más usados para definir el tamaño de cada estrato es el de asignación proporcional, que particiona el tamaño de la muestra en forma proporcional al tamaño de los estratos. Con este procedimiento se obtiene una muestra autoponderada, dado que la fracción de muestreo es la misma en cada estrato. A los estratos más grandes se les asigna mayor tamaño de muestra, y a los más chicos menor tamaño. La fórmula que se utiliza para este criterio es la siguiente:

$$n_i = \left(\frac{N_i}{N}\right)n$$

donde:

N_i =Tamaño de la población en el i-ésimo estrato

N = Tamaño de la población total

n = Tamaño total de la muestra.

Se verá esto más claro en el siguiente ejemplo. Se tiene que del parque vehicular son 15,000 vehículos de los cuales 4,550 tienen calcomanía dos, 8,125 tienen calcomanía uno y 2,325 vehículos tienen calcomanía cero, si se quiere tomar una muestra de 300 vehículos y se divide la muestra en 3 estratos (para cada uno de los diferentes tipos de calcomanía) se tiene lo siguiente:

$$n_1 = \left(\frac{4{,}550}{15{,}000}\right)300 = 91; \quad n_2 = \left(\frac{8{,}125}{15{,}000}\right)300 = 162.5 \approx 162; \quad n_3 = \left(\frac{2{,}325}{15{,}000}\right)300 = 46.5 \approx 47$$

Para vehículos con calcomanía dos se necesitaría una muestra de 91 vehículos, con calcomanía uno de 162 y para vehículos con calcomanía cero 47 vehículos lo que da un total de 300 vehículos.

El primer paso para la selección de una muestra aleatoria estratificada consiste en la especificación clara y detallada de cada estrato, asociando a cada elemento de la población con uno y sólo uno de los estratos.

2.6.1 ESTIMACIÓN DE MEDIA Y VARIANZA ESTRATIFICADA

De la información obtenida de los elementos muestrales, se puede calcular la media estimada $\bar{y}_i$ y la varianza s_i^2 para las observaciones de cada estrato, usando las siguientes fórmulas:

Media:

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

Varianza:

$$s_i^2 = \frac{\sum (y_{ij} - \bar{y}_i)^2}{n_i - 1} \qquad i=1,2,\ldots, L$$

donde y_{ij} es la j-ésima observación del estrato i.

2.6.1.1 Estimación de la media poblacional para una muestra aleatoria estratificada

La varianza s_i^2 es un estimador de la correspondiente varianza del estrato σ_i^2. El estimador $\bar{y}_{est}$ de la media poblacional μ, basado en un muestreo aleatorio estratificado, se da a continuación:

$$\bar{y}_{est} = \frac{\sum_{i=1}^{L} N_i \bar{y}_i}{N}$$

que es el estimado de la media poblaciones en el muestreo estratificado y para la varianza estimada del estimador tenemos:

$$\hat{\sigma}^2_{\bar{y}_{est}} = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

Si se quisiera usar el muestreo estratificado para estimar el total de la población τ, entonces el estimador es:

$$\hat{\tau} = N\bar{y}_{est}$$

Es decir, para obtener el estimador estratificado del total poblacional multiplicamos a la media estratificada por el total N de las unidades de la población, y la varianza estimada del estimador es:

$$\hat{\sigma}_{\hat{t}}^2 = N^2 \hat{\sigma}_{\bar{y}_{est}}^2$$

El muestreo estratificado se usa principalmente para estimar la media y el total de la población. Pero también se puede utilizar para encontrar la proporción $\hat{p}_i$ en el estrato *i*. Las proporciones muestrales de los estratos pueden combinarse para producir un estimador de la proporción poblacional.

2.6.1.2 Estimación de proporción poblacional en una muestra aleatoria estratificada

Estimador

$$\hat{p}_{est} = \frac{1}{N} \sum_{i=1}^{L} N_i \hat{p}_i$$

Con una varianza estimada del estimador:

$$\hat{\sigma}_{\hat{p}_{est}}^2 = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i} \right) \left(\frac{\hat{p}_i \hat{q}_i}{n_i - 1} \right) \qquad \hat{q}_i = 1 - \hat{p}_i$$

Del mismo ejemplo arriba visto se puede estimar la proporción de vehículos que pasaron la verificación en el primer intento. Los datos se muestran en la siguiente tabla:

Tabla 2.2. Vehículos que pasaron la verificación en el primer intento

Estrato	Tamaño de la muestra	# vehículos que pasaron verificación	$\hat{p}$
1	n_1=91	70	.77
2	N_2= 162	100	.62
3	n_3= 47	35	.74

La estimación de la población total que paso la verificación en el primer intento es:

$$\hat{P}_{total} = \frac{1}{15{,}000}\left[(4550)(.77)+(8125)(.62)+(2325)(.74)\right] = .69$$

Lo que quiere decir que de el total de la población que pasó la verificación al primer intento tiene una proporción del 69 %.

Una vez analizado la forma de seleccionar la muestra, el tamaño de la muestra, los tipos de escala y la forma de agrupar los datos se analizará la metodología para la creación de las poblaciones sintéticas.

2.7 PRINCIPIOS DE SIMULACIÓN

En esta sección se verá como se puede generar la población sintética con ayuda de la simulación de una manera muy rápida y sencilla, si se desea mayor información acerca de técnicas de simulación se recomienda revisar la bibliografía referente a este tema. Lo que se propone para la creación de poblaciones sintéticas es que una vez que se tiene la distribución de probabilidad, mediante técnicas de simulación como la de Monte Carlo crear la población sintética.

Se puede decir que la simulación es una técnica por medio de la cual se imita la operación o desempeño de un sistema productivo real a lo largo del tiempo. Esta simulación genera una historia artificial del sistema modelado, y con base en dicha historia ficticia se intenta extraer conclusiones válidas sobre el comportamiento o las características del sistema real.

Se verá brevemente como se clasifican los modelos de simulación. Una primera clasificación de los modelos usados en simulación distingue a los mismos en dos grandes tipos: determinísticos y estocásticos. En los primeros se manejan variables e interrelaciones que no incluyen ninguna medida de probabilidad, mientras que en los segundos se incluye al menos alguna de las variables o relaciones como aleatoria. El uso de representaciones algebraicas, matriciales o de ecuaciones diferenciales es común en los sistemas determinísticos, mientras que las representaciones como procesos estocásticos, cadenas de Markov, o sistemas de colas es común en los sistemas estocásticos.

Ahora se analizará una de las metodología de la simulación que nos servirá para la creación de poblaciones sintéticas que es la técnica de Monte Carlo[15]. Este término fue utilizado por Von Neumann y Ulan en los años cuarenta para resolver problemas nucleares durante la segunda Guerra Mundial. En la técnica de Monte Carlo, la experimentación artificial o datos son generados al usar algún generador de números aleatorios[16] y la distribución de probabilidad acumulada de interés. El proceso o técnica es relativamente fácil, para obtener la muestra aleatoria artificial de alguna población descrita por una función de probabilidad se necesita:

1. Agrupar o tabular los datos de interés, por ejemplo, una función de distribución acumulada con los valores de la variable en el eje de las *x* o abscisas y las probabilidades de 0 a 1 en el eje de las *y* u ordenadas.
2. Elegir un número aleatorio decimal entre 0 y 1 por algún generador de números aleatorios.
3. Proyectar horizontalmente el punto de las *y* correspondiente con este número decimal aleatorio hasta la intersección con la línea de la curva de la acumulada.
4. Proyectar hacia abajo este punto de intersección en la curva con el eje de las *x*.
5. Escribir el valor de la x correspondiente con este punto. Éste valor de x es tomado como el valor de muestreo.
6. Repetir los pasos del 2 al 6 hasta tener el número de muestras elegido.

Como puede verse esta técnica es muy sencilla y se puede utilizar para generar casi cualquier tipo de muestra siempre y cuando se conozca la distribución de probabilidad. Uno de los problemas más difíciles que enfrenta un analista de simulación es obtener un modelo que sea una representación apropiada del sistema real estudiado; dicho de otro modo, su principal preocupación es la de obtener un modelo válido.

La cuestión de la calidad que tenga el modelo como representación del sistema real tiene dos aspectos claramente separados. Por una parte está la validación del mismo. Verificar es construir correctamente el modelo: validar es construir el modelo correcto.

---

[15] Con esta técnica se pueden resolver problemas estocásticos, problemas determinísticos que no se pueden resolver analíticamente, como es el caso de algunas integrales .

[16] El generador de números aleatorios puede ser una tabla de números aleatorios, una ruleta, una subrutina en una computadora, o cualquier otra fuente de números aleatorios distribuidos uniformemente.

La verificación del modelo tiene que ver con la construcción de un modelo computacional correcto. Esto es, verificar el modelo significa compara el modelo conceptual real, que incluye supuestos sobre sus variables, los rangos de valores que toman, las interrelaciones lógicas y matemáticas, etc., con la representación en computadora que implementa dicho modelo.

Aunque conceptualmente la verificación no tiene gran complicación, el esfuerzo real que se debe efectuar para depurar y dejar listo un programa de cómputo que funcione correctamente, es una labor larga y difícil.

La validación del modelo se ocupa de determinar si el modelo conceptual en que se basa la simulación es una representación precisa del sistema estudiado. Cuando se tiene un modelo conceptual válido, los resultados obtenidos de la simulación con el mismo deben ser similares a los que se obtendrían experimentando directamente con el sistema que se estudia.

La aceptación de un modelo de simulación y sus resultados como válidos por parte de los usuarios y decisores que habitualmente trabajan con el sistema real, corresponde a lo que en la práctica de simulación se conoce como tener un modelo creíble. Esta cualidad, deseable en un modelo de simulación, es el principal motivo por el cual se han difundido ampliamente y con gran popularidad las técnicas de animación para ilustrar los resultados del modelo; ya que la animación resulta una forma muy eficaz de comunicar los detalles operativos y el comportamiento del modelo, a los interesados en él.

En sistemas estocásticos, la validación del modelo no debe confundirse con el análisis de resultados. Esta última tarea, más bien tiene que ver con la capacidad del modelo de estimar eficientemente las medidas de interés en el sistema; es una problema estadístico relacionado con determinar el tamaño de las corridas y el número de réplicas que sustenten un nivel de confianza determinado previamente.

La calibración de un modelo es la técnica iterativa que consiste en comparar el modelo con el sistema real, efectuando ajustes o aún cambios mayores al modelo, comparando el

modelo revisado con la realidad, haciendo ajustes adicionales y comparando de nuevo, y así sucesivamente.

Una posible crítica a la calibración es que el modelo se ha validado solamente para un conjunto de datos determinado; es decir, que el modelo se ajusta a ese conjunto particular de datos. Una forma de mejorar el procedimiento es recopilar un nuevo conjunto de datos del sistema, o en todo caso, reservar una parte de datos del conjunto original para usarse al final de la etapa de validación. De este modo, luego que el modelo se ha calibrado con los datos originales, una validación final, es efectuada con un segundo conjunto de datos del sistema. Si persisten discrepancias importantes, se debe reiniciar el proceso de calibración hasta lograr resultados aceptables. Los datos históricos usados en el sistema real, son muy adecuados para este propósito. Esta idea de usar un conjunto de datos para calibrar y otro conjunto independiente (pero homogéneo como el primero) de datos para la validación final es una forma común de operar en estudios econométricos o de bioestadística. Una recomendación adicional sobre la validación, es evaluar el posible aumento en la precisión del modelo que logrará contra el costo del esfuerzo requerido para mejorar la precisión del ajuste. Se tiene que establecer un nivel tolerable de error en los resultados del modelo.

Una guía para el proceso de validación fue introducida por Thomas H. Naylor y J.M. Finger en su articulo "Verification of Computer Simulation Models", (revista Management Science, octubre de 1967). Las ideas son las siguientes:

- Construir un modelo con un aspecto sumamente razonable: Esto significa que el modelo que se construya deberá lucir a los ojos de los usuarios habituales y los expertos conocedores del sistema real como muy razonable. Esto refleja el hecho de que un buen modelo no es una mera ocurrencia o abstracción de un modelador, sino la síntesis de la experiencia y la intuición sobre la operación del sistema bajo estudio que han proporcionado los expertos que lo conocen. Es por tal razón que el desarrollo de un buen modelo implica que el modelador establezca un intercambio continuo y regular de ideas con los expertos que conocen el sistema.

- Comprobar empíricamente las suposiciones del modelo: Esto significa evaluar cuantitativamente los supuestos del modelo en la fase inicial. Así por ejemplo, si se han supuesto ciertas distribuciones de probabilidad para describir algunas variables de

entrada al modelo, se recomienda usar técnicas de bondad de ajuste, como la chi-cuadrada o la de Kolmogorov-Smirnov. Es este caso se utilizará la prueba chi-cuadrada por el tamaño de la muestra

- Determinar si los resultados del modelo son representativos: la gran prueba que un modelo se simulación debe pasar es que lo resultados que genere sean similares a los que se producen o cabría esperar en el sistema real. Un procedimiento que se puede utilizar es la prueba de Turing, donde se comparan los resultados del modelo con el sistema real, invitando a un grupo de expertos conocedores del sistema. Los expertos reciben varios conjuntos de datos para examinar, unos generados por el modelo y otros tomados del sistema real. Si los expertos son capaces de distinguir entre los datos reales y los otros entonces se deberán hacer modificaciones al modelo simulado. Esta parte faltaría por realizarse.

2.8 MÉTODO PARA LA CREACIÓN DE POBLACIONES SINTÉTICAS

Antes de empezar a describir detalladamente el método para la creación de poblaciones se verá que se entiende por una población sintética y para que se va a utilizar. Una población sintética también se le puede llamar artificial o virtual, intenta ser una representación lo más parecido posible a la población real, ya que se analiza el comportamiento de la población real, este tipo de poblaciones se utilizan cuando se necesita tener una muestra o población, pero es imposible o muy costoso contar con la población real, por ejemplo en el uso de gasolinas de los vehículos si se desea saber que porcentaje de vehículos utilizan gasolina Magna y que porcentaje Premium del total de la flota vehicular, el tomar el total de la población podría resultar muy caro pero si se cuenta con una muestra representativa de la población y si se utiliza el concepto de población sintética podrían reducirse los costos de la investigación.

Se describirá brevemente como surgió esta método y para que fue utilizada. El método para la creación de poblaciones sintéticas fue utilizado por Beckman y Baggerly (1996) para crear poblaciones sintéticas de individuos y viviendas usando datos de censos de 1990 dados en el "Public Use Microdata Sample" (PUMS) que es una colección de tablas demográficas, donde la información que se tenía era de tres tipos: por familias; no familias que son personas que viven solas o que viven en la misma casa pero no tienen ningún parentesco y;

personas que viven en viviendas o habitaciones como prisiones o internados. Lo que se buscaba era que a partir de los datos del censo saber cuantas viviendas había con determinado número de personas. El algoritmo básico para la construcción de una población sintética se basó en datos de censos. La muestra dada en el PUMS fue representativamente (no necesariamente aleatoria) del 5% de la muestra de viviendas

Se construyeron tablas de entrada múltiple o de contingencia. Para propósitos matemáticos se asumió que todos los bloques de censos tienen la misma estructura de correlación. Una vez que se tenía las tablas de entrada múltiple o marginales se les aplicó la metodología de Iterative Proportional Fitting (IPF) para estimar la proporción en las tablas, aunque no es indispensable utilizar esta metodología, también se puede utilizar el método de máxima verosimilitud o el de mínima chi-cuadrada los cuales no se describirán en este trabajo. Con estas proporciones se aplicó una fórmula y se obtuvo la distribución de probabilidad, para posteriormente mediante alguna técnica de simulación obtener la población sintética.

El número de viviendas es generado por cada tipo de demografía (teniendo un grupo específico de demografía) y se determinó para cada región del censo. Este número o población sintética puede ser obtenido de dos maneras, por multiplicar el número de total de objetos por las probabilidades en las tablas de entradas múltiples estimadas si es que la estimación se hizo con base a proporciones, o de sacar un número aleatorio según estas probabilidades.

A continuación se verá de manera detallada los principales elementos para la creación de las poblaciones sintéticas como es el método de ajuste.

2.8.1 Método de ajuste IPF (Iterative Proportional Fitting)

La herramienta primaria usada para completar la tabla de entrada múltiple fue el método del IPF (Iterative Proportional Fitting). IPF ha sido usado en modelos de transporte. En algunos círculos esto es conocido como el modelo "Fratar". Este fue usado por Dugway (1993) para sintetizar familias empleando una técnica similar a la presentada en el artículo de Beckman y Baggerly.

Cuando se tiene una tabla de contingencia y se necesita hacer un ajuste de la proporción de las celdas donde π_{ij} denota la probabilidad, el problema es encontrar el estimador $\hat{\pi}_{ij}$ para

ajustar la proporción de la celda a la probabilidad de la celda. Deming-Stephan (1940) fueron los primeros en utilizar esta metodología en el que encontramos el estimador de máxima verosimilitud que tiene las siguientes ecuaciones:

$$\hat{m}_i^{(A)} = \sum_j \hat{m}_{ij} = n_i^{(A)}$$

$$\hat{m}_j^{(B)} = \sum_i \hat{m}_{ij} = n_j^{(B)}$$

$$\log\left(\hat{m}_{ij} / g_{ij}\right) = \hat{\lambda} + \hat{\lambda}_i^{(A)} + \hat{\lambda}_j^{(B)}$$

$$\sum \hat{\lambda}_i^{(A)} = \sum \hat{\lambda}_j^{(B)} = 0$$

La solución no la veremos aquí ya que esto no tiene solución directa, los autores utilizaron el método de aproximaciones sucesivas por multiplicadores de lagrange. La estimación de mínimos cuadrados fue la siguiente:

Los estimadores $\hat{\lambda}$, $\hat{\lambda}^{(A)}$ y $\hat{\lambda}^{(B)}$ no aparecen explícitamente en el algoritmo.

$$m_{ij}^0 = g_{ij}$$

$$m_{ij}^t = \begin{cases} m_{ij}^{(t-1)} n_i^A / m_{ij}^A & t \quad impar \\ m_{ij}^{(t-1)} n_j^B / m_{ij}^B & t \quad par... \end{cases}$$

donde g_{ij} son los valores iniciales

Everitt (1983) propone que estos valores iniciales sean 1, aunque si se tiene una matriz inicial diferente de 1 será recomendable, ya que el resultado depende de la matriz inicial.

El término de ajuste proporcional iterativo surge porque en cada iteración se hace un ajuste proporcional de renglón o columna. La iteración continúa hasta que los cambios relativos entre cada iteración en cada estimación m_{ij}^t es pequeña.

Bishop (1975) muestra que este algoritmo tiene las siguientes propiedades.

1. Siempre converge
2. La regla para detener el proceso depende del grado de precisión que se necesite de las estimaciones, es decir, de cómo se establezca el criterio de convergencia.
3. Las estimaciones solo dependen de los cálculos anteriores.
4. Cualquier grupo de valores iniciales puede ser elegidos para hacer la aproximación
5. Si existen estimaciones directas el procedimiento encuentra los valores en un ciclo o iteración.

El siguiente ejemplo fue tomado de Beckman y Baggerly (1996, p. 422), se tiene el total de las marginales de una tabla, y además se tiene otra tabla que es la matriz inicial y se verá como se van haciendo los ajustes.

Tabla inicial

	B_1	B_2	***Total***
A_1	*45*	*108*	***153***
A_2	*63*	*37*	***100***
Total	***108***	***145***	***253***

Total de marginales

	B_1	B_2	***Total***
A_1			***3105***
A_2			***1955***
Total	***2205***	***2855***	***5060***

Es importante notar que las dos tablas contienen el mismo tipo de datos, solo que en una no se conocen los valores de las celdas solo el total de las marginales, es decir, se tiene una pequeña muestra de la población donde se conoce el comportamiento de esta que es la tabla inicial pero además se conocen los totales del universo sin conocer el comportamiento de la población total, lo que se pretende es que a partir de la pequeña muestra de la población se pueda estimar el comportamiento de la población total. Al aplicar el método IPF que nos ayuda a obtener estimaciones por celda se obtiene lo siguiente:

Primer iteración

	B_1	B_2	***Total***
A_1	*913*	*2192*	***3105***
A_2	*1232*	*723*	***1955***
Total	***2205***	***2855***	***5060***

Resultado final

	B_1	B_2	***Total***
A_1	*949*	*2156*	***3105***
A_2	*1256*	*966*	***1955***
Total	***2205***	***2855***	***5060***

En la práctica los autores encontraron que este procedimiento converge entre 10-20 iteraciones.

Si se calcula la estimación para cada celda con el procedimiento de multiplicación de marginales resulta lo siguiente:

Tabla inicial

	B_1	B_2	*Total*
A_1			***3105***
A_2			***1955***
Total	***2205***	***2855***	***5060***

Resultado final

	B_1	B_2	*Total*
A_1	*1353*	*1752*	***3105***
A_2	*852*	*1103*	***1955***
Total	***2205***	***2855***	***5060***

Los resultados de aplicar los métodos de IPF y el de multiplicación de marginales, por ejemplo en la celda (A_1,B_1) con el método IPF dio 949 y con el de multiplicación de marginales fue 1353 una diferencia de 404, además que para el método de multiplicación de marginales no es importante la matriz inicial.

Ahora, se verá un ejemplo, donde se conoce el comportamiento de la población total para contrastar los dos método para el ajuste de tablas, es decir, IPF y multiplicación de marginales, se obtuvo de Deming y Stephah (1943, pp. 433-434) que es la tabla de una muestra artificial de frecuencias de individuos de raza blanca que asisten a la escuela, clasificadas por edad y por distrito, del estado de Nueva Inglaterra, 1930. Nos presentan también una pequeña muestra que es nuestra tabla inicial para poder hacer las estimaciones de frecuencia por celda de la población. Lo que a continuación se hará es comparar las dos metodología utilizando la prueba chi-cuadrada para ver si las estimaciones de la proporción de las tablas siguen la distribución de frecuencia propuesta.

Tabla 2.3 Personas de raza blanca que asisten a la escuela en Nueva Inglaterra, población total

	Edad				
Distrito	**7 a 13**	**14-15**	**16-17**	**18-20**	**Total**
Maine	3623	781	557	313	**5274**
New Hampshire	1570	395	251	155	**2371**
Vermont	1553	419	264	116	**2352**
Massachusetts	10538	2455	1706	1160	**15859**
Rhode Island	1681	353	171	154	**2359**
Connecticut	3882	857	544	339	**5622**
Total	**22847**	**5260**	**2493**	**2237**	**33837**

Tabla 2.4 Personas de raza blanca que asisten a la escuela en Nueva Inglaterra, población inicial

	Edad				
Distrito	**7 a 13**	**14-15**	**16-17**	**18-20**	**Total**
Maine	50	11	8	4	**73**
New Hampshire	32	8	5	3	**48**
Vermont	32	5	5	2	**47**
Massachusetts	84	19	14	9	**126**
Rhode Island	35	7	4	3	**49**
Connecticut	52	11	7	5	**75**
Total	**285**	**64**	**43**	**26**	**418**

Tabla 2.5 Estimación de la población total por multiplicación de marginales

	Edad				
Distrito	**7 a 13**	**14-15**	**16-17**	**18-20**	**Total**
Maine	3561.0	819.8	544.4	34837	**5274**
New Hampshire	1600.9	368.6	244.8	156.7	**2371**
Vermont	1588.1	365.6	242.8	155.5	**2352**
Massachusetts	10708.1	2465.3	1637.1	1048.5	**15859**
Rhode Island	1592.8	366.7	243.5	156.0	**2359**
Connecticut	3796.0	874.0	580.4	371.6	**5622**
Total	**22847**	**5260**	**2493**	**2237**	**33837**

Tabla 2.6 Estimación de la población total por IPF

	Edad				
Distrito	**7 a 13**	**14-15**	**16-17**	**18-20**	**Total**
Maine	3597.4	813	568.2	295.4	**5274**
New Hampshire	1573.0	404.0	242.6	151.4	**2371**
Vermont	1594.3	409.5	245.9	102.3	**2352**
Massachusetts	10524.5	2445.6	1731.5	1157.4	**15859**
Rhode Island	1677.5	344.7	189.2	147.6	**2359**
Connecticut	3880.2	843.2	515.6	383.0	**5622**
Total	**22847**	**5260**	**2493**	**2237**	**33837**

Al aplicar la prueba chi-cuadrada a los dos casos se obtiene lo siguiente:

Para el método IPF

$$X^2 = \left(\frac{3623^2}{3597.4} + \frac{781^2}{813} + \ldots + \frac{339^2}{374.8}\right) - 33{,}837 = 17.27$$

y para multiplicación de marginales

$$X^2 = \left(\frac{3623^2}{3561} + \frac{781^2}{819.8} + \ldots + \frac{339^2}{371.6}\right) - 33{,}837 = 82.02$$

Se tiene que X^2=22.31 con (r-1(c-1)=(6-1)(4-1)=15 grados de libertad y un nivel de confianza α=.10 , se rechaza la hipótesis de que la tabla de contingencia tiene esa distribución de probabilidad si $X^2 > 22.31$, al comparar con las dos metodologías se puede ver que con el método IPF no se rechaza la hipótesis, es decir, se acepta que las observaciones tienen la distribución planteada, pero con la metodología de multiplicación de marginales se concluiría que no se tiene evidencia suficiente para decir que la distribución de la tabla es la que se esperaba. Por esto es que en los ejemplos posteriores se utilizará la metodología de IPF.

Al encontrar las proporciones por celda se encontró la distribución de probabilidad de la población que se obtiene dividiendo cada celda entre el total de la población, lo que queda:

Tabla 2.7 Distribución de probabilidad por IPF

	Edad				
Distrito	**7 a 13**	**14-15**	**16-17**	**18-20**	**Total**
Maine	0.106	0.024	0.017	0.009	**0.156**
New Hampshire	0.046	0.012	0.007	0.004	**0.07**
Vermont	0.047	0.012	0.007	0.003	**0.07**
Massachusetts	0.311	0.072	0.051	0.034	**0.469**
Rhode Island	0.05	0.01	0.006	0.004	**0.069**
Connecticut	0.115	0.025	0.015	0.011	**0.166**
Total	**0.675**	**0.155**	**0.103**	**0.066**	1

Lo cual se puede leer como la probabilidad de que se seleccione a una persona de raza blanca que tenga entre 7 y 13 años y además viva en el distrito de Maine es de 10.6%.

Con la metodología de IPF se obtiene el estado estacionario de las tablas, es decir, que si se repitiera la selección de la muestra infinidad de veces, el promedio de datos por celda no cambiaría, tiende a comportarse de esta manera.

2.8.1.1 Criterio de Convergencia

Un criterio de convergencia nos ayuda a determinar en que momento debemos detener las iteraciones. Ratkowsky 1990 dice que " la convergencia depende de la parametrización del modelo y de los estimadores iniciales". Birkes 1993 establece un criterio conveniente para determinar cuándo detener una iteración que es el siguiente:
Detener cuando dos estimaciones sucesivas, es decir m_{ij}^t y m_{ij}^{t+1} satisfacen la condición de que las diferencias relativas $|m_{ij}^{t+1} - m_{ij}^t| / |m_{ij}^t|$ es menor que 10^{-4}. Esto garantiza que las dos

estimaciones sucesivas casi concuerdan en 4 dígitos significativos (o de *k* dígitos significativos si 10^{-4} es reemplazado por 10^{-k}).

Cabe mencionar que el criterio de convergencia depende en muchas ocasiones del grado de precisión que se desee en las estimaciones. Una vez que se tiene las estimaciones de la tabla de contingencia se necesita obtener las distribución de probabilidad, a continuación se verá como se puede obtener.

2.8.2 Obtener la función de distribución de probabilidad

Los autores reconocen que las probabilidades se pueden obtener de considerar la distancia entre la tabla original indicada por *p* , y de la tabla de entrada múltiple que se obtuvo de la metodología IPF que se indica con la letra c[17].

Sin embargo Beckman , Baggerly y McKay (1996) recomienda no usar la formula de D si alguno de los resultados de la tabla original es exactamente igual o muy cercano al de la tabla de entrada múltiple que se obtuvo de utilizar el método IPF. Llaman a esta función "*non 0-1 loss function*" que es α=0 y k=0. En caso de que esto no suceda, ellos sugieren que los valores de k y α deberán ser los más cercano posible a cero. Cuando no se utiliza la fórmula de distancia para obtener la distribución de probabilidad se toman los valores que

[17] La siguiente función es usada para calcular las probabilidades:

$$D(p,c) = w_p \prod_{i \in J} \left(1 - \left|(d_i^p - d_i^c)/r_i\right|^k\right) \cdot \prod_{i \in \bar{J}} \left(1 - \Delta\left(d_i^p, d_i^c\right)\right)$$

donde, se tiene que :

1. J es el grupo de variables ordinales tales como {ingreso, edad, trabajo} y ~J es el grupo de variables categóricas tales como {tipo de familia, raza}
2. d^p_i es el valor de la i-ésima demografia o característica p de la tabla original
3. d^c_i es el valor de la i-ésima demografia o característica de la tabla de entrada múltiple estimada.
4. r_i es el rango de demografia i en la tabla original.
5. w_p es el peso asociado de la caracteristica p para la tabla original.
6. $$\Delta\left(d_i^p, d_i^c\right) = \begin{cases} \alpha & d_i^p = d_i^c \\ 1-\alpha & d_i^p \neq d_i^c \end{cases}$$
7. k es una constante positiva arbitraria. Valores de k cerca de 0 darán más peso a estas demografias d^p_i , las cuales están para la construcción de tablas demográficas d^c_i.

Para obtener la población sintética, la fórmula está dada por: Pr{seleccionar viviendas p} = $D(p,c) / \sum_j D(j,c)$

anteriormente se encontraron con la metodología IPF, ya que si se recuerda se obtuvo la distribución de probabilidad.

Como ya se tiene la distribución de probabilidad ya se puede generar la población sintética ayudándose de alguna técnica de simulación.

2.8.3 Generación de la población Sintética

Beckman y Baggerly en su artículo proponen dos pasos para la creación de poblaciones sintéticas, que son; 1) obtener la estimación de las proporciones de las tablas de frecuencia, utilizando cualquier técnica de estimación y; 2) obtener la población sintética, ya sea multiplicando las proporciones por el total de la población o utilizando alguna técnica de simulación. Después de haber analizado el método propuesto por los autores se vio que no eran suficientes los dos pasos sino que para poder llegar a estos se necesitaba más cosas por lo que se proponen cinco pasos, que son los siguientes.

1. *Construcción de la muestra*, este paso los autores lo dan por entendido, pero para que a través de la muestra se pueda reproducir el universo con la precisión que se requiere en cada caso, es necesario el diseño muestral. Los métodos de muestreo están constituidos con la finalidad de construir modelos reducidos de la población con resultados extrapolables al universo del que se extraen. Cuando el muestreo viene avalado por la teoría y cada elemento tiene la misma probabilidad igual o independiente de figurar en la muestra, en este supuesto, las estimaciones son insesgadas y se puede calcular el error de muestreo que permite determinar la precisión de las estimaciones.

2. *Recoger la información y poner los datos en tablas de distribución de frecuencia*. Una vez determinadas las características de la población se pueden recoger los datos y agruparlos en tablas de frecuencia, solo se necesita determinar como se agruparan estos, dependiendo de las características que se quieran estudiar.

3. *Encontrar el estado estacionario de las tablas*, en este paso lo que se hace es obtener los estimadores, con los cuales se puede plantear el supuesto de que la población tiene determinado comportamiento, en este paso, que es el primero que proponen los autores, se pueden utilizar diferentes metodologías, ya sea la de IPF, multiplicación de

marginales, etc. Para este trabajo se utilizó la metodología IPF por lo demostrado en este capítulo.

4. *Obtener la distribución de probabilidad*, de las tablas de frecuencia. Este paso aunque no es mencionado por los autores, si analizamos los principios básicos de simulación se puede ver que para poder utilizar cualquier técnica de simulación es necesario contar con una distribución de probabilidad, sin ésta no se podría hacer la simulación.

5. *Generar la población sintética* aplicando alguna técnica de simulación, como las descritas anteriormente en este capítulo. Con la simulación se puede aplicar algún proceso de validación, evaluando los supuestos con alguna técnica de bondad de ajuste, y de ser posible utilizar la prueba de Turing.

En el siguiente capítulo se verá como se puede aplicar este método, en el caso particular de emisión de contaminantes de los automóviles particulares que circulan en la ZMCM.

III

CREACIÓN DE LA POBLACIÓN SINTÉTICA DE LA FLOTA VEHICULAR DE LA ZMCM.

Como se vio en el capítulo uno, una de las principales preocupaciones en la ZMCM es la contaminación atmosférica, siendo el ozono uno de los contaminantes más abundantes por lo que se ha intentado reducir la formación de éste, sus principales precursores son los NO_x, HC y CO que en su mayoría son generados por vehículos particulares. Una de las medidas que ha tomado el gobierno es el programa de verificación vehicular en el cual establece ciertas normas con el fin de otorgar calcomanías a los vehículos, dependiendo de sus emisiones y suspender su circulación dos, uno o ningún día. Por esto el propósito es determinar cuántos vehículos de los que circulan en la ZMCM modelos de 1991-1998 están dentro de la Norma Oficial para cada uno de los principales precursores del ozono y cuántos fuera de la norma, a partir de un método para la creación de poblaciones sintéticas, ya que sólo se contaba con una muestra de 120 vehículos.

La flota vehicular de la ZMCM está distribuida por modelos de la siguiente manera:

Tabla 3.1. Padrón vehicular 1991-1998

Modelo	No. de Vehículos
1991	188,369
1992	218,930
1993	202,286
1994	190,247
1995	114,634
1996	68,062
>1997	233,124
Total	**1,215,652**

Fuente: Secretaría de Transporte y Vialidad (1998) Gobierno del Distrito Federal

3.1 METODOLOGÍA DE LA INVESTIGACIÓN

En el capítulo anterior se vieron los pasos para poder generar la población sintética. El primer paso es la construcción de la muestra: Lo primero que se necesita es el marco muestral y conocer las características de la población. En este caso el marco muestral es el padrón vehicular y la manera en la que están divididos los datos es por modelo, además como se necesita saber la cantidad de vehículos que están dentro de la norma, entonces se propone dividir los datos por estratos dependiendo del modelo lo cual nos da la clasificación por renglones y por columnas se determinó por las emisiones de contaminantes de los vehículos, es decir, dentro de la norma y fuera de la norma.

Este estudio es retrospectivo, puesto que, los datos ya existían y se adecuaron a las necesidades de este caso. El estudio se hizo a partir de los registros de las emisiones de 120 vehículos - información proporcionada por el departamento de Motoquímia del Instituto Mexicano del Petróleo. Se conoce que la muestra fue tomada de manera aleatoria pero la muestra fue controlada por lo que se supone un muestreo por cuotas, además no se determinó el tamaño de muestra que se necesitaba, sino que se utilizó toda la información referente a los modelos de 1991 a 1998. Con estos datos se puede ir al segundo paso que es el poner los datos en tablas de distribución de frecuencia.

3.1.1 CONSTRUCCIÓN DE TABLAS DE CONTINGENCIA

Con la información de las mediciones, las cuales están tomadas en gr/km y no en ppm (partes por millón) como lo marca la norma o como se miden en los centros de verificación, se agruparon los datos primero por modelos resultando 7 renglones, por ejemplo, el primer renglón para los modelo 1991, etc., se dividió la información en dos intervalos o grupos para determinar las columnas, en aquellos que estaban dentro de la norma establecida para vehículos nuevos, NOM 042 (véase capítulo 1 tabla 7) y los vehículos que no estaban dentro de la norma, es decir aquellos que rebasaban los límites establecidos, resultando dos columnas, con lo que se formaron las tablas de contingencia necesarias para la creación de poblaciones sintéticas.

Para HC se obtuvo:

Tabla 3.2 Para emisiones de HC (gr/km)

	Dentro de la norma 0.00 - 0.25	*Fuera de la norma >0.25*	*Total*
1991	1	8	**9**
1992	2	10	**12**
1993	13	19	**32**
1994	8	4	**12**
1995	2	2	**4**
1996	7	3	**10**
>1997	39	2	**41**
Total	**72**	**48**	**120**

En ésta tabla, de los 120 vehículos, 72 están dentro de la norma, es decir, con emisiones menores a 0.25 gr/km y 48 vehículos están fuera de ésta. La proporción de vehículos que se encuentran dentro de la norma es $p = 72/120 = 0.6$.

Para CO se tiene:

Tabla 3.3 Emisiones de CO (gr/km)

	Dentro de la norma 0 - 2.11	*Fuera de la norma >2.11*	*Total*
1991	1	8	**9**
1992	1	11	**12**
1993	2	30	**32**
1994	1	11	**12**
1995	0	4	**4**
1996	7	3	**10**
>1997	31	10	**41**
Total	**43**	**77**	**120**

En este caso disminuyó el número de vehículos dentro de la norma con solo 43 y 77 fuera de la norma. Beckman y Baggerly (1996) sugieren que si en alguna de las celdas existe un valor cero se colapse la tabla (unir renglones o columnas según sea el caso), para evitar errores en el procedimiento IPF por lo que se colapsaron los años de 1994 y 1995 de la tabla 3.3, resultando:

Tabla 3.4 Tabla de CO colapsada

	Dentro de la norma 0 - 2.11	***Fuera de la norma >2.11***	***Total***
1991	1	8	**9**
1992	1	11	**12**
1993	2	30	**32**
1994-1995	1	15	**16**
1996	7	3	**10**
>1997	31	10	**41**
Total	**43**	**77**	**120**

Que nos da una proporción de vehículos dentro de la norma para emisiones de CO igual a $\hat{p} = {}^{43}/_{120} = 0.36$

Y por último para las emisiones de NO_x se tiene:

Tabla 3.5. Para emisiones de NO_x (gr/km)

	Dentro de la norma 0 - 0.62	***Fuera de la norma >0.62***	***Total***
1991	1	8	**9**
1992	2	10	**12**
1993	2	30	**32**
1994	6	6	**12**
1995	1	3	**4**
1996	9	1	**10**
>1997	39	2	**41**
Total	**60**	**60**	**120**

Para el caso de emisiones de NO_x se tiene que la mitad de los vehículos estuvieron dentro de la norma y la otra mitad fuera de ésta, por lo que la proporción de vehículos dentro de la norma es de $p = 0.5$.

Una vez que se tienen agrupados los datos en tablas de contingencia, sigue el paso 3 que es el encontrar el estado estacionario, en este caso se aplicó el método IPF para el ajuste de tablas, por lo explicado en el capítulo 2 de dar un mejor ajuste que el método de multiplicación de marginales.

3.1.2 ANÁLISIS DE DATOS Y OBTENCIÓN DE ESTIMADORES

Como se necesita conocer el total de las marginales tanto de renglón como de columna para poder aplicar el método IPF, ya se tiene el total por renglón, que es el que se obtiene del padrón vehicular (tabla 3.1), el de la columna se obtiene de la fórmula del capítulo dos de estimar las proporciones de la población en muestreo estratificado, como se agruparon los datos por estratos (modelo), se utilizaron los datos obtenidos en las tablas de emisiones (tablas 3.2 a 3.5) para poder calcular la proporción de la población total.

La primer tabla en la que se hizo el cálculo es la tabla 3.2 donde se obtuvo la proporción de vehículos dentro o fuera de la norma para cada uno de los modelo o renglón, por ejemplo, para el renglón 1 columna 1 en la tabla 3.2 se tiene un valor de 1 y si se necesita saber la proporción de vehículos modelo 1991 dentro de la norma se divide esta cantidad entre el total del renglón que es 9 resultando 0.11111, aplicando el mismo criterio en todos los casos, resulta la siguiente tabla:

Tabla 3.6 Para emisiones de HC (gr/km)

	Dentro de la norma 0.00-0.25	***Fuera de la norma >0.25***
1991	0.11111	0.88889
1992	0.16667	0.83333
1993	0.40625	0.59375
1994	0.66667	0.33333
1995	0.5	0.5
1996	0.7	0.3
>1996	0.95122	0.04878

Para estimar la proporción de vehículos que están dentro de la norma para emisiones de HC, puesto que la población está estratificada, se aplica la fórmula vista en el capítulo 2 página 49 para estimar la proporción y resulta:

$$\hat{p}_{est} = \frac{1}{1{,}215{,}652}\left[(188{,}369)(0.1111) + (218{,}930)(0.16667) + \ldots (68{,}062)(0.7) + (233{,}124)(0.95122)\right]$$
$$= .48792$$

Donde 188,369 es la suma del primer renglón del total de la flota vehicular (tabla 3.1) y $\hat{p}_i$ = 0.1111 es la proporción de vehículos dentro de la norma modelo 1991 (tabla 3.6). Por lo tanto, la suma estimada de la primer columna, es decir, vehículos dentro de la norma, es la multiplicación de la proporción estimada por el total de la flota vehicular, lo que nos da (0.48792) * 1,215,652 = 593,141 y el total de la segunda columna es la diferencia del total de la población menos el total de vehículos dentro de la norma 1,215,652 - 539,141 = 622,511, es decir, la estimación de vehículos fuera de la norma o con emisiones superiores a 0.25 gr/k de HC. Si se quiere calcular el error de muestreo de la estimación de la proporción de vehículos dentro de la norma para emisiones de HC es con la siguiente fórmula:

$$e_p = \sqrt{\left(\frac{N-n}{N-1}\right)\left(\frac{P(1-P)}{n}\right)}$$

donde:

N es la población total

n es la población de la muestra

P es la proporción

Al aplicar la fórmula se tiene que

$$e_p = \sqrt{\frac{(1{,}251{,}652-120)}{(1{,}215{,}652-1)} \bullet \frac{.48792(1-.48792)}{120}} = 0.00208$$

Un 0.2 % de error de muestreo.

En el segundo caso para estimar la proporción de vehículos que están dentro de la norma para emisiones de CO se tiene:

Tabla 3.7 Proporción para emisiones de CO (gr/km)

	Dentro de la norma 0-2.11	***Fuera de la norma >2.11***
1991	0.11111	0.88889
1992	0.08333	0.91667
1993	0.0625	0.9375
1994-1995	0.0625	0.9375
1996	0.7	0.3
>1996	0.75610	0.24390

$$\hat{p}_{est} = \frac{1}{1{,}215{,}652}\left[(188{,}369)(0.1111) + (218{,}930)(0.09091) + \ldots (68{,}062)(0.7) + (233{,}124)(0.75610)\right]$$
$$= .24249$$

De manera similar al cálculo anterior se obtuvieron estos datos. La suma estimada de la primer columna (dentro de la norma) de la tabla 3.7 es (0.24249) * (1,215,652) = 294,780 y el total de la segunda columna es 1,215,652 - 294,780 = 920,872.

El error de muestreo de la estimación de la proporción de vehículos dentro de la norma para emisiones de CO es:

$$e_p = \sqrt{\frac{(1{,}251{,}652 - 120)}{(1{,}215{,}652 - 1)} \bullet \frac{.24249(1 - .24249)}{120}} = 0.00153$$

Que es un 0.15 % de error de muestreo.

Y por último, para estimar la proporción de vehículos que están dentro de la norma para emisiones de NO_x se tiene:

Tabla 3.8 Para emisiones de NO_x (gr/km)

	Dentro de la norma 0-0.62	***Fuera de la norma >0.62***	***Total***
1991	0.11111	0.88889	**1**
1992	0.16667	0.83333	**1**
1993	0.0625	0.9375	**1**
1994	0.5	0.5	**1**
1995	0.25	0.75	**1**
1996	0.9	0.1	**1**
>1996	0.95122	0.04878	**1**

$$\hat{p}_{est} = \frac{1}{1{,}215{,}652}\left[(188{,}369)(0.1111) + (218{,}930)(0.16667) + \ldots (68{,}062)(0.9) + (233{,}124)(0.95122)\right]$$
$$= .39226$$

La suma estimada del total de la primer columna (dentro de la norma o menor de 0.62 gr/km) de la tabla 3.8 es (0.39226) * (1,215,652) = 476,851 y el total de la segunda columna es 1,215,652 - 476,851 = 738,801.

Para las emisiones de NO_x el error de muestreo es:

$$e_{\hat{p}} = \sqrt{\frac{(1{,}251{,}652 - 120)}{(1{,}215{,}652 - 1)} \bullet \frac{.39226(1 - .39226)}{120}} = 0.00199$$

El error de muestreo para las emisiones de NO_x es de 0.2 %. Con estos resultados se obtienen las siguientes tablas:

Tabla 3.9. Total de marginales para cada uno de los contaminantes de los vehículos de la ZMCM

a. Marginales para la emisión de HC

	0-0.25	*>0.25*	*Total*
1991			**188,369**
1992			**218,930**
1993			**202,286**
1994			**190,247**
1995			**114,634**
1996			**68,062**
>1997			**233,124**
Total	**593,141**	**622,511**	**1,215,652**

b. Marginales para la emisión de CO

	0-2.11	>2.11	Total
1991			188,369
1992			218,930
1993			202,286
1994-1995			304,881
1996			68,062
>1997			233,124
Total	294,780	920,872	1,215,652

c. Marginales para la emisión de NO_x

	0-0.62	>0.62	Total
1991			188,369
1992			218,930
1993			202,286
1994			190,247
1995			114,634
1996			68,062
>1997			233,124
Total	476,851	738,801	1,215,652

3.2 Obtención de Tablas ajustadas

Como ya se tienen las tablas iniciales (tabla 3.2, 3.4 y 3.5) y se conocen los totales de las marginales de la población (tabla 3.9) se puede aplicar el método IPF para estimar la proporción de cada celda. Es importante notar que el total de las marginales del renglón es igual, lo que cambia es la proporción de vehículos que están dentro de la norma para cada uno de los contaminantes o el total de las marginales de las columnas. Una vez que se tienen los totales de las marginales se aplicó el método IPF para cada una de las tablas, para tal efecto se utilizó un programa computacional, que contiene dicho método (ver anexo 2), estos pasos se podían hacer de manera manual, pero como se puede ver, al menos se necesita hacer 15 iteraciones, por lo tanto para facilitar cálculos y ahorrar tiempo se elaboraron los programas. Los resultados arrojados por el programa son los siguientes:

Tabla 3.10. Tablas ajustadas para cada uno de los contaminantes de los vehículos de la ZMCM

a. Resultado de IPF para la emisión de HC

	0-0.25	*>0.25*	*Total*
1991	20,930	167,439	**188,369**
1992	36,488	182,442	**218,930**
1993	82,179	120,107	**202,286**
1994	126,831	63,416	**190,247**
1995	57,317	57,317	**114,634**
1996	47,643	20,419	**68,062**
>1997	221,752	11,372	**233,124**
Total	**593,141**	**622,511**	**1,215,652**

b. Resultado de IPF para la emisión de CO

	0-2.11	*>2.11*	*Total*
1991	20,930	167,439	**188,369**
1992	18,244	200,686	**218,930**
1993	12,643	189,643	**202,286**
1994-1995	19,055	285,826	**304,881**
1996	47,643	20,419	**68,062**
>1997	176,265	56,859	**233,124**
Total	**294,780**	**920,872**	**1,215,652**

c. Resultado de IPF para la emisión de NO_x

	0-0.62	*>0.62*	*Total*
1991	20,930	167,439	**188,369**
1992	36,488	182,442	**218,930**
1993	12,643	189,643	**202,286**
1994	95,123	95,124	**190,247**
1995	28,659	85,975	**114,634**
1996	61,256	6,806	**68,062**
>1997	221,752	11,372	**233,124**
Total	**476,851**	**738,801**	**1,215,652**

Para emisiones de HC el método IPF converge en 20 iteraciones, en el segundo de emisiones de CO el método IPF converge en 18 iteraciones y para la emisión de NO_x el método converge en 24 iteraciones, el criterio de convergencia para detener el proceso fue $|m_{ij}^{t} - m_{ij}^{t-1}| < 10^{-5}$. Con la tabla 3.10 se puede calcular la distribución de probabilidad, que es el paso 4 del método para la creación de poblaciones sintéticas.

3.3 Obtención de la distribución de Probabilidad

En el capítulo dos se vio que no es conveniente aplicar la fórmula (propuesta por Beckman y Baggerly) para obtener la distribución de probabilidad de una tabla de contingencia, solo en algunos casos, por lo que se obtuvo la distribución de probabilidad de las tablas dividiendo el valor de cada celda entre el total de la población.

Tabla 3.11. Distribución de probabilidad de cada contaminante

a. Frecuencia relativa para HC

	0-0.25	***>0.25***	***Frecuencia relativa***	***Frecuencia relativa acumulada***
1991	0.01722	0.13774	**0.15495**	**0.15495**
1992	0.03002	0.15008	**0.18009**	**0.33504**
1993	0.0676	0.0988	**0.1664**	**0.50144**
1994	0.10433	0.05217	**0.1565**	**0.65794**
1995	0.04715	0.04715	**0.09430**	**0.75244**
1996	0.03919	0.0168	**0.05599**	**0.80823**
>1997	0.18241	0.00935	**0.19177**	**1**
Total	**0.48792**	**0.51208**	**1**	

b. Frecuencia relativa para la emisión de CO

	0-2.11	*>2.11*	*Frecuencia relativa*	*Frecuencia relativa acumulada*
1991	0.01722	0.13774	**0.15495**	**0.15495**
1992	0.01501	0.16509	**0.18009**	**0.33504**
1993	0.01040	0.156	**0.1664**	**0.50144**
1994-1995	0.01567	0.23512	**0.25080**	**0.75244**
1996	0.03919	0.0168	**0.05599**	**0.80823**
>1997	0.145	0.04677	**0.19177**	**1**
Total	**0.24249**	**0.75751**	**1**	

c. Frecuencia relativa para la emisión de NO_x

	0-0.62	*>0.62*	*Frecuencia relativa*	*Frecuencia relativa acumulada*
1991	0.01722	0.13774	**0.15495**	**0.15495**
1992	0.03002	0.15008	**0.18009**	**0.33504**
1993	0.01040	0.156	**0.1664**	**0.50144**
1994	0.07825	0.07825	**0.1565**	**0.65794**
1995	0.02359	0.07072	**0.09430**	**0.75244**
1996	0.05039	0.00560	**0.05599**	**0.80823**
>1997	0.18241	0.00935	**0.19177**	**1**
Total	**0.39226**	**0.60774**	**1**	

Ya con la distribución de probabilidad de las tablas, se puede generar la población sintética, que es el quinto y último paso de este método.

3.4 GENERACIÓN DE LA POBLACIÓN SINTÉTICA POR CONTAMINANTE

Con ayuda de la técnica de simulación Monte Carlo que nos ayuda a clasificar la información se generaron dos números aleatorios, el primero sirvió para hacer la clasificación por renglón, es decir, por año, considerando la frecuencia relativa acumulada de la tabla 3.11 y el segundo número para hacer la clasificación por columna, dentro de la norma o fuera de la

norma, por ejemplo: para el caso de las emisiones de HC si el primer número aleatorio es 0.250 cae en la clasificación del segundo renglón, es decir, para modelos 1992 y para saber si esta dentro de la norma o no, se hace uso de la probabilidad condicional, que es la probabilidad de que esté dentro de la norma dado que es del año 1992, para este caso, en números menores a 0.16667 se clasifican dentro de la norma, y si el número es mayor fuera de ella; el número fue 0.599 cae en la clasificación de fuera de la norma. Se creó un programa (ver anexo 3) ya que se necesitaban muchas iteraciones para generar la población, se simularon 50 poblaciones sintéticas para el caso de emisiones de CO y 100 para HC y No_x ,con el fin de validar los resultados con la prueba de bondad de ajuste chi-cuadrada, los resultados fueron los siguientes:

Tabla 3.12. Población sintética para emisión de HC de vehículos de la ZMCM

	0-0.25	*>0.25*	*Total*
1991	20915	167386	**188301**
1992	36509	182494	**219003**
1993	82116	120122	**202238**
1994	126913	63392	**190305**
1995	57317	57318	**114634**
1996	47625	20413	**68038**
>1997	221761	11373	**233134**
Total	**593156**	**622497**	**1215653**

Al aplicarle la prueba de bondad de ajuste chi-cuadrada, para ver si la generación de poblaciones sintéticas era válida, ya que la hipótesis que se plantea es que las frecuencias observadas sean iguales a las frecuencias estimadas (donde la frecuencia observada es la que resulta de la generación de poblaciones sintéticas y la frecuencia esperada es la que resulta del método IPF), se obtuvo $X^2 = 2.25$ contra el valor crítico de tablas de X^2 con $(r - 1)(c - 1) = (7 - 1)(2 - 1) = 6$ grados de libertad, y con un nivel de significación $\alpha = 0.05$, $X^2 = 12.59$ por lo que en este caso no se rechaza la prueba, ya que el valor estimado es menor que el valor crítico de las tablas, con lo que se puede decir que la población sintética, sí es válida para el caso de la emisión de HC en la ZMCM.

Tabla 3.13. Población sintética para la emisión de CO de vehículos de la ZMCM

	0 - 2.11	*>2.11*	*Total*
1991	20910	167565	**188475**
1992	18246	200678	**218924**
1993	126332	189539	**202171**
1994-1995	19067	285831	**304898**
1996	47697	20414	**68111**
>1997	176232	56842	**233074**
Total	**408484**	**807169**	**1215653**

En el caso de emisión de CO el valor de X^2 = 2.25 y el valor crítico de tablas es $X^2_{95,5}$ = 11.07, por lo que no se rechaza la hipótesis de que los datos de emisión de HC siguen la distribución de probabilidad planteada.

Tabla 3.14. Población sintética para la emisión de NO_x de vehículos en la ZMCM

	0 - 0.62	*>0.62*	*Total*
1991	20926	167445	**188371**
1992	36512	182428	**218940**
1993	12629	189710	**202339**
1994	95100	95145	**190245**
1995	28656	85952	**114608**
1996	61256	6821	**68077**
>1997	221694	11378	**233072**
Total	**476773**	**738879**	**1215652**

Por último, para el caso de la emisión de NO_x el valor que se estimó es de X^2 = 0.125 y el valor crítico en tablas para $X^2_{95,6}$ = 12.59 por lo que al igual que los dos casos anteriores la hipótesis se acepta.

3.5 DISCUSIÓN DE RESULTADOS Y CRITERIOS DE VALIDACIÓN

Una vez aplicados los cinco pasos para la generación de poblaciones sintéticas, se analizan los resultados, en la tabla 3.15 se muestra el resultado del cálculo de la proporción de vehículos dentro de los límites de la norma para cada uno de los contaminantes, además de los errores de cálculo en cada uno de los casos.

Tabla 3.15. Resultados del cálculo de la proporción

	% de vehículos dentro de la norma	*% de error*
HC	48.79	0.21
CO	24.25	0.15
NO_x	39.23	0.19

En la tabla 3.15, en todos los casos, el error en la estimación es pequeño, ya que el mayor es de sólo 0.21 %, por lo que puede decirse que para todos los casos los cálculos de la estimación de vehículos son aceptables.

Estos datos sirvieron para completar las tablas de las marginales (tabla 3.9), y con ellos poder aplicar el método IPF para hacer los ajustes a las tablas, y una vez con las tablas ajustadas se pudo obtener la distribución de frecuencia y crear las poblaciones sintéticas. Los resultados de la generación de las poblaciones sintéticas, para cada uno de los contaminantes se pueden ver en las tablas 3.12, 3.13 y 3.14.

En el caso de la generación de las poblaciones sintéticas se aplicó la prueba de bondad de ajuste para ver si la distribución observada era igual a la esperada. Los resultados pueden verse en la tabla 3.16.

Tabla 3.16. Resultados de la creación de poblaciones sintéticas

	X^2	$X^2_{v,\alpha}$	α
HC	2.25	12.59	0.05
CO	2.25	11.07	0.05
NO_x	0.125	12.59	0.05

Se puede observar que en todos los casos se cumple el supuesto de que el conjunto de datos se apega a la distribución, con un nivel de confianza del 95 %, por lo que se puede decir que la simulación es válida, teniendo en cuenta la naturaleza de la muestra, y las limitantes de ésta. Faltaría solo validar los resultados con el método de Turing, pero debido a la diferencia de las mediciones, en la muestra y a la manera en la que se miden las emisiones en los centros de verificación esto no fue posible.

CONCLUSIONES

Al escuchar el término de método para la creación de poblaciones sintéticas uno podría pensar que es un proceso o un método bastante complicado, pero como se pudo ver en los capítulos anteriores, no es difícil utilizar este método, ya que básicamente se necesitan cinco pasos para poder crear las poblaciones sintéticas que son los siguientes:

1. *Elementos teóricos conceptuales para la construcción de la muestra.*
2. *Recoger la información y poner en tablas de distribución de frecuencia.*
3. *Buscar algún modelo estacionario, para obtener los estimadores.*
4. *Obtener la distribución de probabilidad*
5. *Generar la población sintética.*

En términos generales se necesita agrupar el marco muestral en grupos homogéneos que reúnan las condiciones o características necesarias para el estudio de la población. Al hablar de condiciones que debe reunir la muestra hay que hacer referencia a la teoría de probabilidades y a los procesos de selección y estimación, ligados a la misma, ya sea muestreo estratificado, por conglomerados, por cuotas, etc., e incluso habrá ocasiones en las que se necesite un muestreo multietápico. Esto quiere decir que, las muestras cuyos resultados vienen avalados por la teoría, son las muestras probabilísticas en las que cada elemento del universo tiene una probabilidad igual o independiente de figurar en la muestra, bajo este supuesto las estimaciones son insesgadas y se puede calcular el error de muestreo que nos ayuda a determinar la precisión de las estimaciones.

La precisión de los resultados y la posibilidad de extrapolarlos depende del tamaño de la muestra y de los proceso de selección y estimación que se apliquen. Algunas bibliografías sugieren utilizar un tamaño de muestra del 5 % de la población, pero habrá ocasiones en las que ésto no sea posible, por ejemplo, si se quiere crear una población sintética del total de la población de la República Mexicana que es de aproximadamente cien millones de habitantes la muestra sería de 5 millones de habitantes lo cual resulta muy difícil de

conseguir, ya que es una muestra muy grande, por lo que, si se tiene un buen estudio y se selecciona una muestra representativa de la población no se necesita un tamaño de muestra tan grande.

Una vez que se eligió la técnica de muestreo para la selección de la muestra se procede a recoger la información y agrupar los datos en tablas de contingencia. Para este trabajo se tomo una muestra de 120 vehículos y se supuso un muestreo estratificado, como se explicó en el tercer capítulo.

Con las tablas de contingencia se pueden estimar las p[illegible]iones, com[illegible] se vio en el capítulo dos los estimadores se pueden obtener con el m[illegible] IPF éste método no es el única que se puede utilizar, también está el de multiplicación de marginales, el de Newton-Rapson, etc.

En el método IPF los resultados dependen de la matriz inicial, si se utiliza un matriz inicial de 1's el método converge en las *n* dimensiones de la matriz, es decir, si se tiene una matriz de dos dimensiones el método converge en dos iteraciones, y los resultados son los mismos que al utilizar el método de multiplicación de marginales, por eso hay que tener cuidado en la matriz inicial, ya que de ella depende los resultados finales. Para este trabajo se utilizó el método IPF, porque al aplicar la prueba de bondad de ajuste chi-cuadrada al método IPF no se rechaza la hipótesis planteada y con el de multiplicación de marginales si se rechaza.

Una vez que se tiene la matriz estacionaria se puede obtener la distribución de probabilidad que en este caso se determina con la proporción de cada celda. Esta distribución de probabilidad nos sirve para crear la población sintética, cabe mencionar que para este trabajo se utilizó el método de simulación Monte Carlo por que su uso es muy sencillo.

El método Monte Carlo ayuda a resolver problemas analíticos como es el encontrar la solución de integrales, con él se encontró el valor de *pi* (π), y se utiliza para resolver complejos fenómenos físicos, entre muchas aplicaciones que tiene.

La técnica de validación que se utilizó fue la de la chi-cuadrada, también se puede utilizar otros métodos complementarios como el de Turing, o el comparar los resultados simulados con los reales si es que se cuenta con estos.

Para poder decir que el método para la creación de poblaciones sintéticas es válido, se vio que depende de muchos factores como los que se mencionaron anteriormente, que van desde una adecuada técnica de muestreo, un método de convergencia, una distribución de probabilidad, una técnica de simulación e incluso un adecuado método de validación.

Con esto puedo decir que se cumplió con el propósito de este trabajo que fue el de utilizar el método de creación de poblaciones sintéticas para estudiar el comportamiento de la emisión de contaminantes de los automóviles que circulan en la ZMCM. Considero que si se elige adecuadamente la técnica de muestreo y las técnicas de validación la población sintética será muy parecida a la población real.

El método para la creación de poblaciones sintéticas sirve para ver el comportamiento de una población, ya que no solo se utiliza una muestra de la población sino que se pueden extrapolar los resultados y ver el comportamiento de la población total.

Por último recomendaría que se hiciera un estudio en donde las mediciones fueran las mismas que se utilizan en los centros de verificación, y con las norma que regulan a los vehículos en circulación, para que el estudio sea más confiable, ya que en este caso las mediciones fueron en gr/km y en los centros de verificación está en ppm, por lo que se tuvo que utilizar la norma para vehículos nuevos (en esta norma las mediciones están en gr/km).

También utilizar no sólo dos columnas, para vehículos dentro de la norma o fuera de esta sino el utilizar como columnas las calcomanías que se emiten que son; doble cero, cero, uno y dos. Y damos una idea más clara de cómo está en términos de contaminación dividida la flota vehicular en la ZMCM. Si se pudiera, incrementar el tamaño de la muestra y hacer el estudio de la flota vehicular y no tener que recurrir a estudios retrospectivos. Otro aspecto que se podría utilizar es el de analizar por marca, puede resultar que una marca contamine más que otra. Todo esto depende en gran parte de cuál es el marco muestral y de cómo se delimite el universo.

ANEXO 1

CARACTERÍSTICAS DE LOS ESTIMADORES

Las características para los estimadores que se verán se referirán a los estimadores puntuales; es decir, dado un parámetro (sea μ) se estima con un valor de $\hat{x}$.

a) *Estimador insesgado:* En general, cuando el valor esperado de un estadístico empleado como estimador es igual al parámetro de la población que se va a estimas, se dice que es estimador es insesgado.

En símbolos podemos decir:

$\hat{\theta}$ es un estimador insesgado de θ si $E(\hat{\theta}) = \theta$

b) *Estimador eficiente:* La eficiencia se define relacionándola con el estimador más pequeño de la varianza. Si se encontrara un estimador con varianza inferior a la de cualquier otro estimador, se utilizaría éste como base de la medida de la eficiencia; en términos de eficiencia se dice que este estimador de varianza más pequeña es un "estimador eficiente".

c) *Estimador consistente:* A medida que aumentamos el tamaño de una muestra la imagen de la población que nos proporciona será, en general, mejor y, por tanto nuestras estimaciones pordrán ser mejores; dicho de otra manera, parece lógico que procuremos escoger nuestros estimadores de forma que sus cualidades mejoren a medida que aumentemos el tamaño de la muestra, o sea que se aproximen cada vez más a θ a medida que aumentamos n (tamaño de la muestra). Esta propiedad se expresa diciendo que un estimador $\hat{\theta}$ de un parámetro θ es consistente cuando $\underset{n\to\infty}{P}\{(\hat{\theta}-\theta)<\varepsilon\}$ tiende a 1, siendo ε tan pequeño como queramos.

d) *Estimador suficiente*: Esta propiedad fue estudiada por Sir R. A. Fisher. Un estadístico suficiente (tal como $\bar{x}$) es un estimador que utiliza toda la información que posee una muestra sobre el parámetro que se estima. Por ejemplo, $\bar{x}$, es un estimador suficiente de la media de la población μ. Esto significa que no hay otro estimador de μ, tal como la media de la muestra, que pueda añadir más información sobre el parámetro μ que se estima.

Fuentes de información:

- ✓ Yamane, T. (1979): *Estadística*. México: Harla. Tercera edición, pags. 107-112.
- ✓ Jódar, B. (1981): *Análisis estadístico de experimentos, principios básicos*. España: Alhambra, pags. 78-80.

ANEXO 2

PROGRAMA IPF

Este programa fue desarrollado en lenguaje de programación C, y va haciendo los ajustes de la matriz primero por columna y posteriormente por renglón, hasta encontrar una diferencia establecida entre una iteración y otra.

Los datos que se necesitan para que pueda correr este programa son:

a) una matriz inicial, los valores pueden ser cualesquiera, y;
b) b) los totales de renglón y de columna de la matriz final en la que se desconocen los valores de las celdas.

```
#include<stdio.h>
#include<math.h>
#include<conio.h>
#include<stdlib.h>
int c,r,i,j,t,flag;
double
mEnt[15][15],mPro[2][15][15],sRen[15],sCol[15],sRenP[15],sColP[15],total;
char s;
void main()
{
  clrscr();

  textbackground(BLACK);
  textcolor(BLUE);
  printf(" N£mero de renglones ");
  scanf("%d",&r);
  printf(" N£mero de columnas  ");
  scanf("%d",&c);

/* Matriz inicial se piden los valores inciales*/
for (i=0; i<r;i++)
{
    for (j=0; j<c; j++ )
    {
       printf("[%d,%d] :",i,j);
       scanf("%lf",&mPro[0][i][j]);
    }
}

 /* se realizan las sumas y se piden total de marginales */

 total=0;
 /* por renglón se piden los valores de cada renglón de la matriz final */
      for (i=0; i<r;i++)
 {
       printf("renglon %d :",i);
       scanf("%lf",&sRen[i]);
       total+=sRen[i];
```

```
}
/* por columna se piden los valores por columna de la matriz final */
for (j=0; j<c;j++)
{
      printf("columna %d :",j);
      scanf("%lf",&sCol[j]);
}


for(i=0; i<r; i++)
{
     for (j=0; j<c; j++)
     {
          sRenP[i]+=mPro[0][i][j];
          sColP[j]+=mPro[0][i][j];
     }
}

printf("\n la suma total es %lf \n",total);
t=0;
do
{
     flag=0;
     printf("%d",t);
     for (i=0; i<r;i++)
     {
          for (j=0; j<c; j++)
          {
            if ( (t%2)==0 )
            {
              /* Fórmula para hacer el ajuste por renglón */
                 mPro[1][i][j]=(mPro[0][i][j]*sRen[i])/sRenP[i];
                 printf(" \t %7.1f",mPro[1][i][j]);
            }
            else
            {
              /*  Fórmula para hacer el ajuste por columna */
               mPro[1][i][j]=(mPro[0][i][j]*sCol[j])/sColP[j];
               printf(" \t %7.1f",mPro[1][i][j]);
            }
          }
          printf("\n");
      }
     t++;
     for(i=0; i<r; i++)
       sRenP[i]=0;

     for (j=0; j<c; j++)
       sColP[j]=0;

     for(i=0; i<r; i++)
     {
       for (j=0; j<c; j++)
       {
           sRenP[i]+=mPro[1][i][j];
           sColP[j]+=mPro[1][i][j];
```

```
            /* criterio de convergencia para detener las iteraciones */
            if ( 0.00001 < ( fabs(mPro[1][i][j]-mPro[0][i][j]) ) )
            {
                  flag=1;
            }
            mPro[0][i][j]=mPro[1][i][j];
        }
    }
  }while (flag ==1);
}
```

ANEXO 3

PROGRAMA PARA CREACIÓN DE POBLACIONES SINTÉTICAS

Este programa fue desarrollado en lenguaje de programación Java(debido a que es un mejor generador de números aleatorios que el lenguaje de programación C), y sirve para crear las simulaciones de las poblaciones sintéticas, se utiliza la distribución de probabilidad de las tablas de contingencia y se generan dos números aleatorios, uno para hacer la clasificación por renglón y el otro para hacer la clasificación por columna.

Se utilizó este programa debido a que la población es de más de un millón de vehículos y si se toma en cuenta que se generaron dos números aleatorios para la clasificación, se puede imaginar que un programa facilitaría el trabajo para la creación de las poblaciones sintéticas.

```
import java.util.*;
import java.awt.*;

class ranhc
{
  ranhc()
  {
    // Declaración de variables
    System.out.println("incio de proceso: "+new Date() );
    double [][] matriz = new double[7][2];
    short i,j,l;
    int k;
    double al1,al2;
    for (l = 0; l < 100; l++)
    {
      RandomChooser obR = new RandomChooser();

      // generación de números aleatorios
      for(k = 0; k < 1215652; k++)
      {
        al1 = Math.random();
        al2 = Math.random();
        short shIJ[] = obR.chooser(al1,al2);
        matriz[ shIJ[0] ][ shIJ[1] ]++;
      }
      System.out.println("Iteracion : "+l );
    }
    for (i=0;i<7;i++)
    {
      for (j=0; j<2;j++)
      {
         System.out.println("     "+Math.round(matriz[i][j]/100) );
      }
      System.out.println("   " );
    }
    System.out.println("fin de proceso: "+new Date() );
```

```
}
public static void main (String[] arg)
{
  new ranhc();
}
class RandomChooser
{
   short[] shReg = {0,1};
   RandomChooser()
   {
   }
    // Le asigna los índices al valor de la celda, dependiendo del renglón
    o columna que se obtuvo

   short[] chooser(double al1, double al2)
   {
      shReg[1] = 1;
      if ( al1 <= 0.15495)
      {
        shReg[0] = 0;
        if (al2 <= 0.11111)
           shReg[1] = 0;
      }
      else if ( al1 <= 0.33505)
      {
        shReg[0] = 1;
        if (al2 <= 0.16667)
          shReg[1] = 0;
      }
      else if (al1 <= 0.50145)
      {
        shReg[0] = 2;
        if (al2 <= 0.40623)
          shReg[1] = 0;
      }
      else if ( al1 <= 0.65794)
      {
        shReg[0] = 3;
        if (al2 <= 0.66667)
          shReg[1] = 0;
      }
      else if ( al1 <= 0.75224)
      {
        shReg[0] = 4;
        if (al2 <= 0.5)
          shReg[1] = 0;
      }
      else if (al1 <= 0.80823)
      {
        shReg[0] = 5;
        if (al2 <= 0.69999)
          shReg[1] = 0;
      }
      else
      {
        shReg[0] = 6;
```

```
            if (al2 <= 0.95121)
              shReg[1] = 0;
          }
          return shReg;
        }
    }
}
```

BIBLIOGRAFÍA

Libros

1. Abad, A., Servin L.(1982): Introducción al muestreo. México: Limusa. Segunda edición.

2. Berenson, M., Levine D. y Rindskopf D. (1988): Applied Statistica a first course. New Jersey: Prentice Hall.

3. Birkes, D., Dodge, Y. (1993): Alternative Methods of Regression. U.S.A.: John Willey & Sons.

4. Bishop, Y. (1967): Multidimensional contingency tables: cell estimates. Massachusetts: Harvard University Cambridge.

5. Deming, W. (1960): Sample Design in Bussines Research. New Jersey: John Willey & Sons.

6. Everitt, B. (1987): Introduction to optimization Methods and their application in Statistics. New York: Chapman and Hall.

7. Fernández, Ma. José, (1992): Resolución de problemas de estadística aplicada a las ciencias Sociales. Madrid: Síntesis.

8. Freund, E., Smith R. (1989): Estadística. México: Prentice Hall Hispanoamericana. Cuarta edición.

9. Holguin, F., Hayashi L. (1977): Elementos de Muestreo y Correlación. México: Textos Universitarios.

10. Knapp , R. y Miller M. (1992): Clinical epidemiology and biostatistics. U.S.A: National Medical Series from William & Wilkins.

11. Mendenhall W., Wackerly D.y Scheaffer R. (1994): Estadística Matemática con aplicaciones. México : Grupo editorial Iberoamérica. Segunda edición.

12. Mendenhall W., Reinmuth E. (1981): Estadística para administración y economía. México: Grupo editorial Iberoamérica.

13. Mood A., GrayBill F. y Boes D. (1974): Introduction to the Theory of Statisticas. New Jersey: Mc Graw Hill. Tercera edición.

14. O'Muircheartaingh, A. y Clive P. (1977): Expliring Data Structures .N.J.: John Willey & Sons. Vol I.

15. O'Muircheartaingh, A. y Clive P. (1977): Modeling Fitting .N.J.: John Willey & Sons. Vol II.

16. Papacostas, c. S. And Prevedouros, P. D. (1993): Trasportation Engineering and Planning. N.J.: Prentice Hall, Englewood Cliffs. Second edition.

17. Ratkowsky, D. (1990): Handbook of nonlinear regression model. New York: Marcel Dekker, Inc.

18. Rodríguez, J. (1991): Métodos de muestreo. Madrid: Centro de investigaciones Sociológicas. Colección de cuadernos Metodológicos.

19. Selltiz C., Jahoda M., Deutsch M. y Cook S.W. (1965): Métodos de la investigación en las relaciones sociales. México: Rialp S.A..

20. Siegel, S., Castellan N. (1988): Nonparametric Statistics for the behavioral sciences. U.S.A: Mc Graw-Hill. Segunda edición.

Revistas

1. Beckman, R., Bagerly, K. and McKay, M. (1996): *Creating synthetic baseline populations.* Traspn. Res. A. 30, 415-429.

2. Deming, W. E. and Stephan, F. (1940): *On a least squares adjustment of a sampled frecuency table when the expected marginal tables are known.* Annals Mathl Stats 11, 427-444.

3. Ireland C. T. and Kullback S. (1968): *Contingency tables with given marginals.* Biometrika 55, 179-188.

4. Little R. J. and Wu M. (1991): *Models for contingency tables with know marginals when target and sampled poputations differ.* J. Stat Assoc. 86, 87-95.

Internet

1. http://wwwitch.unl.edu/zeng/joy/mclab/mcintro.html

2. http://www2.ncsu.edu/eos/info/mat310_info/Monte/calcInt.html

3. http://www3.uniovi.es/user_html/Psicología/metodos/tutor.7/p2.html

4. http://www.coesc.gob.mx/imeca.html

5. http://www.rds.org.mx/INEPub/pmcaum/capi2.html

6. http://www.txinfinet.com/mader/ecotravel/mexico/ecologia/97/1197df2.html

7. http://www.ine.gob.mx/indicadores/espanol/ca2_32.html