UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

DISEÑO MUESTRAL PARA ESTIMAR VENTAS DE ARTÍCULOS COMERCIALIZADOS EN AUTOSERVICIOS

REPORTE DE TRABAJO PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

DAVID IVÁN PORRAZ PULIDO

TUTOR:

ACT. JAIME VAZQUEZ ALAMILLA

2012

AGRADECIMIENTOS

A mi madre por su ejemplo y apoyo desde siempre.

A mi padre en paz descanse, se que te hubiera gustado verlo.

A mi familia: Ana Luz, Luz Aline, María Elizabeth, Octavio, Daniel, Vianney, Danna Sofia y Juan.

A Verónica Atencio y Adriana Rivera, por todo el apoyo en mi carrera.

A mis amigos: Loyda Vergara, Diana Cortez, Diana Moran, Arturo Pedraza, Karina Pedraza, Luis y Javier Juárez, José Luis Guerra, Julián Morales y Ana Álvarez por estar siempre ahí.

A Grisel Campistrano, por su apoyo y colaboración en la recuperación de información perdida.

A Diana Cabezas por todo su impulso para iniciar este proyecto y colaboración en la colecta y captura de información.

A mi tutor Act. Jaime Vázquez Alamilla, por su invaluable apoyo y motivación para poder concluir este trabajo y llevarlo a buen término.

A mis sinodales, Dra. María del Pilar Alonso Reyes, M. en C. José Antonio Flores Díaz, M. en C. José Salvador Zamora Muñoz y Mat. Margarita Elvira Chávez Kano, por sus comentarios y aportaciones para realizar todos los ajustes que fueron necesarios así como la oportunidad con que fueron brindados

A Nielsen, por la oportunidad de desarrollo y crecimiento que me ha brindado desde hace más de 10 años.

A la Universidad Nacional Autónoma de México, por la formación académica con enfoque de trascendencia y aportación que me fue brindada desde el Colegio de Ciencias y Humanidades plantel Sur y culminó en la Facultad de Ciencias, ha sido un privilegio pisar sus aulas.

PREFACIO

Dentro de este documento podrá verse el desarrollo de una solución implementada en la empresa donde se usó y que sigue teniendo vigencia a pesar de que fue aplicada hace ya algunos años (2005). El lector podrá encontrar un ejemplo práctico de aplicación de las técnicas de muestreo que, junto con las poderosas herramientas informáticas, permiten la realización de este tipo de ejercicios confiables para la presentación de una población objetivo con particularidades como: universo pequeño, gran cantidad de variables de estudio (artículos de venta) y necesidad de mantener el panel por un lapso de tiempo prolongado.

El trabajo está distribuido de la siguiente manera:

Capítulo 1: Introducción

En este primer capítulo podrá encontrarse la justificación del trabajo en relación a la profesión del actuario así como la descripción del marco contextual en donde se desarrolla la empresa Nielsen, sus objetivos, los productos que ofrece a los Clientes y finalmente se plantea la problemática en cuestión basada en el servicio scantrack ante las limitantes de colaboración de algunas cadenas que perciben problemas de confidencialidad por el nivel y completez de la información que proporcionan.

Capítulo 2: Marco teórico

Este capítulo pretende dar una breve revisión de conceptos de muestreo involucrados en la solución planteada. Inicia en conceptos básicos de muestreo, estimadores, el importante concepto de representatividad de una muestra, afijación de muestra y por último, concluyendo con un breve pasaje por el concepto de estimadores de razón que son aplicados en la solución propuesta.

Capítulo 3: Aplicación práctica

Finalmente, en este capítulo se describen las características de la solución planteada a un ejercicio práctico y fueron realizados comparativos con otras posibles soluciones basadas en variantes de la afijación de la muestra y los niveles de estratificación, encontrando de manera práctica que la afijación seleccionada da mejores resultados que el resto, avalando así lo mostrado en el marco teórico.

1 INTRODUCCIÓN

1.1 JUSTIFICACIÓN DEL TEMA

1.1.1 Actividades desarrolladas en la institución donde se realizó el trabajo

Nielsen es la empresa donde se realizó la aplicación práctica de lo que se mostrará en este documento, en particular, dentro del área de estadística, la cual tiene la responsabilidad de garantizar precisión, veracidad y expedición del reporte a clientes. Para esto, se intenta asegurar la manipulación de información a través de procesos y metodologías estadísticas gracias a las cuales se pretende contar con un producto final confiable.

Dentro del servicio, el ejecutivo de estadística scantrack, diseña, da mantenimiento y controla la muestra de autoservicios participantes, garantizando representatividad en todos y cada uno de los territorios de reporte. Entre otras cosas, colabora en la validación y proyección de la información proporcionada por los colaboradores, con la finalidad de asegurar la consistencia de la misma. Para esto se hace uso de varias técnicas como son: series de tiempo, análisis de regresión y diversos métodos de imputación de información para tratar faltantes de información.

Nielsen es una empresa dedicada a la investigación de mercados y líder en el ramo en México. Incluye como clientes a los principales fabricantes y detallistas de productos de consumo masivo a nivel mundial. Dentro de ella son generados una gran variedad de productos dentro de los cuales sobresalen dos grandes rubros: continuos y ad-hoc.

Los primeros son toda una gama de servicios de realización periódica y miden la dinámica de mercado dentro del comercio detallista y los segundos responden a una necesidad particular de un cliente e implican un diseño personalizado para medir alguna variable de interés.

Los estudios continuos representan la gran fortaleza de la empresa y el trabajo propuesto está centrado en uno de ellos conocido como scantrack. Aproximadamente la quinta parte de los

ingresos son generados por este servicio. Es el segundo de mayor importancia en la compañía a nivel Latinoamérica en términos de rentabilidad y diversificación de subproductos. Consiste en la medición de la dinámica comercial del canal de tiendas de autoservicio de la iniciativa privada con tecnología de “scanner” en sus carriles de salida, como lo son las tiendas de cadena de Walmart, Soriana, Comercial Mexicana, etc.

La fuente de información es proveniente de las cadenas colaboradoras y consiste en datos consolidados semanalmente de los movimientos de venta realizados en cada periodo (ventas en pesos y ventas en unidades por artículo-negocio).

Semanalmente se generan reportes con la información proporcionada, esto lo hace un servicio táctico con la mayor oportunidad de entrega en relación a la mayoría de los otros productos proporcionados por la compañía. Adicionalmente, reporta información desglosada a nivel artículo y región.

1.1.2 Justificación e importancia del tema en relación con la práctica de la profesión

La investigación de mercados es una de las opciones laborales donde el actuario puede aplicar sus conocimientos. Gran parte de las técnicas y metodologías requeridas están relacionadas con la estadística, la cual es abordada con particular profundidad matemática en la formación del actuario, no obstante, cada una de las ramas del conocimiento adquiridas en esta carrera permiten desempeñarse con visión analítica y gran versatilidad en la resolución de problemas de distinta índole. En la investigación de mercados, los desafíos más frecuentes están relacionados, por una parte, con el tratamiento y manipulación de información para extrapolar e inferir el comportamiento de una variable de interés, y por otra, con la interpretación y análisis de las mismas dentro de un entorno comercial de una población específica. Dicho entendimiento, ayuda a tener una mejor perspectiva en la toma de decisiones e implementar mejores estrategias de mercadeo para incrementar ventas, tener mayor captación de clientes potenciales, mayor conocimiento de marca, mayor penetración en el mercado, mayor capacidad de negociación, mayor competitividad, vigencia y crecimiento.

El tema a desarrollar, es una aplicación de las técnicas de muestreo probabilístico, las cuales son muy útiles y relevantes en la inferencia estadística y dentro de la investigación de mercados, permitiendo conocer el comportamiento de una población objetivo sin necesidad de recabar información de la variable de estudio en todos y cada uno de los elementos de la misma. Aplicarlo adecuadamente, es una forma económica de obtener información de variables que se requieren conocer para entender el mercado, permitiendo al usuario tomar mejores decisiones en la implementación de estrategias aplicables a la población objetivo. De esta forma se justifica la relación entre el tema propuesto y la práctica de la profesión, al involucrar la implementación del muestreo probabilístico para resolver un problema de mercadeo, siendo esta técnica aplicada de manera natural por el actuario al ejercer su profesión ya que la formación estadística permite hacer uso y aprovechamiento de sus beneficios responsablemente.

1.1.3 Descripción general del trabajo

Para la empresa, una de sus mayores responsabilidades y compromisos es proporcionar información precisa de ventas y participación de mercado (“share”) con sus respectivas consideraciones, alcances y limitaciones sobre el comportamiento de los consumidores mexicanos en los distintos canales de venta dentro de los cuales un fabricante puede colocar su producto para el usuario final. Esta fuente permite principalmente a los fabricantes y detallistas la toma de mejores decisiones en cuanto a: estrategias de negociación y comercialización.

Scantrack reporta información principalmente del canal de autoservicios. Cuando nació el servicio se recibía información de todos y cada uno de los autoservicios de las cadenas participantes e implicaba invertir recursos en asegurar la calidad de la información, sin embargo, los principales colaboradores detallistas decidieron restringir el envío de información, limitándolo a un grupo de tiendas de autoservicio seleccionadas por Nielsen. Las cadenas toman la determinación por razones de confidencialidad entre sus competidores para no estar en desventaja competitiva en las negociaciones con sus principales proveedores de productos comercializados en sus establecimientos. Bajo este escenario, nace la problemática

de carencia de información, resuelta por el proyecto propuesto, determinando un diseño de muestra que permite generar estimadores eficientes, suficientes y consistentes (Mood Alexander, 1974: 288-307) para lograr inferir los volúmenes de venta de cada artículo comercializado bajo cierto margen de error y grado de certeza.

1.2 MARCO DE REFERENCIA

1.2.1 Antecedentes de The Nielsen Company

En sus orígenes, la empresa Nielsen fue establecida en Estados Unidos en 1923 por Arthur Charles Nielsen Sr., conocido como uno de los fundadores de la investigación de mercados moderna. Dentro de varias innovaciones en la mercadotecnia enfocada al consumidor y en investigación de medios, el Sr. Nielsen creó la técnica de medición de venta detallista que brinda a los clientes el primer tipo de información fidedigna y objetiva sobre su desempeño competitivo y el impacto de sus programas de mercadotecnia y ventas en ingresos y ganancias. La información da un significado práctico al concepto de participación de mercado, convirtiéndose en uno de los indicadores críticos de desempeño de los negocios.

Nielsen abrió su primera oficina internacional en el Reino Unido en 1939 y después de la Segunda Guerra Mundial, progresivamente extendió sus operaciones en Europa Occidental, Australia y Japón. A la fecha, la presencia de Nielsen se manifiesta en más de 100 países.

Nielsen en México

Nielsen se estableció en México en 1967 y desde entonces apoya a sus clientes aportando elementos para la medición del desempeño de sus negocios. La cartera de servicios que ofrece Nielsen México es la más completa en Latinoamérica y sus resultados financieros la posicionan como la empresa de investigación de mercados más importante del país.

¿Qué se hace en Nielsen?

Nielsen ofrece información de mercados a través de una gran diversidad de servicios. Generalmente se mide el desempeño de fabricantes y sus productos dentro de los principales canales de consumo que existen en México, se entregan reportes acerca del comportamiento

de variables de interés[1] comercial dentro de una región en particular siendo medidas a partir de la lectura de las mismas en una muestra.

La información para la medición se obtiene a través de diferentes fuentes, ya sea en el mejor de los casos electrónicamente o visitando regularmente a los elementos que la muestra. Adicionalmente, en Nielsen se cuenta con diversas herramientas para el procesamiento de información, sistemas de análisis, así como un equipo especializado en servicio a clientes para apoyarlos en encontrar las mejores estrategias de crecimiento partiendo de la interpretación de los resultados obtenidos que ayudan a resolver preguntas como las siguientes:

- ¿la categoría de producto está creciendo o decreciendo?
- ¿existe una contracción en el mercado?
- ¿qué sectores están creciendo o declinando?
- ¿existe algún competidor que esté impulsando la categoría?
- ¿cómo varían los niveles de precio y ofertas entre las marcas?
- ¿cuáles son mis participaciones de mercado y tendencias?
- ¿cómo se están desempeñando mis marcas con respecto a mi competencia?
- ¿cuáles son los pronósticos de largo plazo sobre los consumidores y las categorías?
- ¿se puede optar por un incremento de precios sin perder ventas?
- ¿qué tan sensitivos al precio son los consumidores?
- ¿qué estrategias de precio maximizarán las ventas?
- ¿cuáles son los puntos de sensibilidad de los precios?
- ¿cómo se desempeñan mis tiendas comparadas con el mercado?
- ¿qué otras marcas son alternativas importantes?
- ¿cómo se influencia la lealtad del consumidor con productos de marca genérica?
- ¿qué detallista representa la mayor oportunidad de venta?
- ¿qué canales alternos de compra están aprovechando los consumidores para sus marcas?
- ¿existe suficiente distribución para iniciar la campaña publicitaria?

[1] Generalmente ventas y participación de mercado de un producto(s) de consumo masivo en los principales canales de comercialización, a decir: autoservicios, farmacias, tiendas tradicionales, minisuperes, tiendas de conveniencia, etc.

- ¿es correcta la mezcla de productos en la categoría?
- ¿dónde se encuentran los faltantes y oportunidades de distribución?
- ¿qué tan efectiva es mi fuerza de ventas?
- ¿existen problemas de agotamiento?

De esta forma, la información que proporciona a sus clientes está enfocada en profundizar el conocimiento de los siguientes rubros:

- Medir su desempeño de mercado

    - Mediante la medición de ventas de productos de consumo masivo recabadas en puntos de venta de tiendas detallistas de diversos formatos y tamaños, Nielsen proporciona a los clientes información referente al desempeño de sus productos en comparación con la competencia, así como cambios en el mercado e impactos de estos en sus ingresos.

    - Provee de información relativa a actividades promocionales; precios, exhibidores especiales, niveles de distribución de inventarios, entre otros indicadores

- Analizar la dinámica del mercado

    - Cuenta con diversos paneles de monitoreo periódico, los cuales permiten a los clientes, tanto fabricantes como detallistas, dar respuesta a diversas interrogantes que se plantean día a día para alcanzar un mejor desempeño de su negocio

Productos y servicios

Para que los clientes puedan responder preguntas como las planteadas en párrafos previos y así puedan tomar decisiones, acciones de mercadotecnia e implementar estrategias de venta que les permitan incrementar sus utilidades y rentabilidad, Nielsen cuenta con una serie de servicios que en la historia de la empresa la han posicionado a nivel mundial, como líderes dentro de la investigación de mercados. No serán presentados todos los servicios que maneja[2], solamente

[2] Para mayor información, ver página WEB www.nielsen.com.mx

aquellos que tienen que ver con el presente trabajo y que pueden ampliar el contexto sobre lo que se está presentando.

- **Retail Measurement Services (RMS)**

  Los servicios de Retail Measurement brindan información sobre el desempeño de los productos, participación del mercado, distribución, precio, entre otros indicadores, que son considerados como el estándar global de la industria en México, por ello, los reportes que proporciona deben contar con altos estándares de calidad, con la precisión y oportunidad necesaria para que los clientes que los usan puedan tomar decisiones acertadas basados en información altamente confiable.

  En el mercado, donde los consumidores pueden encontrar el mismo producto a través de múltiples canales, los fabricantes cuentan a través de los servicios de Nielsen, con una cobertura muy amplia de las ventas al detalle, esencial para apoyar sus programas de mercadotecnia y ventas. Por otra parte, también apoya a los comerciantes detallistas a comprender su situación competitiva en el mercado.

  Los servicios de RMS integran información proveniente de una gran variedad de canales de consumo como son:

  - Autoservicios
  - Farmacias
  - Tiendas de conveniencia
  - Tiendas de abarrotes
  - Estanquillos
  - Licorerías
  - Mayoristas

  Dentro de los servicios RMS sobresalen principalmente Retail Index y Scantrack, dentro de este último tiene aplicación práctica el presente trabajo.

- **Retail Index**

  - **Descripción**

    Es una herramienta con valor estratégico, que ayuda a las áreas de mercadotecnia y ventas a conocer “la gran visión” del comercio del país y con ello diseñar mejor sus estrategias de negocio. A través de su análisis e interpretación es posible identificar amenazas y oportunidades en la comercialización de productos. Determina los factores claves que afectan la demanda del consumidor al correlacionar las condiciones de presencia, promoción, precios, niveles de inventarios e identificar las causas que afectan el desempeño competitivo de la marca en el corto y mediano plazo.

  - **Metodología**

    Brevemente, los datos de Retail Index se basan en la medición de las ventas al consumidor en una muestra de tiendas diseñada para inferir y extrapolar la variable de interés dentro de cada canal de distribución medido. La muestra es auditada mensualmente mediante visitas de auditores que físicamente cuentan las existencias y registran las compras de la tienda de todos los productos pertenecientes a categorías reportadas por Nielsen. La medición continua de estas variables permite obtener una aproximación de las ventas al consumidor, entre otras variables.

- **Scantrack**

  - **Descripción**

    Es un servicio continuo de entrega cada 4 semanas con cortes semanales en el interior del reporte. Aprovecha los datos de venta generados por los scanners de los autoservicios, farmacias, mayoristas y tiendas de conveniencia,

proporcionando información de ventas, precio y distribución de los diferentes productos, marcas y fabricantes participantes en las regiones de reporte.

Al provenir el insumo directamente de las cadenas, es necesario cuidar la confidencialidad de la información y para dicho efecto se garantiza que en los mercados de reporte no se mostrarán explícitamente los volúmenes de venta desplazados por una cadena en particular, siempre estarán conviviendo con otras y la importancia de una no puede ser preponderante.

Cada una de las categorías de producto reportadas cuenta con una definición perfectamente acotada permitiendo al interior identificar productos que son competidores entre sí, es decir, un producto de una categoría no compite con el de otra aunque eso no quiere decir que la decisión de compra no esté relacionada entre categorías.

Con scantrack pueden realizarse análisis para medición del desempeño de los productos y marcas de los diferentes fabricantes dentro de las tiendas de autoservicio. Con este servicio, Nielsen dice cómo ha sido la evolución de las ventas, la evolución de los precios y lo más importante, las participaciones de mercado que les permiten conocer su posicionamiento en relación a la competencia.

Este servicio se concibió como censal para la medición de ventas y muestral para la medición de actividad promocional (este tema no se aborda en este trabajo), sin embargo, en los últimos años ha cambiado a ser un servicio muestral para la medición de ventas y de ahí se deriva la necesidad de crear un diseño muestral permita que se representen correctamente las variables de ventas y participación de mercado en todas y cada una de las ciudades reportadas. De acuerdo al planteamiento de la propuesta, el objetivo está centrado en resolver este tema.

- **Canales**

  Los canales de reporte manejados son autoservicios, farmacias y mayoristas de la iniciativa privada cuyos negocios cuentan con dispositivos de scanner y que adicionalmente colaboran con Nielsen.

- **Nivel de reporte**

  La información proporcionada por Nielsen puede visualizarse hasta nivel código de barras[3] en cada uno de los mercados de reporte (regiones de interés).

- **Variables manejadas**

  Las principales variables reportadas en este servicio son:

  - Ventas: corresponden a los volúmenes en pesos, unidades de uso o convertidas desplazas en una semana por todos los negocios participantes en cada mercado de reporte.
  - Precio: corresponde promedio ponderado del mismo por unidad de uso o convertida.
  - Distribución numérica: es el porcentaje de negocios en el universo que vendieron el producto en análisis en una semana.

- **Metodología**

  La información se obtiene del registro de ventas a través de los "scanners" instalados en las cajas de salida de las tiendas.

  Las cadenas entregan para cada tienda y código de barras, las ventas en unidades de uso y las ventas en pesos resultantes del ejercicio de cada semana.

[3] Código numérico que identifica de manera única a un producto, generalmente es conformado por código de país + código de fabricante + código de marca + código de producto + dígito verificador. Por ejemplo: 7503007038500, el cual corresponde al producto "Polvo para preparar Chocolate" de la marca "Nutra Systems".

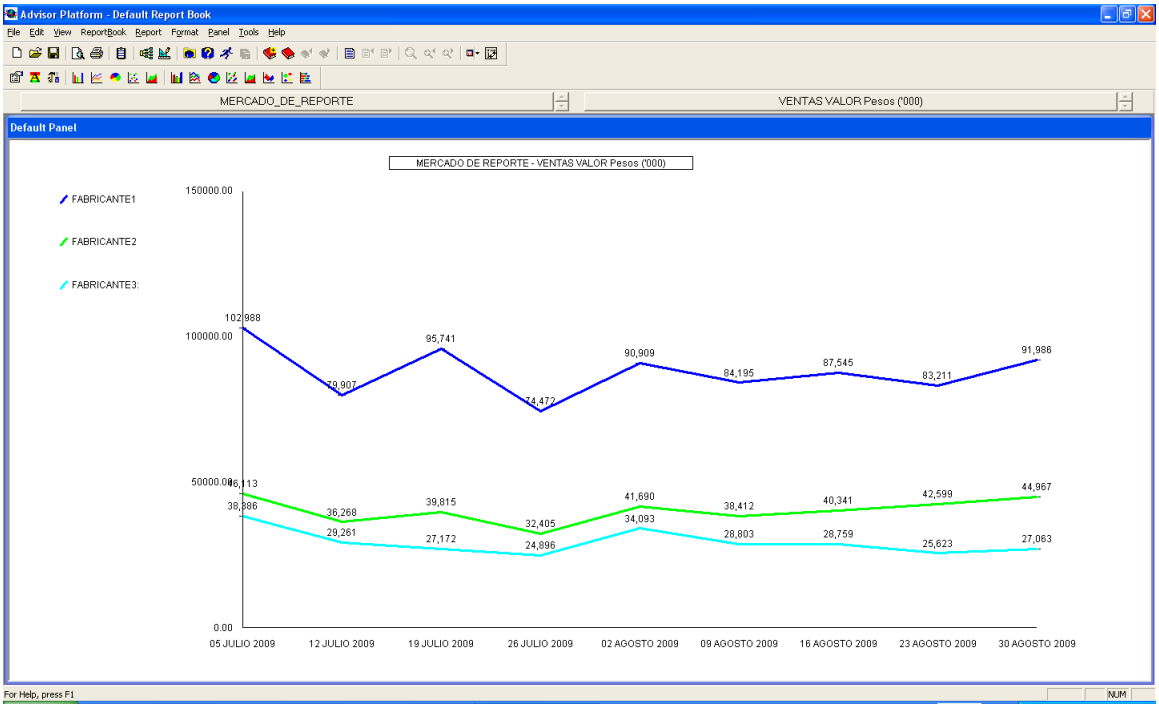
Con esta información, se deduce el precio promedio ponderado que es obtenido para fines prácticos a partir del cociente entre las ventas valor y unidad.

Posteriormente, la información es sometida a diferentes validaciones para valorar la calidad de la misma, mediante procesos automáticos se valora el precio de venta de cada uno de los artículos y en caso de encontrar anomalías es ajustado en función de la historia de precios manejada del artículo en periodos previos o en función del comportamiento observado en el mercado en la semana en cuestión. Adicionalmente, se realiza una validación de los volúmenes de ventas y número de artículos desplazados semanalmente mediante diferentes pruebas que permiten identificar si existe alguna anomalía en la información, ya sea que los colaboradores estén filtrando parte de los datos en algunos negocios, que la consolidación sea errónea o simplemente exista algún fenómeno particular en la plaza donde se desempeña el negocio que justifique su comportamiento.

Finalmente, existen una serie de procesos que se encargan de la puesta a punto de:

- estructuras que los clientes verán en la base de datos que analizarán
- identificación de artículos nuevos
- mantenimiento del listado que concentra los artículos existentes
- altas y mantenimiento de mercados de reporte en las bases de datos de cada cliente
- actualización de los universos de tiendas, añadiendo los negocios que fueron identificados como nuevos al ser proporcionados por las cadenas en su información semanal o al ser identificados como aperturas en los listados que las mismas aportan.

Ejemplo de información reportada en el servicio scantrack



(Figura 1.1)

1.2.2 Descripción de la problemática

Recordando el antecedente del servicio de acuerdo a lo descrito en secciones previas, el producto scantrack nació como un servicio de medición censal, sin embargo, varias cadenas importantes deciden dejar de compartir la información del censo al considerar que puede perjudicarles por estar descubiertas ante su competencia en plazas donde son sumamente importantes y donde el comportamiento del mercado está sujeto a las estrategias que ellos trabajan. Por esta razón, contemplan la necesidad de enviar solamente una muestra de negocios a elegir por Nielsen y aplican esta estrategia en sus envíos semanales de información. Esta muestra cuenta en los convenios de colaboración con un tope de fracción de muestreo, a decir, 60%.

Ante este contexto, se abordó este tema haciendo uso de las técnicas de muestreo.

Parte del reto era encontrar una buena muestra que permitiera garantizar niveles de precisión adecuados a nivel producto, mercado de reporte y esto se avalaría al momento de incorporar la inferencia resultante con los entregables censales previos, en principio no se debería ver rompimiento de tendencia. Así pues, al presentarse un trabajo basado en la teoría del muestreo, por ello se considera necesario describir algunas de las técnicas que condujeron a lograr el resultado esperado sin afectar a los clientes.

2 MARCO TEÓRICO

2.1 CONCEPTOS GENERALES DE MUESTREO

En Nielsen, como en cualquier empresa de investigación de mercados, el muestreo es una técnica elemental cuando se desean conocer las características de una población objetivo[4] sin necesidad de realizar una importante inversión para obtener la información deseada de todos los elementos que la componen. Es por ello que en este capítulo se comentan sobre los diferentes conceptos que lo describen, ya que el sustento de este proyecto está altamente relacionado con esta importante rama de la estadística.

En el muestreo, a través de la medición de la variable de interés en una parte de la población es posible investigar y concluir ciertos rasgos de la misma.

Dentro de los elementos primordiales que se deben plantear al realizar la medición de la variable de interés en esa fracción de la población llamada muestra, es necesario considerar lo siguiente:

- ✓ ¿Qué se quiere conocer?
- ✓ ¿Cuál es la población?
- ✓ ¿De qué forma será obtenida la información?

En el presente estudio, las respuestas a estas tres interrogantes iniciales son las siguientes:

- ✓ Se quiere conocer el volumen de ventas totales de todos y cada uno de los artículos comercializados en una cadena de tiendas que desea cambiar su esquema de colaboración de censal a muestral
- ✓ La población son todas las tiendas que conforman la cadena
- ✓ La información se obtiene a través de envío de datos electrónicos semanales que realiza la cadena detallista

[4] También conocida como Universo (N)

2.1.1 Marco de muestreo

La población debe contar con un medio físico que identifique directa o indirectamente a todos los elementos de una población. Ese medio físico se llama marco de muestreo. Puede ser un directorio, un archivo, un mapa, etc. Así, el marco es el medio físico que identifica a todos los elementos de una población.

En este estudio, el marco de muestreo es un consolidado de los catálogos de negocios proporcionados por las cadenas colaboradoras y contienen datos de ubicación geográfica como son: dirección, estado, municipio y código postal.

2.1.2 Selección de muestra

Partiendo del marco de muestreo, el cual identifica todos los elementos de la población, se deben determinar de qué forma serán considerados los de la muestra (unidades de muestreo) para hacer la medición deseada. Al respecto, es conveniente señalar algunas de las diversas formas que se tienen para tomarlos:

- **A juicio**

  Selección de la muestra necesaria con base en la experiencia y conocimiento de la población

- **Por cuotas**

  La muestra debe cumplir con algunas proporciones conocidas referentes a variables de la población. Por ejemplo, sexo y edad

- **Probabilístico**

  Los elementos son escogidos basados en una probabilidad de selección conocida y mayor a cero para todas y cada una de las unidades poblacionales.

Para resolver el problema planteado, el ejercicio realizado está desarrollado en una selección probabilística que nos permite garantizar que la inferencia realizada contará con los niveles de precisión y confianza deseados. Sin embargo, para entender mejor la importancia de aplicar este tipo de selección, es necesario conocer el concepto de representatividad de una muestra, para ellos se introduce el siguiente apartado que busca explicarlo.

2.1.3 Representatividad de las muestras

Este término es sumamente importante en el muestreo y se refiere al hecho de que los atributos de la muestra se infieren a la población, es decir, en una buena muestra, su comportamiento y su composición, es tal, que "representa" adecuadamente al universo, permitiendo aplicar un proceso de generalización ó extrapolación como aproximación al comportamiento de la población objetivo. Esta es una de las bondades que da como consecuencia el muestreo probabilístico aplicado de forma adecuada. En la medida que se pueda construir una muestra que contenga estructuralmente atributos conocidos similares a los de la población, se garantiza que la información obtenida de la muestra puede ser extrapolada a la población. Dicho de otra forma y en sentido inverso, las variables observadas en un conjunto de unidades de una población dada (muestra), pueden ser generalizadas a elementos con características similares a los estudiados (población).

Una muestra es representativa de la población, cuando las distribuciones marginales y conjuntas de variables que caracterizan a la población son muy similares entre la muestra y la población.

Pues bien, se aplica una selección probabilística que busca que la muestra sea representativa del universo, sin embargo, es necesario fijar un tamaño de muestra y la selección

probabilística tiene variantes por lo que es necesario hablar del concepto de “Diseño de muestra”

2.1.4 Diseños de muestra

A la forma en que es seleccionada la muestra y a la determinación del número de elementos que deben seleccionarse se llama diseño muestral. Se nombran algunas opciones y en apartados posteriores se profundizará en la descripción de los que están involucrados en el presente trabajo:

- **Muestreo aleatorio simple (MAS)**

  Selección aleatoria de los elementos muestrales con probabilidades de selección en cualquier extracción iguales y sin reemplazo.

- **Muestreo sistemático**

  Se ordenan los elementos de la población y son seleccionados 1 de cada K, si el orden es aleatorio equivale a un MAS.

- **Muestreo con probabilidad proporcional al tamaño**

  Los elementos de la población tienen probabilidades distintas de selección en función del comportamiento de una variable conocida en el universo. Un elemento con valores grandes en dicha variable tiene una mayor probabilidad de ser seleccionado en la muestra

- **Muestreo estratificado**

  La población es segmentada en conjuntos excluyentes llamados estratos. Dentro de cada uno de ellos existe una selección aleatoria, sistemática o con probabilidades proporcionales.

En el diseño de una muestra, se persigue como objetivo primario el que ésta pueda ser representativa del universo para que a través de ella se pueda inferir el comportamiento del universo en las variables de interés.

A partir de este punto se llega a dos grandes interrogantes:

- ¿Cómo lograr que la muestra sea representativa de la población objetivo?
- ¿Cualquier inferencia a partir de la muestra nos dará un resultado cercano al que se leería en el universo?

Respondiendo a estas interrogantes, lamentablemente existen fenómenos para los cuales resultará difícil encontrar una relación adecuada entre los atributos de la muestra y los de la población o, más aún, los que definen al universo e influyen sobre su comportamiento pueden ser demasiados y muy posiblemente no se cuente con información de los más relevantes. Esto implica que no todas las inferencias son seguras, ya que siempre existe la posibilidad de equivocarse. Aplicando el muestreo probabilístico, se busca que la probabilidad de equivocarse sea pequeña, difícilmente se puede lograr que la muestra sea representativa a menos que se cuenten con los elementos necesarios y suficientes que describan a la población y por tanto permitan cotejar sus características relevantes con la muestra con la finalidad de verificar la igualdad de estructuras y garantizar representatividad.

No cualquier inferencia dará un resultado aproximado a lo que realmente sucede en la población objetivo, pero para ello es que se usan las técnicas estadísticas que nos permiten saber qué tan aproximada es la inferencia y valorar si es posible concluir algo respecto a la variable de interés en el universo.

Para garantizar esta aproximación, el muestreo guarda su sustento estadístico en diferentes teoremas y leyes de probabilidad que permiten alcanzar convergencia estadística, a decir, las leyes de los grandes números y el teorema central del límite (Méndez Ignacio, 2004: 12, 26).

Se han descrito diferentes tipos de selección de la muestra, sin embargo es necesario hablar de las características que puede tener dentro de un diseño estadístico dado. A continuación serán descritos dos tipos que nos revelan algunas particularidades que deben ser definidas para extrapolar la información que se está procesando.

2.1.5 Muestras autoponderadas

Una muestra autoponderada es aquella en la que todos los elementos de la población tienen la misma probabilidad de ser seleccionados. Si el universo puede ser dividido en segmentos, implícitamente esta probabilidad contempla tanto la que tiene cada segmento, como la que cada unidad de muestreo tiene de ser elegida dentro de la partición a la que pertenece.

2.1.6 Muestras no autoponderadas de poblaciones finitas

Este concepto se refiere al caso en que los elementos la población no tienen la misma probabilidad de ser seleccionados dentro de una muestra aleatoria, se usan cuando se requiere minimizar varianzas en los estratos conformados del universo, cuando se tiene un alto índice de "no respuesta" en algún segmento y es necesaria una muestra más grande dentro del mismo y por tanto los elementos cuentan con una mayor probabilidad de selección, cuando los costos para acceder a cierto conjunto de unidades del universo es elevado y por tanto es necesario diferenciar la probabilidad de selección de los mismos para ajustarse al presupuesto, etc. En estos casos, se asume que los promedios o proporciones muestrales no están cerca (no convergen) a los valores poblacionales. Lo anterior ocurre porque la muestra no es representativa al contar con una proporción mayor de elementos de ciertos estratos o segmentos, sin embargo, para obtener estimadores de medias (o algunas otras características), se hacen ajustes en el cálculo a través de los llamados "factores de expansión" que son el inverso de las probabilidades de selección de las unidades muestrales. Esto equivale a restaurar la representatividad de la muestra en forma analítica.

2.2 ESTIMADORES Y ESTIMACIÓN

En este apartado se detallará más sobre las técnicas de muestreo para entender cómo es el proceso que está implícito las mismas así como los tipos de inferencias a partir de estimadores que se pueden obtener al aplicarlas.

2.2.1 Definición de estimador

Un estimador es cualquier estadística de muestra que se utilice para estimar un parámetro de población. Por ejemplo, la media de una muestra conformada por los elementos $\{u_1, u_2, ..., u_n\}$, puede ser un estimador de la media de una población conformada por los elementos $\{u_1, u_2, ..., u_N\}$ (N>n).

2.2.2 Definición de estimación

Cuando es observado un valor numérico específico del estimador, se hace referencia a ese dato como una **estimación**. En otras palabras, una estimación es una cifra específica observada de una estadística. Se hace una estimación si se toma una muestra y se calcula el valor que toma el estimador.

Mediante el muestreo, lo que se desea hacer es una inferencia, esto es, una estimación de un parámetro de un universo finito. Así, la población es un conjunto *P* de N unidades $P = \{u_1, u_2, ..., u_N\}$ y a cada unidad se le asocia una variable de interés $y(u_i) = y_i$.

La muestra es un subconjunto de n unidades de la población, ésta se obtiene con probabilidades conocidas para todos y cada uno de los elementos de la población. El tamaño de la muestra se denota como *n*.

En la muestra de n unidades se determinan *n* valores de la variable *y*, esto es: y_1, y_2, ... y_n.

Con los mismos se construyen estimadores de los parámetros de interés en la población objetivo, generalmente se desea conocer la media (μ) y la desviación estándar (σ^2) para la variable de interés. Algunos ejemplos de los estimadores que se pueden construir (y los más comunes) son los siguientes:

$\bar{y} = \hat{\bar{Y}} = \frac{\sum_1^n y_i}{n}$, será el estimador de la media de la población (μ) (2.1)

$s^2 = \hat{S}^2 = \frac{\sum_1^n (y_i - \bar{y})}{n-1}$, corresponde al estimador insesgado de la varianza (S^2)[5] (2.2)

$Y = N\frac{\sum_1^n y_i}{n}$, representará el estimador del total poblacional ($Y = N\bar{y}$) (2.3)

2.2.3 Cualidades de un buen estimador

Algunas estadísticas son mejores estimadores que otras. Afortunadamente, se puede evaluar la calidad de estas mediante el uso de cuatro criterios:

- Insesgado

  Un estimador es insesgado si su sesgo es nulo, es decir, la diferencia entre su esperanza y el verdadero valor del parámetro a estimar es cero.

  Si una muestra $T = \{t_1, t_2, ..., t_n\}$ procede de una población de media μ, para que T sea insesgado debe cumplir que $E(T) = \mu$.

[5] Por motivos de notación, las expresiones se presentan generalmente usando S^2 y no σ^2 donde el denominador es N y no N-1 como en la primera. Esto tiene la ventaja de que los resultados se presentan de forma más simple (Cochran , 1998: 47).

Ésta es una propiedad deseable para un buen estimador. El término insesgado se refiere a que la media de la distribución de muestreo de las medias de muestras tomadas de una misma población es igual a la media de la población misma. Se puede señalar que una estadística tiene ésta característica, si tiende a tomar valores que están por encima del parámetro de la población que se está estimando con la misma frecuencia y la misma extensión con la que tiende a asumir valores por debajo del parámetro de población.

- Eficiencia

Se dice que un estimador es más eficiente o más preciso que otro estimador, si la varianza del primero es menor que la del segundo. De ésta forma, si T_1 y T_2 son ambos estimadores de μ y $Var(T_1) < Var(T_2)$, diremos que T_1 es más eficiente que T_2. Un estimador es más eficiente (más preciso), por tanto, cuanto menor es su varianza.

- Consistencia

La definición de está propiedad más común exige que:

1. $E(T) \to \mu$ cuando $n \to \infty$
2. $Var(T) \to 0$ cuando $n \to \infty$

Así, una estadística es un estimador consistente de un parámetro de población si al aumentar el tamaño de la muestra, se tiene casi la certeza de que el valor de la estadística se aproxima bastante al parámetro de la población. Si un estimador es consistente, se vuelve más confiable si se tienen tamaños de muestra más grandes.

- Suficiencia

Un estimador es suficiente si utiliza una cantidad de la información contenida en la muestra que ningún otro podría extraer sobre el parámetro de la población que se está estimando[6].

[6] Para mayor referencia sobre las propiedades de los estimadores, ver Mood 1974, páginas 288-307.

2.3 DISEÑOS MUESTRALES

En este apartado serán detallados diferentes tipos de diseños muestrales que se describieron de forma muy general previamente. El término diseño muestral es referirse a la forma de selección de muestra y determinación de tamaño de muestra por lo que es necesario profundizar en estos temas.

2.3.1 Muestreo aleatorio simple (MAS)

Una muestra aleatoria simple se refiere a la extracción aleatoria de cierto número de elementos de la población objetivo, esta se puede hacer de dos formas: con reemplazo, donde el mismo elemento puede ser seleccionado más de una vez en la muestra, y sin reemplazo, donde todas las unidades elegidas son diferentes.

En general, la teoría del MAS contempla el supuesto de una selección sin reemplazo, de trabajar una con reemplazo, podrían existir duplicados en la muestra, es decir, elementos que fueron elegidos más de una vez y no aportan información adicional. Por otra parte, este tipo de selección simplifica la construcción y valoración de estadísticos.

Una muestra aleatoria de tamaño n se selecciona de manera que cada una de las muestras posibles de dicho tamaño tienen la misma probabilidad de ser elegidas, de esta forma. Todas las posibles combinaciones de tamaño n a partir de N elementos representan todos los conjuntos distintos que pueden ser elegidos $\binom{N}{n}$.

La probabilidad de selección de cada una de las muestras (S), está dado por $\binom{N}{n}^{-1}$ ya que todas las posibles tienen la misma posibilidad de ser seleccionadas.

$$P(S) = \binom{N}{n}^{-1} = \frac{n!(N-n)!}{N!} . \qquad (2.4)$$

Para realizar la extracción de una muestra aleatoria simple, es necesario contar con un marco muestral, es decir, una lista de los elementos que pertenecen a la población.

Con MAS se puede estimar la media poblacional $\bar{Y}$ a partir de la muestral dada por:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{2.5}$$

Este estimador es insesgado y su varianza está dada por (Cochran 1998: 46-47):

$$V(\bar{y}) = \frac{S^2}{n}(1-f), \tag{2.6}$$

donde f es la fracción de muestreo $f = \frac{n}{N}$.

La expresión $(1-f)$ es conocida como "factor de corrección por finitud". Si la población es infinita, este tiende a cero, como su nombre lo indica, es de importancia en el tratamiento de poblaciones finitas, ya que para pequeñas, la fracción de muestreo (f) es mayor, es decir, se cuenta con mayor información de la población y por lo tanto la varianza debe ser menor, de esta forma, cuando el tamaño de muestra seleccionada sea igual al del universo la varianza será cero.

Generalmente la varianza de la población σ^2 es desconocida y un procedimiento para obtener un valor aproximado de la misma es estimarla mediante la de la muestra de tamaño n:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2. \tag{2.7}$$

El estimador insesgado (Cochran, 1998: 50) de la varianza del estimador de la media es:

$$\hat{V}(\bar{y}) = \frac{s^2}{n}(1 - f)\,. \tag{2.8}$$

A partir del estimador de la varianza se llega a lo que es conocido como el error estándar (ee) obteniendo la raíz cuadrada del primero.

$$ee(\bar{y}) = s_{\bar{y}} = \sqrt{\hat{V}(\bar{y})} = \sqrt{\frac{s^2}{n}(1 - f)}\,. \tag{2.9}$$

Éste es un indicador muy importante al mostrar qué tan precisos pueden ser los estimadores que se está construyendo, proporciona una aproximación de la variabilidad que puede manifestarse en la distribución del estimador de la media y en base a ello determinar qué tan alejada puede estar la estimación del valor real, es decir, definir qué tan precisa es la aproximación a la media de la variable de estudio.

En términos relativos, se habla del error estándar relativo (eer), éste resulta ser el coeficiente de variación calculado a partir del error estándar de la distribución de los estimadores de la media y la estimación puntual de la misma:

$$eer(\bar{y}) = cv(\bar{y}) = \frac{ee(\bar{y})}{\bar{y}}\,. \tag{2.10}$$

Estos resultados se aplican a la estimación del total de la población (Y), así

$$Y = \sum_{i=1}^{N} y_i = N\bar{Y}\,, \tag{2.11}$$

y para estimarlo, se usa el estimador insesgado:

$$\hat{Y} = N\frac{\sum_{i=1}^{n} y_i}{n} = N\bar{y}, \tag{2.12}$$

de esta forma, la varianza del estimador es:

$$V(\hat{Y}) = N^2 V(\bar{y}) = N^2(1-f)\frac{S^2}{n}. \tag{2.13}$$

Al igual que en el tratamiento de la varianza de la media cuando es desconocida la varianza de la variable de estudio, se cuenta con una aproximación de la varianza a partir del estimador de S^2, llegando a la siguiente expresión;

$$\hat{V}(\hat{Y}) = N^2(1-f)\frac{s^2}{n}. \tag{2.14}$$

2.3.1.1 Intervalos de confianza

Como fue señalado previamente, el *ee* o *eer* (según sea el caso) nos proporciona una idea importante de variabilidad del estimador a través de diversas muestras y por tanto nos ayudará para establecer cual será la precisión esperada de la inferencia, esto toma sentido cuando se analiza la inferencia en términos de intervalos de confianza, donde un elemento determinante en su construcción es precisamente el *ee*.

No es suficiente con informar cuál es el resultado puntual y saber así cuáles serían los posibles valores que puede tomar el estimador, es necesario indicar adicionalmente la precisión o exactitud de las estimaciones. Para dicho efecto, se usan los intervalos de confianza (IC) que nos permiten proporcionar esta información adicional.

Generalmente cuando se presenta un resultado de alguna inferencia a partir de estimadores, se dice que la estimación cuenta con una precisión *e* a un nivel de confianza $(1-\alpha)\%$ siendo este último un valor cercano a 100%, el cual indica la confianza de que el intervalo construido

a partir del valor del estimador y la precisión deseada contenga el valor real del parámetro a estimar. Dicho de otra forma, la construcción de un intervalo de confianza parte de la necesidad de garantizar que la diferencia entre el estimador y el valor real del parámetro sea menor que una precisión *e* dado, así, siendo $\hat{\theta}$ un estimador de θ, se espera que $P(\left|\hat{\theta}-\theta\right|<e)=1-\alpha$.

Para construir un IC se usa el teorema central del límite. Esto será válido si la distribución de los estimadores converge a una distribución normal cuando *n* tiende a infinito.

Un intervalo de confianza al $100*(1-\alpha)\%$ en una muestra de tamaño n para la media de la población es:

$$\left[\bar{y}-z_{\alpha/2}\sqrt{1-f}\,\frac{s}{\sqrt{n}},\bar{y}+z_{\alpha/2}\sqrt{1-f}\,\frac{s}{\sqrt{n}}\right]. \qquad (2.15)$$

En términos del error estándar (EE) es:

$$\left[\bar{y}-z_{\alpha/2}ee(\bar{y}),\bar{y}+z_{\alpha/2}ee(\bar{y})\right]. \qquad (2.16)$$

Donde $z_{\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estándar.

Aunque se pueden medir muchas variables, con frecuencia es recomendable centrarse en una o dos que sean de interés fundamental en el estudio y utilizarlas para estimar el tamaño de muestra necesario.

A manera de recomendación, el estadístico debe preguntarse lo siguiente para asegurar la adecuada selección de variables a medir en la muestra;

- ¿Qué se espera de la muestra?
- ¿Cuánta precisión se necesita?

- ¿Cuáles son las consecuencias de los resultados de la muestra?
- ¿Cuál es la cantidad de error tolerable?

Son interrogantes muy concisas que tienen la finalidad de definir de manera acotada los alcances de la muestra en cuestión. Es importante resaltar los alcances y limitaciones de los resultados de la muestra debido a que de ahí se deriva el grado de responsabilidad del estadístico encargado del diseño muestral.

2.3.1.2 Muestreo Aleatorio Simple (precisión y cálculo del tamaño de muestra)

Tomando como base lo visto previamente en la sección de intervalos de confianza, la precisión de la muestra para un estimador de la media en términos relativos puede ser expresada en los siguientes términos:

$$P\left(\left|\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right| \leq er\right) = 1-\alpha\,. \tag{2.17}$$

Donde *er* corresponde a la precisión relativa o margen de error relativo, por tanto $e\bar{Y}$ corresponde al error absoluto y $1-\alpha$ se refiere al nivel de confianza con el cual se espera que la diferencia relativa entre el estimador y el valor real del parámetro en cuestión sea menor que *er*. Generalmente los valores comunes para estos conceptos son; *er* = 0.03 y $\alpha = 0.05$, sin embargo estos variarán, como ya se mencionó, en función de las necesidades del estudio y por tanto la precisión con la que sea necesaria generar los resultados del mismo.

Asumiendo normalidad en la distribución de $\bar{y}$ y usando los resultados anteriores de la sección de intervalos de confianza, se tiene que:

$$e = z_{\alpha/2}\sqrt{1-f}\,\frac{s}{\sqrt{n}}\,. \tag{2.18}$$

Y al despejar n, se tiene que:

$$n = \frac{z_{\alpha/2}^2 s^2}{e^2 + \frac{z_{\alpha/2}^2 s^2}{N}} = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{2.19}$$

donde

$$n_0 = \frac{z_{\alpha/2}^2 s^2}{e^2}. \tag{2.20}$$

El valor n_0 es el tamaño de muestra para una muestra aleatoria simple sin reemplazo.

Para calcular un tamaño de muestra con el cual se pueda obtener una precisión relativa dada, es sustituido $er\bar{Y}$ en vez de *e* dado en al ecuación previa, lo cual produce lo siguiente:

$$n = \frac{z_{\alpha/2}^2 s^2}{(er \cdot \bar{y})^2 + \frac{z_{\alpha/2}^2 s^2}{N}} = \frac{z_{\alpha/2}^2 cv^2}{er^2 + \frac{z_{\alpha/2}^2 cv^2}{N}} = \frac{n'_0}{1 + \frac{n'_0}{N}}, \tag{2.21}$$

donde

$$n'_0 = \frac{z_{\alpha/2}^2 \cdot cv^2}{e^2}. \tag{2.22}$$

Para alcanzar una precisión relativa dada, el tamaño de muestra se puede determinar al utilizar sólo el coeficiente de variación.

2.3.2 Muestreo Aleatorio Estratificado (MAE)

El MAE consiste en dividir a la población en L estratos y de cada uno de ellos seleccionar una muestra probabilística.

Este tipo de muestreo es aplicado por las siguientes razones:

- Estadística: cuando la población está constituida por unidades heterogéneas entre sí y se tienen antecedentes de los grupos de unidades más homogéneos es recomendable crear estratos y obtener estimadores de los parámetros de interés para cada uno de ellos. Esto permite contar con una mayor precisión en la inferencia al agrupar elementos de la población que son semejantes garantizando una reducción en la variabilidad dentro de cada estrato

- Marco: en ocasiones se dispone de uno parcial, es decir, de tan sólo una porción de la población, en este caso, el mismo se usa para esta parte y los complementos de la población se analizan con marcos más imprecisos. Posiblemente llegará a utilizarse un diseño de muestra diferente dependiendo del marco muestral

- Costo: se pueden identificar claramente los costos de ubicar y levantar la información de las unidades para cada estrato suponiendo mismos costos para unidades del mismo estrato

Dadas estas razones, es común que los estratos formen unidades homogéneas, con un mismo tipo de marco y con costos de localización y captación de información semejantes.

Al dividir la población de N unidades de muestreo en L estratos, es necesario conocer el tamaño del universo de cada estrato, de forma que se pueda establecer la siguiente relación: $N = N_1 + N_2 + N_3 + ... + N_L$, donde N_i es el universo conocido del estrato i.

En el MAE, es tomada una muestra aleatoria simple de manera independiente en cada estrato, de modo que se seleccionen de forma aleatoria n_h observaciones de las unidades de población en el estrato *h*.

2.3.2.1 Notación, estimadores de media y varianza, intervalos de confianza

La notación para describir el MAE es la siguiente:

N_h, número total de unidades en el estrato *h*.

n_h , número de unidades en la muestra del estrato *h*.

y_{hj} , es el valor de la variable observada de la unidad j en el estrato *h.*

$W_h = \frac{N_h}{N}$, corresponde a la ponderación del estrato *h*.

$f_h = \frac{n_h}{N_h}$, fracción de muestreo en el estrato *h*.

Considérese que para la notación correspondiente a la muestra, se selecciona una muestra aleatoria simple dentro de cada estrato de tal forma que si n_h es el tamaño de muestra en el estrato *h*, la muestra total es $n = n_1 + n_2 + n_3 + ...n_L$.

En la siguiente tabla resumen se podrán observar algunas características poblacionales en comparación con sus respectivos estimadores que surgen a partir de la muestra de tamaño *n*.

Características poblacionales vs. muestrales en MAE

Característica poblacional (cp)	Descripción de la cp	Estimador de la cp	Descripción del estimador de la cp
$Y_h = \sum_{j=1}^{N_h} y_{hj}$	Total poblacional en el estrato h	$\hat{Y}_h = \sum_{j=1}^{n_h} \frac{N_h y_{hj}}{n_h}$	Estimador del total poblacional en el estrato h
$Y = \sum_{h=1}^{L} Y_h$	Total poblacional del universo	$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h$	Estimador del total poblacional
$\bar{Y}_h = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$	Media poblacional del estrato h	$\bar{y}_h = \frac{\sum_{j=1}^{n_h} y_{hj}}{n_h}$	Estimador de la media poblacional en el estrato h
$\bar{Y} = \frac{Y}{N} = \frac{\sum_{h=1}^{L}\sum_{j=1}^{N_h} y_{hj}}{N}$	Media poblacional del universo	$\bar{y}_{st} = \frac{\hat{Y}}{N}$	Estimador de la media poblacional
$S_h^2 = \sum_{j=1}^{N_h} \frac{\left(y_{hj} - \bar{Y}_h\right)^2}{N_h - 1}$	Varianza poblacional en el estrato h	$s_h^2 = \sum_{j=1}^{n_h} \frac{\left(y_{hj} - \bar{y}_h\right)^2}{n_h - 1}$	Estimador de la varianza poblacional en el estrato h

Tabla 2.1

De acuerdo a la tabla previa, para estimar la media poblacional se usa:

$$\bar{y}_{st} = \frac{\hat{Y}}{N} = \frac{\sum_{h=1}^{L}\sum_{j=1}^{n_h} \frac{N_h y_{hj}}{n_h}}{N} = \frac{\sum_{h=1}^{L} N_h \sum_{j=1}^{n_h} \bar{y}_h}{N} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h \,. \qquad (2.23)$$

Al cociente $\frac{N_h}{N}$ se le suele llamar "peso de muestreo" y representa la importancia que tiene el estrato *h* en el universo respecto al número de elementos que conforman la población, comúnmente es denotado por W_h que representa el promedio ponderado de las medias de los estratos de la muestra.

En cuanto a las propiedades de estos estimadores:

- $\hat{Y}_{st}$ y $\bar{y}_{st}$ **son estimadores insesgados** de Y y $\bar{Y}$ dado que:

$$E\left[\sum_{h=1}^{L}\frac{N_h}{N}\bar{y}_h\right]=\sum_{h=1}^{L}\frac{N_h}{N}E[\bar{y}_h]=\sum_{h=1}^{L}\frac{N_h}{N}\bar{Y}_h=\bar{Y}\,. \tag{2.24}$$

- **Estimador de la varianza de los estimadores**. Debido a que se está aplicando un muestreo a partir de estratos excluyentes, se conoce $\hat{V}(\bar{y}_h)$ de cada uno de los estratos partiendo de la teoría del muestreo aleatorio simple por lo que el estimador insesgado de la varianza para el estimador de la media en MAE se obtiene de la siguiente expresión:

$$\hat{V}(\bar{y}_{st})=\sum_{h=1}^{L}W_h^2\hat{V}(\bar{y}_h)=\frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2s_h^2}{n_h}-\frac{1}{N^2}\sum_{h=1}^{L}N_h^2s_h^2\,. \tag{2.25}$$

El estimador de la varianza del estimador del total se calcula mediante la siguiente expresión:

$$\hat{V}(\hat{Y}_{st})=N^2\hat{V}(\bar{y}_{st}). \tag{2.26}$$

De igual forma al tratamiento descrito en el MAS, el error estándar de un estimador es la raíz cuadrada de la varianza del estimador: $ee(\bar{y}_{st})=s_{\bar{y}_{st}}=\sqrt{\hat{V}(\bar{y}_{st})}$.

- **Intervalos de confianza para muestras estratificadas.** Un intervalo de confianza aproximado del $100(1-\alpha)\%$ para la media es::

$$[\bar{y}_{st}-z_{\alpha/2}ee(\bar{y}_{st}),\bar{y}_{st}+z_{\alpha/2}ee(\bar{y}_{st})]=[\bar{y}_{st}-e,\bar{y}_{st}+e], \tag{2.27}$$

donde *e* es conocida como la precisión y representa la mitad de la longitud del intervalo de confianza.

Y sólo tiene sentido su uso si se dan algunas de las siguientes 2 condiciones:

1. Si los tamaños de las muestras dentro de cada estrato son grandes
2. Si el diseño del muestreo tiene una gran cantidad de estratos

2.3.3 Afijación de la muestra

A partir de un tamaño de muestra n, se le llama afijación de la muestra al proceso de distribuirlo en los diferentes estratos que conforman la población. En esta sección se abordarán algunos métodos existentes para la afijación de la muestra en cada uno de los estratos.

2.3.3.1 Afijación proporcional

Dentro de este tipo de afijación, se respeta la estructura del universo en cuanto al tamaño de cada uno de los estratos en la muestra, de esta forma, se espera que $\frac{n_h}{n} = W_h$ para cada estrato *h*.

En la afijación proporcional, la probabilidad de selección es igual para todos los elementos de la población independientemente del estrato en el cual se encuentren. Cuando se realiza una selección de muestra con estas propiedades, también se dice que se está hablando de una muestra autoponderada, donde todas y cada una de las unidades de muestreo representan al mismo número de elementos de la población. Es importante señalar que en una muestra aleatoria simple también todos los elementos tienen la misma probabilidad de selección, sin embargo, bajo MAE, se tiene la ventaja de que se garantiza que todos los diferentes conjuntos (estratos) que conforman la población se tienen representados y en MAS no.

Para el cálculo del tamaño de muestra en el estrato *h* a partir de *n* en una muestra estratificada con afijación proporcional, se parte de la siguiente relación $\frac{N_h}{N} = \frac{n_h}{n}$, así, para el cálculo de n_h simplemente se despeja la variable de la igualdad dada, y:

$$n_h = n\frac{N_h}{N} = nW_h . \tag{2.28}$$

Afijación proporcional (varianza de los estimadores, cálculo del tamaño de muestra)

Varianza de los estimadores bajo afijación proporcional

Para encontrar la varianza de los estimadores, simplemente se parte de las expresiones mostradas previamente sustituyendo el tamaño de muestra n_h así, para la estimación del total se tenía que:

$$\hat{V}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L}\hat{V}\left(\hat{Y}_h\right) = \sum_{h=1}^{L}\left(1-\frac{n_h}{N_h}\right)N_h^2\frac{s_h^2}{n_h} . \tag{2.29}$$

Y bajo afijación proporcional:

$$\hat{V}_{prop}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L}\left(1-\frac{(n\frac{N_h}{N})}{N_h}\right)N_h^2\frac{s_h^2}{(n\frac{N_h}{N})} = \sum_{h=1}^{L}\left(1-\frac{n}{N}\right)N_h(\frac{N}{n})s_h^2 . \tag{2.30}$$

Y para la estimación de la media se tiene que:

$$\hat{V}\left(\bar{y}_{st}\right) = \sum_{h=1}^{L}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{s_h^2}{n_h}. \tag{2.31}$$

Y bajo afijación proporcional, la estimación de la media es:

$$\hat{V}_{prop}(\bar{y}_{st}) = \hat{V}_{prop}\left(\frac{\hat{Y}_{st}}{N}\right) = \frac{1}{N^2}\left(\sum_{h=1}^{L}\frac{NN_h s_h^2}{n} - \sum_{h=1}^{L} N_h s_h^2)\right). \quad (2.32)$$

Cálculo del tamaño de muestra

Es sabido que la precisión (*e*) dada por la mitad de la longitud del intervalo de confianza a partir de una muestra estratificada esta dado por $e = z_{\alpha/2} ee(\bar{y}_{st})$ y también se sabe que $ee(\bar{y}_{st}) = \sqrt{\hat{V}_{prop}(\bar{y}_{st})}$ por lo que

$$e^2 = z_{\alpha/2}^2 \frac{1}{N^2}\left(\sum_{h=1}^{L}\frac{NN_h s_h^2}{n} - \sum_{h=1}^{L} N_h s_h^2)\right), \quad (2.33)$$

de donde

$$n = \frac{N\sum_{h=1}^{L} N_h s_h^2}{N^2 \frac{e^2}{z_{\alpha/2}^2} + \sum_{h=1}^{L} N_h s_h^2}. \quad (2.34)$$

2.3.3.2 Afijación de Neyman

Si las varianzas S_h^2 son más o menos iguales a lo largo de todos los estratos, la afijación proporcional es probablemente la mejor distribución para una mayor precisión. Cuando las S_h^2 son pequeñas, la afijación de Neyman produce un menor costo en comparación a otras opciones de afijación de muestra. De manera intuitiva, se basa no tan sólo en el número de unidades de muestreo de cada estrato sino también en la variabilidad existente en cada uno de ellos, de tal forma que la probabilidad de selección disminuye si la unidad de muestreo se encuentra en un estrato cuya variabilidad y/o universo sean pequeños. A mayor variabilidad

mayor necesidad de muestra se tendrá. De esta forma, la distribución del tamaño de muestra en cada estrato a partir de la n total, está dado de la siguiente manera:

$$n_h = n\frac{N_h S_h^2}{\sum_{h=1}^{L} N_h S_h^2}. \tag{2.35}$$

2.3.3.3 Comparación entre MAS, MAE (Proporcional) y MAE (Neyman)

En general, para universos grandes dentro de cada estrato dentro de los cuales se puede ignorar el factor de corrección por finitud y para los cuales $N_h \approx N_h - 1$, se puede asumir lo siguiente:

$$V(\bar{y}_{mas}) = V_{prop}(\bar{y}_{st}) + \frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh}(\bar{y}_h - \bar{Y})^2}{nN}, \text{ es decir,} \tag{2.36}$$

$$V(\bar{y}_{mas}) \geq V_{prop}(\bar{y}_{st}). \tag{2.37}$$

De igual forma, se puede establecer que:

$$V_{prop}(\bar{y}_{st}) = V_{ney}(\bar{y}_{st}) + \frac{\sum_{h=1}^{L} N_h}{nN}\left(S_h - \frac{\sum_{h=1}^{L} N_h S_h}{N}\right)^2, \tag{2.38}$$

es decir,

$$V_{prop}(\bar{y}_{st}) \geq V_{ney}(\bar{y}_{st}). \tag{2.39}$$

Concluyendo por transitividad que:

$$V(\bar{y}_{mas}) \geq V_{prop}(\bar{y}_{st}) \geq V_{ney}(\bar{y}_{st}). \tag{2.40}$$

2.3.4 Estratificación por particiones sucesivas

Parte de los objetivos de la conformación de estratos es poder identificar y controlar las fuentes de variación que describen la variable de estudio dentro de la población objetivo maximizando la precisión de las inferencias y minimizando el costo, el cual puede estar directamente asociado al tamaño de muestra necesario.

La explicación que se verá en el siguiente párrafo, trata de detallar el procedimiento y lógica de un método para estratificar llamado particiones sucesivas.

Dentro del MAE, uno de los elementos fundamentales para lograr buenos resultados en las inferencias estadísticas bajo este muestreo es contar con métodos eficientes que permitan optimizar los estratos, por definición, se pretende conformar conjuntos de elementos que sean homogéneos al interior y heterogéneos entre los mismos. De esta forma, en la medida es posible conformar conjuntos para los cuales se puede garantizar la mayor igualdad posible al interior de cada partición y la mayor diferencia que sea alcanzable al exterior (entre estratos) entonces es posible lograr optimizar la definición de estratos de manera sustancial.

Hablando simplemente en términos de la suma de cuadrados, es posible establecer la siguiente expresión:

$$\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2=\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2+\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\bar{Y}_h-\bar{Y})^2\ . \qquad (2.41)$$

La cual indica que la variabilidad total puede verse como la suma de todas las variabilidades dentro de cada uno de los estratos existentes más la suma de las variabilidades entre estratos.

De resultados previos se tiene la siguiente relación entre la variabilidad de MAS y MAE prop.

$$V(\bar{Y}_{mas})=V_{prop}(\bar{Y}_{est})+\frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh}(\bar{Y}_h-\bar{Y})^2}{nN}\ . \qquad (2.42)$$

Ahora bien, el objetivo será maximizar la variabilidad entre estratos, de esta forma se minimiza la variabilidad al interior de cada uno de los estratos y por tanto se logra una estratificación eficiente en el diseño.

Procedimiento de particiones sucesivas [7].

Se explicará de forma empírica cómo es la minimización de la suma de cuadrados dentro de estratos a partir del procedimiento llamado particiones sucesivas.

- Se parte de una serie ordenada descendente o ascendente de los valores de la variable de estratificación, la cual se espera este correlacionada con la variable de estudio.

- Se supone de arranque, la existencia de dos grandes estratos, uno que tiene todos los elementos de la población y otro que no tiene ningún elemento. En este caso se está hablando de un solo estrato.

- Se calcula la aportación de cada uno de los dos grupos a la suma de cuadrados, en el paso inicial o final, uno de los dos estratos incluye la suma de cuadrados total y el otro tiene contribución nula

$$\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^2 . \qquad (2.43)$$

- En cada paso se suman las aportaciones de los dos grupos a la suma de cuadrados:

$$\sum_{i=1}^{N_1}(y_{1i}-\bar{Y}_1)^2+\sum_{i=1}^{N_2}(y_{2i}-\bar{Y}_2)^2 . \qquad (2.44)$$

[7] Sánchez Villareal Francisco, 2004: 2 (particiones sucesivas)

- Se observará que la suma de los cuadrados disminuye desde los extremos hasta un punto, generalmente alejado del centro del grupo de observaciones, en el cual la suma de cuadrados es mínima.

- Este punto será la frontera para definir los dos estratos, en caso de que el investigador note que los estratos resultantes siguen siendo demasiado grandes o que al interior manejan una alta variabilidad ($CV \geq 0.3$) puede decidir realizar nuevamente el procedimiento para los estratos resultantes que desean ser afinados hasta llegar al conjunto ideal de estratos.

En el siguiente ejemplo se muestra un conjunto inicial de 7 tiendas con coeficiente de variación inicial tiene un valor de 0.549. Al aplicarse el proceso de particiones sucesivas usando la variable de ventas, se sugiere una partición en la opción "D" ya que justo en este corte la suma de cuadrados es mínima, como primer subconjunto son tomadas las primeras 4 tiendas para las cuales se puede observar un volumen considerablemente menor en relación a las 3 restantes que conformarán el segundo subconjunto. Como resultado, el primer subconjunto tiene un CV de 0.17 y el segundo de 0.25, resultando ambos por debajo de la cota establecida.

Ejemplo: particiones sucesivas

Tienda	CIUDAD	Estrato	Ventas	CV INICIAL	CV	Orden	SUM C A /1M	SUM C B /1M	OPCIÓN	SUM OA+OB /1M
8287	11	1100	**11306**	0.549	0.177	1	677.5	0.0	A	677
7950	11	1100	**12887**		0.177	2	502.2	0.6	B	503
2912	11	1100	**15129**		0.177	3	320.4	4.9	C	325
7915	11	1100	**16977**		0.177	4	111.0	14.0	D	**125**
1709	11	1101	26607		0.256	5	22.2	115.4	E	138
1724	11	1101	35417		0.256	6	0.0	366.6	F	367
1672	11	1101	44848		0.256	7	0.0	840.9	G	841

Tabla 2.2

La gráfica que se muestra a continuación muestra el comportamiento de la suma de cuadrados por opción. La opción "A" contiene la tienda más pequeña como subconjunto uno y las siete restantes como subconjunto 2. La opción "B" tendrá las 2 más pequeñas como subconjunto uno y las seis restantes como subconjunto dos y así sucesivamente.

Claramente la opción"D" es la mejor alternativa aunque se encuentra muy cerca de la opción "E". Gráficamente se visualiza de la siguiente manera:

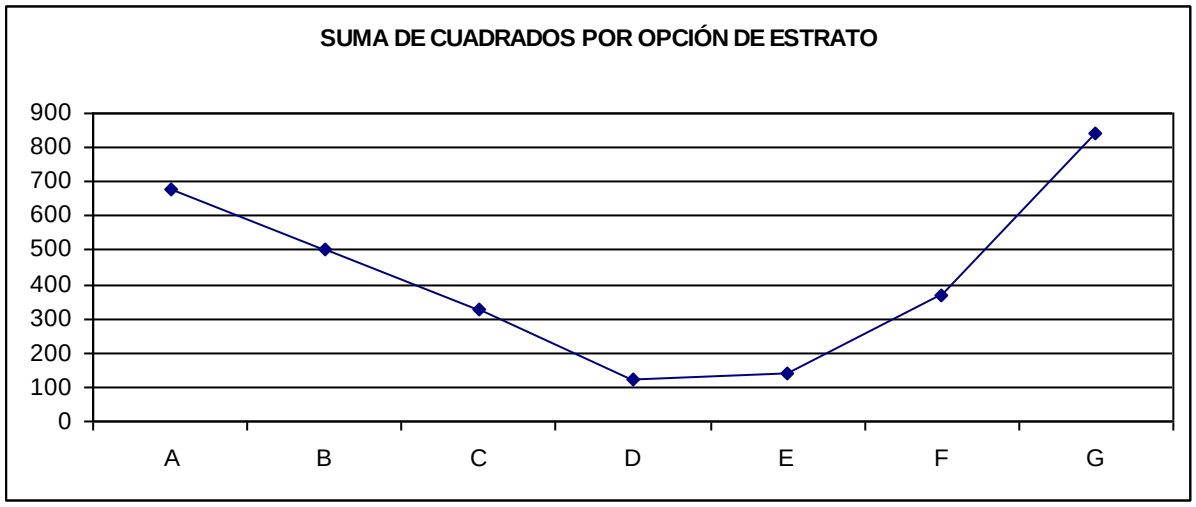

SUMA DE CUADRADOS POR OPCIÓN DE ESTRATO
900
800
700
600
500
400
300
200
100
0
A
B
C
D
E
F
G

Figura 2.1

2.4 ESTIMADORES DE RAZÓN

En principio, el estimador de razón permite realizar una inferencia sobre la razón que existe entre dos subconjuntos de una estructura dada que maneja la población objetivo, por ejemplo, si se desea saber cuánto representa el monto total de ventas de una marca en un mercado específico para tener una idea de su participación de mercado ó si se desea saber cuánto representan las ventas promocionadas de un producto respecto a sus ventas totales con la finalidad de evaluar las dimensiones de las ventas promocionadas, entonces este estimador es de gran utilidad.

Adicionalmente, es importante señalar que estos estimadores incrementan la precisión en las inferencias al lograr estimadores más eficientes.

Cuando se opta por usar estimadores de razón, se parte del hecho de contar con una variable auxiliar x_i, fuertemente correlacionada[8] con y_i para cada unidad de la muestra y se hace uso de dicha característica con la finalidad de mejorar la precisión, generalmente, cuando es este caso, el total de la población X se conoce, no así cuando se desea estimar propiamente una razón. A veces x_i resulta ser un valor previo de y_i y en ocasiones también se usan estos estimadores para inferir sobre el cambio relativo de un momento $t-1$ a t del total poblacional Y, así, el estimador de razón es construido de la siguiente manera:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^{n} y_i / n}{\sum_{i=1}^{n} x_i / n}, \tag{2.45}$$

donde tanto x_i como y_i son variables, con x_i conocida para todos los valores de la población, así, el estimador está conformado por el cociente de dos resultando ligeramente sesgado.

[8] Condición necesaria, correlación positiva y en lo posible cercana a 1

Partiendo de x_i conocida para toda la población $X = \sum_{i=1}^{N} x_i$ representa el total poblacional y así el estimador del total poblacional para Y ($\hat{Y}_R$) será:

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} X = \frac{\sum_{i=1}^{n} y_i / n}{\sum_{i=1}^{n} x_i / n} \sum_{i=1}^{N} x_i = \sum_{i=1}^{n} y_i \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i} \quad . \tag{2.46}$$

Donde al cociente $\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}$ conformado a partir de la información conocida, comúnmente se le llama factor de expansión y tiene sentido el nombre ya que el mismo aplicado a la muestra de la variable de interés nos ayuda a representar el complemento del universo derivado de conocer el comportamiento del mismo en una variable conocida altamente correlacionada con la estudiada.

Media poblacional

$$\bar{y}_R = N^{-1} \sum_{i=1}^{n} y_i \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} \bar{X} \, . \tag{2.47}$$

En estos casos, X y por tanto $\bar{X}$ son conocidos y éstos son los estadísticos que sirven para mejorar la precisión en la inferencia de los totales y las medias poblacionales.

Cuando se desea inferir sobre el crecimiento de un momento a otro, x_i con $i = 1,2.....n$ es la medición del periodo previo, a veces resulta ser la medición previa de un censo. Si no es el caso, entonces x_i es una variable conocida correlacionada altamente con y_i, por ejemplo; x_i

puede ser la venta total de un negocio B y y_i las ventas del producto B_j en la tienda B, si se asume que las variables están altamente correlacionadas y que a mayor venta total de un negocio mayor venta del producto en cuestión, entonces la relación $\frac{Y}{\bar{y}}$ se espera sea muy similar a $\frac{X}{\bar{x}}$ y por tanto, a partir de conocer las x_i puede inferirse estructuralmente el comportamiento de las y_i y por consecuencia su total. Es por esta razón que ambas variables deben estar altamente correlacionadas, si no se da esta relación tan estrecha puede caerse en el error de inferir el comportamiento de y_i a partir de una estructura muy diferente de la que realmente tiene esta variable y por tanto llegar a un error grande en la inferencia.

2.4.1 Sesgo del estimador de razón

Se ha señalado que los estimadores de razón son sesgados aunque el sesgo puede ser despreciable para poblaciones grandes y adicionalmente, la varianza reducida del estimador de razón compensa la presencia de sesgo.

El sesgo del estimador de razón adopta la siguiente conformación[9]:

$$E(\hat{R}) - R = \left(\frac{N-n}{N}\right)\frac{1}{n\bar{X}^2}\left[RS_x^2 - \rho S_y S_x\right], \qquad (2.48)$$

donde ρ es el coeficiente de correlación entre *x* y *y*.

Notar que el sesgo se aproxima a cero cuando el tamaño de la muestra es próximo al del universo (lo cual no es relevante) y cuando se da la relación $RS_x - \rho S_y = 0$. Esto se cumple cuando la razón entre la variabilidad de ambas muestras (y_i y x_i) es R, esto es, R es constante para todas las unidades de muestreo y por tanto $\frac{S_y}{S_x} = R$ y la correlación es 1.

[9] Ver Sánchez Villareal Francisco, 2004: 66-68.

2.4.2 Varianza aproximada de los estimadores de razón

Los estimadores de razón son consistentes y adicionalmente, conforme *n* se hace grande, su distribución es normal sujeta a ciertas restricciones sobre el tipo de población que se está muestreando. En muestras de tamaño moderado, la distribución presenta una tendencia hacia asimetría positiva. No se poseen fórmulas exactas para el sesgo y la varianza de muestreo de la estimación sino sólo aproximaciones que son válidas en muestras grandes donde se puede suponer igualdad entre la media muestral y la poblacional de la variable auxiliar.

Por otra parte, para el cálculo de la varianza debe considerarse la presencia de covarianzas, así, las expresiones para obtenerla para los estimadores de razón bajo MAS son las siguientes:

Varianza para el estimador de razón:

$$V(\hat{R}) = \frac{(1-f)}{n\bar{X}^2}\left(\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}\right). \tag{2.49}$$

La covarianza muestral entre y_i y x_i es

$$S_{yx} = \frac{\sum_{i=1}^{n}(y_i - \hat{Y})(x_i - \hat{X})}{n-1}. \tag{2.50}$$

De esta forma, el estimador de la razón cambia a la siguiente expresión:

$$V(\hat{R}) = \frac{(1-f)}{n\bar{X}^2}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right). \tag{2.51}$$

2.4.3 Condiciones bajo las cuales es más precisa la estimación de razón que la obtenida por un MAS

En muestras grandes, con muestreo aleatorio simple, la estimación de razón $\bar{y}_R$ tiene una varianza menor que la estimación $\bar{y} = N\bar{y}$ obtenida por MAS sí:

$$\rho > \frac{1}{2}\frac{\left(\frac{S_x}{\bar{X}}\right)}{\left(\frac{S_y}{\bar{Y}}\right)}. \tag{2.52}$$

Es decir, el coeficiente de correlación es mayor que la mitad del cociente de los coeficientes de variación de x_i y y_i.

2.4.4 Estimaciones de razón en muestreo aleatorio estratificado

Cuando la población está estratificada, existen dos alternativas para estimar la razón, la primera es mediante el *estimador combinado* y la segunda mediante el *estimador separado*.

A grandes rasgos, el primero estima la razón a partir de los cocientes de los estimadores de la media para las x_i y y_i obtenidos de manera independiente y el segundo, estima la razón en cada uno de los estratos y por tanto el estimador resultante es obtenido de ponderar (según la afijación) los diferentes estimadores de los estratos manejados.

Se explicará posteriormente con mayor detalle a qué se refiere el párrafo anterior.

2.4.4.1 Estimador combinado

El estimador combinado $\hat{R}_C$ surge a partir del cociente de los estimadores de cada una de las variables obtenidas de manera independiente, así:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{2.53}$$

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h \tag{2.54}$$

$$\hat{R}_C = \frac{\bar{y}_{st}}{\bar{x}_{st}} \tag{2.55}$$

Varianza

Se puede aproximar con el supuesto de igualdad entre $\bar{x}_{st} = \bar{X}$, así, el estimador de la varianza según el caso se conforma de la siguiente forma[10]:

Estimador de razón combinado

$$\hat{V}(\hat{R}_C) = \frac{1}{\bar{X}^2}\left(V(\bar{y}_{st}) + \hat{R}^2 V(\bar{x}_{st}) - 2\hat{R}Cov(\bar{y}_{st}, \bar{x}_{st})\right) = \frac{1}{\bar{X}^2 N^2}\left(\sum_{i=1}^{n} \frac{N_h^2 s_{dh}^2}{n_h} - \sum_{i=1}^{n} N_h s_{dh}^2\right). \tag{2.56}$$

[10] Ver Sánchez Villareal Francisco, 2004: 70

Estimador del total combinado

$$\hat{V}(\hat{Y}_{RC}) = \sum_{i=1}^{n} \frac{N_h^2 s_{dh}^2}{n_h} - \sum_{i=1}^{n} N_h s_{dh}^2 \text{, donde } \hat{Y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X \text{ .} \quad (2.57)$$

Estimador de la media combinado

$$\hat{V}(\bar{y}_{RC}) = \frac{1}{N^2}\left(\sum_{i=1}^{n} \frac{N_h^2 s_{dh}^2}{n_h} - \sum_{i=1}^{n} N_h s_{dh}^2 \right) \text{, donde } \bar{y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} \text{ .} \quad (2.58)$$

Varianza bajo afijación de Neyman

Estimador de razón

$$\hat{V}(\hat{R}_{Cney}) = \frac{1}{n\bar{X}^2 N^2}\left(\sum_{h=1}^{L} N_h s_{dh} \right)^2 - \frac{1}{\bar{X}^2 N^2} \sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \hat{R}_{Cney} = \frac{\bar{y}_{ney}}{\bar{x}_{ney}} \text{.} \quad (2.59)$$

Estimador de la media

$$\hat{V}(\bar{y}_{RCney}) = \frac{1}{nN^2}\left(\sum_{h=1}^{L} N_h s_{dh} \right)^2 - \frac{1}{N^2} \sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \bar{y}_{RCney} = \frac{\bar{y}_{ney}}{\bar{x}_{ney}} \bar{X} \text{ .} \quad (2.60)$$

Estimador del total

$$\hat{V}(\hat{Y}_{RCney}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h s_{dh} \right)^2 - \sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \hat{Y}_{RCney} = \frac{\bar{y}_{ney}}{\bar{x}_{ney}} X \text{ .} \quad (2.61)$$

Varianza bajo afijación de proporcional

Estimador de razón

$$\hat{V}(\hat{R}_{C\,prop}) = \frac{1}{n\bar{X}^2 N}\sum_{h=1}^{L} N_h s_{dh}^2 - \frac{1}{\bar{X}^2 N^2}\sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \hat{R}_{Cprop} = \frac{\bar{y}_{prop}}{\bar{x}_{prop}} . \qquad (2.62)$$

Estimador de la media

$$\hat{V}(\bar{y}_{RC\,prop}) = \frac{1}{nN}\sum_{h=1}^{L} N_h s_{dh}^2 - \frac{1}{N^2}\sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \bar{y}_{RCprop} = \frac{\bar{y}_{prop}}{\bar{x}_{prop}}\bar{X} . \qquad (2.63)$$

Estimador del total

$$\hat{V}(\hat{Y}_{RC\,prop}) = \frac{N}{n}\sum_{h=1}^{L} N_h s_{dh}^2 - \sum_{h=1}^{L} N_h s_{dh}^2 \text{, donde } \hat{Y}_{RCprop} = \frac{\bar{y}_{prop}}{\bar{x}_{prop}} X . \qquad (2.64)$$

Tamaño de muestra

Invariablemente, para el cálculo del tamaño de muestra nos apoyaremos de la siguiente relación al igual en las anteriores explicaciones para este mismo concepto:

$$ee(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})} \quad . \qquad (2.65)$$

De esta forma, los tamaños de muestra resultantes serán determinados de la siguiente manera para cada una de las afijaciones correspondientes:

Tamaño de muestra bajo afijación de Neyman

El tamaño de muestra dentro de cada estrato estará dado de la siguiente forma,

$$n_h = n\frac{N_h S_{dh}}{\sum_{h=1}^{L} N_h S_{dh}}. \tag{2.66}$$

Y el tamaño de muestra total será calculado como sigue en función del estimador usado:

Estimador de la razón

$$n = \frac{\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2}{ee^2(\hat{R}_{Cney})\overline{X}^2 N^2 + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.67}$$

Estimador de la media

$$n = \frac{\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2}{ee^2(\bar{y}_{RCney})N^2 + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.68}$$

Estimador del total

$$n = \frac{\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2}{ee^2(\hat{Y}_{RCney}) + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.69}$$

Tamaño de muestra bajo afijación proporcional

En el caso de esta afijación, el tamaño de muestra en cada estrato se calculará de la misma forma a lo que se vio en MAE:

$$n_h = n\frac{N_h}{\sum_{h=1}^{L} N_h}. \tag{2.70}$$

Y el tamaño de muestra total será calculado con las siguientes expresiones según el estimador usado:

Estimador de razón

$$n = \frac{N\sum_{h=1}^{L} N_h S_{dh}^2}{ee^2(\hat{R}_{Cprop})\bar{X}^2 N + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.71}$$

Estimador de la media

$$n = \frac{N\sum_{h=1}^{L} N_h S_{dh}^2}{ee^2(\bar{y}_{RCprop})N^2 + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.72}$$

Estimador del total

$$n = \frac{N\sum_{h=1}^{L} N_h S_{dh}^2}{ee^2(\hat{Y}_{RCprop}) + \sum_{h=1}^{L} N_h S_{dh}^2}. \tag{2.73}$$

2.4.4.2 Estimador separado

El estimador separado de razón, se obtiene como la suma ponderada de las estimaciones en los estratos, es decir, se realiza el cálculo de la razón dentro de cada uno y para obtener el

estimador global se ponderan las razones de cada estrato, de esta forma, los estimadores se expresan de la siguiente forma:

Estimador separado de razón

$$\hat{R}_S = \sum_{h=1}^{L} W_h \hat{R}_h \,, \tag{2.74}$$

donde $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$.

Así, la estimación de la media bajo un estimador separado de razón es:

$$\bar{y}_{RS} = \sum_{h=1}^{L} W_h \bar{y}_{RS_h} = \sum_{h=1}^{L} W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h \tag{2.75}$$

Para el cálculo los varianzas y tamaños de muestra para las diferentes afijaciones de muestra, se pueden usar las mismas expresiones que fueron mencionadas en la explicación del estimador combinado, solamente que S_{dh}^2 tendrá como $\hat{R}_h$ el estimador del estrato y no el estimador de la población, a diferencia del combinado, donde $R_h = \frac{\hat{Y}_h}{\hat{X}_h}$.

3 APLICACIÓN PRÁCTICA

Dentro de este apartado se explicará el ejercicio práctico, las consideraciones que se tuvieron sobre el mismo y los resultados obtenidos.

Recordando el objetivo principal, es necesario realizar la selección de una muestra para representar la información de una cadena detallista que actualmente colabora con Nielsen de manera censal, es decir, proporcionando información de todos sus negocios, pero no desea que su información sea develada totalmente. Así, a partir de la selección de muestra propuesta, ésta pasará por un proceso de autorización del detallista para proceder a enviar solamente la información mínima requerida.

Como consideraciones adicionales se tienen los siguientes elementos que fueron tomados en cuenta durante todo el desarrollo:

- Confidencialidad de la información

    - No es permitido develar información de los colaboradores y para ello existen ciertas reglas entre Nielsen y los colaboradores para proteger los intereses de estos. Por esta razón, no puede ser mencionado en este trabajo el nombre de la cadena analizada así como las zonas donde está presente

- Fracción de muestreo

    - Generalmente los universos de las cadenas detallistas son pequeños, esto redunda en que la fracción de muestreo necesaria en ocasiones pueda ser grande (mayor a 0.5) y los detallistas son conscientes de ello, sin embargo no aceptan fracciones de muestreo mayor a 0.6 para representar correctamente la información de la cadena.

- Error estándar relativo (*eer*)

    - Tanto para clientes detallistas como fabricantes, es garantizado un *eer* de 0.1 en el reporte de las marcas importantes.
    - De esta forma, la precisión deseada depende del nivel de confianza con el que se trabaje, generalmente es manejado un 90%, es decir, en la conformación de intervalos de confianza, el cuantíl de la normal estándar toma el valor de 1.64 ($Z_{1-\frac{\alpha}{2}} = 1.64$).

- Marcas importantes

    - Se refiere a aquellas que cuentan con dos características principales:

        1. Están presentes en más del 80% de los negocios del mercado de reporte
        2. Representan más del 5% de las ventas dentro del mercado-categoría de producto en el que se desempeñan

- Mercados de reporte

    - La muestra seleccionada debe representar correctamente la información de las marcas importantes, con el *eer* establecido para la variable ventas valor ($ pesos) en los siguientes mercados de reporte:

        - Las 6 Áreas Nielsen (consiste en división geográfica de la república Méxicana realizada por Nielsen)

        - Total México.

        - Debido a que la cadena trabajada no está presente en tres de las áreas Nielsen y debido a que no se puede develar el nombre de la cadena por

principios de confidencialidad, para fines prácticos se crearon los siguientes cuatro mercados de reporte sobre los cuales se tratará en los siguientes apartados donde se explicará el desarrollo del análisis:

1. Cadena_CM - Area Nielsen A
2. Cadena_CM - Area Nielsen B
3. Cadena_CM - Area Nielsen C
4. Cadena_CM - Total México

3.1 DESARROLLO DEL ANÁLISIS

Parte de la complejidad de este ejercicio, radica en la necesidad de representar adecuadamente cada una de las marcas importantes en cada mercado de reporte con una misma muestra, esto es, se necesita encontrar una muestra que represente adecuadamente a todas.

El diseño muestral aplicado consistió en muestreo aleatorio estratificado con estimador separado de razón y afijación de Neyman. Cabe responder las siguientes preguntas para explicar porqué se trabajó el diseño de esta forma:

¿Porqué utilizar MAE?

Cómo se ha observado en la parte teórica expuesta en capítulos anteriores, el MAE tiene ganancias importantes en precisión vs. MAS. Da la ventaja de representar de manera independiente las distintas dinámicas comerciales que pueden darse en diferentes regiones del país en cuanto a las ventas de los negocios que se están midiendo.

¿Qué criterio de estratificación se usó?

Básicamente se usó un criterio de cercanía de los negocios para definir estratos bajo el supuesto de dinámicas comerciales diferentes en cada una de las regiones y semejantes al interior de cada región. Adicionalmente se realizó una partición adicional tomando en cuenta la

variable ventas totales[11] para distinguir aún más los establecimientos en función del tamaño de sus ventas. Esto último de trabajó mediante el proceso de particiones sucesivas.

¿Porqué usar el estimador de razón separado?

El estimador de razón es usado para obtener una ganancia en precisión adicional. Se cuenta con una variable conocida que son las ventas totales de cada uno de los negocios y esta información es usada como variable auxiliar. Adicionalmente se hizo uso del estimador separado debido a que el comportamiento de la razón en cada estrato es diferente y por tanto da mejores resultados aplicarlo así.

Es importante señalar que una condición que debe tener la variable auxiliar es que debe estar fuertemente correlacionada con las variables que se desean medir. A este respecto, la variable de ventas totales de cada negocio muestra una fuerte correlación con los volúmenes desplazados a nivel categoría y producto. En la mayor parte de los casos, esta correlación es mayor a 0.8.

¿Porqué usar afijación de Neyman?

Haciendo referencia a la estratificación usada, una característica importante que se observó en los negocios es que frecuentemente llegan a ser diferentes en cuanto a los volúmenes de ventas, esta característica frecuentemente incrementa la variabilidad dentro de estratos y por tanto la necesidad de un mayor tamaño de muestra para lograr una buena representación. Un primer paso para mejorar el tratamiento estos casos fue estratificar por volumen de ventas mediante la técnica de particiones sucesivas. Para la asignación de tamaños de muestra fue relevante lograrla en función de la variabilidad y número de tiendas existentes en los estratos finales, se mostrará que esto da una ganancia importante en necesidades de muestra.

[11] Se refiere a las ventas en pesos que tuvo cada uno de los establecimientos en una semana en particular.

3.2 DESCRIPCIÓN DEL PROCESO

3.2.1 Marco muestral

Ya que actualmente la cadena colabora con Nielsen proporcionando el censo de negocios, se cuenta con un marco de tiendas que contiene la totalidad de establecimientos manejados por el detallista en análisis, cada uno con su respectiva ubicación geográfica en términos de estado, municipio, ciudad Nielsen[12] y por tanto Área Nielsen.

Los universos que maneja la cadena están dados de acuerdo a la siguiente tabla por área Nielsen:

Universos por área Nielsen.

Area	Total
A	144
B	442
C	92
Total	678

Tabla 3.1

3.2.2 Estratificación

Como primer criterio de estratificación se usó la ciudad Nielsen para realizar una agrupación "natural" en función de la plaza o región dentro de la cual se desempeñan cada uno de los negocios, adicionalmente se tuvieron las siguientes consideraciones y criterios:

- Los municipios que no pertenecen a una zona metropolitana, es decir, una ciudad Nielsen, son asociados a una región llamada Resto Área Nielsen. Para los negocios en esta situación

[12] Ciudad Nielsen corresponde a zona metropolitana (suma de municipios) con alta densidad poblacional y por tanto comercial.

se genera una estratificación a nivel Estado-Área Nielsen, con la finalidad de realizar una mejor agrupación de estos negocios.

- Para todos los estratos resultantes, ya sea de agrupar por ciudad Nielsen (áreas metropolitanas) o Estados (Resto Área Nielsen) se aplica un proceso de particiones sucesivas para crear estratos que al interior sean aún más homogéneos en términos de niveles de ventas. Al aplicar este procedimiento, los estratos candidatos a ser separados son aquellos que al interior manifiestan un coeficiente de variación mayor al 30% (CV>30%).

- Se anexa listado mostrando el estrato inicial y final después de aplicar el proceso de particiones sucesivas. Los CV resultaron disminuidos y esto ayuda de manera positiva en las necesidades de muestra.

Coeficiente de variación por estrato (inicial y final)

Area Nielsen	Estrato Inicial	Estrato_final	Media	S	N	CV
AREA A	Est_10	Est_1000	32259	7063	27	22%
		Est_1001	55281	7139	16	13%
	Est_10 Total		**40825**	**13262**	**43**	**32%**
	Est_11	Est_1100	14075	2491	4	18%
		Est_1101	35624	9122	3	26%
	Est_11 Total		**23310**	**12787**	**7**	**55%**
	Est_12	Est_120000	13257	4427	6	33%
		Est_120001	29863	3588	14	12%
		Est_1201	47704	5482	7	11%
	Est_12 Total		**30798**	**12867**	**27**	**42%**
	Est_13	Est_1300	26369	7846	13	30%
		Est_1301	51071	12259	15	24%
	Est_13 Total		**39602**	**16207**	**28**	**41%**
	Est_14	Est_140000	13397	3850	7	29%
		Est_140001	29167	5222	5	18%
		Est_1401	70522	0	1	0%
	Est_14 Total		**23857**	**16539**	**13**	**69%**
	Est_15	Est_15	28732	4955	4	17%
	Est_15 Total		**28732**	**4955**	**4**	**17%**
	Est_16	Est_160000	12094	0	1	0%
		Est_160001	21319	0	1	0%
		Est_1601	39224	1284	3	3%
	Est_16 Total		**30217**	**12790**	**5**	**42%**
	Est_17	Est_1700	21099	5389	6	26%
		Est_1701	39354	6256	8	16%
	Est_17 Total		**31530**	**10960**	**14**	**35%**
AREA B	Est_20	Est_2000	22504	6383	64	28%
		Est_2001	42546	8240	61	19%
	Est_20 Total		**32284**	**12439**	**125**	**39%**
	Est_21	Est_210000	16045	4387	59	27%
		Est_210001	26993	3793	65	14%
		Est_210100	44289	6868	42	16%
		Est_210101	86212	19780	5	23%
	Est_21 Total		**29195**	**15695**	**171**	**54%**
	Est_22	Est_2200	18180	4730	5	26%
		Est_2201	31889	3688	5	12%
	Est_22 Total		**25034**	**8258**	**10**	**33%**
	Est_23	Est_23	30351	8304	3	27%
	Est_23 Total		**30351**	**8304**	**3**	**27%**
	Est_24	Est_2400	23194	5763	21	25%
		Est_2401	40326	7532	9	19%
	Est_24 Total		**28334**	**10115**	**30**	**36%**
	Est_25	Est_2500	11373	1595	7	14%
		Est_2501	21100	3322	13	16%
	Est_25 Total		**17695**	**5516**	**20**	**31%**
	Est_26	Est_26	16753	3806	5	23%
	Est_26 Total		**16753**	**3806**	**5**	**23%**
	Est_27	Est_270000	17106	4137	6	24%
		Est_270001	31376	3272	8	10%
		Est_2701	50053	8604	7	17%
	Est_27 Total		**33524**	**14442**	**21**	**43%**
	Est_28	Est_2800	15247	4096	15	27%
		Est_2801	29244	4062	7	14%
	Est_28 Total		**19701**	**7773**	**22**	**39%**
	Est_29	Est_2900	18765	5277	15	28%
		Est_2901	34720	7695	7	22%
	Est_29 Total		**23842**	**9661**	**22**	**41%**
	Est_80	Est_8000	20637	4300	10	21%
		Est_8001	47291	2892	2	6%
	Est_80 Total		**25079**	**11114**	**12**	**44%**
AREA C	Est_30	Est_3000	17920	3773	2	21%
		Est_3001	34694	2691	2	8%
	Est_30 Total		**26307**	**10047**	**4**	**38%**
	Est_31	Est_3100	23760	5037	42	21%
		Est_3101	42137	7677	17	18%
	Est_31 Total		**29055**	**10230**	**59**	**35%**
	Est_32	Est_32	22496	6722	13	30%
	Est_32 Total		**22496**	**6722**	**13**	**30%**
	Est_33	Est_3300	20470	4066	8	20%
		Est_3301	34684	6054	6	17%
	Est_33 Total		**26561**	**8734**	**14**	**33%**
	Est_35	Est_35	37364	0	1	0%
	Est_35 Total		**37364**	**0**	**1**	**0%**

Tabla 3.2

3.2.3 Cálculos de tamaño de muestra

3.2.3.1 Identificación de marcas importantes

Para realizar el cálculo de tamaño de muestra, se identificó inicialmente si cada una de las marcas participantes en cada categoría de producto eran o no importantes en cada mercado de reporte. Esto es, para calcular un tamaño de muestra que resulte de generalizar los diferentes tamaños encontrados para estas marcas ya que en las mismas Nielsen tiene el compromiso de reportar con una precisión establecida.

Resumen de marcas importantes por área Nielsen

Area_nielsen	%Num Marcas	%Ventas
A	15%	59%
B	13%	57%
C	16%	60%
Total México	14%	58%

Tabla 3.3

En el cuadro anterior puede observarse la importancia de las marcas seleccionadas en dos rubros, en número de marcas y en ventas, que en promedio representan un 58% de las ventas, no así en el de número de marcas dado que están alrededor del 14%, siendo este un porcentaje bajo, es decir, son pocas marcas que acumulan un gran volumen de ventas.

3.2.3.2 Calculo de tamaño de muestra por marca-mercado

Después de identificar si la marca es o no importante dentro del mercado de reporte se realizó el cálculo de tamaño de muestra para cada marca de cada categoría de producto en cada mercado, apoyados de la siguiente relación usando el estimador separado de razón con afijación de Neyman.

$n = \dfrac{\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2}{EE^2(\hat{Y}_{RSney}) + \sum_{h=1}^{L} N_h S_{dh}^2}$, donde S_{dh} es la estimación de la desviación estándar en cada uno de los estratos. Al ser un estimador separado de razón entonces el cálculo está dado de manera diferenciada en cada uno de los estratos.

En el siguiente cuadro se puede observar un ejemplo de los cálculo de tamaño de muestra mencionados (TM). Por señalar un caso, se nota que para la marca 6 de la categoría de pañales en el mercado "A" el tamaño de muestra es considerablemente menor (TM=13) que para el resto de casos. Nótese que es la marca más importante ya que tiene un "share" de 53%. También tiene una muy buena distribución de 92%. En contraparte, la marca 4 que no pertenece el conjunto de relevantes, tiene un "share" y una distribución muy baja, sin embargo requiere un tamaño de muestra muy grande (TM=52).

Ejemplo: cálculo de tamaño de muestra a nivel marca

MARCA	CATEGORÍA	MERCADO	DII	"SHARE"	TM	MARCA IMPORTANTE
MARCA 6	PAÑALES	1	92%	53%	13	1
MARCA 7	PAÑALES	1	93%	24%	20	1
MARCA 5	PAÑALES	1	91%	6%	21	1
MARCA 2	PAÑALES	1	90%	5%	24	1
MARCA 1	PAÑALES	1	83%	2%	25	0
MARCA 3	PAÑALES	1	66%	0%	39	0
MARCA 8	PAÑALES	1	29%	0%	38	0
MARCA 4	PAÑALES	1	35%	0%	52	0

Tabla 3.4

3.2.3.3 Generalización del tamaño de muestra

Al tener los tamaños de muestra para cada una de las marca-mercados se realizó un proceso de generalización. En una estrategia inicial, fue contemplado que bastaba con obtener los valores máximos por estrato, sin embargo se observó que en ocasiones resultaban en fracciones de muestreo muy elevadas debido a casos especiales de marcas que requerían de un tamaño de muestra radicalmente mayor a las del promedio. Ante esta situación, se optó

por generar un intervalo de confianza al 95% con los distintos tamaños de muestra resultantes en cada uno de los estratos. Esto permitió eliminar los valores atípicos que generalmente eran muy cercanos al valor del universo en cada estrato. De esta forma, se tomó el límite superior (LS) del intervalo de confianza dado por:

$$LS_h = \bar{y}_{TM_h} + (Z_{1-\frac{\alpha}{2}} * {}^{s_{TM_h}}\!/\!\sqrt{k_h}) . \qquad (3.1)$$

Donde,

k_h : total de tamaños de muestra calculados para el estrato h.

TM_{ih} : i-ésimo tamaño de muestra calculado para el estrato i.

$\bar{y}_{TM_h} = \sum_{i=1}^{k_h} TM_{ih} / k_h$: estimador de la media del tamaño de muestra necesario en cada estrato h.

s_{TM_h} : desviación estándar observada para los tamaño de muestra calculados en cada estrato.

Alternativamente se realizaron diferentes ejercicios probando tamaños de muestra resultantes variando los errores estándar, las afijaciones y la estratificación, los resultados de estos comparativos son los siguientes:

Comparativo de tamaños de muestra

		CON PARTICIONES SUCESIVAS				SIN PARTICIONES SUCESIVAS			
		Afijación Neyman		Afijación Proporcional		Afijación Neyman		Afijación Proporcional	
Area Nielsen	EER	Est Razón	Est Simple	Est Razón	Est Simple	Est Razón	Est Simple	Est Razón	Est Simple
A	2.5%	102	106	131	133	114	118	132	135
B	2.5%	276	296	370	376	330	342	374	378
C	2.5%	78	78	87	88	82	82	88	88
A	5%	77	83	111	113	94	100	111	119
B	5%	184	197	254	266	223	241	262	270
C	5%	62	64	78	80	67	71	79	81
A	10%	39	45	82	84	56	67	83	89
B	10%	79	88	113	123	97	111	120	127
C	10%	42	46	56	60	46	51	56	61
A	15%	23	28	63	65	34	44	63	69
B	15%	40	46	59	65	50	59	63	67
C	15%	28	32	38	42	31	35	38	44
A	20%	15	18	49	51	22	29	49	54
B	20%	24	27	35	39	30	35	38	41
C	20%	19	22	27	30	22	25	27	31

Tabla 3.5

En el siguiente cuadro puede verse la ganancia derivada del proceso de realizar una mejor estratificación aplicando el procedimiento de particiones sucesivas.

Ganancia por particiones de estratos

		Afijación Neyman		Afijación Proporcional	
Area Nielsen	EER	Est Razón	Est Simple	Est Razón	Est Simple
A	2.5%	11%	10%	1%	1%
B	2.5%	16%	13%	1%	1%
C	2.5%	5%	5%	1%	0%
A	5%	18%	17%	0%	5%
B	5%	17%	18%	3%	1%
C	5%	7%	10%	1%	1%
A	10%	30%	33%	1%	6%
B	10%	19%	21%	6%	3%
C	10%	9%	10%	0%	2%
A	15%	32%	36%	0%	6%
B	15%	20%	22%	6%	3%
C	15%	10%	9%	0%	5%
A	20%	32%	38%	0%	6%
B	20%	20%	23%	8%	5%
C	20%	14%	12%	0%	3%

Tabla 3.6

Es de notar que la ganancia con afijación proporcional es realmente muy pequeña, no así en afijación de Neyman donde los niveles son importantes y con un número menor de tiendas puede obtenerse la misma precisión por el simple hecho de partir los estratos reduciendo la variabilidad en los mismos.

Comparando los tamaños de muestra obtenidos usando el estimador simple y el de razón se tiene que:

Ganancia por usar estimador de razón

Area Nielsen	EER	Neyman	Prop
A	2.5%	4%	2%
B	2.5%	7%	2%
C	2.5%	0%	1%
A	5%	7%	2%
B	5%	7%	5%
C	5%	3%	3%
A	10%	13%	2%
B	10%	10%	8%
C	10%	9%	7%
A	15%	18%	3%
B	15%	13%	9%
C	15%	13%	10%
A	20%	17%	4%
B	20%	11%	10%
C	20%	14%	10%

Tabla 3.7

Notar que la ganancia realmente parecen mínimas cuando los EER son pequeños, sin embargo, siempre es una ganancia positiva, lo cual indica que el estimador de razón es mejor opción que el simple. Más adelante se mostrará el comportamiento de los errores relativos que resultan de comparar el uso de la muestra seleccionada vs. lo que realmente manejó la cadena, este análisis ayudará a corroborar si es acertado este comentario.

Finalmente, al valorar la ganancia obtenida por usar afijación de Neyman vs. afijación proporcional es muy importante corroborarlo para todas las áreas Nielsen y para todos los EER trabajados. Esto también puede asociarse como un efecto de mejorar la estratificación mediante el proceso de particiones sucesivas, precisamente, en los resultados anteriores se mostró que por el hecho de dividir los estratos no existía ganancia importante con la afijación proporcional, esto es, la partición tiene sentido si se aplica afijación de Neyman.

Ganancia por usar afijación de Neyman

Area Nielsen	EER	GANANCIA
A	2.5%	22%
B	2.5%	25%
C	2.5%	10%
A	5%	31%
B	5%	28%
C	5%	21%
A	10%	52%
B	10%	30%
C	10%	25%
A	15%	63%
B	15%	32%
C	15%	26%
A	20%	69%
B	20%	31%
C	20%	30%

Tabla 3.8

3.2.3.4 Determinación del tamaño de muestra final

A partir del proceso aplicado de generalización fue posible obtener de manera inmediata el tamaño de muestra requerido para cada uno de los estratos, entendiendo que estarían abarcadas en cada caso aproximadamente el 95% de marcas al tomar el intervalo de confianza señalado. Dentro del proceso para cálculo de tamaños de muestra se asignaron algunas consideraciones en la aplicación que se encarga de calcularlos para garantizar que estratos que inicialmente tenían asignado un tamaño de muestra cero por no tener variabilidad tuvieran una muestra asignada, al menos un negocio. Así también, si el tamaño de muestra excedía el de la población total del estrato entonces de manera forzada el estrato toma como cálculo final sugerido el universo. Así, después de analizar los cuadros mostrados anteriormente y las tomando en cuenta consideraciones señaladas, el tamaño de muestra obtenido por EER es el siguiente:

Tamaños de muestra finales

Area Nielsen	EER	n
A	2.5%	102
A	5%	77
A	10%	39
A	15%	23
A	20%	15
B	2.5%	276
B	5%	184
B	10%	79
B	15%	40
B	20%	24
C	2.5%	78
C	5%	62
C	10%	42
C	15%	28
C	20%	19
TOTAL	2.5%	456
TOTAL	5%	323
TOTAL	10%	160
TOTAL	15%	91
TOTAL	20%	58

Tabla 3.9

Analizando el comportamiento de las fracciones de muestreo por mercado y en función del EER, es de notar que en casi todos los mercados de reporte con un EER de 5% se tiene una fracción del muestreo por abajo del 60%.

En el mercado C es donde la fracción de muestreo es mayor, este efecto se justifica porque tiene el universo más pequeño, caso contrario al mercado B que cuenta con el más grande.

La restricción en la fracción de muestreo se da en el mercado de referencia de total cadena, en este caso, con un EER del 5% la fracción de muestreo está muy por debajo del tope establecido, por esta razón, es bastante aceptable que se determine el tamaño de muestra final como el resultante después de fijar el EER de 5%.

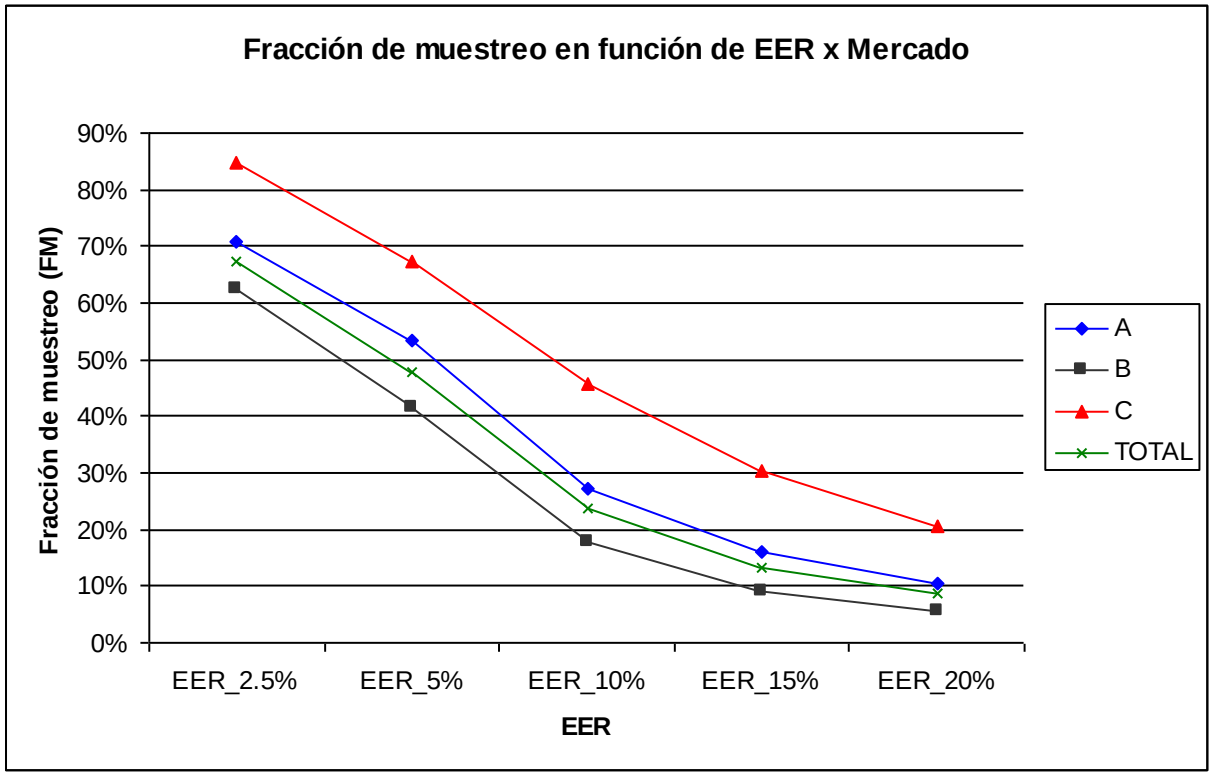


Figura 3.1

Así, extrayendo los tamaños de muestra definitivos son:

Tamaños de muestra definitivos

Area Nielsen	n
A	77
B	184
C	62
TOTAL	323

Tabla 3.10

3.2.4 Selección de la muestra

En SAS fue aplicado un procedimiento para realizar la selección aleatoria por estrato cumpliendo con todos los tamaños de muestra requeridos en los cálculos previos.

3.2.4.1 Pruebas sobre la muestra seleccionada

Con la finalidad de probar la eficiencia y la precisión de la muestra seleccionada, se realizó un ejercicio que consiste en simular cuál sería el resultado de inferir los totales por marca-mercado usando solamente la información disponible de la muestra seleccionada, a partir del estimador de razón y también usando el estimador estratificado simple para comparar los resultados.

A partir de dicha inferencia, se analizó la distribución de errores en las marcas principales y se compararon los resultados del estimador de razón y el estimador estratificado simple, sólo con la finalidad de ilustrar gráficamente el comportamiento.

Así, los resultados de la distribución de frecuencias de errores relativos es la siguiente:

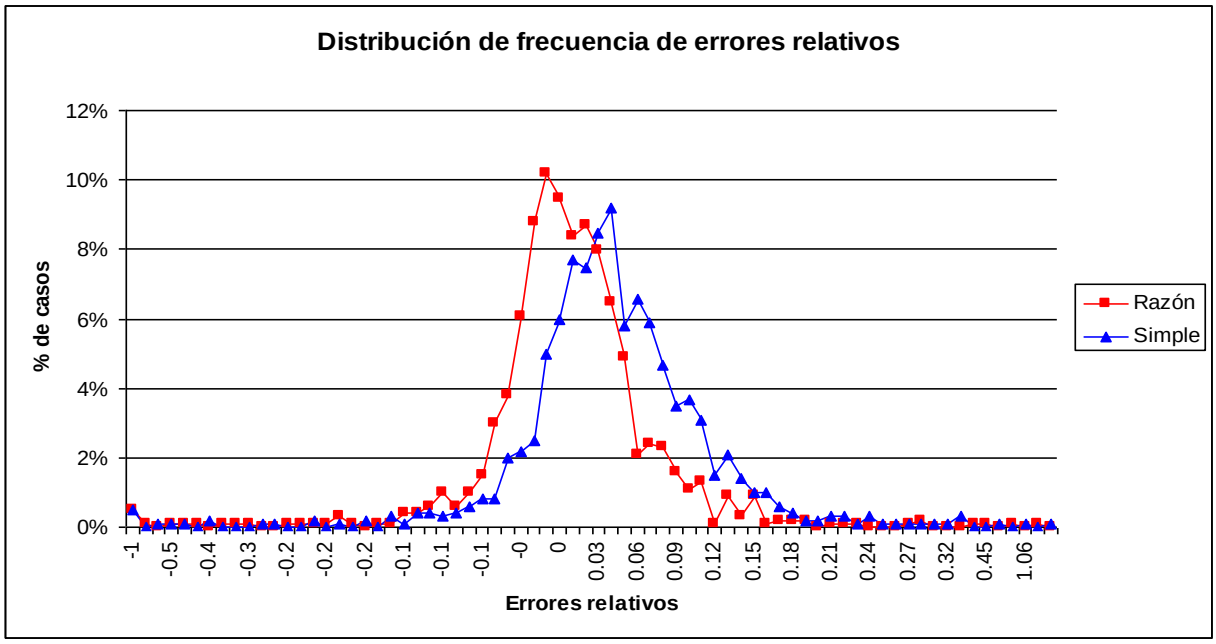


Figura 3.2

Comparando los estimadores, puede observarse que el simple parece un poco más sesgado hacia valores positivos y en contraparte, el de razón manifiesta una distribución concentrada fuertemente en los valores mínimos.

Analizando los resultados en términos absolutos, los errores se representan gráficamente la siguiente forma:

Figura 3.3

Por otro lado, en términos acumulados el comportamiento de los mismos es el siguiente:

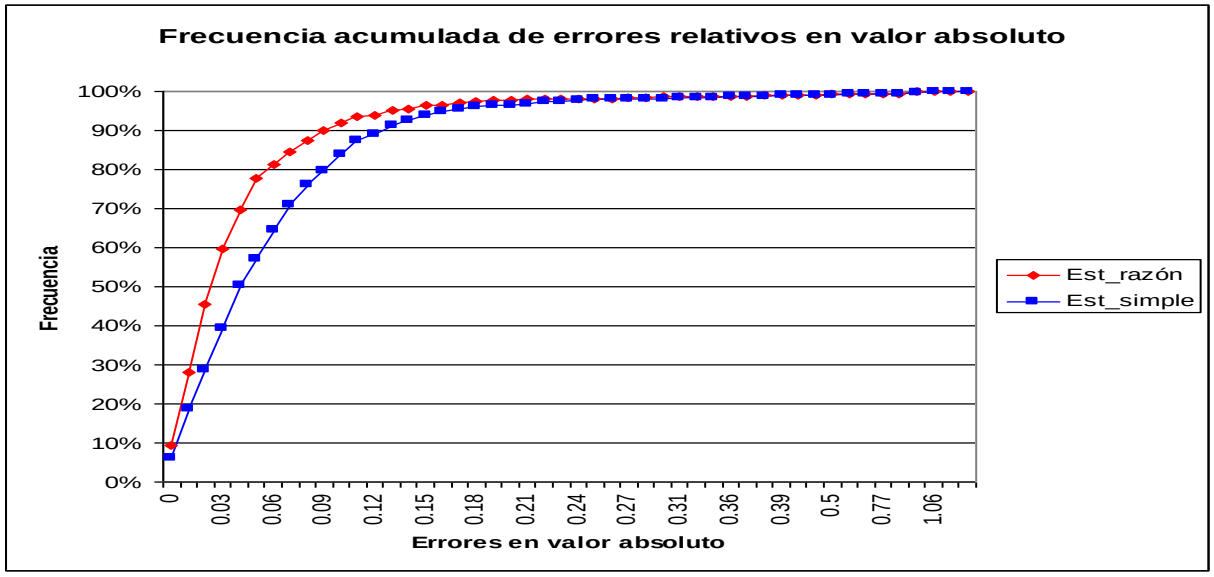


Figura 3.4

Notar que el porcentaje de marcas que alcanzan un error menor o igual al 10% representan aproximadamente el 92% como se plasma en el siguiente gráfico:

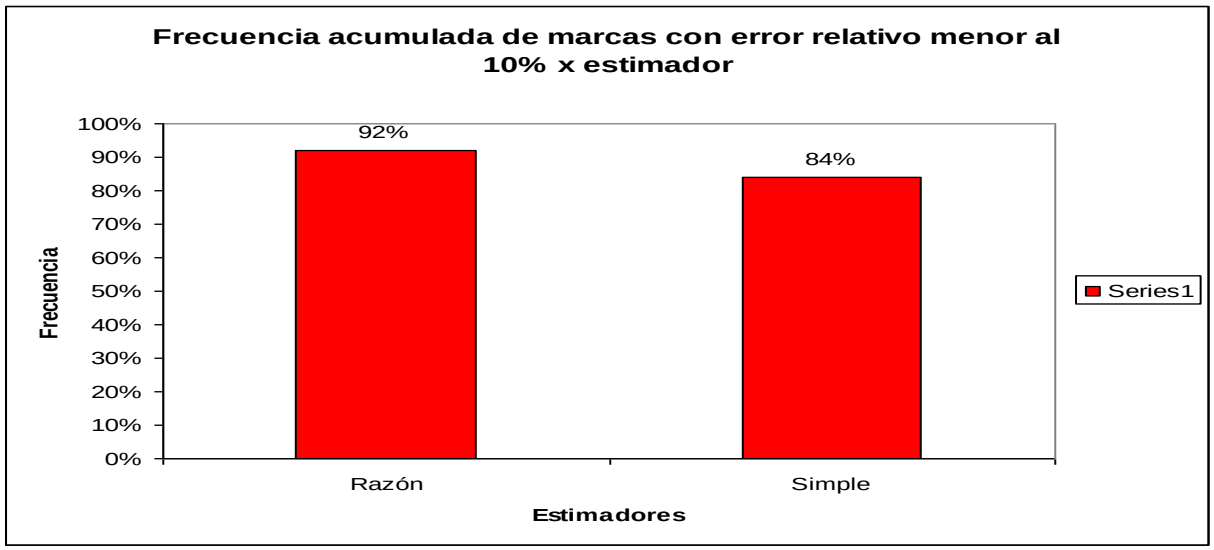


Figura 3.5

3.2.4.2 Mantenimiento de variable auxiliar

La variable auxiliar son las ventas totales de cada uno de los negocios, así, los estimadores para cada una de las marcas mercados son: $\hat{Y}_{hi} = \hat{R}_{hi} * \hat{X}_h = \frac{\bar{y}_{hi}}{\bar{x}_h} * \hat{X}_h$, es decir, las ventas totales de la marca i en el estrato h estarán representadas por el cociente entre la estimación del promedio de ventas de la marca i en el estrato h y la estimación de la media de las ventas totales en el estrato h. Lo anterior multiplicado por la estimación de las ventas totales en el estrato h.

Para estimaciones de los totales de ventas por negocio, no es necesario contar con esta información correspondiente a la misma semana que se está procesando, esto puede provenir de información previa, haciendo uso también del estimador de razón se puede obtener una estimación de las ventas totales del periodo actual *t* para cada estrato *h*.

$\hat{X}_{h_t} = \frac{\bar{x}_{h_t}}{\bar{x}_{h_{t-1}}} * \hat{X}_{h_{t-1}}$ y a partir de la información de periodos previos pueden inferirse los totales

De esta forma:

$\hat{X}_{h_t} = \frac{\bar{x}_{h_t}}{\bar{x}_{h_{t-1}}} * \hat{X}_{h_{t-1}}$ y a partir de la información de periodos previos pueden inferirse los totales correspondientes al periodo actual. En este caso, la correlación es mucho más alta, en niveles aproximados de 90%.

CONCLUSIONES

El tamaño de muestra ofrecido al colaborador fue de 323 tiendas y representa una fracción de muestreo de 47.6%, la cual se encuentra por debajo de la cota establecida en los convenios de colaboración, obteniendo un EER de 5%, el cual inclusive garantiza que marcas que no cumplan los criterios de distribución numérica e importancia en ventas tendrán una estimación razonablemente buena.

Debe ser tomado en cuenta que la precisión establecida con los clientes fabricantes es de 10% y aplica para los mercados totalizadores por región donde el colaborador está participando junto con otro grupo de detallistas que generalmente comparten información censal, lo que se traduce en que dentro del mismo mercado totalizador la precisión será mayor.

Este resultado no hubiese sido posible si no se hubiera implementado el proceso de particiones sucesivas para mejorar la estratificación y la afijación de Neyman, ya que estos elementos van de la mano en la solución, siendo un elemento distintivo en la afijación seleccionada la asignación de muestra en función del universo pero sobretodo de la variabilidad al interior de cada estrato. De esta manera, al reducir considerablemente la variabilidad haciendo eficiente la estratificación, se reducen así también las necesidades de muestra.

Sin duda resultó interesante realizar los comparativos vs. la afijación proporcional ya que se pudo corroborar de manera muy tangible los beneficios teóricos de la solución generada.

Consideraciones adicionales: con la finalidad de dar un buen mantenimiento, es necesario que semana a semana se cuente con una actualización del universo para aplicarla dentro de los marcos muestrales disponibles.

Por otro lado, debe cuidarse que la muestra seleccionada siga siendo representativa del universo vigente. Para ellos se sugiere revisión al menos anual del diseño muestral para mantener vigente la proyección.

Posibles riesgos: si la cadena decide realizar prácticas de comercialización diferenciadas entre las tiendas de la muestra y la no muestra es altamente probable que se subestimen los fenómenos de las prácticas aplicadas en la no muestra y se sobreestimen los de la muestra.

Alternativa de solución del problema: aprovechando que antes de realizar el cambio de esquema de censo a muestra se tiene disponible toda la información del colaborador, puede abordarse el tema de otra forma, realizando por estrato un ejercicio iterativo de búsqueda del mejor y menor conjunto de tiendas tales que su extrapolación al universo en cada estrato brinda un error absoluto dentro de rangos aceptables. Esta sería una solución más bien computacional, sin embargo viable al considerar un conjunto reducido de unidades de muestreo.

BIBLIOGRAFÍA

1. Canavos C (1988). Probabilidad y Estadística: Aplicaciones y Métodos México: McGraw-Hill Interamericana.

2. Cochran G. (1998). Técnicas de muestreo (14ª reimp.). México: Compañía editorial continental.

3. Garza T. (1990). Elementos de cálculo de probabilidades. México: Dirección General de Publicaciones, UNAM.

4. Kish L. (1979). Muestreo de encuestas (2ª reimp.). México: Editorial Trillas.

5. Lohr L. (1999). Muestreo: Diseño y análisis. México: Internacional Thomson Editores.

6. Méndez I., Eslava G. y Romero P. (2004). Conceptos Básicos de Muestreo (Serie Monografías Vol 12, No 27). México: Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM

7. Mood M., Graybill A. y Boes C. (1974). Introduction to the theory of statistics (3ª ed.). Singapore: McGraw-Hill, Inc.

8. Pérez C. (2000). Técnicas de muestreo estadístico: Teoría, práctica y aplicaciones informáticas. México: Alfaomega grupo editor.

9. Sánchez Villareal F. (2004). Introducción al muestreo probabilístico. Pragma S.A de C.V.