DISEÑO DE MUESTREO PARA ESTIMAR EL TIEMPO DE FUNCIONAMIENTO BAJO ESPECIFICACIONES TÉCNICAS DE MEDIDORES DE GAS INSTALADOS EN COLOMBIA

JHOAN ANDRES BOLAÑOS RUIZ

UNIVERSIDAD DEL CAUCA

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y DE LA EDUCACIÓN

DEPARTAMENTO DE MATEMÁTICAS

POPAYAN

2019

DISEÑO DE MUESTREO PARA ESTIMAR EL TIEMPO DE FUNCIONAMIENTO BAJO ESPECIFICACIONES TÉCNICAS DE MEDIDORES DE GAS INSTALADOS EN COLOMBIA

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARCIAL PARA OPTAR AL TITULO EN MATEMATICAS

JHOAN ANDRES BOLAÑOS RUIZ

DIRECTOR

Dr. YILTON RIASCOS FORERO

UNIVERSIDAD DEL CAUCA

FACULTAD DE CIENCIAS NATURALES, EXACTAS Y DE LA EDUCACIÓN

DEPARTAMENTO DE MATEMÁTICAS

POPAYAN

2019

Nota de aceptación

_____

_____

_____

_____

Director: _____

Dr. Yilton Riascos Forero

Jurado: _____

Mg. Edwin Rengifo Cañizales

Jurado: _____

Jhon Alejandro Delgado

Lugar y fecha de sustentación: Popayán, 19 de julio de 2019

AGRADECIMIENTOS

Doy gracias a Dios por darme la oportunidad de tener personas en este mundo tan maravillosas que me apoyan y me ayudan en todas las formas posibles. Mis padres, hermanos, esposa e hija.

A mi director el profesor Yilton Riascos Forero por su paciencia y enseñanzas que me han ayudado a culminar este logro en mi vida.

Tabla de Contenido

Introducción

Se va a estudiar ciertos conceptos de la teoría de muestreo, el muestreo es uno de los métodos básicos de la estadística para obtener información que permite satisfacer necesidades establecidas en proyectos de investigación de diferente cobertura, se constituye actualmente en una metodología científica para la toma de decisiones mediante el uso de inferencia estadística.

Tres aspectos, son de suma importancia en la obtención de una muestra probabilística desde el punto de vista práctico: primero como se obtienen los datos de la muestra, la cual debe estar regida por la aleatoriedad con el propósito de evitar sesgos de selección; en segundo lugar, el tamaño de dicha muestra soportada sobre la base de supuestos teóricos para la distribución, comportamiento del estimador asociado al parámetro investigado, el error asumido con la estimación debido a la diferencia entre la estimación y el parámetro objeto de estimación y el nivel de confianza en términos de probabilidad que se desea tenga el cumplimiento de dicho error. Finalmente en tercer lugar los costos involucrados en la obtención de dicha muestra, pues nadie niega la importancia de trabajar con muestras significativas y que me brinden toda la información para el análisis que se necesita.

Una buena parte de las decisiones investigativas, están basadas en datos estadísticos que provienen del análisis, observación o medición de unas cuantas unidades estadísticas, llevadas al laboratorio o seleccionadas para ser encuestadas utilizando diversas estrategias teóricas y practicas

Capítulo 1

1. Antecedentes

En este capítulo se presenta los inicios de cómo se empezó a trabajar con el gas natural, para así ver como se introduce el gas natural en Colombia, además, como es la distribución de gas y los medidores que manejan las compañías gaseras, se presenta la distribución de redes en todo el País. También, con el propósito de tener un mejor conocimiento sobre el desgaste de los medidores en el tiempo en Colombia, se lleva a cabo un estudio previo de muestreo basado en una muestra óptima arrojo información precisa y confiable en el tiempo de los medidores tipo domiciliar en Colombia.

1.1 Historia del gas[1]

El gas natural es hoy en día una fuente de energía que circula bajo el suelo de la mayor parte de las ciudades del mundo civilizado; aporta comodidad doméstica y provee a la industria de la energía que necesita. Paradójicamente, el gas natural que ahora llamamos "la energía del futuro", es conocido por la humanidad hace miles de años. Los hombres primitivos observaban las llamaradas que se producían en los pantanos cuando caía un rayo. Desde entonces, el tercer estado de la materia, el gaseoso, no ha dejado de inspirar curiosidad y temor, por lo misterioso e intangible de su naturaleza.

Los primeros descubrimientos de yacimientos de gas natural fueron hechos en Irán entre los años 6.000 y 2.000 a.C. Estos yacimientos de gas, probablemente encendidos por primera vez mediante algún relámpago, sirvieron para alimentar los "fuegos eternos" de los adoradores del fuego de la antigua Persia. También se menciona el uso del gas natural en China hacia el 900 a.C. Precisamente en China se reporta la perforación del primer pozo conocido de gas natural, de 150 metros de profundidad, en el 211 a.C. los chinos perforaban sus pozos con varas de bambú y primitivas brocas de percusión, con el

[1] Tomado del trabajo de Guerrero Suárez & Llano Camacho, 2002.

propósito expreso de buscar gas en yacimientos de caliza. Quemaban el gas para secar las rocas de sal que encontraban entre las capas de caliza. En el siglo VII en Japón se descubrió la existencia de un pozo de gas.

Las civilizaciones griega y romana, así como la Edad Media, conocieron los efectos de la combustión del gas. En el siglo XVI Paracelso, alquimista y médico suizo, produjo por primera vez gas combustible (hidrógeno) por contacto de ácidos con metales y lo llamó "espíritu salvaje"; Juan Bautista van Helmot lo denominó "ghost" (fantasma, espíritu) de donde se derivó, por deformación de esta palabra, el nombre de "gas".

En el siglo XVII Robert Boyle, químico y físico irlandés, obtuvo vapor de agua, alquitrán gas por destilación o carbonización de la hulla. Así mismo, en Gran Bretaña, William Murdock consiguió en 1792 alumbrar con gas su casa y sus talleres. El gas lo obtenía en una retorta vertical de hierro estañado y se conducía por tubería a unos veinte metros de distancia. En 1797 se instaló luz, a partir del gas, en la Avenida Pall Mall de Londres, y a partir de entonces se desarrolló rápidamente la industria del gas en Inglaterra. En Alemania, Guillermo Augusto Lampidus, farmacéutico y químico, alumbró en 1811 con gas un sector de Freiberg en donde era profesor de química en la escuela de minas.

También en Alemania, en 1828, se alumbraron las calles de Dresden en un gran acontecimiento, en presencia del Rey de Sajonia. Gracias al aporte del austriaco Carl Auer (con el mechero que lleva su nombre), a partir de 1895, el gas de alumbrado adquirió gran importancia en las principales ciudades del mundo. Su aplicación como fuente de luz y calor se desarrolló aceleradamente por su facilidad de transporte por tuberías y la sencillez de la regulación y control de la llama, en una época en que no existía la electricidad.

De acuerdo con lo anterior, en principio el gas que comenzó a utilizarse en las ciudades europeas fue de origen manufacturado, obtenido de la destilación o carbonización de la hulla. Este gas preparó el camino tecnológico a la posterior utilización del gas natural.

Los Estados Unidos fueron los pioneros de la exploración y explotación del gas natural. En 1821, los habitantes de Fredonia (cerca de Nueva York), hicieron un pozo de nueve metros de profundidad y condujeron el gas por tuberías de madera y de plomo a varias casas para su alumbrado.

A lo largo del siglo XIX, el uso del gas natural permaneció localizado porque no había forma de transportar grandes cantidades de gas a través de largas distancias, razón por la que el gas natural se mantuvo desplazado del desarrollo industrial por el carbón y el petróleo.

A comienzos de 1900 el gas manufacturado es implementado en Argentina, país con mayor historial en Latinoamérica en este tema. A partir de 1930 comenzaron a explotarse en los Estados Unidos los yacimientos de gas, independientemente de los petrolíferos. Hasta entonces el gas natural que acompañaba el petróleo era quemado o reinyectado en los pozos para mantener la presión de extracción del petróleo.

Un importante avance en la tecnología del transporte del gas ocurrió en 1890, con la invención de las uniones a prueba de filtraciones. Sin embargo, como los materiales y técnicas de construcción permanecían difíciles de manejar, no se podía llegar con gas natural más allá de 160 kilómetros de su fuente. Por tal razón, la mayor parte del gas asociado se quemaba en antorchas y el no asociado se dejaba en la tierra.

El transporte de gas por largas distancias se hizo practicable a fines de la segunda década del siglo XX por un mayor avance de la tecnología de tuberías. En Estados Unidos, entre 1927 y 1931 se construyeron más de diez grandes sistemas de transmisión de gas. Cada uno de estos sistemas se construyó con tuberías de unos 51 centímetros de diámetro y en distancias de más de 320 kilómetros.

Después de la Segunda Guerra Mundial se construyeron más sistemas de mayores longitudes y diámetros. Se hizo posible la construcción de tuberías de 142 centímetros de diámetro. Pero el gran auge en la historia del gas natural no llega, prácticamente, hasta 1960. Entonces los grandes descubrimientos y la explotación de importantes yacimientos en diferentes partes del mundo, especialmente en Europa Occidental, Rusia y norte de África, dan progresivamente una auténtica dimensión mundial a la industria del gas natural.

1.2 Gas natural en Colombia

La utilización del gas natural en Colombia se remonta al descubrimiento de los campos de Santander. Con excepción de los campos de gas libre, el gas asociado fue considerado en el país como un subproducto de la explotación del crudo, y era quemado en las teas (un tipo de antorcha) de los campos petroleros. Desde 1961, la conciencia sobre el valor del gas se empieza a plasmar en la legislación, y es por primera vez a través de la Ley 10 de 1961, que se prohíbe de forma explícita su quema, posteriormente se ratifica mediante el decreto 1873 de 1973.

En 1973 se inicia la construcción en la Costa Atlántica del primer gasoducto para atender las necesidades del sector industrial para esa zona del país, extendiéndose a todos sus departamentos. Con el objeto de sustituir energéticos de alto costo, en 1986 se estableció el primer plan nacional de uso general del gas natural, llamado "Programa de gas para el cambio". El bajo volumen de reservas de esa época y la coyuntura en que se desenvolvían los energéticos, los cuales estaban subsidiados, limitaron el desarrollo de este plan.

En 1990 surge una vez más la necesidad de crear la cultura del gas. Con el documento oficial "Lineamientos del cambio", se da pie para que se adelanten una serie de estudios, los cuales confirman los beneficios económicos que se derivarían para el país a partir de la utilización de este combustible.

Hacia finales de 1991, el Consejo Nacional de Política Económica y Social (CONPES) aprobó el programa para la masificación del consumo de gas, con base en el estudio que había adelantado en cooperación con la Comunidad Económica Europea, en el cual se identificaron los principales proyectos del plan de masificación del gas. En este documento el CONPES esbozó una política macroeconómica y energética integral, en la que se establecieron las facilidades para los particulares en la construcción de gasoductos troncales, mediante el esquema de concesión. Igualmente se presentó la posibilidad de la distribución a cargo de empresas privadas o mixtas.

La entonces Comisión Nacional de Energía aprobó, en mayo de 1992, el sistema de transporte de gas, separándolo en troncal, subsistemas y distribución, para garantizar un suministro adecuado a los futuros usuarios. En 1993, se elaboró el documento Minminas Ecopetrol DNP-2646- UINF-DIMEN, a través del cual se expresó nuevamente la necesidad de promocionar una matriz energética más eficiente y conveniente para el país, mediante sustitución de energéticos de alto costo. En el mismo año se expidió el Decreto 408 de marzo 3, en el cual el CONPES aprobó las estrategas para el desarrollo del Plan Gas, que contemplaban la conformación de un sistema de transporte de gas natural, donde Ecopetrol ejercería, directamente o por contrato, la construcción de los gasoductos utilizando esquemas de BOMT (siglas en inglés del esquema de financiación en donde un inversionista privado Construye (B), Opera (O), Mantiene (M) y Transfiere (T o similares), para conectar los campos de producción con los centros de consumo en el país, estableciendo el marco normativo y tarifario, designando a los entes respectivos para garantizar la penetración del gas natural.

Se vio también la necesidad de crear un sistema de transporte de gas independiente de los productores, comercializadores y distribuidores, que garantizase el acceso abierto en igualdad de condiciones a todos los usuarios. Así se llegó, después de varios años de debate, a la creación de la Empresa Colombiana de Gas, Ecogás, el 20 de agosto de 1997, como una Empresa Industrial y Comercial del Estado, con autonomía presupuestal y administrativa, cuya misión es administrar y controlar, operar y explotar comercialmente los sistemas de gasoductos en el interior del país.

Con estas políticas, la masificación del uso del gas se hace una realidad que permitirá modificar el patrón de consumo de todos los sectores y establecer una oferta adecuada de energía. Es así como por motivos de interés social y con el fin de que la cobertura de los servicios públicos se pueda extender a personas de menores ingresos, la Ley 142 de 1994 faculta al Ministerio de Minas y Energía (MME) para conformar áreas de servicio exclusivo para la distribución domiciliaria de gas combustible y suscribir contratos de concesión especial en los que se incluyen cláusulas de exclusividad que establecen que ninguna otra empresa podrá prestar el servicio de distribución en esa área.

En la actualidad la red de distribución de gas en Colombia cubre gran parte del territorio como se observa en las siguientes figuras.

Figura 1. Mapa de redes de distribución de gas en Colombia

Fuente: Ecogas

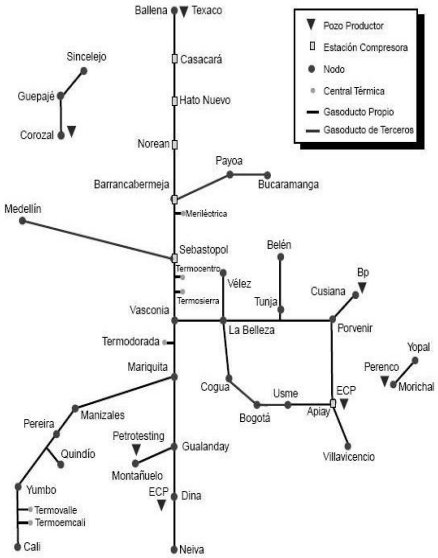


Figura 2. Diagrama de flujo de la red Nacional de Gasoductos de Colombia

Fuente: Ecogas

1.3 Distribución de gas en Colombia

En el país existen más de quince empresas distribuidoras de gas natural, cada una de ellas debe disponer de un sistema de tuberías para tomar el gas de las City Gate y entregarlo a los consumidores finales. Se destacan como mercados objetivos los sectores

1. Residencial
2. Comercial
3. Industrial
4. El gas comprimido para vehículos (GNCV).

Particularmente en relación al gas del mercado residencial, las empresas prestadoras del servicio, entre las que se encuentran Alcanos, Gases de Occidente, EPM, Surtigas, entre otras, emplean regularmente los medidores G1.6 tipo diafragma para registrar el consumo de los usuarios. Muchos de estos medidores son fabricados por la empresa Metrex S.A.,

Figura 3. Medidor para Gas G1.6 Tipo Diafragma

El cuidado y mantenimiento de estos equipos de medición son responsabilidad de las empresas en primer lugar y de los usuarios en segunda instancia, por lo que la empresa debe realizar revisión de instalaciones domésticas cada cinco años.[2]

Los medidores G1.6 tipo diafragma, al analizar su proceso de producción, deben responder,

[2] Resolución CREG 067 de 1995

en funcionamiento, a especificaciones establecidas en la norma técnica colombiana 2728 para medidores de gas tipo diafragma (ver [7]),

Capítulo 2

Planteamiento del Problema

Es evidente que la instalación del servicio de gas natural domiciliar en Colombia mejoró la calidad de vida de las personas, al permitir que la reducción del costo económico que implicaba el consumo de energía, para las mismas actividades, pudiera ser destinada a otros gastos familiares.

Este servicio es prestado por compañías gaseras del país y su costo es estipulado a partir del consumo que registran los medidores que estas compañías instalan en las viviendas de los usuarios.

Estas compañías gaseras deben cumplir exigencias de normatividad nacional y funcionamiento para estos medidores. Por información previa de estudios de medidores en otros países se sabe que, en intervalos de cinco años, se realicen inspecciones a las redes internas de gas domiciliar, que consiste en un procedimiento que verifica todos los requisitos de una vivienda segura.

A pesar de esto, se presentan inquietudes en relación con el tiempo de funcionamiento adecuado (bajo especificaciones) que dura trabajando un medidor domiciliar, lo que implica dudas acerca del cobro de facturación del consumo de gas.

Por ello, las compañías gaseras en Colombia están interesadas en conocer el tiempo máximo de funcionamiento de un medidor de gas antes de salirse de las especificaciones. Este interés implica la realización de un diseño experimental que permita, a partir de la selección de una muestra, estimar las características de instalación de los medidores que puedan afectar este funcionamiento. Para el estudio participa la compañía Metrex encargada de reponer los medidores que son extraídos y cuatro compañías gaseras que son: Alcanos con una

participación del 18.1%, Surtigas con una participación del 27.1%. Gases de Occidente con una participación 20.5% y EPM con una participación del 24.3%.[3]

Esta investigación se interesa por establecer las condiciones y características del diseño de muestreo que garantice el éxito del diseño experimental que se requiere.

Para ello se van a estudiar conceptos de la teoría de muestreo, debido a que éste es uno de los métodos básicos de la estadística para obtener información que permita satisfacer necesidades establecidas en proyectos de investigación de diferente cobertura y se constituye actualmente en una metodología científica para la toma de decisiones mediante el uso de inferencia estadística.

Se va abordar la problemática de cuantos medidores se deben revisar de la población de medidores para determinar el funcionamiento de los medidores de gas domésticos instalados en Colombia, para saber en qué tiempo están por fuera de especificaciones y empiezan a presentar fallas en su funcionamiento en el tiempo. Se va a revisar una muestra de los medidores retirados y se les hará una prueba la cual consiste en dejar un periodo de ocho horas en el laboratorio para que los medidores estén a temperatura ambiente, luego, se colocan de diez medidores en la maquina donde se le hacen tres pruebas de aproximadamente una hora y esta verifica a distintos tipos de caudal si el medidor está o no bajo especificaciones según la norma técnica y se decidirá si el medidor esta bueno o malo.

Metodológicamente, se responderá la siguiente pregunta: ¿Cual diseño de muestreo probabilístico resulta adecuado para estimar el tiempo de funcionamiento y las características del parámetro tiempo medio de funcionamiento de la población de medidores de gas domiciliar instalados en Colombia?

En este estudio se pretende llevar al cabo un diseño de muestreo estadistico que permita recolectar los datos para estimar el tiempo de un medidor de gas antes de estar fuera de

[3] A partir de aquí se va a llamar a las empresas gaseras como: Empresa 1 (Alcanos), Empresa 2 (Surtigas), Empresa 3 (Gases de Occidente), Empresa 4 (EMP)

especificaciones, teniendo en cuenta los reportes de las empresas productoras. La altura sobre el nivel del mar a la que se encuentra instalado el medidor porque es una variable que parece tener incidencia en el funcionamiento, así que esta será considerada dentro del estudio e inicialmente se establecerá rangos de variación para esta variable y se **utilizará** dentro del diseño para estratificar los valores que se tengan.

Finalmente entonces se identificar el tipo de muestreo que se va a utilizar de los cuatro tipos de muestreo que se conocen; muestreo aleatorio simple, muestreo aleatorio sistemático, muestreo por conglomerado, muestreo aleatorio estratificado y posteriormente dentro de esto el tipo de estimación que se utilizará si es para promedio, si es para proporciones o si es para totales. Luego con base en esto definir el tamaño de la muestra, cuantos elementos de la muestra se van a tomar en total y cuantos por cada uno de los estratos que se establecen por piso térmico bajo las condiciones que se trasmiten adecuadas para esto.

A continuación se hace una breve demostración de algunos tipos de muestreos estadísticos

Capítulo 3

3. Muestreo aleatorio simple

De acuerdo a las características que este método presenta, entenderemos el *Muestreo Aleatorio Simple* (*MAS*) como una selección aleatoria de elementos con la misma probabilidad de ser elegidos para formar parte de la muestra, en otras palabras, es un muestreo equiprobable y puede hacerse con o sin restitución (o reemplazamiento) en un conjunto.

En la práctica, un *MAS* se realiza unidad por unidad, enumerando las unidades de observación de uno a *N* (donde *N* es el total de unidades del conjunto), y seleccionando un conjunto de números aleatorios ya sea utilizando una tabla o mediante un programa de computación que los genere.

En cada extracción, el proceso debe otorgar la misma oportunidad de selección a todos y cada uno de los números que no hayan sido elegidos. Las unidades que se relacionen con estos *n* números constituyen la muestra y si en todas las extracciones subsecuentes se descarta cada vez un número extraído, este método es llamado *muestreo aleatorio sin restitución*.

En este caso, la probabilidad de selección de una muestra específica está dada mediante cálculos estadístico-matemáticos por $\frac{(N-n)!}{N!}$, ya que $\frac{N!}{n!(N-n)!} = NCn$ es el número total de subconjuntos (muestras) de tamaño n que pueden ser seleccionados de un conjunto (población) de tamaño N. En forma similar puede calcularse la probabilidad de que una unidad cualquiera de la población esté presente en la muestra. Esta probabilidad puede obtenerse dividiendo el número de muestras posible que contendrían la unidad específica considerada, por el número posible de muestras, esto es $\frac{(N-1)C(n-1)}{nCn} = \frac{n}{N}$.

A menudo un parámetro importante a estimar en una población es la *media poblacional*. En el *MAS* sin reemplazo la media muestral es definida como $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ la cual es un

estimador insesgado de la media poblacional $\bar{y}$. Igualmente, la *varianza de la media muestral* que se define como, $V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$ donde $S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{n-1}$ es, así mismo, un estimador insesgado de la varianza poblacional finita S^2.

Otro estimador de la $V(\bar{y})$ se obtiene reemplazando S^2 por su respectivo estimador s^2, siendo el estimador de la varianza un estimador insesgado. Y lo denotamos como $V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$ y $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$.

El factor $(N-n)/N$ que aparece en la fórmula de la varianza se denomina *correcciones debidas a población finita o corrección por finitud* (*cpf*). Se debe tener presente que, siempre y cuando la fracción de muestreo n/N sea pequeña, este factor toma valores cercanos a la unidad y el tamaño de la población como tal no tiene un efecto directo en el error estándar de la media de la muestra. Mientras que, si la fracción de muestreo es grande, la información que se tiene de la población también será mayor y, por lo tanto, la varianza es menor. Para poblaciones grandes, el tamaño de la muestra extraída es el que determina la precisión del estimador (y no el porcentaje de población muestreada).

Existen algunas situaciones donde el objetivo principal es estimar el total poblacional de los valores de una variable (o parámetro) y un estimador insesgado del total poblacional Y, está dado por $\hat{Y} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i$, donde N/n se conoce como el *factor de expansión* y $\sum_{i=1}^{n} y_i$ es el *total muestral*.

De la misma manera que la varianza de la media muestral es un estimador insesgado, la varianza del estimador del total $\hat{Y}$ es insesgado y puede deducirse fácilmente a partir de esta, obteniendo como resultado $V(\hat{Y}) = N^2 V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{N^2 S^2}{n} = N(N-n)\frac{S^2}{n}$. Cuyo estimador se define como $v(\hat{Y}) = N(N-n)\frac{s^2}{n}$.

Sin embargo, en muchas de las investigaciones que se llevan a cabo que buscan conocer las preferencias que las personas tienen por determinado candidato o producto, o la cantidad de elementos de una población que pueden clasificarse en un grupo específico. Esto implica que la investigación debe centralizarse en la estimación de la proporción poblacional del atributo deseado o del número de elementos que poseen dicho atributo. Por esta razón estudiaremos la *proporción poblacional P,* y el número total de unidades en la población con el atributo deseado *A*, que pueden estimarse con p y $\hat{A}$ respectivamente, y que se definen como $p = \frac{a}{n}$ y $\hat{A} = Np = \frac{N}{n}a$, donde a es el número de elementos en la muestra con el atributo deseado. Las varianzas de estimación de la proporción muestral y de estimador del total de unidades con la característica están dadas por $V(p) = \frac{N-n}{N-1}\frac{PQ}{n}$, $Q = 1 - P$ y $V(\hat{A}) = N^2\frac{N-n}{N-1}\frac{PQ}{n}$, y sus respectivos estimadores están dados por $V(p) = \frac{N-n}{N}\frac{pq}{n-1}$ que es también insesgado y $V(\hat{A}) = N(N-n)\frac{pq}{n-1}$.

Similarmente que el *MAS* sin restitución, el *MAS con restitución o reemplazo* también es practicable cuando en cada extracción todos los miembros de la población (*N*) tienen la misma oportunidad de extracción, sin que importe el número de veces que se extrajeron antes. Además, en el muestreo con restitución las fórmulas que se presentan de las varianzas y varianzas estimadas realizadas a partir de la muestra son más simples, motivo por el cual, se utilizan en planes de muestreo[4] más complicados, aunque parece ser inútil tener dos o más veces la misma unidad dentro de la muestra.

En vista de que las *n* selecciones de las unidades son independientes y, al igual que el *MAS* sin reemplazo, cualquier posible sucesión de *n* unidades, distinguidas por el orden de selección y donde puede haber selecciones repetidas, tienen la misma probabilidad de ser seleccionada, esto da lugar a que el número total de muestras posibles sea N^n ya que cualquiera de las *N* unidades puede aparecer en cualquiera de las *n* selecciones. Por lo tanto, la probabilidad de selección de una sucesión específica de *n* unidades es $1/N^n$. De manera similar puede calcularse la probabilidad de que una unidad cualquiera de la población sea seleccionada al menos una vez y se expresa como sigue $1 - \left(\frac{N-1}{N}\right)^n$.

[4] Según Ospina B.D. el plan de muestreo es la metodología utilizada para seleccionar la muestra de la población.

Para realizar estimaciones en este tipo de muestreo respecto a la media, el total, la proporción y varianzas muestrales y poblacionales se pueden determinar de forma similar al *MAS* sin restitución teniendo en cuenta ciertos requerimientos que influyen en el desarrollo de estas. Por ejemplo, una consecuencia inmediata es que, la varianza de la media muestral $V(\bar{y})$ en un muestreo sin restitución, es solamente $(N-n)/(N-1)$ veces su valor en el muestreo con restitución.

Es de gran importancia recordar y tener siempre presente que en el muestreo de poblaciones finitas se asume que el número total de elementos de la población es un valor N conocido, de los cuales se seleccionan aleatoriamente *n*, donde generalmente se tiene la relación $n < N$. Así, la selección aleatoria de los elementos se vuelve una condición indispensable para poder hacer uso correcto de los procesos de inferencia estadística. Dado el caso de trabajar con poblaciones finitas, usualmente la varianza de los valores yi se define como $\sigma^2 = \sum_{i=1}^{N} \frac{(y_i-\bar{Y})^2}{N}$.

Por motivos de notación, los resultados se presentan en términos de una expresión ligeramente diferente, en donde el divisor (*N-1*) se usa en lugar de *N*. Entonces se tiene la siguiente expresión $S^2 = \sum_{i=1}^{N} \frac{(y_i-\bar{Y})^2}{N-1}$.

Esta convención la usan quienes enfocan la teoría del muestreo por medio del análisis de la varianza. Su ventaja es que la mayoría de los resultados toman una forma ligeramente más simple.

Otro resultado interesante que se tiene, es que las fórmulas para los errores estándar de la estimación de la media y del total de la población se usan principalmente para tres propósitos: uno, para comparar la precisión obtenida por el muestreo aleatorio simple con otros métodos de muestreo; otro, para estimar el tamaño de la muestra que se necesita en una encuesta que esté siendo planeada; y el otro, para estimar la precisión realmente obtenida en una encuesta que se haya terminado. Dichas fórmulas involucran a la varianza de la población S^2, que en

general no será conocida en la práctica pero puede ser estimada a partir de los datos de la muestra.

Aunque, generalmente se presupone que las estimaciones de la media de la población $\bar{y}$ y del total $\hat{Y}$ se distribuyen normal alrededor del valor correspondiente de la población. Las razones[5] para esta suposición y sus limitaciones se consideran a continuación:

- Se han hechos muchos estudios en la teoría de probabilidades sobre la distribución de las medias de muestras aleatorias.
- Se ha probado que para cualquier población que tiene una desviación estándar finita, la distribución de la media muestral tiende a la normalidad conforme *n* aumenta.

Este conocimiento deja algo que desear. No es fácil contestar a una pregunta tan directa y aparentemente simple como lo es "¿Qué tan grande debe ser *n* en esta población para que la aproximación normal sea suficientemente exacta?". Una buena práctica de muestreo tiende a hacer la aproximación normal más válida, es decir, un diseño de muestreo bien estructurado. La falla en la aproximación normal ocurre principalmente cuando la población contiene algunas unidades extremas que dominan el promedio de muestra cuando están presentes. Además, estos extremos también tienen un efecto más serio, al aumentar la varianza de la muestra y disminuir la precisión.

Al planear una encuesta por muestreo, siempre se alcanza una etapa en donde hay que tomar una decisión respecto al tamaño de la muestra. No se podría discutir este asunto sin antes contestar dos preguntas fundamentales: ¿Qué tan grande debe ser la muestra? y ¿Qué tan exacta desea la estimación? Análisis apresurados de la situación estudiada pueden conducir a definir tamaños de muestra insuficientes que no proporcionan estimaciones con la precisión y confiabilidad requeridas, disminuyendo la utilidad de los resultados, o, en el otro extremo, tamaños muy grandes que, aunque pueden cumplir con los objetivos trazados, desborden el presupuesto asignado. La decisión no siempre puede tomarse satisfactoriamente, a menudo

[5] Cochran. 1980. Pág. 66.

no disponemos de suficiente información para saber si el tamaño de la muestra seleccionada, es el óptimo. La teoría del muestreo proporciona un marco dentro del cual podemos analizar, mejor, respecto a este problema.

A continuación se enuncian algunas pautas necesarias, involucradas en la solución de la elección del tamaño de la muestra[6]:

- Debe existir algún enunciado respecto a lo que se espera de la muestra. Este puede darse en términos de límites de error deseados o bien en términos de alguna decisión o acción que debe tomarse una vez que se conocen los resultados de la muestra.
- Se debe encontrar una ecuación que relacione n con la precisión deseada de la muestra. La ecuación variará según el contenido del enunciado de precisión y el tipo de muestreo propuesto. Esta ecuación tendrá como parámetros ciertas propiedades desconocidas de la población, que deben estimarse para obtener resultados específicos.
- Con frecuencia sucede que los datos estipulan para ciertas subdivisiones mayores de la población y que los límites de error deseados se establecen para cada subdivisión. De ser así, se hace un cálculo separado para el valor n en cada subdivisión y el n total se encuentra por adición.
- Generalmente se mide más de un atributo o característica en una encuesta por muestreo. Si se estipula un grado de precisión para cada atributo, los cálculos conducirán a una serie de valores conflictivos de n, uno para cada atributo. Por lo tanto, debe encontrarse un método para reconciliar estos valores.
- Finalmente, debe apreciarse el valor elegido de n, para que sea consistente con los recursos de muestreo disponibles. Esto exige una estimación del costo, trabajo, tiempo y materiales que se necesitan para obtener la muestra del tamaño propuesto o encontrado.

[6] Cochran. 1980. Pág. 105.

En el caso del *MAS* sin reemplazo la fórmula más utilizada y de más fácil manejo es la varianza del estimador de la media poblacional. Esta fórmula es válida, tanto para estimar la media como para estimar el total poblacional, siempre y cuando se esté considerando una población finita. Si el tamaño es desconocido (o se asume infinito) sólo se podrá estimar la media poblacional mas no el total.

Usando la fórmula para la varianza de la media muestral y después de manipulaciones algebraicas, se obtiene $n = \frac{\frac{S^2}{V(\bar{y})}}{1+\frac{S^2}{N\,V(\bar{y})}}$, donde cada uno de los términos ya ha sido definido. Desafortunadamente la fórmula anterior presenta varios problemas, siendo el principal de ellos el desconocimiento tanto de S^2 como de $V(\bar{y})$ que son parámetros poblacionales. Al ser desconocidos, es necesario asignarles valores previos. S^2 se estima a menudo con base en estudios anteriores de la misma población o de poblaciones similares, en un conocimiento aproximado de la distribución original de la variable bajo estudio o a través de una muestra piloto. En el caso de $V(\bar{y})$ es menos complicado de determinar ya que corresponde a la varianza del estimador de la media que se utiliza. El investigador, por tanto, puede establecer de antemano un valor para este parámetro (que no sería otra cosa que fijar la variabilidad que se desea para el estimador que en este caso es la media muestral).

Asumiendo que el *Teorema del Límite Central*[7] puede aplicarse, es necesario ser prudentes, especialmente en el caso del *MAS* sin reemplazo[8]. Si las observaciones originales $y_1, y_2, ..., y_n$ son una sucesión de variables aleatorias independientes e idénticamente distribuidas con media y varianza finitas, la distribución de $\frac{\bar{y}-\bar{Y}}{\sqrt{V(\bar{y})}}$ se aproxima a una distribución normal estándar cuando n aumenta. El resultado es válido si $V(\bar{y})$ se reemplaza por un estimador razonable y n es suficientemente grande. Esto indica que, en el caso del *MAS* con reemplazo, la aproximación puede usarse, siempre y cuando *n* sea razonablemente grande. Por lo tanto, $\bar{y}$ se distribuye aproximadamente normal con media $\bar{Y}$ y varianza $V(\bar{y})$. Asumiendo que se conoce $V(\bar{y})$, un intervalo de confianza apropiado para $\bar{Y}$ sería:

[7] Sea X una variable aleatoria con media μ y desviación típica $\sigma > 0$, definida en una población cualquiera. Si n es grande $(n \geq 30)$ y el tamaño de la población es grande comparado con n, entonces la media muestral $\bar{X}$ está aproximadamente distribuida como la normal con media μ y desviación típica $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

[8] Citado por Ospina B. D. 2001. Pág. 39.

$$\bar{y} \pm zE(\bar{y})$$

Donde z es el correspondiente cuantíl de la distribución normal estándar y $E(\bar{y})$ es el *error absoluto máximo en la estimación de la media* (equivalente a la mitad del intervalo de confianza establecido, el cual se denota como ε. $E(\bar{y})$ puede entonces reemplazarse por $\frac{\varepsilon}{2}$ y $V(\bar{y})$ por $\frac{\varepsilon^2}{2^2}$. Sustituyendo este último valor en la fórmula para n se llega a

$$n = \frac{\frac{z^2 S^2}{\varepsilon^2}}{1 + \frac{1}{N}\left(\frac{z^2 S^2}{\varepsilon^2}\right)}$$

Esta fórmula indica que el tamaño de muestra es una función de cuatro factores:

- La confiabilidad para las estimaciones. Esto es, qué tan a menudo se espera que las estimaciones obtenidas de muestras aleatorias simples independientes se encuentren como máximo a una distancia ε de la verdadera media poblacional $\bar{Y}$.
- La variabilidad de la población, representada por S^2, es la varianza poblacional.
- La precisión en las estimaciones, representada por ε, es el error absoluto máximo admisible en la estimación de la media.
- El tamaño de la población N.

De las cantidades mencionadas, z y ε son establecidas previamente por el investigador, ya que corresponden a la confiabilidad y precisión mínimas deseadas para la estimación, y *N* se supone que es conocida. El no contar con el verdadero valor de S^2, sugiere utilizar la estimación S^{*2} de la varianza poblacional, cualquiera que haya sido el procedimiento para ello. La fórmula para n es entonces

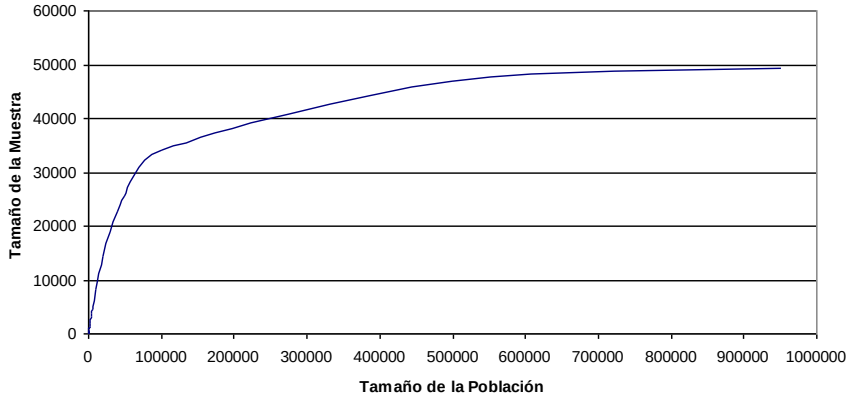
$$n = \frac{\frac{z^2 S^{*2}}{\varepsilon^2}}{1 + \frac{1}{N}\left(\frac{z^2 S^{*2}}{\varepsilon^2}\right)}$$

A groso modo, de la fórmula anterior para *n* podemos concluir que esta es bastante sensible a pequeños cambios en el error absoluto máximo y en menor intensidad al nivel de confiabilidad y a la variabilidad de la población. El tamaño poblacional sólo tiene verdadera importancia cuando su valor es pequeño. Además, como es de suponerse, a un mayor error corresponde un menor tamaño de muestra, pero a un nivel de confiabilidad mayor, una variabilidad más grande o una población más numerosa, el tamaño de muestra aumenta.

Si N es suficientemente grande, o el muestreo es con reemplazo, n puede aproximarse por n_0, como $n_0 = \frac{z^2 S^2}{\varepsilon^2}$. Así, n puede ser expresado como función de n_0 y obtenemos $n = \frac{n_0}{1+\frac{n_0}{N}}$.

El valor de n_0 se considera entonces como una cota superior para *n* cuando se toma *N* como única variable. Esto, por lo tanto, desvirtúa una opinión bastante generalizada que considera que el tamaño de muestra debe ser proporcional al tamaño de la población o de que la muestra debería crecer indefinidamente si la población también lo hace (Como se presenta en el gráfico 1).

Gráfico 1. Convergencia del tamaño de la muestra



Azorín, F[9]. (1969) demuestra que existe un valor *N* a partir del cual no tiene sentido incrementar *n*. Este valor viene dado por $N = \frac{z^2 S^2}{\varepsilon^2}\left(\frac{z^2 S^2}{\varepsilon^2} - 1\right)$, donde S^2 puede ser reemplazado por S^{*2}, para fines prácticos.

Pero, no siempre el interés del proceso de estimación está centrado en controlar el error máximo absoluto. En muchos casos es más conveniente tratar de controlar el *error máximo relativo*. Una justificación razonable, considera que el establecimiento del error como una cantidad absoluta no puede suponerse que sea suficiente, puesto que, al no compararse esta cantidad con otra de interés, no puede afirmarse que sea grande o pequeña. Esto es válido cuando se desconoce en alto grado la distribución poblacional de la variable y el resultado de la muestra se vuelve bastante incierto. Asumiendo aceptable la aproximación normal, los tamaños de muestra estimados son frecuentemente mayores cuando se desea controlar el error relativo que el error absoluto.

Por otro lado, la eficiencia de un diseño específico, que se denota *EFD*[10] o *deff*[11], es la eficiencia relativa que el estimador, usando este diseño, tiene con respecto al estimador del *MAS*, en la estimación de un parámetro. En forma general se denota de la siguiente manera $EFD = \frac{V(\hat{\theta})}{V(\hat{\theta}_{MAS})}$.

La fórmula anterior puede ser utilizada para estimar el tamaño de muestra necesario de acuerdo con el nuevo diseño, para lograr la misma eficiencia que con el *MAS*.

Utilizando la metodología empleada anteriormente, en la *determinación del tamaño de muestra para proporciones* en el *MAS*, cuando se desea controlar el error máximo absoluto se llega a $n = \frac{\frac{z^2 PQ}{\varepsilon^2}}{\frac{N-1}{N} + \frac{1}{N}\left(\frac{z^2 PQ}{\varepsilon^2}\right)}$, donde el principal problema es el desconocimiento de las proporciones poblacionales P y Q, para ello se recurre a una estimación previa P^2 con base

[9] Citado por Ospina B. D. 2001. Pág. 81.
[10] Citado por Ospina B. D. 2001. Pág. 84-85.
[11] Cochran. 1980. Pág. 119.

en el conocimiento que se tenga de la población. Así, una fórmula más apropiada para determinar el tamaño es $n = \frac{\frac{z^2 P^* Q^*}{\varepsilon^2}}{\frac{N-1}{N} + \frac{1}{N}\left(\frac{z^2 P^* Q^*}{\varepsilon^2}\right)}$.

El problema de la estimación del tamaño de muestra en proporciones es, generalmente, más fácil de solucionar que en el caso de la media, ya que la proporción es un número real entre 0 y 1, lo cual permite, en el peor de los casos, establecer un tamaño de muestra suficiente, cuando la proporción se hace igual a $1/2$, valor donde $n_0 = \frac{z^2 P^* Q^*}{\varepsilon^2}$ tiene su máximo. Esto implica que en ningún caso el tamaño de muestra debe ser mayor a

$n_0 = \frac{z^2}{4\varepsilon^2}$.

Por ello, controlar el error máximo relativo es de gran importancia cuando existe un desconocimiento considerable de la proporción poblacional. En esas situaciones establecer un error máximo absoluto puede ser un inconveniente pues se puede fallar fácilmente por exceso o por defecto, originando valores para *n* demasiado grandes o pequeños (en este último caso se pueden originar, eventualmente, intervalos de confianza inconsistentes, donde el límite inferior es negativo). En el caso del error relativo ello no ocurre siempre que se tenga cuidado con la estimación previa de *P*.

De forma análoga, se puede derivar fácilmente la fórmula para *n* en términos del error relativo, cuya expresión es $n = \frac{\frac{z^2 Q^*}{\delta^2 P^*}}{1 + \frac{z^2 Q^*}{N\delta^2 P^*}}$.

Desafortunadamente este tamaño de muestra es muy sensible a ligeros cambios en la estimación previa de P^*, lo que implica hacer un esfuerzo adicional para tener estimaciones válidas de este parámetro. Si el tamaño de la población es muy grande, la anterior fórmula puede aproximarse por $n_0 = \frac{z^2 Q^*}{\delta^2 P^*}$, cuya función es decreciente en P, y por tanto, no tiene máximo.

Así, la especificación de la precisión deseada se puede establecer, al definir la cantidad de *error tolerable*[12] en las estimaciones muestrales. En ocasiones, es difícil saber qué tanto error debería tolerarse, particularmente cuando los resultados se destinan a varios fines. Si la muestra se toma con un propósito bien definido como sería una decisión entre "si" o "no", entonces, la precisión requerida se puede enunciar usualmente de una manera más específica, en términos de las consecuencias de los errores de decisión. Una estimación inicial del tamaño requerido de la muestra se hace separadamente para cada una de estas características importantes. En tal caso, un método para determinar el tamaño de la muestra es la especificación de los márgenes de error para las características que se consideran vitales en la encuesta por muestreo.

Al completar la estimación de *n* para cada característica, se concreta la situación. Puede ser que los n requeridos estén bastante próximos. Si el valor más grande de *n* está dentro de los límites del presupuesto, se toma este valor. Pero con frecuencia, hay suficiente variación entre los n y, por lo tanto, no se selecciona el más grande, ya sea por consideraciones presupuéstales o porque este valor dará una precisión global sustancialmente más elevada que la considerada en un principio. En este caso, el estándar de precisión deseado puede relajarse un poco para algunas características, lo que permite utilizar un valor más pequeño de n. Hay casos en que los n requeridos, para diferentes características, son tan discordantes que algunos se deben abandonar en la investigación; ya que con los recursos disponibles, la precisión esperada para estas características es totalmente inadecuada. Además, algunas características requerirán tipos diferentes de muestreo en comparación con otras.

En ocasiones, puede desarrollarse un enfoque más lógico para determinar el tamaño de la muestra, cuando se va a tomar una decisión práctica, basada en los resultados de la muestra. Se puede presuponer que la decisión estará más sólidamente fundamentada, si la estimación muestral tiene un error, pequeño, en lugar de uno elevado.

Otra consideración en este tipo de problemas a tener en cuenta es que, el conocimiento aproximado de la varianza poblacional es el principal problema a enfrentar cuando se desea

[12] Error tolerable es por definición $e = |\hat{\theta} - \theta|$, donde $\hat{\theta}$ y θ son el estimador y el parámetro poblacional, respectivamente.

estimar un tamaño de muestra y controlar el error máximo absoluto. Existen varias opciones para estimar esta varianza; entre estas se describen a continuación las más recomendadas[13]:

1. Revisión bibliografía de estudios anteriores sobre la misma población o poblaciones similares. Si en tales estudios se presentan estimaciones de varianza, el buen juicio del investigador decidirá cual de las estimaciones es la más apropiada. Es posible que ella necesite ajustarse teniendo en cuenta el tipo de variable analizada, el tiempo transcurrido y el hecho de que la población estudiada sea la misma u otra similar.
2. Selección de una muestra piloto de tamaño n_1 (generalmente debe ser menor de 30) y estimación de la varianza poblacional S^2 con la varianza de esta muestra, la cual representa a S^{*2} en la fórmula respectiva. Si la muestra no se selecciona en una forma completamente aleatoria, los elementos que forman parte de ella no deben ser parte de la muestra definitiva. Muchas veces los elementos pueden seleccionarse intencionalmente atendiendo criterios de expertos en el tema. El trabajo piloto se limita a una parte de la población que se puede manejar convenientemente, o que revelará la magnitud de ciertos problemas.
3. Selección de una muestra aleatoria simple de tamaño n_1 y cálculo de la varianza de esta muestra. Esta varianza se toma como estimación de la varianza poblacional. Sin embargo, el hecho de estimar la varianza a partir de una muestra aleatoria pequeña conlleva un margen de incertidumbre que se supone puede corregirse multiplicando el valor obtenido por un factor dependiente de n_1. Este proporciona las estimaciones más confiables de S^2, pero no se usa frecuentemente porque tarda la consumación de la encuesta. La fórmula final para n queda expresada por $n = \frac{z^2 S^2}{\varepsilon^2}\left(1 + \frac{2}{n_1}\right)$.

   La cantidad $\frac{2}{n_1}$ es el precio que se paga por el desconocimiento de S^2. Una desventaja tanto de este procedimiento como del anterior, es la demora que se puede presentar en la recolección de toda la información. No obstante, según Ospina B.D. este procedimiento se considera como el más aceptable, entre los descritos aquí.

4. Una alternativa similar a la del numeral anterior es presentada por *Desu* y *Raghavarao* (1990) y es una modificación del procedimiento bietápico de Stein (1945). La diferencia con el caso anterior es que se hace uso de la distribución *t* antes

[13] Citado por Ospina. 2001. Pág. 30

que de la distribución normal en la fórmula de cálculo de *n*. Esta fórmula se convierte en:

$$n = \max\left\{n_1, \left[\frac{t^2_{n_1-1}S^{*2}}{\varepsilon^2}\right] + 1\right\}$$

Donde t_{n_1-1} es el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribución t con $n_1 - 1$ grados de libertad.

5. Determinación tentativa, o con base en supuestos adecuados, de la estructura de la población para escoger la distribución teórica que mejor podría representarla (normal, exponencial, uniforme, etc.). La identificación de una distribución apropiada permite hacer uso de sus propiedades para obtener una estimación más realista de la varianza. En ocasiones es posible hacer una estimación útil de la varianza poblacional S^2, a partir de una información relativamente escasa respecto a la naturaleza de la población.

Deming (1960)[14] muestra cómo algunas distribuciones matemáticas simples son útiles en la estimación de S^2, a partir de cierto conocimiento del intervalo donde se encuentre y de una idea general de la forma de la distribución. Cuando el desconocimiento es total se debe recurrir a la distribución uniforme.

3.1 Muestreo aleatorio estratificado

Una población homogénea con N unidades, $\{U_i\}, i = 1,2,3, \dots, N$, se divide en L grupos

$$\{U_{hi}\}, h = 1,2,3, \dots, L\ ; i = 1,2,3, \dots, N_h$$

Con N_h tamaño del estrato h.

Obtener una muestra de una población homogénea, muy sencillo y tranquilizante en tanto permita que la conclusión obtenida con base en la dicha muestra, es muy cercana a la realidad, vale decir la estimación del parámetro en dicha población, es bastante cercana al mismo y un MAS, es lo suficientemente bueno para lograr los objetivos establecidos con el muestreo.

[14] Citado por Cochran. 1980. Pág. 114.

Situaciones como lo anterior, no son muy comunes, en la realidad casi siempre la población que se investiga tiene bastante variabilidad y se requiere que ésta afecte lo menos posible, a las estimaciones que se obtengan de dicha población, mediante una muestra de ella.

Se denomina *estratificación*, a un proceso que se asigna las unidades poblacionales a cada grupo en que se ha dividido una población, de acuerdo con unos criterios prefijados con anticipación. Cada grupo se llamará *estrato.* El proceso de muestreo, una vez encasilladas las unidades poblacionales (estratificada), posibilitará la realización en cada estrato de un muestreo independiente, lo cual, facilitará la aplicación de diferentes métodos de muestreo acorde a la información disponible, el costo y las razones que motivaron la estratificación de la población.

Los criterios para estratificar una población, así como el número de estratos a considerar, dependerá de los objetivos de la investigación, de la información disponible y de la estructura de la población. Se debe tener siempre presente, que las variables utilizadas para realizar la estratificación deben estar altamente correlacionadas con las variables de estudio en la investigación.

El proceso de estimación más frecuente en el muestreo estratificado consiste en considerar cada estrato como una subpoblación y realizar, en primera instancia, estimaciones acerca de los parámetros correspondientes a cada una de esas subpoblaciones. Una vez se han hecho las estimaciones para estos subgrupos, se procede a combinarlas para obtener las estimaciones globales de los parámetros de interés. La estimación dentro de cada estrato puede llevarse a cabo de acuerdo a diferentes procedimientos, lo importante es que las muestras seleccionadas en cada uno de estos sean independientes para poder obtener fórmulas directas de estimación para los parámetros poblacionales.

Una ventaja del MAE con respecto al MAS tiene que ver con la eficiencia de sus estimadores, los cuales, en general, para un tamaño de muestra global *n*, poseen varianzas menores que los del MAS. Esto, sin embargo, no implica que se tenga que estratificar siempre.

Así que, para implementar este muestreo existen muchas razones, dentro de las cuales se consideran como principales las siguientes:

1. Los datos de interés deben tener una precisión determinada para algunas subdivisiones de la población.
2. Por conveniencia administrativa.
3. El problema de muestreo puede tener marcadas diferencias en diversas partes de la población.
4. La estratificación puede dar lugar a una ganancia en la precisión de las estimaciones de características de la población total. Quizá sea posible dividir una población heterogénea en subpoblaciones, en las que cada una sea internamente homogénea.

La principal característica de la teoría del muestreo estratificado es que se ocupa de las propiedades de las estimaciones de una muestra estratificada y de la mejor elección para los tamaños de muestras n_h en los estratos que deben dar la precisión máxima.

En este desarrollo se presupone que ya se construyeron los estratos. Aunque este asunto presenta varios problemas tales como ¿Cuál es la mejor característica para la construcción de los estratos? ¿Cómo se determinan los límites de error entre estratos? ¿Cuántos estratos debería haber? Dado el número de estratos, las ecuaciones para determinar los mejores límites entre ellos bajo asignación proporcional y de Neyman, han sido obtenidas por Dalenius (1957). Otros investigadores han encontrado algunos métodos de aproximación más rápidos.

Si queremos estimar la proporción de unidades en la población que pertenecen a una clase definida D, la estratificación ideal se obtiene al colocar en el primer estrato toda unidad que pertenezca a D y en el segundo toda unidad que no pertenezca. Si esto falla, tratamos de construir estratos tales que la proporción de unidades P_h en la clase D varíe tanto como sea posible de estrato a estrato.

Así como en el caso del MAS, en este tipo de muestreo un estimador insesgado de la media poblacional $\bar{Y}$ del MAE está dado por $\bar{y}_{st} = \frac{\sum_{k=1}^{L} N_h \bar{y}_h}{N} = \sum_{k=1}^{L} W_h \bar{y}_h$, donde la media muestral $\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ es un estimador insesgado de la media verdadera $\bar{Y}_h$en el estrato h y $W_h = N_h/N$ es la proporción de la población que pertenece a dicho estrato (o tamaño relativo de cada estrato).

La diferencia entre la media muestral, que puede escribirse como, $\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n}$ y $\bar{y}_{st}$, es que las estimaciones a partir de estratos individuales reciben sus ponderaciones correctas. Es claro que $\bar{y}$ coincide con $\bar{y}_{st}$ cuando en cada estrato $\frac{n_h}{n} = \frac{N_h}{N}$ o $\frac{n_h}{N_h} = \frac{n}{N}$.

Esto significa que la fracción de muestreo es la misma en todos los estratos. En otras palabras, la cantidad de unidades en la muestra y en cada estrato es proporcional al tamaño del propio estrato. Así, esta estratificación se describe como *estratificación con asignación proporcional* de los números n_h y da lugar a una muestra autoponderada, donde cada unidad de la muestra tiene el mismo peso y representa el mismo número de unidades de la población.

Si las muestras de cada estrato son independientes, la varianza del estimador de la media poblacional está dada por $V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h)$, donde $V(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$ es la varianza muestral en cada estrato h. Una forma alternativa para esta varianza es:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_k} - \sum_{h=1}^{L} \frac{W_h S_h^2}{N}.$$

Estimadores insesgados para $V(\bar{y}_{st})$ y $V(\bar{y}_h)$ se obtienen reemplazando las varianzas verdaderas por sus correspondientes estimadores.

Un estimador insesgado del total poblacional Y es $\hat{Y}_{st} = N\bar{y}_{st}$ cuya varianza de este estimador está dada por $V(\hat{Y}_{st}) = N^2 V(\bar{y}_{st})$.

Similarmente, como se definió en el MAS la proporción poblacional P, el número total de elementos en la población con el atributo deseado A y las varianzas con sus respetivos estimadores insesgados, se definen análogamente para el MAE.

En el MAS el único procedimiento para mejorar la eficiencia del estimador de la media (y de otros relacionados con ella) consiste en el aumento del tamaño de muestra ya que todos los otros términos son conocidos. En el MAE la situación es diferente ya que la varianza del estimador de la media es una función no sólo de los tamaños y varianzas verdaderas de los estratos sino de los tamaños de muestra asignada a cada uno de ellos. Como esta es una decisión que generalmente debe tomar el investigador, es importante conocer la influencia que ella tiene en la bondad del estimador utilizado. Para esto, existen varios procedimientos de asignación o afijación de la muestra a los diferentes estratos. Entre ellos:

1. **Afijación uniforme**. Si todos los estratos tienen aproximadamente el mismo tamaño y no hay ninguna información disponible acerca de la variabilidad existente dentro de los estratos, lo más sencillo y aconsejable es asignar a cada uno de los estratos el mismo tamaño de muestra. Esta afijación también es conocida como *afijación igual* y está dada por $n_h = \frac{n}{L}$, donde n_h es el número de unidades en la muestra. Este tipo de afijación da la misma importancia a todos los estratos, en cuanto al tamaño de la muestra. Favorece a los estratos pequeños y perjudica a los grandes en cuanto a precisión.

2. **Afijación Proporcional.** Cuando los tamaños de los estratos son diferentes, es común darle a todas las unidades en la población la misma probabilidad de formar parte de la muestra. Para que ello se cumpla, es necesario que el tamaño de muestra correspondiente a cada estrato sea proporcional al tamaño de dicho estrato, esto se representa como $n_h = nW_h$.

3. **Afijación óptima.** Las dos afijaciones anteriores pocas veces producen buenos resultados pues ellas no tienen en cuenta la *homogeneidad* dentro de los estratos, razón, que en último caso, justifica la estratificación. La posición más conveniente

consiste en balancear la *variabilidad* dentro de los estratos con el tamaño de ellos. La asignación resultante teniendo en cuenta estos dos aspectos (homogeneidad y variabilidad) se denomina *afijación óptima* y está dada por $n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$.

Obsérvese que, el tamaño de muestra correspondiente a cada estrato es directamente proporcional al producto del tamaño del estrato por su variabilidad, representada esta por la desviación estándar poblacional.

La afijación óptima produce "la mejor estimación" en el sentido que da una menor varianza del estimador de la media (y, por tanto, del total y de la proporción). Como las desviaciones estándar poblacionales deben estimarse de antemano, una fórmula más apropiada para n_h sería reemplazando S_h por S_h^*, donde esta última es una estimación "previa" de la desviación estándar para el estrato h.

Para calcular el tamaño de muestra en un MAE tenemos:

Para estimar una media μ,

$$n = \frac{\sum_{i=1}^{k} N_i^2 \sigma_i^2 / W_i}{\frac{N^2 B^2}{Z_{\alpha/2}^2} + \sum_{i=2}^{k} N_i \sigma_i^2}$$

Donde B es el sesgo.

Para estimar una proporción p,

$$n = \frac{\sum_{i=1}^{k} N p_i q_i / W_i}{\frac{N^2 B^2}{Z_{\alpha/2}^2} + \sum_{i=2}^{k} N_i p_i q_i}$$

Para estimar un total τ,

$$n = \frac{\sum_{i=1}^{k} N_i^2 \sigma_i^2 / W_i}{\frac{B^2}{Z_{\alpha/2}^2} + \sum_{i=2}^{k} N_i \sigma_i^2}$$

Un factor a menudo determinante en la asignación de la muestra es el costo de recolección de la información. Para tal propósito, la función de costo frecuentemente considerada es $C = C_0 + \sum_{h=1}^{L} C_h\, n_h$, siendo C el presupuesto total asignado a la recolección de la información, C_0 el costo fijo que no depende del número de elementos a seleccionar y C_h el costo de muestrear un elemento perteneciente al estrato h. Generalmente los costos de transportes entre estratos se incluyen en C_0 y los de transporte entre elementos de un mismo estrato en C_h. El costo como es de suponerse afecta negativamente el tamaño de muestra para cada estrato, entre mayor sea el costo de muestreo por elemento, menor es el número de elementos a seleccionar del estrato.

En general, si los costos por unidad son los mismos en todos los estratos, dos reglas de trabajo útiles son: (a) la ganancia en precisión del MAE sobre el MAS es pequeña o modesta, a menos que la proporción de unidades en D en el $h - ésimo$ estrato, varíe mucho de estrato a estrato; (b) la asignación óptima con una n fija, producirá estimadores con menor varianza que los producidos por asignación proporcional cuando existe variabilidad entre las varianzas de los estratos.

3.2 Muestreo aleatorio sistemático

Este método es una forma modificada del *MAS* puesto que comprende la selección de elementos de una población de manera sistémica en lugar de al azar. Una concepción para este tipo de muestreo se puede establecer en términos de una muestra sistemática y podemos enunciarla como sigue.

Una *muestra sistemática (MS)* es una selección aleatoria de elementos de una población ordenada que se hace con base en un intervalo constante de longitud k del marco muestral, en el cual se realiza inicialmente un *MAS* para obtener la primera unidad de la muestra y luego mediante el proceso iterativo de sumar k veces a la posición anterior de la unidad seleccionada, se va conformando el resto de la muestra hasta completar n.

También, podemos definir una *MS* como un *MAS* de una unidad conglomerada, tomada de una población de k unidades conglomeradas, al considerarse $N = nk$. Así, la operación de elegir una muestra sistemática aleatoriamente localizada, es solo la de elegir una de estas grandes unidades de muestreo al azar. Por lo tanto, el muestreo sistemático viene a ser la elección de una sola unidad de muestreo compleja, que constituye la muestra total.

El *MS* es frecuentemente comparado con el *MAS* y con el *MAE*. En muchas ocasiones, la población está ordenada con respecto a una variable específica (edad, tamaño, tiempo, etc.), en otros casos no existe ningún criterio conocido y se puede asumir que, con respecto a la variable de interés en el estudio, los elementos están ordenados aleatoriamente. Toda la información previa que se pueda obtener permitirá juzgar acerca de la conveniencia o no de la utilización del *MS*.

Si la población puede considerarse ordenada de manera completamente aleatoria, la eficiencia del *MS* es equivalente a la del *MAS* y debe utilizarse, pues generalmente proporciona economía en tiempo y dinero. Si, por el contrario, la correlación entre la variable de estudio y la de ordenamiento es alta, la eficiencia del *MS* será mayor que la del *MAS* y en muchos casos similar a la proporcionada por el *MAE*.

Específicamente, cuando la población tiene tendencia lineal con respecto a la variable de estudio, se ha demostrado que el *MS* es más eficiente que el *MAS* pero menos que el *MAE* con una observación por estrato. Cuando la población presenta variaciones periódicas en la ordenación de sus elementos, el *MS* no es aconsejable pues puede ocurrir que el intervalo aleatorio sea múltiplo o submúltiplo de la longitud del periodo considerado, lo cual llevaría a seleccionar siempre elementos muy similares en lo que a la variable se refiere.

Sin embargo, existen algunas situaciones donde el *MS* es la alternativa apropiada respecto al *MAS* y al *MAE*. Entre ellas tenemos[15]:

[15] Muestreo Estadístico Métodos Básicos. Klinger A. 1991.

1. Cuando no existe previamente un marco muestral y este se va complementando con el tiempo.
2. Es más fácil sacar una muestra y a menudo, más fácil hacerlo sin cometer errores.
3. Intuitivamente, el *MS* puede ser más preciso que el *MAS*. En efecto, estratifica la población en *n* estratos, que consisten de las k primeras unidades, las segundas k unidades, etc. Por lo tanto, podemos esperar que la *MS* sea tan precisa como la *MAE* correspondiente con una unidad por estrato. La diferencia radica en que en el *MS* cada elemento seleccionado en la muestra ocupa la misma posición relativa dentro del estrato, mientras que en el *MAE* los elementos de cada estrato se obtienen aleatoriamente.
4. La *MS* se reparte más uniformemente sobre la población, y este hecho, algunas veces ha dado al *MS* una precisión mayor que la del *MAE*.

Ahora bien, si el tamaño de la población N es conocido, para una muestra sistemática de n elementos, k debe ser menor o igual que N/n y si N es desconocido no podemos seleccionar exactamente a k, y se debe suponer el valor de k necesario para obtener el tamaño de muestra n requerido y por lo tanto, las diferentes muestras sistemáticas de la misma población finita pueden variar de tamaño. El sesgo producido cuando $N \neq nk$ es despreciable si el tamaño de la muestra es grande.

De esta manera, el *MS* es preciso cuando las unidades dentro de una misma muestra son heterogéneas y es impreciso cuando son homogéneas. Si hay poca variación dentro de una *MS* en relación con la de la población, las unidades sucesivas en la muestra repiten más o menos la misma información.

En el *MS*, el principal inconveniente para estimar el tamaño de muestra es el desconocimiento que se tenga acerca del patrón de ordenamiento de la variable de estudio. Si el orden es aleatorio, el problema se reduce a estimar un tamaño de muestra en el *MAS*. No obstante, este supuesto es muy fuerte y a menudo no se cumple. En estos casos, existe una solución práctica que, sin embargo, puede representar un costo adicional considerable. El procedimiento consiste en seleccionar m' muestras piloto sistemáticas, repetidas, del marco muestral para obtener estimaciones iniciales de los parámetros. Con base en estas estimaciones, se determina el número m definitivo de muestras necesarias para estimar los parámetros con la

precisión y confiabilidad requeridas. Realmente el tamaño de las muestras piloto, así como su número debe ser pequeño (entre cinco y diez elementos en cinco o seis muestras es suficiente)[16]. El valor *m* se obtiene a través de la siguiente fórmula:

$$m = \frac{\frac{z^2 v(y_i)\left(\frac{N}{n}\right)}{(\bar{\bar{y}})}}{\delta^2\left(\frac{N}{n}-1\right)+\frac{z^2 v(\bar{y}_i)}{(\bar{\bar{y}})^2}}$$

Siendo $\bar{y}_i$ la media de la i-ésima muestra piloto sistemática, $\bar{\bar{y}}$ la media de estas medias y $var(\bar{y}_i)$ su correspondiente varianza, donde $\bar{\bar{y}} = \frac{1}{m} = \sum_{i=1}^{m} \bar{y}_i$ y $v(\bar{y}_i) = \frac{\sum_{i=1}^{m}(\bar{y}_i - \bar{\bar{y}})}{m-1}$.

Una vez han sido seleccionados los elementos de la muestra, la media muestral en muestreo sistemático lineal $\bar{y}_s$, se calcula de la manera tradicional, esto es $\bar{y}_s = \frac{\sum_{i=1}^{n} y_i}{n}$ el cual a su vez es un estimador insesgado siempre que $N = nk$. Como en el caso de la media, un estimador del total poblacional $\hat{Y}_s = N\bar{y}_s$ es insesgado bajo la condición de que $N = nk$.

Uno de los principales problemas en el *MS* es la estimación de la varianza del estimador de la media poblacional $V(\bar{y}_s)$, ya que el comportamiento de la media muestral es muy diferente dependiendo de la ordenación original de la variable y del intervalo sistemático escogido.

La varianza de la media de una muestra sistemática es $V(\bar{y}_s) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S^2$, donde $S_s^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$ es la varianza entre las unidades que se encuentran dentro de la misma *MS*, cuyo denominador de esta varianza $k(n-1)$, se construye por medio de las reglas usuales en el análisis de la varianza: cada una de las k muestras contribuyen con $(n-1)$ grados de libertad a la suma de cuadrados del numerador.

La media de una *MS* es más precisa que la media de una *MAS* si y solamente si $S_s^2 > S^2$. Este resultado importante que en general se aplica al *MC*, enuncia que el *MS* es más preciso que el

[16] Ospina B.D. 2001. Pág. 153.

MAS si la varianza dentro de las muestras sistemáticas es mayor que la varianza de la población total.

El procedimiento en la estimación de la proporción poblacional es el mismo que el utilizado en el *MAS*, siempre que se pueda asumir aleatoriedad en el ordenamiento de la población. A menudo este supuesto se cumple más fácilmente para las variables dicotómicas que para las discretas y continúas. El estimador para esta proporción viene dado por $P_s = \frac{a}{n}$ donde a es el número de elementos de la muestra con el atributo deseado y la varianza de este estimador es $V(P_s) = \left(\frac{N-n}{N-1}\right)\frac{PQ}{n}$.

Tanto el *MAE* como el *MS* son mucho más efectivos que el *MAS*, pero, el *MS* tiene menos precisión que el *MAE*. El éxito del *MS* con relación al *MAS* o *MAE*, depende mucho de las propiedades de la población. En algunas poblaciones, el *MS* es extremadamente preciso y en otras resulta menos preciso que el *MAS*. Es necesario conocer algo sobre la estructura de la población para usarlo de manera efectiva.

3.3 Muestreo aleatorio por conglomerados

Los métodos de muestreo descritos en los capítulos anteriores, en donde el proceso de selección se lleva a cabo para unidades individuales, no siempre son los más convenientes debido a los altos costos económicos (dinero, tiempo, recursos) que generalmente conllevan, así como a la dificultad en algunos casos de identificar con anterioridad las unidades de estudio. Cuando ello ocurre, es aconsejable recurrir a otras técnicas convenientes como el muestreo por conglomerados que permite seleccionar, antes que unidades elementales, grupos de ellas que sirven como unidades de muestreo, conocidas como *conglomerados*. En muchas situaciones de la vida real, los conglomerados se construyen de unidades que están físicamente cercanas.

En esta sección definiremos el *muestreo por conglomerados* (*MC*) en una etapa simple como un plan de muestreo en el cual se selecciona los conglomerados haciendo uso del *MAS* sin reemplazo y dentro de cada conglomerado seleccionado se escogen todas las unidades elementales que lo componen.

La facilidad que representa la selección de conglomerados, y que redunda en un puesto menor, conlleva un "castigo" que se mide en términos de eficiencia: la varianza de los estimadores en el *MC* es generalmente mucho mayor que en el *MAS* o *MAE* debido a que la similaridad entre unidades de un mismo conglomerado pueden incrementar sustancialmente el error de estimación ya que la mayoría de la información recolectada seria "redundante". Un análisis preliminar tanto de los costos como de la eficiencia esperada para los estimadores, debe conducir a la selección del método apropiado. Si se selecciona el *MC*, es necesario considerar los siguientes factores[17]:

1. Los conglomerados deben estar bien definidos de manera que todo elemento de la población pertenezca a uno y sólo a un conglomerado.
2. Debe existir una estimación razonable acerca del número de elementos de cada conglomerado.
3. Los conglomerados deben ser suficientemente pequeños para que sea posible algún ahorro en los costos.
4. Los conglomerados deben escogerse de manera que se minimice el incremento en el error de muestreo debido al agrupamiento.

Sin embargo, hay dos razones principales para la aplicación del *MC*, aunque la primera intención sea la de usar los elementos como unidades de muestreo, se ha encontrado que para muchas encuestas no se tiene una lista confiable de los elementos de la población y la segunda, el costo de desplazamiento entre conglomerados.

[17] Citado por Ospina B.D. 2001.

Los conglomerados no tienen que definirse idénticamente para toda la población. En la mayoría de las aplicaciones ellos no son iguales a menos que se definan conglomerados de igual tamaño.

Cuando los conglomerados son de igual tamaño, la selección generalmente se hace haciendo uso del *MAS* sin reemplazo aplicando la teoría de este diseño. Pero, cuando los tamaños de los conglomerados son diferentes, existen varias alternativas, de las cuales, las más comunes son la selección de los conglomerados mediante el *MAS* sin reemplazo o con probabilidad proporcional al tamaño (*ppt*).

En la gran mayoría de las investigaciones que hacen uso del muestreo por conglomerados, la población sigue estando compuesta por N conglomerados, pero ellos generalmente tienen diferente tamaño. El conglomerado i consta de M_i unidades elementales. Por lo tanto, el total de unidades elementales de la población es $M_0 = \sum_{i=1}^{N} M_i$.

Como los tamaños de los conglomerados son diferentes, primero se estima el total poblacional y luego se divide por el total de unidades elementales en la población para obtener la estimación de la media poblacional. Simbólicamente se escriben así

$Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ Es el total poblacional, donde $Y_i = \sum_{j=1}^{M_i} y_{ij}$ es el total de la variable y para el i-ésimo conglomerado; y_{ij} es el valor de la j-ésima unidad en el i-ésimo conglomerado. $\bar{Y} = \frac{Y}{N}$ Es la media poblacional por conglomerado. $\bar{Y} = \frac{Y}{M_0}$ Es la media poblacional por unidad, cuyos estimadores insesgados en su respectivo orden vienen dados por

$$\hat{Y}_{con} = \frac{N}{n} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \; ; \; \bar{y}_{con} = \frac{\hat{Y}_{con}}{N} \; y \; \bar{\bar{y}} = \frac{\hat{Y}}{M_0} = \frac{N}{nM_0} \sum_{i=1}^{n} Y_i$$

La fórmula para el tamaño de muestra (número de conglomerados a seleccionar), se obtiene de la varianza de la media de medias, esto es $V(\bar{\bar{y}}) = \frac{N(N-n)}{nM_0^2} S_{con}^2$, de la cual se deriva $n = \frac{\frac{z^2 N^2 S_{con}^{*2}}{\varepsilon^2 M_0^2}}{1+\frac{z^2 N^2 S_{con}^{*2}}{\varepsilon^2 M_0^2}}$, siendo S_{con}^{*2} la estimación preliminar de S_{con}^2.

El procedimiento de selección con *ppt* es un caso particular de otro más general donde los conglomerados son seleccionados con una probabilidad establecida de antemano y con reemplazo. Cuando el muestreo se realiza de esta manera puede asimilarse a un proceso que genera una distribución de probabilidad multinomial, la cual se utiliza para derivar las propiedades de los estimadores.

Para este procedimiento de estimación, considérese el caso general donde p_i es igual a la probabilidad de seleccionar el conglomerado i en cualquier extracción. Por lo tanto,

$\hat{Y}_{p_i} = \frac{Y_i}{p_i}$ es un estimador insesgado del total poblacional.

Al seleccionar una muestra con reemplazo de tamaño n, se tendrá n estimadores independientes de Y. Por tanto, el promedio de estas estimaciones $\hat{Y}_P = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{p_i}$ es también un estimador insesgado para Y. Al igual que el estimador para la media

$$\hat{Y}_{con} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij} = \frac{N}{n}\sum_{i=1}^{n} Y_i \; ; \; \bar{y}_{con} = \frac{\hat{Y}_{con}}{N} \; y \; \bar{\bar{y}} = \frac{\hat{Y}}{M_0} = \frac{N}{nM_0}\sum_{i=1}^{n} Y_i$$

Basado en la teoría multinomial, la varianza del estimador del total es:

$V(\hat{Y}_P) = \frac{1}{n}\sum_{i=1}^{n} p_i \left(\frac{Y_i}{p_i} - Y\right)^2$ y la del estimador de la media por unidad es $V(\bar{\bar{y}}_p) = \frac{V(\hat{Y}_p)}{M_0^2}$

Como los valores de los p_i pueden ser teóricamente cualquier probabilidad, es conveniente seleccionarlos apropiadamente. Cuando se seleccionan conglomerados de unidades, los tamaños de estos conglomerados, *Mi*, están, a menudo, relacionados con sus totales Y_i. Esto

lleva a definir las probabilidades proporcionales a los tamaños de los conglomerados, de la siguiente forma $p_i = \frac{M_i}{M_0}$.

Para el proceso de selección, el primer paso consiste en obtener los tamaños de los conglomerados, M_i, de información obtenida previamente. Seguidamente, las M_i se van acumulando para cada uno de los conglomerados. Una vez se tiene esto, se aplica un rango asociado de números aleatorios a cada conglomerado y se seleccionan n números aleatorios entre uno y M_o que sería el límite superior para el rango correspondiente al último conglomerado. Al ser el muestreo con reemplazo, es posible que uno o más de los conglomerados puedan ser seleccionados.

Los estimadores para el total y la media, así como sus varianzas, se derivan de las fórmulas ya presentadas para el caso general.

Para el total:

$\hat{Y}_p = \hat{Y}_{ppt} = \frac{M_0}{n}\sum_{i=1}^{n}\frac{Y_i}{M_i} = \frac{M_0}{n}\sum_{i=1}^{n}\bar{Y}_i$, su varianza es

$V(\hat{Y}_p) = V(\hat{Y}_{ppt}) = \frac{M_0}{n}\sum_{i=1}^{n} M_i(\bar{Y}_i - \bar{\bar{Y}})^2$ y $v(\hat{Y}_{ppt}) = \frac{M_0^2}{n}\sum_{i=1}^{n}\frac{(\bar{Y}_i - \bar{\bar{Y}})^2}{n-1}$, donde $\bar{Y}_i$ es la media del conglomerado i por unidad.

Para la media:

$$\bar{\bar{y}}_{ppt} = \frac{\hat{Y}_{ppt}}{M_0} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$$

Su varianza es

$$V(\bar{\bar{y}}) = V(\bar{\bar{y}}_{ppt}) = \frac{1}{nM_0}\sum_{i=1}^{n} M_i(\bar{Y}_i - \bar{\bar{Y}})^2 \; y \; v(\bar{\bar{y}}_{ppt}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\bar{Y}_i - \bar{\bar{Y}}_{ppt})^2}{n-1}$$

Debido al supuesto implícito de trabajar con una población infinita, el *cpf* no juega papel alguno en la fórmula de la varianza. Jaeger (1982) propone la siguiente formula $n = \frac{z^2}{\varepsilon^2}\sum_{i=1}^{n}\frac{M_i}{M_0}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$, donde todos los términos son conocidos.

Para fines prácticos, si se desconoce$\sum_{i=1}^{N}\frac{M_i}{M_0}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$, esta cantidad puede estimarse a partir de una muestra piloto con $\sum_{i=1}^{n'}\frac{\left(\bar{Y}_i - \bar{\bar{Y}}_{ppt}\right)^2}{(n'-1)}$, donde n' es el tamaño de la muestra preliminar. En general, el tamaño de muestra resulta mayor de lo presupuestado debido principalmente a la no utilización del factor de corrección. Esto es más frecuente en aquellos casos donde la población consta de un número no muy grande de conglomerados (<100).

En caso de seleccionar un *MAS* de conglomerados de tamaño diferente donde se desea estimar la proporción *P* de elementos con un atributo especifico, el estimador a considerar es una razón, ya que tanto el tamaño de los conglomerados como el número de elementos dentro de ellos que poseen el atributo específico, son variables. El estimador de la proporción y la varianza se definen respectivamente como $P_{con} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i}$, donde A_i es el número de elementos en el *i*-ésimo conglomerado con el atributo deseado y M_i es el tamaño del conglomerado, y $V(P_{con}) \approx \frac{N-n}{Nn\bar{M}^2}\sum_{i=1}^{N}\frac{(A_i - PM_i)^2}{N-1}, donde\ \bar{M} = \sum_{i=1}^{N}\frac{M_i}{N}$ cuyos estimadores se calculan reemplazando los parámetros poblacionales por sus respectivos valores estimados.

Recordemos que los estratos se definían como grupos más o menos homogéneos en cuanto a su composición interna, en cambio en los conglomerados se espera que la composición interna sea lo más heterogénea posible, de tal forma que cada conglomerado represente en lo posible a la población.

Para calcular el tamaño de muestra en un *MC* tenemos:

Para estimar una media μ

$$n = \frac{N\sigma_c^2}{\frac{NB^2M^2}{Z_{\alpha/2}^2} + \sigma_c^2}, donde\ se\ estima\ \sigma_c^2\ a\ partir\ de\ S_c^2\ igual\ a\ S_c^2 = \sum_{i=1}^{n}\frac{(y - \bar{Y}m_i)^2}{n-1}$$

Para estimar una proporción p,

$$n = \frac{N\sigma_c^2}{\frac{NB^2\bar{M}^2}{Z_{\alpha/2}^2} + \sigma_c^2}, donde\ \sigma_c^2\ se\ estima\ por\ S_c^2\ igual\ a\ S_c^2 = \sum_{i=1}^{n} \frac{(a_i - \hat{p}m_i)^2}{n-1}$$

Donde a_i es igual al número total de elementos en el conglomerado i.

Para estimar un total τ,

$$n = \frac{N\sigma_c^2}{\frac{B^2}{Z_{\alpha/2}^2} + \sigma_c^2}\ donde\ \sigma_c^2\ es\ estimada\ por\ S_c^2$$

Para un tamaño de muestra dado, una unidad de muestreo pequeña suele ser más precisa que una unidad grande. Al cotejar costo contra precisión, la unidad mayor puede resultar, por consideraciones económicas, más conveniente. Se puede seleccionar racionalmente entre los dos tipos o tamaños de unidades mediante el conocido principio de elegir la unidad que da la varianza más pequeña para un costo dado, o el menor costo para una varianza prefijada como se mencionó anteriormente.

Capítulo 4

En este capítulo se presenta el diseño de muestreo para las cuatro empresas gaseras de Colombia que participan en el estudio para un diseño de muestreo, se presenta la consolidación de la base de datos final reunida de las bases presentadas por cada empresa participante y el análisis de las variables objetivo de estudio mediante pruebas estadísticas, finalmente, se presenta el tamaño de la muestra necesario de medidores de gas domiciliar.

4. Diseño de muestreo

En la mayoría de los trabajos que requieren de la estadística, es necesario y conveniente realizar un diseño de muestreo. Este diseño tendrá su importancia en la selección de la muestra que él genere y en los resultados obtenidos. En particular, este trabajo requiere del muestreo, motivo por el cual, a continuación se presenta el diseño e implementación de un diseño de muestreo que permitirá seleccionar adecuadamente una muestra representativa y garantizar la validez obtenida a partir de los resultados que se generen de ella.

El diseño muestral incluye tanto el plan de muestreo como los procedimientos de estimación. El plan de muestreo es la metodología utilizada para seleccionar la muestra de la población. Los procedimientos de estimación son los algoritmos o fórmulas usadas para obtener estimaciones de valores poblacionales y su confiabilidad a partir de los datos muestrales. La selección de un diseño muestral particular debe tener en cuenta que variables van a medirse, que estimaciones se requieren, que niveles de confiabilidad y validez se necesitan y cuáles son las restricciones en cuanto a los recursos existentes.

Luego de que se aceptara, por parte de la Universidad del Cauca y de las Empresas prestadoras del Servicio de Gas Domiciliar y la productora de los medidores, el compromiso para la realización de un estudio consistente en determinar la probabilidad de que un medidor de gas domicilia Tipo G1.6, que se encuentre instalado en una vivienda, deje de trabajar bajo condiciones de especificación, considerando el tiempo de instalación del medidor, la altura sobre el nivel del mar a la que se encuentre y el estrato socioeconómico, se dio inicio al proceso de determinación del tipo de estudio y diseño del muestreo.

A partir de la información suministrada por las empresas Gaseras, en los archivos denominados Clientes y Lecturas, se logró la concatenación de información para luego proceder a una complementación con información relacionada a la Altura Sobre el Nivel del Mar (ASNM) para cada uno de los municipios reportados y cálculo del tiempo, en años, de funcionamiento del medidor. Posteriormente se procedió a la depuración de información inconsistente, así como a la eliminación de los registros de medidores que tuvieran menos de un año de funcionamiento, de tal forma que finalmente se obtuvo una población objetivo de 3.345.701 medidores.

Para abordar esta metodología en el desarrollo de este trabajo, se consideró la población objeto de estudio de medidores de gas domiciliar en Colombia, instalados desde los años 1931 y 2017. El listado de dichos trabajos, fue construido basado en los registros de 4 compañías gaseras que son: Empresa 1, Empresa 2, Empresa 3 y Empresa 4. Donde unifico toda la información en una misma base de datos puesto cada empresa dio la información por cuenta propia en bases distintas y cada una de estas tenía información adicional que en otras no la brindaba como por ejemplo dirección de los usuarios. Finalmente para obtener esta base de datos final se organizaron en las siguientes variables de estudio:

- Empresa (Empresas gaseras de Colombia)
- Numero Medidor (Numero de registro de medidor)
- Municipio (Municipio donde se encuentras los medidores de gas domiciliar)
- ASNM (Altura sobre el nivel del mar)
- Fecha de Instalación (Fecha en la cual se instaló el medidor)
- Estrato (Estrato socioeconómico en Colombia)
- Consumo Promedio (Consumo mensual promedio que marca cada medidor)
- Consumo (Ultimo registro de consumo de cada medidor)
- Tiempo en años (Tiempo de instalación de cada medidor)
- ASNM2 (Altura sobre el nivel del mar estratificada)
- GRUPOSA (Tiempo de instalación estratificada)
- GRUPOFINAL (Cadena final estratificada de las variables de estudio)

Para la selección de la muestra de la población se implementó un muestreo aleatorio estratificado, considerando que:

La población de los medidores de gas instalados en Colombia se encuentra clasificadas mediante una cadena de tres números, el primer número es el estrato socioeconómico que va de 1 a 6, el segundo numero altura sobre el nivel del mar (ASNM2) que su rango 1 para los medidores que están menos de 500 msnm, 2 para los que están de 500 a 1000 msnm, 3 de 1000 a 2600 msnm y 4 para más de 2600 msnm, finalmente el tercer numero tiempo en años (GRUPOSA) que tiene un rango 1 para los medidores menos de 8 años, 2 para los medidores entre 8 y 16 años, 3 entre 16 y 24 años y 4 para los mayores a 24 años, un ejemplo es la cadena 234 lo cual nos dice que en ese punto se encuentran los medidores que están en el estrato socioeconómico 2, está ubicado en una altura entre 1000 y 2600 msnm con tiempo de instalación más de 24 años. Así, estas cadenas pueden ser consideradas como estratos para nuestro objetivo de muestreo (GRUPOFINAL).

Los estratos establecidos presentaron distintos tamaños, motivo por el cual, el tamaño de la muestra, no se distribuyó de acuerdo al tamaño de cada estrato (N_h), por ello utilizando la afijación uniforme y está, dada por $n_h = \frac{n}{L}$, donde n_h es el número de unidades en la muestra para estimar una media.

El nivel de confianza α con que se trabajó fue del 90%, confiabilidad que está garantizada por $z = 1.645$, el cuantíl de la distribución normal estándar.

El sesgo B para la estimación de que se acordó fue de $200.000\ m^3$ máximo admisible en la estimación.

Para estimar el tamaño de la muestra que se va a estimar a partir de un muestreo aleatorio estratificado en la población debido a que la característica buscada, determina la concepción de muestreo que los investigadores tienen, y por ello se utiliza la fórmula de afijación uniforme para estimar la media. Por lo tanto, la fórmula que relaciona n con el grado de precisión deseado fue:

$$n = \frac{\sum_{i=1}^{k} N_i^2 \sigma_i^2 / W_i}{\frac{N^2 B^2}{Z_{\alpha/2}^2} + \sum_{i=2}^{k} N_i \sigma_i^2}$$

Donde B es el sesgo.

Quitando los estratos que en los cuales no se puede obtener una muestra por tener muy pocos medidores en ellos o porque no tienen ninguno, se obtienen 86 estratos válidos, teniendo una población final de 3.345.694 medidores de gas. Así, arrojando un resultado de $n = 536{,}1197 \approx 536$. Dado que se trabaja por afijación uniforme se quiere obtener el número de medidores aproximado para cada estrato

$$n_h = \frac{n}{L} = \frac{536}{86} = 6.38$$

De esta manera se ajusta para que de cada estrato se tomen 6 medidores, por lo tanto, se ajusta el sesgo para la estimación B de $200.000\, m^3$ a un sesgo de $203.870\, m^3$ máximo admisible en la estimación.

Arrojando un resultado de $n = 515{,}962 \approx 516$. De esta manera se estableció el tamaño de muestra.

Luego, se procedió a seleccionar los medidores de gas de Colombia que conformarían la muestra. Para esto se aprovechó el hecho de que los medidores se encuentran ordenados según su municipio, Se generó a través del software IBM SPSS STATISTICS 19.

4.1 Resultados

Para determinar los estratos en los cuales se debían separar los medidores, se procedió a identificar las variables relevantes para el estudio, cuya información se encontrara en los archivos o pudiera ser generada a partir de ellos; es así como se señalan las variables ESTRATO, ASNM2 y GRUPOSA.

Para la variable ESTRATO se realizó un análisis de varianza para comparar los consumos de gas, obteniéndose los siguientes resultados:

Tabla 1.
Análisis de la Varianza para la Variable Consumo por estrato

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	7,140E17	5	1,428E17	14582,441	,000
Intra-grupos	3,276E19	3345695	9,793E12		
Total	3,348E19	3345700			

A partir de la tabla se puede observar que se rechaza la hipótesis de que todos los estratos consumen en promedio la misma cantidad de gas. Se hicieron las pruebas de normalidad (Contraste de Kolmogorov-Smirnov) y se llegó a concluir que se rechaza la hipótesis de que los datos sean normales, pero, el hecho de que los residuos no se aproximen a una distribución normal en general no afecta de forma importante al estadístico F, que está basado en el supuesto de normalidad. La razón se debe a que, como están comparando medias, puede ser válida la aplicación del teorema central del límite a datos procedentes de una distribución no normal.[18]

A continuación, se acude a realizar las pruebas Post Anova de Duncan, con las que se obtiene el siguiente resultado:

[18] Tomado de "Análisis de Varianza Universidad Autónoma de Madrid Santiago de la Fuente Fernández"

Tabla 2.

Prueba Post Anova para el Consumo por Estrato

Duncan[a,b]

ESTRATO	N	Subconjunto para alfa = 0.05					
		1	2	3	4	5	6
1	857296	863301,51					
2	1237091		1323902,20				
3	780655			1860216,36			
5	140960				2091942,85		
6	64621					2166110,43	
4	265078						2266542,33
Sig.		1,000	1,000	1,000	1,000	1,000	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 202721,104.

b. Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

Esta tabla, además de permitirnos observar el total de medidores por estrato, señalando que el estrato dos es el que cuenta con la mayor cantidad, muestra que todos ellos conservan diferencias significativas en sus promedios de consumo, que el estrato 1 es el que tiene menor promedio y que los estratos 4, 5 y 6 son los que más consumen, por lo que se procede a utilizar cada uno de ellos como un estrato de clasificación de los medidores en la población.

Para la variable ASNM2 se procedió a realizar un análisis descriptivo de la misma, a partir del cual se establecieron los siguientes rangos 1) menos de 500 msnm; 2) de 500 a 1000 msnm; 3) de 1000 a 2600 msnm y 4) más de 2600 msnm; con ellos se procedió a realizar el análisis de varianza y las pruebas Post Anova, obteniéndose los siguientes resultados:

Tabla 3.

Análisis de la Varianza para la Variable Consumo por ASNM2

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1,346E18	3	4,486E17	46712,473	,000
Intra-grupos	3,213E19	3345697	9,604E12		
Total	3,348E19	3345700			

En la tabla se observa que se rechaza la hipótesis de que los consumos de gas, por ASNM2, sean iguales.

Tabla 4.
Prueba Post Anova para el Consumo por ASNM2

Duncan[a,b]

ASNM2	N	Subconjunto para alfa = 0.05			
		1	2	3	4
2	268503	194178,58			
1	1228026		1041410,16		
3	1452743			1692984,68	
4	396429				2712361,36
Sig.		1,000	1,000	1,000	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 516159,604.

b. Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

De forma similar a lo acontecido con la variable ESTRATO, se puede observar que las categorías para la variable ASNM2 presentan diferencias significativas y además hay más consumo en el grupo 1 que en el grupo 2.

Finalmente, para la variable GRUPOSA, se procede en forma análoga, realizando un análisis descriptivo, determinando las siguientes categorías 1) menos de 8 años; 2) 8 a 16 años; 3) 16 a 24 años y 4) más de 24 años. Los resultados de las tablas de análisis de varianza y pruebas Post Anovas se presentan a continuación.

Tabla 5.
Análisis de la Varianza para la Variable Consumo por GRUPOSA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	4,065E18	3	1,355E18	154131,411	,000
Intra-grupos	2,941E19	3345697	8,791E12		
Total	3,348E19	3345700			

La información de la tabla 5 lleva a la conclusión de rechazar la hipótesis de que los promedios de consumo de gas son iguales en los grupos de tiempo de funcionamiento del

medidor, por lo que se procede a calcular la prueba Post Anova correspondiente (Como se muestra en la tabla 6).

Tabla 6.

Prueba Post Anova para el Consumo por GRUPOSA

Duncan[a,b]

GRUPOS A	N	Subconjunto para alfa = 0.05			
		1	2	3	4
4	117145	14971,35			
3	474603		159466,48		
2	1205925			595465,68	
1	1548028				2629295,20
Sig.		1,000	1,000	1,000	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa el tamaño muestral de la media armónica = 330069,074.

b. Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

Con la información anterior, se procede a determinar los grupos finales a considerar en el diseño del muestreo, los cuales se forman a partir del producto cartesiano de las categorías establecidas en las tres variables del análisis. En consecuencia, se consideran 6 grupos socioeconómicos, 4 grupos de altura sobre el nivel del mar, 4 grupos de tiempo de uso del medidor en años para generar 6*4*4= 96 clases o estratos en los que se clasificarán los 3.345.701 medidores de la base de datos.

Para tener seguridad de que los 96 estratos mantienen las diferencias significativas que requiere el diseño del estudio, se realiza la prueba de análisis de varianza y la distribución de los medidores por grupos, obteniéndose los siguientes resultados:

Tabla 7.

Análisis de la Varianza para la Variable Consumo mes por Grupos Finales

	Suma de cuadrados	Gl	Media cuadrática	F	Sig.
Inter-grupos	6,959E18	86	7,999E16	10091,195	,000
Intra-grupos	2,652E19	3345613	7,927E12		
Total	3,348E19	3345700			

En la tabla 7 se tiene las sumas de cuadrados intergrupales e intragrupales en la primera columna, en la segunda columna los grados de libertad para los dos grupos, en la tercera columna la media cuadrática de estos grupos, para obtener el resultado estadístico F demasiado grande, esto implica que se rechaza la hipótesis de que los grupos presentan medias iguales. En la última columna si tiene el valor de probabilidad el cual es más pequeños que el nivel de significancia por ende se llega a la misma conclusión de rechazo de la hipótesis.

En la tabla 8 la información esta segmentada por altura sobre el nivel del mar, es una malla de todos los estratos y en cada una están los medidores disponibles. Se tiene que los 96 grupos finales que se presentan, se observar que 10 estratos tienen una cantidad insuficiente de medidores para tomar la muestra deseada los cuales se pueden ver marcados en los recuadros amarillos, por lo tanto, se trabajará con un total de 86 estratos.

Tabla 8.

Numero de medidores de gas, según los Grupos Finales

ASNM2			ESTRATO 1	2	3	4	5	6	Total
Menos de 500	GRUPOS A	Menos de 8 años	300124	184300	79327	30955	3161	2361	600228
		8 a 16 años	157794	125545	55223	10486	1923	3463	354434
		16 a 24 años	67428	66173	18404	4024	2028	1116	159173
		Más de 24 años	21855	50136	26002	8354	3665	4179	114191
	Total		547201	426154	178956	53819	10777	11119	1228026
500 a 1000	GRUPOS A	Menos de 8 años	34199	59332	13149	5616	1513	262	114071
		8 a 16 años	25873	49384	18893	3849	1373	124	99496
		16 a 24 años	6709	28289	16185	2016	850	110	54159
		Más de 24 años	258	403	111	5	0	0	777
	Total		67039	137408	48338	11486	3736	496	268503
1000 a 2600	GRUPOS A	Menos de 8 años	101066	206579	147283	64030	32919	15785	567662
		8 a 16 años	77618	243485	184209	64953	46752	22053	639070
		16 a 24 años	5398	65487	106429	29068	25072	12382	243836
		Más de 24 años	0	378	1286	53	77	381	2175
	Total		184082	515929	439207	158104	104820	50601	1452743
Más de 2600	GRUPOS A	Menos de 8 años	49339	111931	68285	24379	10469	1664	266067
		8 a 16 años	8782	41278	39735	13844	8777	509	112925
		16 a 24 años	853	4389	6134	3446	2381	232	17435
		Más de 24 años	0	2	0	0	0	0	2
	Total		58974	157600	114154	41669	21627	2405	396429
Total	GRUPOS A	Menos de 8 años	484728	562142	308044	124980	48062	20072	1548028
		8 a 16 años	270067	459692	298060	93132	58825	26149	1205925
		16 a 24 años	80388	164338	147152	38554	30331	13840	474603
		Más de 24 años	22113	50919	27399	8412	3742	4560	117145
	Total		857296	1237091	780655	265078	140960	64621	3345701

Es importante resaltar en este momento que un análisis semejante al que se acaba de presentar, considerando como variable independiente el consumo mensual, también se realizó para la variable lectura, que corresponde al funcionamiento en el tiempo del medidor, obteniéndose resultados similares, por lo que no se hizo necesaria su inclusión en este informe. También se realizaron otros análisis gráficos sobre los histogramas de las variables.

Cada uno de los grupos finales se constituye por la combinación de las clases de las tres variables determinadas como relevantes para el estudio del funcionamiento de los medidores, de tal forma que si asumimos la numeraciones de las variables del 1 al 6 para el ESTRATO, y del 1 al 4 para ASNM2 (Altura sobre el nivel del mar) y GRUPOSA (Tiempo de Antigüedad), entonces los grupos van en combinaciones del 111 hasta el 644, excluyendo los 10 grupos antes señalados (424,524,624,134,144,244,344,444,544,644).

Las estadísticas descriptivas de las lecturas de los medidores para estos 86 grupos finales se presentan a continuación en la tabla 9.

Tabla 9.
Estadística descriptiva para la variable Consumo por Grupos Finales

GRUPOS FINAL	N	Mínimo	Máximo	Media	Desv. típ.	CV
111	300124	1	11876081	1070083,20	2875417,361	2,687097
112	157794	1	11633445	64591,72	387822,102	6,00420769
113	67428	1	848695	4307,58	15896,789	3,69042223
114	21855	1	99926	4919,47	5378,336	1,0932755
121	34199	1	11873723	347810,48	1702551,742	4,8950559
122	25873	2	33206	2019,97	1472,273	0,72885884
123	6709	9	29204	3506,99	2389,317	0,68130134
124	258	27	9995	6214,17	2908,655	0,46806814
131	101066	1	11876088	2435810,33	4171455,829	1,71255363
132	77618	1	11562512	200333,68	805309,568	4,01984114
133	5398	3	11145101	18005,24	194572,204	10,806421
141	49339	1	11887367	2710854,74	4150394,288	1,53102792
142	8782	5	11403088	90473,28	620399,893	6,85727204
143	853	32	2365013	17899,38	124772,444	6,97076904
211	184300	1	11884871	2044740,68	3708131,492	1,8134972
212	125545	1	11705309	241163,24	727957,125	3,0185244
213	66173	1	3905372	6840,00	50694,869	7,41153056
214	50136	1	99086	4963,22	4510,100	0,90870443
221	59332	1	11847997	469880,37	1860379,968	3,95926301
222	49384	6	44709	1808,25	1558,529	0,86189907
223	28289	5	38368	3077,88	2419,869	0,78621291
224	403	8	33482	6902,98	3456,161	0,50067666
231	206579	1	11890823	3040997,33	4269918,535	1,40411782
232	243485	2	11859233	627087,18	1171900,131	1,86879938
233	65487	2	11655483	93505,81	356376,835	3,81128012
234	378	159969	11637003	1473997,47	941406,494	0,63867579
241	111931	1	11900563	3422051,86	4306838,867	1,25855453
242	41278	1	11633051	674327,30	1552143,811	2,30176624
243	4389	1	11636211	789185,61	844381,949	1,06994088
311	79327	1	11917628	4024427,31	4905153,710	1,63319353
312	55223	1	11840182	845431,74	1380753,647	4,60411393
313	18404	1	11470623	49479,51	227809,301	1,44060738
314	26002	1	807500	4770,66	6872,648	2,75180415
321	13149	1	11550903	899933,38	2476440,407	1,00941978
322	18893	4	28792	1561,82	1576,532	0,89117665
323	16185	20	36281	2649,22	2360,923	0,55220387
324	111	13	31098	7337,77	4051,945	1,19111238

331	147283	1	11882679	3913308,25	4661189,917	1,49263218
332	184209	1	11905256	878422,19	1311161,228	2,54205259
333	106429	1	11625369	176148,97	447779,946	7,11943693
334	1286	49	2352494	15843,50	112796,799	1,15439253
341	68285	1	11923726	3895749,57	4497224,220	1,72487225
342	39735	1	11699590	1123438,91	1937788,603	0,75629349
343	6134	25	11662952	930420,12	703670,681	1,04839158
411	30955	1	11871733	5206943,64	5458915,880	1,65097371
412	10486	1	11771669	1032712,84	1704981,748	3,94153245
413	4024	1	923633	3890,62	15335,005	1,0634587
414	8354	1	99909	4760,23	5062,308	18,8145832
421	5616	2	11834835	24088,20	453209,443	0,94498856
422	3849	1	21566	1376,95	1301,202	0,9924217
423	2016	32	25612	2368,87	2350,918	1,30092557
431	64030	1	11869979	3640173,02	4735594,168	1,45532499
432	64953	6	11862021	1167955,72	1676027,460	0,15124813
433	29068	8	11896228	398286,87	579636,836	1,40976352
434	53	680467	2358858	1227076,11	185592,962	1,46754142
441	24379	1	11873028	3544002,39	4996205,297	0,5828685
442	13844	4	10966402	1318937,89	1935595,978	14,4371804
443	3446	1195	4462627	973932,57	567674,621	8,35666864
511	3161	1	11491045	44767,79	646320,661	1,98782195
512	1923	1	3856467	38075,08	318180,827	1,16299845
513	2028	1	93083	3989,39	7930,197	16,1414871
514	3665	1	98750	4804,61	5587,754	0,9010433
521	1513	2	11658073	33486,17	540516,582	0,83237059
522	1373	17	13061	1469,38	1323,975	1,20685843
523	850	35	20124	2513,64	2092,280	1,40006625
531	32919	1	11925842	3972485,75	4794227,901	1,1318987
532	46752	12	11886758	1227592,25	1718710,474	1,01332301
533	25072	2	11374232	506978,87	573848,723	0,7839695
534	77	241	1071141	523213,66	530184,440	0,7382576
541	10469	1	11869951	6562861,33	5145083,099	0,93307249
542	8777	8	11056501	2655481,97	1960429,755	11,6617739
543	2381	689	3922891	696070,07	649483,830	4,70752335
611	2361	1	955283	1771,33	20656,850	1,05165854
612	3463	1	339177	2086,04	9820,082	1,19665825
613	1116	17	79797	3412,04	3588,301	1,34496517
614	4179	1	99698	4358,80	5215,994	0,90934924
621	262	16	8370	512,53	689,335	0,95113641

622	124	32	13558	1904,54	1731,892	0,99311825
623	110	33	17622	2448,51	2328,867	0,96991361
631	15785	2	11896260	5127903,37	5092614,407	0,79383093
632	22053	18	11921291	1815689,29	1761061,748	0,37321275
633	12382	33	11316835	818570,20	649806,346	1,35653488
634	381	503422	10623825	1337857,53	499305,482	0,71715158
641	1664	1	11852016	3846233,26	5217549,588	0,56871875
642	509	16	6750458	3264521,13	2341156,483	2,687097
643	232	6877	3381740	1076589,44	612276,602	6,00420769

Para realizar la selección de los medidores en los 86 estratos válidos, se va a segmentar el número de medidores por la variable EMPRESA. Así, se tiene la posibilidad de observar donde se va a tomar los medidores necesarios de cada una de las empresas. Esta información se encuentra en la siguiente tabla.

Tabla 10.

Numero de medidores de gas, según los Grupos Finales segmentados por Empresa

EMPRESA	ASNM2			ESTRATO 1	2	3	4	5	6	Total
Empresa 1	Menos de 500	GRUPOS A	Menos de 8	50295	47823	15940	7321	1252	6	122637
			8 a 16	35384	47739	12981	1954	250	1	98309
			16 a 24	5389	15514	3847	268	8	11	25037
			Más de 24	604	2979	504	217	55	7	4366
		Total		91672	114055	33272	9760	1565	25	250349
	500 a 1000	GRUPOS A	Menos de 8	13074	15550	2071	234	1	0	30930
			8 a 16	9179	8884	1247	84	0	1	19395
			16 a 24	1138	1294	150	3	0	0	2585
			Más de 24	258	403	111	5	0	0	777
		Total		23649	26131	3579	326	1	1	53687
	1000 a 2600	GRUPOS A	Menos de 8	13945	25896	13594	6412	1024	263	61134
			8 a 16	10109	20733	7852	3880	612	69	43255
			16 a 24	1266	6480	2698	228	48	4	10724
		Total		25320	53109	24144	10520	1684	336	115113
	Más de 2600	GRUPOS A	Menos de 8	31107	50434	30268	15035	3350	1026	131220
			8 a 16	3743	21875	19875	7664	1419	94	54670
			16 a 24	306	489	51	1	0	0	847
			Más de 24	0	2	0	0	0	0	2
		Total		35156	72800	50194	22700	4769	1120	186739
	Total	GRUPOS A	Menos de 8	108421	139703	61873	29002	5627	1295	345921
			8 a 16	58415	99231	41955	13582	2281	165	215629
			16 a 24	8099	23777	6746	500	56	15	39193
			Más de 24	862	3384	615	222	55	7	5145
		Total		175797	266095	111189	43306	8019	1482	605888
Empresa 2	Menos de 500	GRUPOS A	Menos de 8	30777	11043	2198	248		12	44278
			8 a 16	10269	9773	3690	284		0	24016
			16 a 24	1454	1939	802	6		0	4201
		Total		42500	22755	6690	538		12	72495
	500 a 1000	GRUPOS A	Menos de 8	19452	40073	9513	5366	1506	262	76172
			8 a 16	16108	40491	17646	3765	1373	123	79506
			16 a 24	5372	26916	16035	2013	850	110	51296
		Total		40932	107480	43194	11144	3729	495	206974
	1000 a 2600	GRUPOS A	Menos de 8	43771	51808	30934	22580	11095	5684	165872
			8 a 16	57596	99505	65565	29239	22442	7136	281483
			16 a 24	3979	51727	82834	16573	11365	2943	169421

			Más de 24	0	0	1164	0	39	0	1203
		Total		105346	203040	180497	68392	44941	15763	617979
	Más de 2600	GRUPOS A	Menos de 8	1587	1504	81	1			3173
			8 a 16	4095	2072	360	5			6532
			16 a 24	236	292	170	2			700
		Total		5918	3868	611	8			10405
	Total	GRUPOS A	Menos de 8	95587	104428	42726	28195	12601	5958	289495
			8 a 16	88068	151841	87261	33293	23815	7259	391537
			16 a 24	11041	80874	99841	18594	12215	3053	225618
			Más de 24	0	0	1164	0	39	0	1203
		Total		194696	337143	230992	80082	48670	16270	907853
Empresa 3	Menos de 500	GRUPOS A	Menos de 8	176257	70637	20804	7440	1889	2343	279370
			8 a 16	103187	39341	10155	4146	1646	3460	161935
			16 a 24	60557	48379	12527	3750	2020	1105	128338
			Más de 24	21251	47157	25498	8137	3610	4172	109825
		Total		361252	205514	68984	23473	9165	11080	679468
	500 a 1000	GRUPOS A	Menos de 8	247	19					266
			8 a 16	586	9					595
			16 a 24	199	79					278
		Total		1032	107					1139
	1000 a 2600	GRUPOS A	Menos de 8	1098	670	32				1800
		Total		1098	670	32				1800
	Más de 2600	GRUPOS A	Menos de 8	479	188	0				667
			8 a 16	540	188	5				733
			16 a 24	283	580	5				868
		Total		1302	956	10				2268
	Total	GRUPOS A	Menos de 8	178081	71514	20836	7440	1889	2343	282103
			8 a 16	104313	39538	10160	4146	1646	3460	163263
			16 a 24	61039	49038	12532	3750	2020	1105	129484
			Más de 24	21251	47157	25498	8137	3610	4172	109825
		Total		364684	207247	69026	23473	9165	11080	684675
Empresa 4	Menos de 500	GRUPOS A	Menos de 8	42795	54797	40385	15946	20	0	153943
			8 a 16	8954	28692	28397	4102	27	2	70174
			16 a 24	28	341	1228	0	0	0	1597
		Total		51777	83830	70010	20048	47	2	225714
	500 a 1000	GRUPOS A	Menos de 8	1426	3690	1565	16	6		6703
		Total		1426	3690	1565	16	6		6703
	1000 a 2600	GRUPOS A	Menos de 8	42252	128205	102723	35038	20800	9838	338856
			8 a 16	9913	123247	110792	31834	23698	14848	314332
			16 a 24	153	7280	20897	12267	13659	9435	63691

		Más de 24	0	378	122	53	38	381	972
	Total		52318	259110	234534	79192	58195	34502	717851
Más de 2600	GRUPOS A	Menos de 8	16166	59805	37936	9343	7119	638	131007
		8 a 16	404	17143	19495	6175	7358	415	50990
		16 a 24	28	3028	5908	3443	2381	232	15020
	Total		16598	79976	63339	18961	16858	1285	197017
Total	GRUPOS A	Menos de 8	102639	246497	182609	60343	27945	10476	630509
		8 a 16	19271	169082	158684	42111	31083	15265	435496
		16 a 24	209	10649	28033	15710	16040	9667	80308
		Más de 24	0	378	122	53	38	381	972
	Total		122119	426606	369448	118217	75106	35789	1147285

4.2 Selección de los medidores de la base de datos

Utilizando un procedimiento aleatorio, se seleccionan los registros de los medidores que deberán ser retirados de su domicilio para introducirlos en el estudio, aparecen para cada empresa. Se presenta secuencialmente la lista que contiene los medidores requeridos, de tal forma que se inicie retirando los primeros de ellos.

Es absolutamente imprescindible que los medidores sean recolectados como aparecen en la lista que se adjunta, debido a que un cambio en el procedimiento estaría involucrando fallas en la aleatoriedad desarrollada y por lo mismo fallas en los resultados del estudio, por lo que se solicita mucha responsabilidad e idoneidad en la realización de esta parte de la investigación.

De esta forma la lista de los elementos a introducir en el estudio es la siguiente:

Tabla 11.

Lista de medidores a involucrar en el estudio según Grupos Finales

		Empresa 1	Empresa 2	Empresa 3	Empresa 4	Total
1	111	0	0	4	2	6
2	112	1	0	4	1	6
3	113	2	0	4	0	6
4	114	3	0	3	0	6

5	121	0	0	4	2	6
6	122	0	0	6	0	6
7	123	0	0	6	0	6
8	124	3	3	0	0	6
9	131	0	0	4	2	6
10	132	0	0	4	2	6
11	133	0	0	4	2	6
12	134	0	0	0	0	0
13	141	0	1	3	2	6
14	142	0	0	5	1	6
15	143	0	0	6	0	6
16	144	0	0	0	0	0
17	211	1	2	2	1	6
18	212	1	1	2	2	6
19	213	2	1	1	2	6
20	214	3	0	3	0	6
21	221	3	0	0	3	6
22	222	3	3	0	0	6
23	223	3	3	0	0	6
24	224	3	3	0	0	6
25	231	2	2	1	1	6
26	232	2	2	0	2	6
27	233	2	2	0	2	6
28	234	0	0	0	6	6
29	241	1	2	2	1	6
30	242	1	1	2	2	6
31	243	2	1	1	2	6
32	244	0	0	0	0	0
33	311	2	2	1	1	6
34	312	1	2	2	1	6
35	313	1	1	2	2	6
36	314	3	0	3	0	6
37	321	3	0	0	3	6
38	322	3	3	0	0	6
39	323	3	3	0	0	6
40	324	3	3	0	0	6
41	331	2	2	0	2	6
42	332	2	2	0	2	6
43	333	2	2	0	2	6
44	334	0	3	0	3	6
45	341	2	2	0	2	6
46	342	2	2	0	2	6
47	343	2	2	0	2	6
48	344	0	0	0	0	0
49	411	2	1	1	2	6

50	412	2	2	1	1	6
51	413	0	3	3	0	6
52	414	3	0	3	0	6
53	421	6	0	0	0	6
54	422	3	3	0	0	6
55	423	0	6	0	0	6
56	424	0	0	0	0	0
57	431	2	2	0	2	6
58	432	2	2	0	2	6
59	433	2	2	0	2	6
60	434	0	0	0	6	6
61	441	3	0	0	3	6
62	442	3	0	0	3	6
63	443	0	0	0	6	6
64	444	0	0	0	0	0
65	511	3	0	3	0	6
66	512	3	0	3	0	6
67	513	0	0	6	0	6
68	514	0	0	6	0	6
69	521	0	6	0	0	6
70	522	0	6	0	0	6
71	523	0	6	0	0	6
72	524	0	0	0	0	0
73	531	2	2	0	2	6
74	532	2	2	0	2	6
75	533	2	2	0	2	6
76	534	0	3	0	3	6
77	541	3	0	0	3	6
78	542	4	0	0	2	6
79	543	0	0	0	6	6
80	544	0	0	0	0	0
81	611	0	0	6	0	6
82	612	0	0	6	0	6
83	613	0	0	6	0	6
84	614	0	0	6	0	6
85	621	0	6	0	0	6
86	622	0	6	0	0	6
87	623	0	6	0	0	6
88	624	0	0	0	0	0
89	631	2	2	0	2	6
90	632	2	2	0	2	6
91	633	0	3	0	3	6
92	634	0	0	0	6	6
93	641	6	0	0	0	6
94	642	3	0	0	3	6

95	643	0	0	0	6	6
96	644	0	0	0	0	0
Total		129	129	129	129	516

Conclusiones

Finalizado este trabajo diseño de muestreo para estimar el tiempo de funcionamiento bajo especificaciones técnicas de medidores de gas instalados en Colombia, podemos consignar las siguientes reflexiones como conclusiones:

- Es prioritario establecer el tiempo de funcionamiento, bajo especificaciones técnicas, que dura trabajando un medidor de gas domiciliar, con el fin de determinar la forma en la que se puede estar afectando el valor del consumo real de gas domiciliar.

- La partición de la población de medidores de gas domiciliar, resultó adecuada a las necesidades del estudio, debido a que los análisis de varianza mostraron diferencias significativas entre los grupos determinados, para cada una de las variables (Estrato, ASNM y Tiempo de funcionamiento del medidor).

- Dada la complejidad del fenómeno consumo de gas domiciliar, en términos de las condiciones de Estrato, ASNM y Tiempo de funcionamiento del medidor, se explica el gran número de categorías (96) que finalmente se debió establecer para general el diseño del muestreo.

- Es de destacar la forma como mediante muestreo aleatorio estratificado se logra tener el tamaño de muestra deseado para poder saber dónde se tenían que tomar los medidores y lograr asignar a cada una de las empresas involucradas, un número de medidores igual, considerando las zonas en las cuales prestan sus servicios, e incluso en las zonas comunes a dos empresas.

- Es importante considerar la participación de las empresas gaseras de Colombia para poder llevar a buen término este proyecto, ya que además de tener la información de los usuarios, también tienen acceso a su ubicación física y las condiciones logísticas para llevar acabo la recolección de la información de los usuarios.

Recomendaciones

Conscientes de la problemática abordada y de la necesidad de afrontar un cambio, tanto conceptual como curricular que apunte en la dirección de procurar mejorar las condiciones y conocimientos, partiendo de esta experiencia con la esperanza de contribuir a esta solución, recomendamos:

- Tener en cuenta que el medidor de gas tipo domiciliar se puede clasificar como derecho o izquierdo, puede ser una variable que ayude a mejorar las condiciones del diseño de muestreo, e incluso las condiciones de clasificación para determinar el tamaño de la muestra.

- Se debe procurar una homogenización en la organización de la información sobre los usuarios manejan las empresas gaseras de Colombia, puesto que se encontraron inconsistencias en datos, obligando la modificación de esquemas de trabajo previamente establecidos, lo que significó consumo de tiempo.

- Resultaría conveniente realizar seminarios-taller con el propósito de dar a conocer, a partir de los resultados encontrados, la importancia del muestreo en investigaciones de este tipo, así como de ejemplos de su aplicación.

Bibliografía

[1] Raj, D. (1979). *La estructura delas encuestas por muestreo.* México, México: Fondo de Cultura Económica.

[2] Cochran, W. (1980). *Técnicas de Muestreo.* México, México: CECSA.

[3] Lohr, S. (2000). *Muestreo: Diseño y Análisis.* México, México: International Thomson Editores.

[4] Ospina, D. (2001). *Introducción al muestreo.* Bogotá, Colombia: Universidad Nacional de Colombia.

[5] Klinger, R. A. (2011). *Muestreo Estadístico: Métodos Básicos.* Cali, Colombia: Programa Editorial Universidad del Valle.

[6] Cochran, W. (1983). *Planning and Analysis of Observational Studies.* New York, USA: John Wiley & Sons.

[7] NTC 2728. (2005). Norma Tècnica Colombiana:Medidores de gas tipo diafragma. Bogotà, Colombia: ICONTEC. 10

[8] Guerrero Suárez, F., & Llano Camacho, F. (2002). Gas Natural en Colombia - GAS e.s.p. Universidad Icesi, Departamento de Administración. Cali: Universidad Icesi.