MUESTREO ESTADÍSTICO PARA DOCENTES Y ESTUDIANTES

Primera Edición

ÁNGEL GÓMEZ DEGRAVES
KARINÉ GÓMEZ MARQUINA

2019

Índice General

Índice General……..…. 1
Índice de Tablas..…….. 2
Índice de Figuras…….……. 3
Introducción…….. 5
Sinopsis……..…….. 6
El Censo ……..…….. 8
Razones para utilizar Muestreo…….. 9
Conceptos y definiciones en Muestreo Estadístico…….…… 14
Marco de Muestreo…….. 17
Unidades Estadísticas……..…….. 18
Muestra Representativa…… 20
Margen de Error o Error de Estimación……. 22
Etapas de una Encuesta por Muestreo…….. 23
Tipos de Errores en las Encuestas por Muestreo…….. 24
Factores que Afectan el Diseño de la Muestra Probabilística…..…… 26
Fases de un Diseño de Muestreo…… 28
Introducción al Tamaño de la Muestra…….…….. 30
Muestreo Probabilístico……. 34
Muestreo Aleatorio Simple…… 36
Muestreo Aleatorio Estratificado…… 45
Muestreo Aleatorio Estratificado (Desproporcional)…… 54
Muestreo Sistemático…….. 59
Muestreo por Conglomerados……... 68
Tamaño del Conglomerado y la Correlación Intraclase…….. 72
Muestreo por Conglomerados en Dos Etapas…….. 80
Estimación en MPC (Monoetápico) con Probabilidades Proporcionales al Tamaño (PPT)……..……. 87

Generalidades del Muestreo Polietápico…… 89
Estimadores de Razón, Regresión y Diferencia en Muestreo Aleatorio simple…… 90
Muestreo No Probabilístico (MNP)…….. 102
Aspectos teórico-metodológicos clave del Muestreo Estadístico en las Investigaciones…….. 109

Índice de Tablas

Tabla 1. Notación de Muestreo Aleatorio Simple…… 37

Tabla 2. Notación en Muestreo Aleatorio Estratificado…… 46

Tabla 3. Datos Necesarios para el Cálculo de la Afijación en los Estratos…….. 55

Tabla 4. Cálculo de La Afijación de la Muestra en Cada Estrato……. 56

Tabla 5. Tiempo en Horas Que se ve Televisión por Semana…….. 57

Tabla 6. Resumen del Tiempo en Horas Que se ve Televisión Por Semana…….. 57

Tabla 7. Unidades Primarias y Secundarias en Muestreo por Conglomerados (MPC)…… 71

Tabla 8. Notación para Muestreo por Conglomerados de Tamaños Desiguales…….. 73

Índice de Figuras

Factores que Afectan el Diseño de la Muestra…… 26

Fases de un Diseño de Muestreo…… 29

Distribución Normal Estándar…….. 33

Relación Entre la Variable de Interés Y, y la Variable Auxiliar X…….. 99

Línea Recta con Pendiente…… 13

Tipos de Muestreo…… 103

Introducción

Al realizar cualquier estudio, sea observacional o experimental, cuantitativo o cualitativo, la mayor parte de las veces los investigadores no cuentan con la infraestructura financiera, técnica, ni de recurso humano para la enumeración completa de la población, y se opta por los métodos de muestreo probabilístico, no probabilístico o combinaciones de estos.

Es así, como cobra muchísima importancia el uso de estos métodos, con el fin de obtener información confiable de la población a partir de una muestra y es aquí, donde las investigaciones por muestreo, o el uso del muestreo en investigaciones, evidencia su alto potencial para realizar inferencias estadísticas con un margen de error medido en términos de probabilidades conocidas a priori.

La teoría del muestreo proporciona estrategias en relación a la selección de muestras y estimaciones de parámetros poblacionales que ofrezcan la mayor cantidad de información al menor coste.

Es la teoría de las técnicas para escoger o seleccionar muestras. Para formular Inferencias Inductivas con una medida de riesgo en términos de probabilidad, es necesario relacionar la población estadística con la muestra, donde se hace coincidir en lo posible la distribución de probabilidades de la población con la distribución de frecuencias de la muestra.

El muestreo, además de ser ciencia Estadística, es un arte, donde no solo los elementos se seleccionan al azar con una medida de probabilidad, en el Muestreo Probabilístico, sino que requiere pericia por parte de investigador en el diseño de la muestra, ya que debe considerar: Claridad en los objetivos del estudio, que la definición de la Población objetivo sea muy clara y dimensionada, obtener información a priori o auxiliar, la técnica de selección de elementos de investigación, la selección de los estimadores apropiados, elegir un tamaño adecuado de la muestra con una precisión (margen de error) y un nivel de

confianza, predefinido y acorde con la cantidad de recursos que el equipo investigador posee para realizar el estudio.

La teoría del **Muestreo Estadístico (ME)** nos permite hacer afirmaciones probables sobre los valores de las características de la población, llamadas parámetros, mediante el uso de técnicas estadísticas inferenciales, de tal manera que obtengamos una muestra aleatoria lo suficientemente representativa de esa población.

Muy importante es mencionar que en ME, la fuente de aleatoriedad proviene de la extracción de la muestra y para usarlo, no necesitamos conocer la Distribución de Probabilidades de la variable de interés a investigar, ventaja ésta muy importante, en comparación con otras técnicas estadísticas que se usan frecuentemente.

Lo que en realidad se persigue con la utilización del **ME** es simplemente, hacer inferencia sobre valores de la Población Objeto de estudio, llamados parámetros (θ), con un mínimo de variabilidad en los estimadores de esos parámetros. Es decir, que la estimación que procede del estimador esté lo más cerca posible del valor del parámetro. Esto se conoce como acuaricidad, y en muchos libros se le llama precisión.

Sinopsis

Es bien sabido que todo proceso de Investigación por encuestas, mediante censo o muestreo, se necesita de estadísticas confiables, se cuenta con un universo o colectivo, de entidades como son las empresas, instituciones, organizaciones, asociaciones, personas, objetos, procesos, donde no es posible realizar todas las mediciones en todas las entidades por razones de tiempo, recursos, coste, dificultad de acceso a la entidad, entre otros. Nos referimos en este caso mayormente a los estudios en las Ciencias Sociales y Humanas, donde en la actualidad se utiliza mucho las muestras complejas.

En el caso de las Ciencias Exactas, igualmente el muestreo tiene su rol vital, como en el caso de los estudios experimentales, donde se desea tener un determinado número de repeticiones o unidades experimentales, que permita hacer inferencias, bien sea estimar parámetros o contrastar hipótesis sobre los valores de parámetros. En este caso, existen variadas e innumerables formas de determinar el número de muestras o replicas para cada tratamiento, de diferentes autores, donde los esquemas de muestreo son diferentes, como es el caso del uso del Bootstrap para determinar un número de réplicas en un experimento, este es solo un ejemplo.

Lo cierto es que el Muestreo Estadístico se considera hoy como una disciplina transversal que toca a gran parte de las investigaciones, cuantitativas o cualitativas, observacionales o experimentales.

Ningún texto de Muestreo Estadístico puede abordar todas sus facetas, no existe dimensión para hacerlo, es por ello, que nosotros en este texto, nos enfocamos en las estrategias más usuales y de importancia en el Diseño de Muestras Estadísticas.

Este texto aborda el Muestreo Estadístico de la manera más sencilla posible, sin dejar de incluir los contenidos teóricos y particularidades de cada esquema de Muestreo Estadístico, tratando el tema del muestreo probabilístico, no probabilístico y los esquemas mixtos, de mucho uso en la actualidad.

Los autores tratamos de hacer un libro que no se aleje de los aspectos teóricos del ME, pero le damos mucha fuerza a los aspectos prácticos y aplicaciones.

El usuario de este texto va, desde Docentes de Universidades, personas de diferentes pregrados, hasta aspirantes a Master en el uso de las técnicas, de manera que sus resultados tengan rigor y validez científica.

Deseamos aclarar que este texto va dirigido al Muestreo Estadístico o Probabilístico, no al Muestreo No Probabilístico, sin embargo, este tipo de

muestreo, tratamos de profundizar lo que nos fue posible, debido a su magnitud, por su importancia en investigaciones cualitativas.

El Censo

Consiste en la enumeración completa o exhaustiva de todos los elementos de la población a investigar.

Entre las características importantes de un censo están:

- Muchas veces no se puede hacer censo por el alto coste de la toma de datos.
- No se puede hacer censo cuando la población es infinita.
- No se realiza censo cuando el proceso de medición es destructivo.
- El censo requiere una organización enorme, encuestadores, supervisores, es mayor la escala de operaciones, esto aumenta los errores sistemáticos o sesgos, no debidos al muestreo.
- El censo no provee una medida del margen de error al cual están sujetos los resultados.
- Un censo es la única posibilidad cuando se necesitan datos locales para cada subdivisión del país o áreas geográficas, para muchas variables, aun cuando en la actualidad se utiliza combinado con el muestreo.

La información de un Censo se utiliza mucho como información auxiliar o clave para el diseño de la muestra actual.

El Censo, actualmente se utiliza en combinación con él ME, utilizándose para obtener información sobre uso de los servicios, el tamaño de la población, su composición y distribución espacial, información sobre los hogares y su constitución y estructura socioeconómica, entre otras variables. Unas variables se determinan con el Censo y otras con el Muestreo. Bien es cierto que con el Censo

hay más detalle de la información final en subgrupos o zonas geográficas muy pequeñas en relación a la población general, que podría no tenerse con él ME.

La mayor parte de las veces en los estudios se utiliza el ME por muchas razones, que vamos a considerar:

Razones para utilizar Muestreo

- Se ahorra dinero (si se compara con un Censo). Aún en poblaciones grandes, es posible obtener resultados precisos y confiables, lo que implica menor coste.
- Se ahorra tiempo. Gran parte de las veces, hacer contacto con todas las UA, resulta muy costoso.
- Se ahorra trabajo de campo y procesamiento de datos.
- Permite concentrar la atención en casos individuales, hay más detalle, mayor calidad de datos e información precisa. Este es el caso cuando la entrevista es larga, como el caso típico de buscar información sobre gastos familiares en encuesta de hogares, por ejemplo.
- Cuando el proceso de la toma del dato es destructivo, en este caso, no se puede usar el Censo.
- Cuando la población es homogénea, que cualquier muestra puede ser lo suficientemente representativa.
- Cuando las poblaciones son muy grandes o infinitas y el Censo excede las posibilidades del investigador.
- Cuando el proceso de medir u observar es costoso, aun cuando la población es pequeña, esto sucede mucho cuando se hacen experimentos.
- El muestreo permite hacer una mejor preparación del personal, bien sea coordinadores, supervisores, encuestadores y otros. Se realiza un mejor control en la supervisión, entrenamiento y se puede contratar mayor número de gente con experiencia en Diseños Muestrales y sus particularidades.

Por nombrar solo algunos campos de aplicación del ME, no probabilístico y sus

combinaciones, se mencionan:

- Encuestas de hogares por muestreo (Encuestas Complejas)
- Encuestas de población activa, desempleo, indicadores de salud, entre otras.
- Opinión de intención de voto
- Encuestas de audiencia en radio y televisión
- Encuestas en Redes Sociales
- Estudio del mercado
- Comportamiento del consumidor de un bien o servicio
- Análisis de la satisfacción del cliente
- Estudio de segmentación del mercado
- Posicionamiento de un bien o servicio
- Grado de importancia de la mezcla de mercado, por parte del consumidor
- Estudio de opinión de los consumidores, sobre un bien o servicio
- Grado de importancia de atributos comerciales de bienes y servicios
- Medición de la calidad del servicio percibido
- Cambios de tendencia de las ventas
- Métodos para rechazar o aceptar un cliente potencial
- Segmentación de zonas geográficas
- Predicciones de ventas por zonas, bienes, servicios, sucursales, canales de comercialización
- Control estadístico de un sistema de calidad
- Aplicación de herramientas estadísticas de calidad
- Mejoramiento de los procesos de una empresa
- Métodos para aceptar o rechazar proyectos de Investigación- Inversión- Desarrollo
- Relacionar características del mercado con características demográficas, geográficas, psicográficas y sociales

- Determinar el perfil del clima organizacional
- Estudios del recurso humano en empresas
- Encuesta en investigaciones cuantitativas y cualitativas
- Validez, Confiabilidad y Dimensionalidad de instrumentos de investigación
- Elaboración de escalas
- Diseño y Análisis de experimentos (específicos para la Ciencias Sociales)
- Análisis estadístico de proyectos
- Evaluación de programas

En las instituciones gubernamentales se realiza el censo de población, generalmente cada diez años; sin embargo, en los últimos años se está combinando el censo con el muestreo. En las investigaciones demográficas se obtienen datos del número de personas y su distribución por edad, sexo, estado civil y otra serie de variables de población, con fines de planificar políticas de estado.

Algunas variables se miden en el censo y otras con un muestreo, esto es frecuente cuando las unidades de análisis son los hogares o familias, y por la complejidad del diseño muestral, el esquema recomendado es el muestreo estratificado aleatorio Polietápico, o en varias etapas. Se realiza la estratificación por Comunidad autónoma, Provincia y Municipio, y las estimaciones se realizan por dominio de estudio o en general para toda la población.

En instituciones o empresas donde se realizan investigaciones en laboratorios, generalmente, no se conoce el tamaño de la muestra adecuada como en control de calidad, estudios biotecnológicos, cultivos hidropónicos, estudios en viveros, estudios donde se toman muestras y se hacen duplicados o triplicados en la industria química y farmacéutica. En estos casos, últimamente ha dado buenos resultados las técnicas Bootstrap para la determinación de un numero óptimo de tamaño de la muestra.

En las instituciones o empresas agrícolas que desean realizar estudios en el área agrícola, se utilizan esquemas de muestreo más complejos.

En la parte agrícola, los esquemas de muestreo tienen alta complejidad, ya que se utilizan marcos múltiples. Además de tener como objetivo la estimación de rendimientos de cultivos (cosecha), hay otro aspecto de interés como son los estudios de incidencia y prevalencia de enfermedades en ganadería, estudios sobre la presencia de enfermedades y plagas en cultivos. Este tipo de estudio requiere de esquemas polietápicos complejos con el mayor uso de variables auxiliares o complementarias. En este caso, se requiere la pericia de un equipo de muestristas e información auxiliar que permita diseñar una muestra adecuada al problema en estudio.

En investigaciones sobre la estructura y producción agrícola se debe tener en cuenta si se va a utilizar marco de áreas, de lista o combinaciones de ambos (marco múltiple). Los dos casos varían, si la unidad de análisis es la unidad de producción o es un área de terreno dentro de la explotación. En el caso que el dato se tome dentro de un área de terreno, se hace más complejo el diseño de la muestra, donde se requiere en primer término formar estratos en base a variables auxiliares, cada unidad de explotación sería un conglomerado y se selecciona en varias etapas, utilizando al final un esquema sistemático con arranque aleatorio.

En cuanto al combate de plagas, si se desea estimar la densidad (número de individuos por sitio de muestreo) con el fin de establecer un método de control, el investigador debe recurrir a esquemas de muestreo viables que mantengan la precisión y confiabilidad (Gómez e Higuera, 1986). Estos autores coinciden en que éste tipo de estudio, va a permitir dar recomendaciones concretas de control de un determinado insecto o enfermedad. Siendo ésta, la fase final a la cual debe llegar todo proceso de investigación en el manejo de plagas. Los métodos de muestreo son instrumentos clave para estimar la intensidad de población, porcentajes de infestación y la recomendación de medidas de control, ellos son afectados por el patrón espacial de la plaga y aspectos metodológicos instrumentales y personales.

Estos autores proponen los métodos de captura y recaptura para evaluar el grado de parasitismo, etapas de desarrollo y clasificación de especies y la misma intensidad de población, además de los esquemas clásicos de muestreo. Proponen un esquema de muestreo secuencial, que permite tomar decisiones basadas en los resultados acumulados en cada unidad muestreada, donde el tamaño de muestra no es fijo como en los métodos tradicionales probabilísticos. Este esquema se utiliza cuando se debe tomar una decisión sobre el control de la plaga, se aplica o no tratamiento, donde, de acuerdo al número de individuos acumulados en muestras secuenciales, se llega a la decisión de que la densidad de una especie ha llegado a un nivel peligroso, donde se hace necesario tomar medidas de control. En muestreo de insectos es el de mayor uso en la actualidad.

El muestreo se utiliza con éxito en países donde se necesitan datos sobre las características de la fuerza de trabajo con el fin de realizar planificaciones económicas, se estima el total de la población, su clasificación y la cantidad de trabajo ejecutado, éstos son datos delicados, que ameritan la realización de diseños de muestras a través de expertos muestristas.

Generalmente, estos estudios se realizan cada mes. El diseño muestral recomendado es un Muestreo Por Conglomerados (MPC) en varias etapas, con estratificación y varianza máxima del estimador. Se puede usar Probabilidad Proporcional al Tamaño en la selección de las unidades primarias y en otra etapa subsiguiente, se utiliza una selección sistemática con probabilidades iguales. Este esquema de muestreo también es muy útil en encuestas de salud, donde se estudia la incidencia de enfermedades, lesiones, naturaleza y tipo de atención médica recibida (Ras, 1979).

En relación a los estudios de opinión pública y de mercados, donde se pide a los entrevistados, sus opiniones, deseos e intereses, hábitos de compra, intención de voto, y otras variables. El público político como el consumidor de productos y servicios, es muy heterogéneo, esto conlleva a utilizar la estratificación, el uso del MPC y para la selección final de los individuos en los puntos muestrales, se puede

hacer una selección sistemática o muestreo por cuotas. Ras encontró que una muestra aleatoria y la de cuotas, tuvieron resultados similares.

En la actualidad, las empresas están obligadas a realizar estudios del mercado, necesitan explorar el mercado para la toma de decisiones, utilizando la investigación de mercados. Es aquí donde es imprescindible utilizar métodos de muestreo adecuados en la prueba de productos, análisis de la competencia, las Seis (6) p`s (precio, producto, plaza, promoción, personas y procesos), pruebas de empaque, calidad de servicio, instrumentos electrónicos y ortos innumerables aspectos del mercado. En estas investigaciones las empresas requieren esquemas de muestreo mixtos probabilísticos y no probabilísticos como el MPC con estratificación y muestras por cuotas en las etapas finales del muestreo.

Conceptos y definiciones en Muestreo Estadístico

Los aspectos fundamentales a considerar en el ME aparecen en la mayoría de los libros de muestreo (Ras, 1980; Cochran, 1976; Sukhatme y Sukhatme, 1970; Kish, 1972; Seijas, 2006; Scheaffer, Mendenhall y Ott, 1987, Lohr, 2010).

Elemento:
Preferimos llamarlo entidad. Se refiere a cada unidad de análisis (UA) que contiene información sobre el parámetro de la población, de la que se desea hacer la Inferencia Estadística.

Población:
Uno de los aspectos de mayor importancia es definir la Población, término que ha sido muy discutido. Nosotros trataremos de dar elementos que apoyen dicha discusión:

En toda investigación donde se utilice el muestreo, va a requerir un conocimiento previo de aspectos de la población (la paradoja de Friedman), entonces vamos a tomar una muestra para conocer aspectos de la población, pero para ello

debemos conocer aspectos de la población antes de tomar la muestra, esto es cierto.

Al definir la población debe estar bien identificada la unidad de análisis, los criterios de inclusión y exclusión de UA de ellas, tomar en cuenta el período de referencia y el de toma de datos en campo. Se hace necesario dimensionar la población en base a los criterios de interés para el grupo investigador. Es decir, las características de las UA que van a identificar la población son elegidas por el grupo investigador.

¿De qué entidad necesito los datos?, ¿A quién se refiere esta investigación?

¿Cuál es el conjunto de unidades de análisis para el cual serán válidas las conclusiones?

Población objeto…….Población accesible……. Población muestreada.

El grupo investigador comienza con la definición de la Población objeto, pero la mayoría de las veces, hay UA no disponibles o que se excluyen, y aparece la Población accesible y luego, por problemas de Marco Muestral, se llega a la Población muestreada, de la que al final, se obtiene la muestra. Las conclusiones se hacen para la población muestreada.

De la población objeto hay mucha exclusión de elementos : personas que se trasladan, miembros de Fuerzas armadas, personas que viven en hoteles, que fluctúan, van y vienen, están en prisión, elementos que no se localizan, no respuesta (No sabe o ignora/ no responde), no están incluidos en el marco. Se hacen restricciones a la población.

La Profesora Lohr (2010) realiza una excelente presentación para explicar la transformación de la Población Objeto en Población Muestreada, donde presenta un ejemplo muy interesante. Recomendamos estudiarlo.

En cuanto a la definición de población, existe controversia en relación a la definición de Universo y Población, la mayoría las maneja por igual.

Azorín y Sánchez-Crespo afirman que población (Universo) se refiere al conjunto o colectivo finito o infinito de elementos que presentan características observables comunes.

Lohr (2010) y Seijas (1999) aseveran que Universo es el conjunto finito o infinito de elementos o entidades y Población es el conjunto finito o infinito de mediciones u observaciones de una característica en particular. Es decir, que en un Universo puede existir varias Poblaciones Estadísticas.

Según la teoría estadística, se define Población como el conjunto de elementos y la variable aleatoria asociada a la característica que se mide en el elemento o entidad, hay una relación entre el elemento y el valor de la variable.

Kish (1972) dice que el universo es un conjunto hipotético infinito no real de elementos generados por un modelo teórico (experimento), y la población es real, aunque en la teoría estadística se habla de población como modelo teórico, ejemplo la Normal, pero se puede manejar cualquiera de los dos enfoques.

Por otro lado, hay que diferenciar entre población real e hipotética: La real puede ser finita o infinita y la hipotética es infinita.

Población real: Está toda en el momento (existe), es tangible, puede ser finita o infinita, si es finita, se puede listar y generalmente hay marco, si es infinita no tiene marco, pero sus elementos son reales.

Población hipotética o conceptual: no está en el momento, no es tangible, no tiene marco, sus elementos no son reales, ni se puede listar, sus casos o elementos se obtienen por repetición de un proceso o experimento o prueba y se considera el futuro, no está en el momento, no la tenemos. Ejemplo: producción de bombillas en una línea de producción en un tiempo futuro, número de veces que alguien visita un cajero en los próximos veinte años.

Una población puede ser cualquier colección de cosas, sujetos, eventos, circunstancias, personas, empresas, hogares, familias, institutos, escuelas,

alumnos, entre otros. Del mismo modo, una población puede ser representada por un conjunto o colección o por un modelo, en este caso de modelo, en los libros de teoría estadística, se dice que X (Variable aleatoria) sigue una distribución normal y se refiere a la Población Estadística.

Marco de Muestreo

Antes del diseño de la muestra se debe tener el Marco Muestral (**MM**) depurado y actualizado, y toda la información posible sobre la población, sobre todo las variables auxiliares clave.

El MM es un conjunto de medios físicos, donde se selecciona la muestra, puede ser lista, mapas físicos o fotografías aéreas, archivos de computadora, volúmenes, horas de transmisión de programas, bases de datos, fotos, que contenga las UA de la Población de donde se obtendrá la muestra.

Debe haber correspondencia biunívoca entre las unidades de la población y las físicas del marco.

Generalmente, se presentan marcos de muestreo imperfectos, hay que depurarlos, falta de información o hay información de más o traslape.

En cualquier marco se puede presentar: elementos faltantes (error por defecto), unidades repetidas (error por exceso), unidades ocultas, elementos extraños (no corresponden a la población), espacios en blanco (error de laguna, o zonas sin elementos)

Hay 2 tipos de marco de muestreo:

a) Marco amplio: Tiene otro tipo de información auxiliar clave que puede ayudar a estratificar, calcular el tamaño de la muestra, utilizar estimadores de razón, de regresión y de diferencia, o utilizar un esquema de Probabilidad Proporcional a una medida de tamaño .

b) Marco restringido: es el que está formado solamente por las unidades de muestreo, no se consigue información auxiliar. Lo ideal es que el marco coincida con la Población Objetivo o investigada.

El marco puede estar formado por unidades elementales o un conjunto de unidades elementales. Una unidad de muestreo puede ser un elemento o UA o un grupo o conglomerado de UA. Este es el caso de Muestreo por Conglomerados, o en varias etapas (Bietápico o Polietápico). El marco ubica y dimensiona la población.

En un marco puede haber: elementos faltantes, unidades repetidas, unidades ocultas, unidades extrañas, espacios en blanco, entre otros.

Todo MM, antes de utilizarse para seleccionar la muestra final, debe ser evaluado en cuanto a la falta de cobertura horizontal, cobertura múltiple, exceso de cobertura, grupos que se forman y hacer los ajustes correspondientes.

Por otro lado, todo marco debe someterse a validación directa de campo, antes de la selección de la muestra, para que sea lo más útil posible. En la práctica uno de los aspectos más difíciles de cumplir en una Encuesta, es tener un marco adecuado y actualizado para la selección de la muestra, generalmente los MM presentan sesgos y se les llama marcos imperfectos.

Unidades Estadísticas

Un aspecto relevante en muestreo lo constituye la definición de las unidades de análisis, de información y de muestreo, ya que confunde con frecuencia a los investigadores.

Son la base para saber la estrategia de medición. Nos definen la estrategia para la medición:

- *Unidad de Investigación*: Partes que se van a analizar, las de mayor dimensión.
- *Unidad de Análisis*: De la cual se desea la información.
- *Unidad de observación, informante o respuesta:* La que proporciona los datos o las respuestas.
- *Unidad de Muestreo:* Conjunto o UA de las cuales se selecciona la muestra.

Como ejemplos podemos citar:

1. Unidad de investigación: el establecimiento, Unidad de análisis: el empleado, Unidad de observación: una ficha de un empleado en un archivo y Unidad de muestreo: la que se usa para seleccionar la muestra (el nombre o código del empleado).
2. Unidad de investigación: La escuela, Unidad de análisis: El alumno, Unidad de observación: El maestro y Unidad de muestreo: Puede ser la escuela o el alumno.

Muestra*:* La muestra es un subconjunto de la población o universo y según Walpole, Myers y Myers (1999), la muestra aleatoria es una sucesión de variables aleatorias independientes donde cada una de ellas tiene la misma función de probabilidades (f(x)) y su función de probabilidad conjunta es:

F(x1, x2, x3,….., xn)= f(x1). f(x2)…… f(xn).

La muestra debe ser:

Reducida para que haya bajo coste.

Amplia para que exista un error de muestreo bajo.

Nosotros consideramos a la muestra aleatoria como una sucesión de pruebas o ensayos de un experimento aleatorio y al conjunto de valores de la variable en estudio, proveniente de una muestra aleatoria, se le conoce como realización de

un conjunto de variables aleatorias independientes. Al final de una encuesta, no es sencillo obtener una muestra aleatoria de la población objeto de estudio, ya que hay individuos que no participan, abandonan el estudio o no saben, no responden, entre otras causas.

Muestra Representativa

La muestra representativa es una muestra de un tamaño relativamente apropiado que ha sido seleccionada por procedimientos aleatorios y las características que se observan en ella, corresponden a la población de la cual se extrajo (Ras, 1980; Cochran, 1976; Scheaffer, Mendenhall y Ott, 1987), no es posible en ningún caso, tener la certeza del grado de representatividad, sino que hay una probabilidad razonable de esa representatividad.

La representatividad no solo depende de la aleatoriedad ni del tamaño de la muestra, también del diseño muestral, muy particular para cada caso, del uso de información auxiliar y de un MM útil y actualizado, y el termino representativo se utiliza siempre y cuando la muestra represente fielmente la variable objeto de estudio, la cual tiene una distribución probabilística en la población y la distribución de frecuencias en la muestra debe ser espejo o muy similar a la de la población. No existe la muestra representativa, es un ideal, nosotros la llamamos muestra suficientemente representativa.

Esto hace notar lo complejo que resulta la selección de una muestra representativa, para ello, debe tenerse en cuenta: la forma de selección de la muestra, los estimadores a proponer y su precisión, la determinación del tamaño de la muestra que tome en cuenta la precisión o margen de error permitido, el nivel de confianza y la variabilidad de la variable sobre la que se va a realizar la Inferencia Probabilística. Así mismo, prestar atención al marco de muestreo disponible y al conjunto de variables auxiliares clave o covariables que estén correlacionadas con las variables de interés, quienes van a permitir mejorar el diseño muestral, con la formación de estratos, selección de estimadores directos o

indirectos (razón, regresión y diferencia) y elegir un tamaño de muestra adecuado a una precisión dada.

Muchas veces al diseñar una muestra probabilística hay que hacer concesiones, sobre todo si la población estadística es asimétrica, inclusive hay veces que se utilizan elementos con probabilidad uno (1) de pertenecer a la muestra y si no se hace esto, la muestra no será lo suficientemente representativa.

Una muestra probabilística en su estructura se aproxima a un mayor grado de lo que se le llama representatividad cuando se hace menor el valor de la distancia entre la estimación de la muestra y el valor del parámetro de la población, esto se conoce como acuaricidad en la Inferencia Estadística.

Cuando el proceso de selección le asigna una probabilidad de antemano a cada elemento. Esta probabilidad es diferente de cero y se conoce, además el error de muestreo es bajo, existe la acuaricidad y se utiliza un proceso aleatorio en su selección, podemos estar en presencia de una muestra lo suficientemente representativa.

La mejor forma que tenemos de definir una muestra lo suficientemente representativa es aquella donde se utilice una estrategia de ME que permita estimar el valor del parámetro con acuaricidad, el mínimo sesgo, el mínimo Error Estándar del estimador de ese parámetro o el mínimo error de estimación, el cual es un múltiplo del Error Estándar del estimador.

Algo importante que el lector debe conocer es la diferencia entre una Estadística, un Estimador y un Parámetro.

Estadística: Es una variable aleatoria función de las observaciones de la muestra de tamaño n, que no depende de parámetros desconocidos.

Estimador: Es una estadística que se utiliza para estimar el valor de un parámetro en la población.

Estimación: Es el valor del estimador en la muestra.

Otro concepto bien importante en él ME es el Margen de Error:

Margen de error o error de estimación

Máximo error permitido entre el estimador $\hat{\theta}$ y θ, el parámetro.

Donde

$$\left|\hat{\theta}-\theta\right| \leq d = e$$

Es el máximo alejamiento entre el estimador y el parámetro que el grupo investigador puede permitir.

Al planear una Encuesta por Muestreo Probabilístico, si se va a determinar el tamaño de la muestra, el grupo investigador debe especificar el margen de error (e), otros lo llaman d y otros B, llamado error de estimación o cota de error que el grupo está dispuesto a tolerar entre el valor del estimador ($\widehat{\theta}$) y el valor del parámetro (Θ).

Suponga que el estimador es la media $\overline{(y)}$ y el parámetro es μ, entonces el margen de error viene dado por una diferencia entre la media muestral y la poblacional, su valor absoluto.

El grupo investigador debe fijar este margen de error y la probabilidad (1-α) de que se cumpla ese margen de error. Generalmente, se utiliza el margen de error absoluto (e), que viene en las mismas unidades de la variable, ya que también puede usarse el margen de error relativo, el cual se expresa en términos del coeficiente de variación.

El error relativo (er) es adimensional y se presenta como % de la media, del total, de la proporción y porcentaje o del número de elementos en una clase que tienen la característica de interés. El error relativo se usa para comparar precisiones en

muestras probabilísticas al estimar parámetros distintos de la misma población o precisiones del mismo parámetro en distintas poblaciones. Se prefieren los errores relativos menores del 10%. Cuando decimos que el error relativo es del 5% con respecto a μ, T, P, A, u otro, estamos diciendo que er =0.05 (u), para el caso de la media, y así sucesivamente para los otros parámetros.

Fijar el margen de error es una de las cosas más difíciles, su valor depende de varios factores: De la finalidad del estudio o trascendencia de la investigación, en una empresa, Zona geográfica o País, no es lo mismo indagar sobre el grado de satisfacción de clientes con un producto o servicio que estimar la tasa de incidencia de una enfermedad o la tasa de desempleo en un país. Se debe analizar el gasto con el margen de error propuesto y determinar las posibles consecuencias que puede tener asumir ese margen de error. También se puede elegir el margen de error, revisando estudios similares anteriores y de la experiencia en el fenómeno en estudio, por parte del grupo investigador.

Ejemplo hipotético muy sencillo, en un contexto: si se desea estimar el ingreso medio por persona activa en el DF, México, y tenemos el listado de las personas, además conocemos el tamaño de la población N y usaremos un Diseño de Muestreo Aleatorio Simple sin reemplazo en la población finita.

Dado este ejemplo, antes de calcular el tamaño de la muestra, podemos elegir un margen de error absoluto (e), suponga que es ± 150 pesos y un 95% de confianza de que se cumpla con ese error o cota fijada. Se tiene un 95% de confianza de que el margen de error no va a ser mayor de 150 pesos. Si se utiliza una variable continua, y un margen de error absoluto, este margen de error viene en las mismas unidades de la variable (pesos).

Etapas de una Encuesta por Muestreo

A grandes rasgos, presentamos las principales etapas en una Encuesta por Muestreo:

1. Se parte de un tema de investigación.
2. Se plantea el problema de investigación, su sistematización.
3. Se elaboran las preguntas o incógnitas de investigación en forma precisa.
4. Definir los objetivos en forma precisa. Los resultados a obtener, deben responder en forma biunívoca con los objetivos.
5. Definir la población en forma concreta.
6. Definir las variables y datos a ser colectados, documentos de estudios anteriores. Variables auxiliares, entre otras.
7. Fijar el grado de precisión y confiabilidad en la estimación.
8. Actualizar el marco de muestreo.
9. Diseñar el plan o esquema de muestreo, que incluye el uso de variables auxiliares, si las hay, los estimadores directos o indirectos a utilizar, el tamaño de la muestra, el coste, el. tipo y clase de muestreo y combinaciones de ellos.
10. Prueba piloto, para el diseño y ajuste del cuestionario u otro, estimación de varianza de la variable de interés, tiempo probable de aplicación, entre otros.
11. Organizar el trabajo de campo.
12. Ejecución de la encuesta en campo.
13. Procesamiento y análisis de datos.
14. Informe final.

Tipos de errores en las encuestas por muestreo

1. Errores debidos al muestreo: Todo muestreo tiene error debido a que no se estudia toda la población, solo una parte. Este error se puede acotar o disminuir por medio del Diseño de la Muestra, que incluye el tamaño de la muestra. Se representa de varias formas: e, B, d., dependiendo del autor. No se puede eliminar, se puede acotar, decrece si se aumenta el tamaño de la muestra y es menor a medida que la población sea más homogénea.

Hay otras fuentes de error, llamados sistemáticos o sesgos, que pueden reducirse diseñando con mucho cuidado todo el proceso de Investigación, hasta el procesamiento y análisis de los datos.

Se reducen cuidando los detalles (instrucciones claras y precisas, entrenamiento adecuado a los encuestadores, elaboración precisa y adecuada de cuestionarios, mediciones precisas, marco adecuado, entre otros.

Estos errores no se acotan, no se reducen al aumentar el tamaño de la muestra, incluso, pueden aumentar y se conocen como errores no debidos al muestreo:

2. Errores no debidos al muestreo (son más difíciles de controlar):

2.1. Sesgo de selección: Ocurre cuando una parte de la población objeto no está en la población muestreada.

2.2. Sesgo de medición: cuando los datos medidos u observados difieren del verdadero valor, por ejemplo: las personas no dicen la verdad, dan información falsa, hay exageración o subestimación. Las personas no comprenden la pregunta, por ignorancia, que el entrevistador lea mal las preguntas o anote mal la respuesta, no se recuerda exactamente el dato, el orden de la pregunta puede tener efecto sobre la respuesta.

2.3. La no respuesta: la persona no contesta, se niega, no se localiza.

2.4. El uso de esquemas no probabilísticos, si se requiere un Probabilístico.

2.5. Errores de registro de datos, codificación, tabulación, transcripción (son involuntarios) cambio de unidades elementales, se alteran las probabilidades de selección.

2.6. Marco de muestreo deficiente o imperfecto.

2.7. Sesgos en la relación entre el entrevistado y el entrevistador.

En un censo se presentan más errores no debidos al muestreo.

Factores que afectan el diseño de la Muestra Probabilística

Diseñar una Muestra Probabilística (**MP**) que sea lo suficientemente representativa de la población estadística, no es tarea sencilla, la mayoría de las personas creen que diseñar una MP, es calcular el tamaño de la muestra; y no es así, ya que depende de un conjunto de factores que subyacen en el diseño muestral. La representatividad de una muestra está determinada por el diseño muestral, todos los elementos que lo constituyen. Los datos recolectados deben representar una característica de la población y el diseño de la muestra debe direccionar la solución del problema. El diseño contribuye a disminuir el error de muestreo.

En el diseño de la MP, no de la Encuesta, es necesario tomar en cuenta los siguientes factores: El objetivo del estudio, el proceso de selección, el proceso de estimación de parámetros, el tipo de marco de muestreo, el uso de variables auxiliares, el tamaño de la muestra, los errores posibles, y la distribución de frecuencias de la variable objeto de estudio, los cuales se presentan en la **Figura 1.**

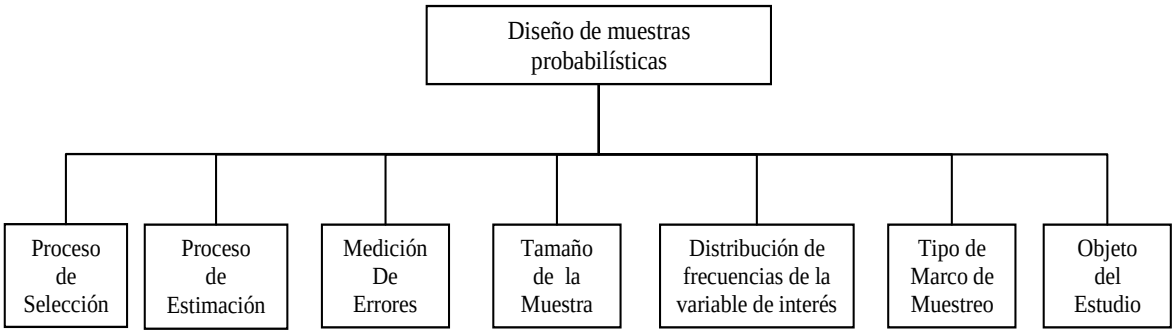

Figura 1. Factores que afectan el Diseño de la Muestra.
Fuente: Elaboración propia.

El comienzo del Diseño de una Muestra es preguntarse por los objetivos del estudio, que se desea encontrar, se desea realizar una investigación cuantitativa o cualitativa, es un estudio experimental u observacional, los objetivos se relacionan con el diseño, al conocer los objetivos, ya se piensa en el diseño muestral más

adecuado, dentro de los objetivos subyace la población objeto sobre las que serán válidas las inferencias. El proceso de selección de la muestra se refiere a la forma de selección de ella, bien sea con un instrumento aleatorio o cuando el investigador se convierte en instrumento de selección; cuando el investigador es el instrumento, se está en presencia de un esquema de muestreo no probabilístico, en el cual no se conoce de antemano la probabilidad de selección, ni el margen de error en la estimación. En gran parte de los DM se utiliza combinaciones de los esquemas probabilísticos y no probabilísticos, sobre todo en esquemas de varias etapas o polietápicos. Así mismo, al utilizar diversos marcos muestrales, se puede utilizar diferentes esquemas de acuerdo a las características de cada marco, siendo el MM, otro aspecto a considerar, ya que se necesita un marco actualizado y preferiblemente ampliado con variables auxiliares.

En relación al proceso de estimación de un parámetro, tiene que ver qué tipo de estimadores se utilizarán, los directos o indirectos (razón, regresión o diferencia), las cualidades de precisión y confiabilidad de cada estimador. Cobra importancia la varianza de estimador, su insesgamiento y otras propiedades como consistencia y suficiencia. El error de muestreo es clave, la variabilidad o desviación estándar del estimador como variable aleatoria, el estimador de la varianza del estimador o error estándar del estimador, el cual debe ser reducido. El hecho es que a medida que aumenta el tamaño de la muestra, disminuye el error estándar del estimador y eso es lo deseable en todo DM.

El muestrista fija el máximo error permitido entre el estimador y el parámetro, el nivel de confianza en la estimación y recurre a la distribución de frecuencias del estimador, para relacionar la muestra con la población. Para realizar un adecuado diseño de muestra, debe identificar las **variables auxiliares clave**, las que tienen un alto grado de importancia en un diseño de muestra. Queremos resaltar el interés en las variables auxiliares clave, ya que van a permitirnos disminuir el Error Estándar de los estimadores y deben tener correlación con la variable de interés , objeto de estudio, las variables auxiliares clave se utilizan para :

1. Estratificar la población, esto se hace con la conformación de los estratos, en los casos de encontrar variables correlacionadas con la variable de interés.
2. La determinación del tamaño de la muestra.
3. La construcción de los estimadores indirectos de regresión, razón y diferencia.
4. El uso de Probabilidades Proporcionales a una medida de tamaño (PPT).
5. Ordenar la población en Muestreo Sistemático (MS).
6. Asignarle distintos % de selección de muestra en los estratos, cuando se tiene una población Asimétrica.
7. Controlar el tamaño de los conglomerados, en Muestreo por Conglomerados.
8. Postestratificar.
9. Usar Muestreo Doble.
10. La utilización de los estimadores de Horvitz-Thompson, en el cálculo de las probabilidades de inclusión.

Fases de un diseño de muestreo

Cada diseño de muestra es un reto, el objetivo como se dijo anteriormente, es obtener una muestra que proporcione el máximo de información al mínimo coste. Consideramos que se debe seguir las siguientes fases, en un diseño de muestra, presentadas en la **Figura 2**.

Estos elementos son suficientemente explicados en los aspectos teóricos del Muestreo Estadístico, siendo el beneficio de estas fases, el que pueden guiar la realización de un diseño de muestra con un procedimiento lógico y científico.

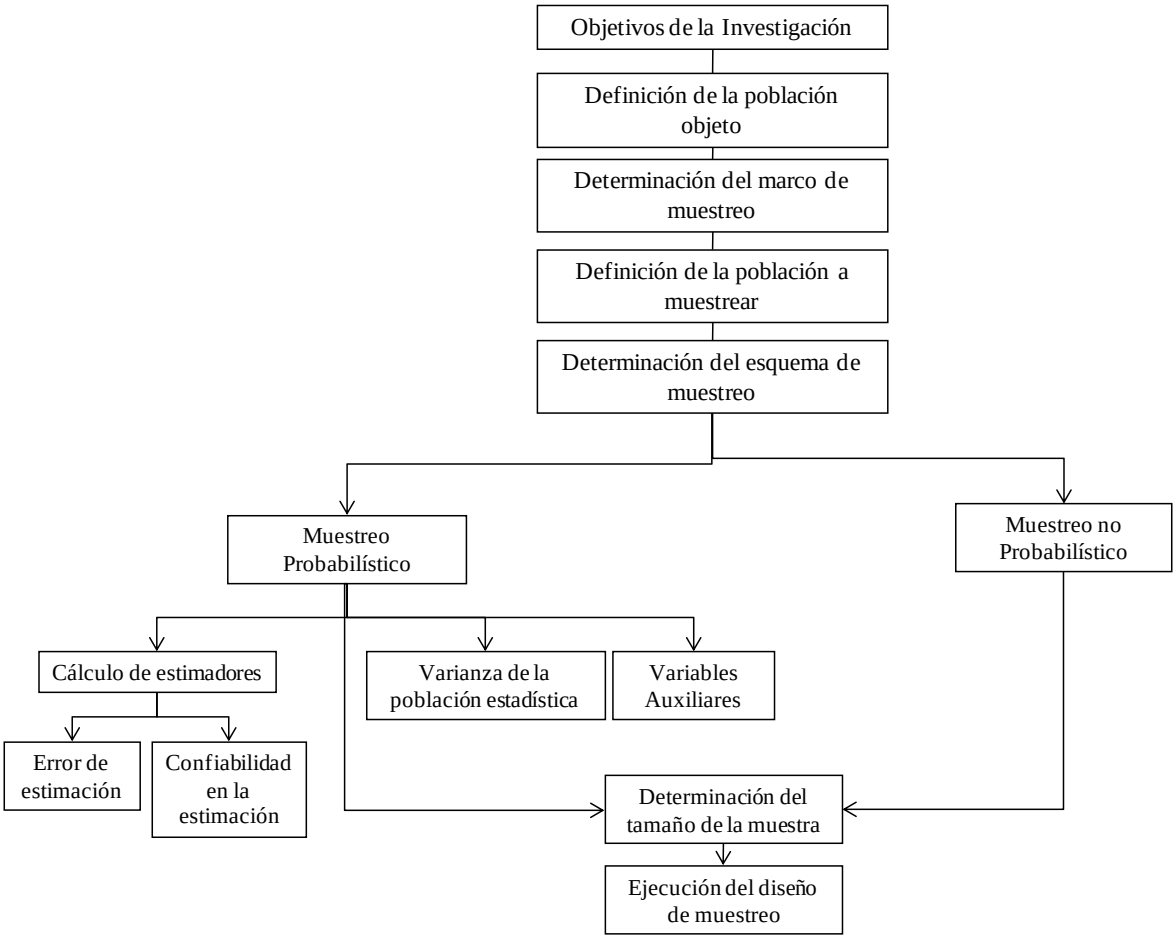

Figura 2. Fases de un diseño de muestreo

Fuente: Elaboración Propia

Hay un aspecto que debemos mencionar, el cual tiene que ver con los pesos muestrales o pesos de diseño. Muchos autores o programas estadísticos para encuestas, requieren que se especifique el peso muestral o peso de diseño, siendo el peso muestral Wi =1/πi, el cual representa el número de unidades en la población que representa cada UA en la muestra, πi es la probabilidad de inclusión de primer orden. Estos pesos muestrales se calculan y difieren en cada esquema muestral. Para el caso del MAS, el peso muestral es N/n, donde N es el tamaño de la población y n es el tamaño de la muestra.

Introducción al tamaño de la muestra

El tamaño de la muestra constituye uno de los aspectos de mayor inquietud, y muchos investigadores piensan solo en él , esto indiscutiblemente no es así, como lo dijimos anteriormente. El tamaño de la muestra es un aspecto dentro del Diseño Muestral.

Una vez definidos el tipo y la clase de muestreo, y el tipo de estimador, se debe elegir el tamaño de la muestra. Muestra grande: se desperdicia tiempo, dinero y talento y muestra pequeña: se tiene información inadecuada por el tiempo y esfuerzo, y se ha hecho un mal gasto.

Varios factores contribuyen en la determinación del tamaño de la muestra: El tamaño de la población (influye si la población es finita y si $n/N>0.05$, según Cochran(1976), la variabilidad de la variable de interés u objeto de estudio (Y), El nivel de acuaricidad o precisión requerido (e), el nivel esperado de no respuesta, el método de muestreo utilizado para el diseño y selección de la muestra, los recursos que se tengan (personal, dinero, tiempo, entre otros), la importancia relativa de las variables a investigar, entre otros.

En efecto, el número de UA o tamaño de muestra necesario en una investigación, no solo depende de los factores ya mencionados, sino del tipo de estudio y finalidad de la investigación, no es igual el cálculo del tamaño de muestra en estudios observacionales que experimentales, hasta el tipo de análisis estadístico a utilizar influye en el tamaño de muestra, se puede requerir un tamaño de muestra diferente si queremos construir un instrumento de investigación como el cuestionario escalar, utilizando un Análisis Factorial Exploratorio o Confirmatorio, o utilizar un Análisis de la varianza (ANOVA) o un análisis de Componentes Principales o un análisis de Correlación Canónica.

Para encontrar el tamaño de la muestra:

1. Debe haber una indicación de lo que se espera. Esto es precisión (límite de error) y confiabilidad en la estimación esperada. Fijar el margen de error (e) y la confiabilidad (1-α).

2. Encontrar una ecuación que conecte **n** a la precisión deseada (e).

La ecuación que utiliza para el cálculo del tamaño de la muestra es en todos los esquemas de muestreo, un múltiplo del Error Estándar o la Desviación Estándar del estimador: K x Error estándar del estimador. De esa ecuación se despeja el tamaño de la muestra (n).

Aclaramos que el Error Estándar de cualquier estimador se usa para:

1. Comparar la precisión de un Muestreo Aleatorio Simple con cualquier otro esquema de muestreo.
2. Estimar el tamaño de la muestra antes de ejecutar la Encuesta.
3. Ya ejecutada la Encuesta, nos permite calcular el intervalo de confianza en la estimación del parámetro de interés.

Todo tamaño de muestra sale de la ecuación del error de estimación o cota de error, el cual es un múltiplo del error de muestreo o error Estándar, no confundir el error estándar o de muestreo con el margen de error en la estimación. Para cada diseño de muestreo hay una ecuación diferente y de cada uno de ellos, se despeja el tamaño de la muestra n, que forma parte del error estándar en la ecuación del margen de error estimado o límite del margen de error estimado como lo llama Scheaffer, Mendenhall y Ott (1987).

4. Esta ecuación contiene parámetros poblacionales, como la varianza o cuasivarianza poblacional de la variable de interés, que debe ser estimada, bien sea por una muestra piloto, preliminar o estudios anteriores, o conociendo aproximadamente el rango de Y, por censo o estudios anteriores, entre otros.

5. Si se fija una función de coste, este coste forma parte de la ecuación. Es difícil obtener una función de coste; por ello, se utiliza la varianza.

6. Comparar tamaños de muestra. El valor de n se calcula basándonos en una o más variables importantes, generalmente correlacionadas con la variable de interés, objeto de estudio. Para cada variable se define el margen de error (absoluto o relativo) y la confianza (1-α) de que se cumpla el margen de error. Luego, se estiman los parámetros desconocidos.
7. Se calculan los diferentes tamaños de muestra, y en consenso, se escoge el tamaño definitivo n. La mayoría de las veces se escoge el máximo tamaño de muestra que conserve una precisión adecuada y se corresponda con los recursos que tiene asignado el grupo investigador, en cuanto a coste, tiempo, esfuerzo, materiales y suministros, para obtener ese tamaño de la muestra.
8. El tamaño de la muestra tiene que ver si el estudio es observacional o experimental. Si es para construir un cuestionario o una escala, para realizar una encuesta de hogares, y depende también del tipo de análisis estadístico que se piensa utilizar.
9. Si hay dominios de estudio, se calculan los tamaños de muestra por dominio.
10. La decisión final sobre el tamaño de muestra dependerá de un compromiso entre la precisión esperada en la estimación y la disponibilidad de recursos (monetarios, humanos y tiempo).

Como ejemplo se calcula el tamaño de la muestra n en un Muestreo Aleatorio Simple (MAS) como sigue, en una población infinita:

Se calcula la probabilidad usando:

$$.Z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim N\ (0,1)$$

Gráficamente en la **Figura 3**,

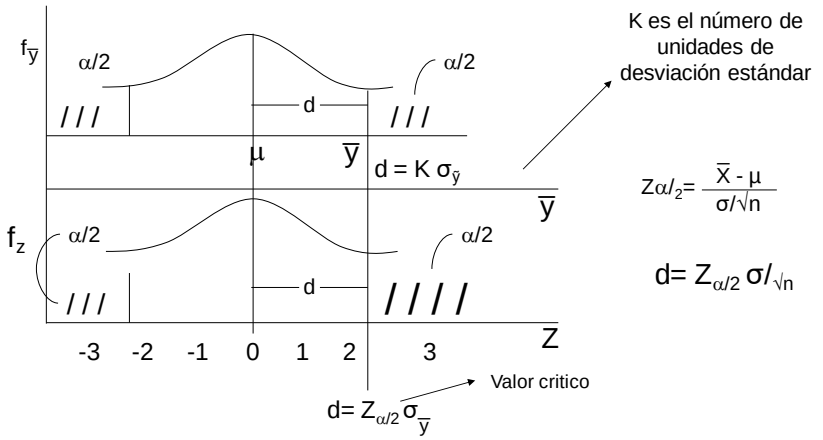

Figura 3. Distribución Normal Estándar

Fuente: Elaboración propia.

Donde:

Margen de error= $(K\sigma_{\bar{y}})$

K = Múltiplo del error de muestreo=$Z_{\alpha/2}$

e= $d = Z_{\alpha/2}\, \sigma_{\bar{y}}$

e= $Z_{\alpha/2}\, \sigma_{y/\sqrt{n}}$

$\sigma_{\bar{y}}$ = Error estándar de $\bar{y}$,

y se despeja n

$$\sqrt{n} = \frac{Z_{\alpha/2}\sigma_y}{d}$$

$n = \frac{Z^2{}_{\alpha/2}\sigma^2 y}{d^2},\ para\ MAS\ en\ una\ población\ infinita.$

Si la población es finita y n/N es >0.05 entonces se usa la ecuación que utiliza el Factor de Corrección por Finitud:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Que viene a ser la misma que

$$n = \frac{N Z^{2}_{\alpha/2} \sigma_y^2}{(N-1)d^2 + Z^{2}_{\alpha/2} \sigma_y^2}$$

Entramos a tratar los diferentes esquemas de Muestreo Probabilístico:

Muestreo Probabilístico

Para que exista una base objetiva para medir el grado de confianza con el cual se está haciendo la estimación de un parámetro, debe conocerse la probabilidad que tiene un elemento de la población de formar parte de la muestra y esto solo se conoce, si utiliza el método de muestreo probabilístico (**MP**), donde también se conoce el margen de error en la estimación.

Con los MP se pueden comparar la precisión de los diferentes métodos de muestreo, mediante el **Efecto del Diseño** (Kish, 1972), además se pueden ejecutar en la realidad de tal forma que sean compatibles con restricciones administrativas, producen resultados con la máxima confiabilidad al mínimo coste y la inspección supervisada estará acorde con las especificaciones predeterminadas.

En ellos cada unidad tiene una probabilidad conocida, diferente de cero, de que cada elemento de la población forme parte de la muestra.

Una vez que se selecciona una muestra probabilística entonces se puede emplear la teoría de las probabilidades, estimando la precisión de los estimadores y midiendo el grado de incertidumbre o grado en que varía la estimación del

verdadero valor del parámetro. En él se fija la precisión (e) de la estimación y se estima la variabilidad del estimador. Además, para aplicar un esquema de muestreo probabilístico, es necesario contar con un MM actualizado para llegar a las UA con una probabilidad conocida.

En el MP, el investigador debe usar estrategias que tiendan a reducir el error estándar de los estimadores y que éste estimador (su estimación) reproduzca el verdadero valor, o muy aproximado, del parámetro poblacional, para obtener una muestra lo suficientemente representativa.

Entre las estrategias a utilizar están: El uso de un proceso de selección aleatoria, la utilización de variables auxiliares para crear estructuras poblacionales como la Estratificación y la Conglomeración, las Probabilidades Proporcionales a una medida de tamaño, el uso de Ponderaciones y de estimadores indirectos de Razón, Regresión y de Diferencia, entre otras. Las Muestras Complejas, de mucho uso actual, toman en cuenta estos aspectos.

El solo hecho de seleccionar una muestra mediante un procedimiento aleatorio, es lo que nos va a permitir determinar la precisión en las estimaciones y aplicar la teoría de las probabilidades, para realizar la Inferencia Estadística.

Los MP en realidad pretenden disminuir la subjetividad originada al seleccionar las UA y evitan los sesgos en la selección de las mismas, pero no garantizan por si solos, lo que se conoce como representatividad, ellos deben apoyarse en un buen Diseño de la Muestra y un MM depurado y actualizado, la pericia del Muestrista, entre otros factores. Sin embargo, los MP tienen como desventaja: Se requiere mayor esfuerzo muestral, consume más tiempo, es más costoso y necesita una mayor infraestructura.

Los elementos fundamentales a considerar en el Muestreo Probabilístico aparecen en la mayoría de los libros de muestreo (Ras, 1980; Cochran, 1976; Sukhatme y Sukhatme, 1970; Kish, 1972; Seijas, 2006; Scheaffer, Mendenhall y Ott, 1987; Lohr, (2010), entre otros.

Muestreo Aleatorio Simple

Generalmente, el Muestreo Aleatorio Simple (**MAS**) no se utiliza como esquema de muestreo, sino que forma parte de otros esquemas.

Es el esquema de selección de muestras más sencillo y forma parte de otros esquemas probabilísticos. Es la base de la mayoría de los diseños muéstrales. Si se selecciona una muestra de tamaño n de un universo de tamaño N, de tal forma que cada muestra posible de tamaño n, tenga la misma probabilidad conocida de ser seleccionada o que cada elemento de la población tenga la misma probabilidad de inclusión en la muestra, entonces, se está en presencia del Muestreo Aleatorio Simple o de una Muestra Aleatoria Simple.

En cada extracción, se asigna igual probabilidad de selección a todas las unidades de muestreo. La probabilidad de que una unidad específica sea seleccionada en cada extracción es igual a la que sea seleccionada en la primera extracción (1/N).

El MAS puede ser con o sin reemplazo. Con reemplazo, coincide con el muestreo de poblaciones infinitas, las observaciones son independientes, si el muestreo se hace sin reemplazo, las observaciones no son independientes, pero si el tamaño de la muestra es muy pequeño comparado con el de la población, esto es n/N es menor de 0.05, entonces se considera la independencia de las observaciones.

Su principal ventaja es su facilidad y simplicidad de uso, especialmente cuando se toma una muestra relativamente pequeña, se usa más en poblaciones pequeñas y homogéneas.

En el MAS se debe tener un marco de muestreo en el cual se identifique todas la UA de la población y no se tiene variables auxiliares que permita utilizar otro esquema de muestreo. Se prefiere en poblaciones homogéneas, poco dispersas y pequeñas. Si la población es dispersa y geográficamente grande, resulta difícil llegar a las UA y puede llegar a obviarse pequeños grupos que pueden ser importantes, en relación a las variables de investigación.

Se utiliza cuando se tiene un marco de muestreo que especifique la manera de identificar cada elemento de la población y no se tienen los valores de la variable de interés (y), ni de otras variables correlacionadas.

Tiene la desventaja de que requiere de un MM con las UA listadas y numeradas, y puede resultar ineficiente si la población es extensa y dispersa.

En la práctica, el MAS es muy raramente utilizado como tal, pero además de ser la base de otros esquemas de ME, se utiliza como referencia para evaluar otros esquemas, esto se hace cuando se calcula el efecto de diseño (Deff), comparándose otro esquema de los probabilísticos, con el MAS.

Los procedimientos más utilizados para extraer una MAS son: las tablas de números aleatorios y procedimientos que vienen en los programas incluidos, para la selección de muestras aleatorias como Excel y SPSS, entre otros.

Seguidamente, mostraremos la notación que sugerimos para la utilización del MAS en la solución de los diferentes problemas o eventos donde se necesite su aplicación. En la **Tabla 1**, puede observarse la notación referida a la población y muestra.

Tabla 1. Notación de Muestreo Aleatorio Simple

Universo	**Muestra**	**Definición**	**Ecuación**
N	n	Número de unidades	
Y_i	Y_i	Valor de la variable en la unidad *i*	
μ	$\bar{y}$	Media	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
T	$\hat{T}$	Total	$\hat{T} = N\bar{y}$

Universo	Muestra	Definición	Ecuación
P	$\hat{P}$	Proporción	$\hat{P} = \frac{\sum^{n} yi}{n}$
σ^2	S^2	Varianza	$\sigma^2 = \frac{\sum_{i=1}^{n} (yi - \mu)^2}{N}$ $S^2 = \frac{\sum_{i=1}^{n} (yi - \bar{y})^2}{n-1}$

Fuente: Elaboración propia.

En el MAS como en cualquier otro esquema, se puede estimar el total, la media, la proporción, el número de casos con una característica o total de una clase, una razón, entre otros parámetros. Aquí nos referimos a los que se estiman comúnmente:

Estimación de la media de la población en Muestreo Aleatorio Simple

El estimador de μ es: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

La varianza del estimador $\bar{y}$ es:

$v(\bar{y}) = \left(\frac{N-n}{N}\right)\left(\frac{\sigma^2}{n}\right)$, donde σ^2 se estima por s^2, la cuasivarianza muestral.

Un límite para el error de estimación es:

$K\sigma\bar{y} = K\sqrt{\left(\frac{N-n}{N}\right)\left(\frac{\sigma^2}{n}\right)}$, donde K= $Z^{\alpha}/_{2}$

Y los límites para el error de estimación a través de un intervalo de confianza del 100 (1-α) % son:

$$\bar{y} - Z_{\alpha/2} \times \sigma_{\bar{y}} \leq \mu \leq \bar{y} + Z_{\alpha/2}\sigma_{\bar{y}} .$$

Donde $z_{\alpha/2}$ es el valor critico para el cual $P\left(z \geq z_{\alpha/2}\right) = {}^{\alpha}/_{2}$, y Z es la variable aleatoria Normal Estándar.

La cantidad $\left(\frac{N-n}{N}\right)$ se denomina factor de corrección por población finita. Cuando n es pequeña en relación a N, esto es: $\frac{n}{N} \leq 0.05$, se puede despreciar el factor de corrección y en tal caso:

$$\hat{V}(\bar{y}) = \frac{S^2}{n},$$

esta ecuación de varianza de la media como estimador, se utiliza para poblaciones infinitas.

Estimación del total (T) de la población en MAS

El estimador del total T de una población es:

$$\hat{\tau} = N\bar{y} = \frac{N}{n}\sum_{i=1}^{n} y_i .$$

Su varianza estimada es:

$$\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2\hat{V}(\bar{y}) = N^2\frac{S^2}{n}\frac{N-n}{N} = N(N-n)\frac{S^2}{n}$$

El limite para el error de estimación es:

1.96. $\sqrt{N^2}$. (s^2 /n). (N-n/N), si se fija una confianza del 95% de que se cumpla la precisión. Recuerde que s^2 es la cuasivarianza muestral.

Estimación de la Proporción en MAS

El estimador de P es:

$$\overline{P} = \frac{\sum_{i=1}^{n} yi}{n}$$

Donde y_i es cualitativa binaria.

La varianza del estimador de P es:

$$V\left(\hat{P}\right) = \left(\frac{N-n}{N}\right)\left(\frac{PQ}{n-1}\right)$$

$$Q = 1 - \underline{P}$$

Un límite para el error de estimación:

$$K\sqrt{V\left(\hat{P}\right)} = K\sqrt{\left(\frac{N-n}{n}\right)\frac{PQ}{n-1}}$$

Donde K= $Z^{\alpha}/_{2}$

Cálculo del tamaño de la muestra en MAS

Depende si la población es finita o infinita, y si los datos son cuantitativos o cualitativos.

Para estimar μ (datos continuos)

$$n_0 = \frac{Z^2 \alpha/2 \sigma^2 y}{d^2}$$

(Para Poblaciones infinitas)

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$ (Para Poblaciones finitas).

Se puede utilizar también la ecuación:

$$n = \frac{N Z^2 \alpha/2 \sigma_y^2}{(N-1)d^2 + Z^2 \alpha/2 \sigma_y^2}$$

donde:

d= margen de error

$Z^{\alpha}/_2 = valor\ crítico\ de\ una\ distribución\ normal\ estandar.$

N= Tamaño de la población.

Para estimar P (datos cualitativos):

$$n_0 = \frac{Z^2 \alpha/2 PQ}{d^2}$$

(Para Poblaciones infinitas)

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad n = \frac{N Z^2 \alpha/2 PQ}{(N-1)d^2 + Z^2 \alpha/2 PQ}$$

(Para poblaciones finitas)

En este caso, donde se estiman porcentajes o proporciones y el total de una clase, en variables de interés, que son de si y no, y no se conoce el valor de la proporción poblacional P, ni hay una aproximación de ella, suele utilizarse lo que se conoce como varianza máxima, utilizando P=Q=0.5, lo que garantiza el mayor tamaño de muestra posible, como solución conservadora, o se puede usar la muestra piloto, muestra preliminar o estudios anteriores, para estimar los valores de P y Q, valores que se colocan en la ecuación para el cálculo de n. Cuando se utiliza varianza máxima P=Q=0.5, puede llegar a obtenerse un tamaño de muestra demasiado grande para el dinero que se dispone en el estudio.

Si se desea estimar el total de una clase, por ejemplo:(A), o número de elementos con la característica de interés, para calcular el tamaño de la muestra, se utiliza la ecuación:

$$n = N^3 Z^{\alpha}/_2 \; P Q / (N-1)\, e2 + Z^{\alpha}/_2 \; N^2 P Q,$$

donde: N= Tamaño de la población.

P= proporción de éxitos o elementos con la característica de interés en la población.

P= proporción de fallas. P+Q=1.

$Z^{\alpha}/_2 = valor\ crítico\ de\ una\ distribución\ normal\ estandar.$

e= margen de error absoluto.

Si se desea estimar el total (T) de una variable de interés en la población, la ecuación es:

$n= Z^{\alpha}/_{2}\ N^2\ \sigma^2 / e^2 + Z^{\alpha}/_{2}\ \sigma^2\ N,$

donde:

N= Tamaño de la población.

σ^2= Varianza de la variable de interés en la población.

e= margen de error absoluto.

Ejercicio para estimar la media y el total en MAS

En una auditoría se examinan las cuentas abiertas con diferentes clientes de una empresa en Madrid. De 1.000 cuentas examinan 300 cuentas. La media muestral de las deudas en las cuentas fue de 1040 € y la cuasivarianza muestral fue de S^2= 45.000 €^{2}. Estime la media y el total de la deuda por cobrar para las N= 1000 cuentas, con un intervalo de confianza del 95%.

Para estimar la media:

El estimador de la varianza de la media, el error de estimación y un intervalo de confianza del 95% es:

La media fue: 1.040 € y los estimadores son:

$$\widehat{V}(\overline{y}) = \frac{S_{n-1}^2}{n}\frac{N-n}{N} = \frac{45000}{300}\ \frac{1000-300}{1000} = 105$$

$$2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{105} = 20,49€$$

$$(1.040 \mp 20,49) = (1.019,51\ ,\ 1.060,49)$$

Para estimar el total:

$$\hat{\tau} = N\bar{y} = 1000 \times 1040 = 1.040.000€$$

$$2\sqrt{\hat{V}(\hat{\tau})} = N2\sqrt{\hat{V}(\bar{y})} = 1000 \times 20,49 = 20.490€$$

Y el intervalo de confianza del 95% para el total es:

$$(1.040.000 \mp 20.490) = (1.019.510\ ,1.060.490)$$

Esto indica que el total de la deuda estará entre 1.019.510 y 1.060.490 €, con una confianza del 95%. Observe que se aproxima el intervalo, ya que se usa como $Z^{\alpha}/_{2}$ el valor de 2 y debe ser 1,96.

Este ejemplo fue obtenido del documento: Técnicas Cuantitativas 3, Grado de Marketing e Investigación de mercados de la Universidad de Granada, España.

Ejercicio de tamaño de la muestra en MAS, con población finita

En una Universidad hay 2.000 estudiantes. Para estimar el porcentaje de estudiantes que consume cannabis, se desea conocer el tamaño de una muestra con una confianza del 95% y con un margen de error del 3%. Por ser una población finita, utilizaremos la ecuación para poblaciones finitas. Como se desconoce el porcentaje de consumo de cannabis, aún por muestra piloto,

preliminar o estudios anteriores de referencia, entonces usamos varianza máxima, donde P=.50 y Q=.50 y el valor de $Z^{\alpha}/_{2}$=1,96.

Si aplicamos la ecuación anterior para poblaciones finitas, obtendremos la estimación del tamaño de la muestra:

$$n = \frac{N Z^{2} \alpha/2 PQ}{(N-1)d^{2} + Z^{2} \alpha/2 PQ}$$

n = 2000 x 1,962x 50 x 50 /((2000-1) x 32) + (1,962 x 50 x 50) =698 estudiantes, debemos seleccionar por un procedimiento aleatorio.

Muestreo Aleatorio Estratificado

El Muestreo Aleatorio Estratificado (MAE) se utiliza cuando los elementos (N) de la población, se dividen en grupos o estratos internamente homogéneos y heterogéneos entre ellos, luego se toma una muestra aleatoria dentro de cada estrato. El tamaño de la muestra en el MAE, va a depender, entre otras cosas, del tipo de afijación o reparto de UA en cada estrato, que se utilice.

El MAE garantiza que cada grupo de UA de la población sea representado en la muestra y facilita la coordinación de los trabajos de oficina y campo en cada estrato.

Con el MAE se obtiene información de la población y para cada estrato, se obtiene mayor precisión ya que a un mismo tamaño de muestra, el error de muestreo es menor y se favorece la administración del muestreo.

El MAE a un mismo coste dado que el MAS, se obtiene un límite de error de estimación más pequeño. Y el objetivo de una encuesta es maximizar la cantidad

de información para un coste dado. Este esquema de muestreo es muy recomendado si la población es heterogénea y no muy dispersa.

Estratificar permite:

1. Producir un límite más pequeño para el error de estimación o máximo error admisible.
2. El coste por observación en la encuesta puede ser reducido mediante la estratificación de los elementos de la población.
3. Se puede obtener estimadores de parámetros poblacionales para subgrupos (estratos) de la población y se obtiene una estimación que engloba todos los estratos. Es decir, los estratos se pueden considerar como dominios de estudio, pero no todos los dominios deben ser estratos.

Condiciones de aplicación del MAE

- Existencia de variables estratificadoras.
- Debe conocerse N_h y N.
- El número de elementos por estratos debe ser al menos 2.
- Un elemento pertenece a un solo estrato.
- $\sum N_h = N$.

En la **Tabla 2** se presenta los parámetros y estimadores:

Tabla 2. Notación en Muestreo Aleatorio Estratificado.

Universo	Muestra	Definición	Ecuación
N	N	Nº de unidades de elementos	$N = \sum_{n=1}^{k} N_h \qquad n = \sum_{h=1}^{k} n_h$

Universo	Muestra	Definición	Ecuación
N_h	N_h	Nº de unidades elementales en h-ésimo estrato	
μ	$\bar{y}_{st}$	Media del muestreo estratificado	$\bar{y}_{st} = \frac{\sum_{h=1}^{k} N_h y_h}{N}$
μ_h	$\bar{y}_h$	Media del estrato h	$\bar{y}_h = \frac{\sum_{h=1}^{\mu_h} y_h}{n_h}$
Y_{hi}	N	Valor de la variable en la unidad y del estrato h	$y_h = \frac{\sum_{h=1}^{k_h} y_h}{n_h}$
σ_h^2	N_h	Varianza del estrato h	$S_h^2 = \frac{\sum^{k}(y_{hi} - \bar{y}_h)^2}{n_h - 1}$
$\bar{\sigma}^2{}_{yst}$	$\bar{y}_{st}$	Varianza del estimador de la media en muestreo estratificado	$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{k} N_h^2 \left(\frac{N_h n_h}{Nn} \right) \frac{\sigma_h^2}{n_h}$
P	$\hat{p}$	Proporción de casos	
P_h	P_h	Proporción de casos en el estrato h con la característica	$W_h = \frac{N_h}{N}$

Universo	Muestra	Definición	Ecuación
W_h	W_h	Ponderación de estrato h	

Fuente: Elaboración propia

.

Estimación de la media y total en MAE

El estimador de μ es:

$$\bar{y}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \bar{y}_h$$

La varianza del estimador $\bar{y}_{st}$ será

$$V\left(\bar{y}_{st}\right) = \frac{1}{N^2} \sum_{h=1}^{k} N_h^2 \left(\frac{N_h n_h}{Nn}\right) \frac{\sigma_h^2}{n_h}$$

$$Varianza\ estimada\ de\ \bar{y}_{st} = \hat{V}\left(\bar{y}_{st}\right) = \frac{1}{N^2} \sum_{h=1}^{k} N_h^2 \left(\frac{N_h n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right)$$

Un límite para el error de estimación y un intervalo del 100 (1-α) % viene dado por:

Un límite para el error de la estimación será:

$$K\sqrt{V\left(\bar{y}_{st}\right)} = K\sqrt{\frac{1}{N^2} \sum_{h=1}^{k} N_h^2 \left(\frac{N_h n_h}{Nn}\right) \frac{\sigma_h^2}{n_h}}$$

El intervalo de estimación:

$$\bar{y}_{st} - \boldsymbol{z}_{\alpha/2} \times \left(\frac{1}{N^2} \sum_{h=1}^{k} N_h \left(\frac{N_h - \boldsymbol{n}_h}{N_h} \right) \times \left(\frac{s_h^2}{n_h} \right) \right) \leq \mu \leq \bar{y}_{st} + \boldsymbol{z}_{\alpha/2} \times \left(\frac{1}{N^2} \sum_{h=1}^{k} N_h \left(\frac{N_h - \boldsymbol{n}_h}{N_h} \right) \times \left(\frac{s_h^2}{n_h} \right) \right)$$

Donde s_h 2 es la cuasivarianza muestral.

Estimación de la proporción poblacional (P) en MAE

El estimador de P es:

$$\hat{P}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \hat{P}_h$$

La varianza del estimador $\hat{P}_{st}$ será

Un límite para el error de la estimación será:

$$\hat{V}\left(\hat{P}_{st} \right) = \frac{1}{N^2} \sum_{h=1}^{K} N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{\hat{p}_h \hat{q}}{n_{h-1}} \right)$$

$$K\sqrt{\hat{V}\left(\hat{p}_{st} \right)} = k \sqrt{\frac{1}{N^2} \sum_{h=1}^{K} N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \left(\frac{\hat{p}_h \hat{q}_h}{n_h - 1} \right)}$$

Recuerde que K= $Z^{\alpha}/_{2}$.

La estratificación puede ser usada con el esquema de MAS y con otros esquemas de muestreo.

Las variables para formar estratos (Estratificar) pueden ser geográficas o variables que estén correlacionadas con la variable de interés (Y).

El MAE se utiliza si se tiene un marco de muestreo donde se conoce N y Nh, y si hay variables auxiliares clave estratificadoras. El número de elementos por estrato debe ser por lo menos 2 y cada elemento pertenece a un solo estrato. Además, el MAE asegura que casos de los pequeños estratos, caigan en la muestra, que usando MAS, podrían no aparecer. Si los estratos son homogéneos con poca variabilidad dentro de estrato, las estimaciones dentro de estratos serán precisas y el combinado de toda la población es preciso, más que MAS en toda la población , la varianza de los estimadores es menor cuando se produce un efecto de la estratificación y es posible en MAE utilizar distintos MM por estrato.

La estratificación es muy utilizada en poblaciones Asimétricas (Skewed) cuando la distribución de frecuencia de la variable de interés no es simétrica, tiende a la derecha o izquierda. Este fenómeno se presenta en variables de interés en empresas grandes, comercios, granjas, entre otras, donde pocas unidades tienen mucha influencia en el valor de la variable de interés y muchas UA aportan poco a ese valor. En estos casos, se forman estratos con variables auxiliares clave y se le proporciona mayor tasa muestral a los estratos de las UA más grandes, incluso puede existir estratos donde se haga censo o se seleccionen la UA con probabilidad uno (1). Esto produce mejoras en la estimación.

No se aconseja utilizar un alto número de estratos (Kish, 1972) debido a que estratos pequeños contribuyen poco a la ganancia con la estratificación y puede aparecer estratos vacíos o con muy pocos elementos. Kish, ha conseguido mayor precisión en los estimadores usando pocos estratos, pero los estratos deben ser muy diferentes entre ellos y debe existir homogeneidad dentro de estrato. En el MAE se utiliza la información auxiliar para formar estratos. Los estratos son homogéneos internamente con respecto a la variable de interés o una variable correlacionada con esta.

Si se utiliza afijación proporcional u óptima, las estimaciones del parámetro tienen precisión tan o mejor que MAS. El MAE asegura mayor representatividad de la muestra, seleccionando una muestra aleatoria en cada estrato. Es útil si se hace necesario tener diferentes esquemas de muestreo en cada estrato, que pudiese reducir el coste total. Cuando los estratos se forman con variables correlacionadas con la variable de interés, se produce un aumento en la precisión de las estimaciones, comparado con MAS, sobre todo si el tamaño de la muestra es grande. La ganancia en precisión está relacionada con la correlación entre la variable objeto de estudio y la/las variable/s auxiliar/es. Una vez formados los estratos, se toma la muestra dentro de cada a estrato usando MAS, MAE, MS o MPC.

Algunas veces se desea formar estratos, pero no se puede ubicar las unidades de muestreo en los estratos hasta después de seleccionar la muestra; es decir, no se conoce a priori el tamaño de los estratos, en este caso se puede utilizar la postestratificación, la cual consiste en estratificar una muestra, no la población. Para postestratificar se toma una muestra aleatoria simple o sistemática y se forman estratos con esa muestra. La efectividad de la postestratificación, después de consultar teóricos (Cochran, 1976; Ras, 1980; Pérez, 2000) se justifica solo si la muestra aleatoria es grande, esto coincide con los resultados de un esquema MAE con afijación proporcional, donde coinciden prácticamente las varianzas de los estimadores de ambos procedimientos de muestreo.

Tamaño de la Muestra para Estimar μ en MAE

1)
$$n_0 = \frac{\sum_{h=1}^{k} w_h \sigma^2_h}{V} \qquad W_h = \frac{N_h}{N}$$

$$V=\frac{d^2}{K^2}=\frac{d^2}{Z^2{}_{\alpha/2}} \qquad n=\frac{n_0}{1+\frac{n_0}{N}} \qquad n_h=n.w_h$$

La afijación proporcional en MAE, tiene mayor efecto si se utiliza cuando las medias de la variable de interés Y, en cada estrato , difieren altamente, en este caso el MAE es superior al MAS, si ello no ocurre, entonces, la afijación proporcional ofrece muy poca ventaja, en cuanto a reducción de la varianza del estimador.

2) Dado el costo del estudio, minimizar varianza del estimador

Hay 2 casos:

2.1) Si deseamos conocer cuál sería el valor del tamaño de la muestra (n) si se desea gastar C (cantidad de Euros), sabiendo que el coste fijo es Co y el coste variable por estrato es C_h, y se conocen y estiman la varianza y tamaños de los estratos σ^2_h y N_h:

$$n=\frac{(c-c_0)\sum_{h=1}^{k}\frac{N_h\sigma_h}{\sqrt{C_h}}}{\sum_{h=1}^{k}N_h\sigma_h\sqrt{C_h}}$$

Luego la afijación o repartición de la muestra en los estratos es:

$$n_h=\frac{N_h\sigma_h\sqrt{C_h}}{\sum_{h=1}^{k}\left(\frac{N_h\sigma_h}{\sqrt{C_h}}\right)}$$

2.2) Si hay un coste total fijo y no varían los costes por unidad en cada estrato y varían las varianzas y tamaños de los estratos, la descomposición óptima será la de Neyman:

.

$$n = \frac{\left(\sum^{k} N_h \sigma_h \right)^2}{N^2 V + \sum_{h=1}^{k} N_h \sigma^2_h} \qquad V = \frac{d^2}{Z^2 \alpha/2}$$

Y la afijación es

$$n_h = \frac{N_h \sigma_h}{\sum^{k} N_h \sigma_h} \cdot n$$

Recuerde que d=e es el margen de error o cota de error.

La afijación de Neyman produce menor Error estándar que la afijación proporcional cuando los tamaños de los estratos y las varianzas de Y (variable objeto de estudio) dentro de cada estrato, varían sustancialmente de estrato a estrato, en este sentido. Neyman toma en cuenta que la tasa de muestreo es proporcional al producto del tamaño del estrato por la Desviación Estándar de Y en el estrato.

3) Dada la varianza del estimador $\bar{y}_{st}$, minimizar coste

Se tiene información sobre los costes variables en cada estrato c_h y s_h y N_h, y se fija el error de estimación y la confianza en esa estimación:

$$n = \frac{\left(\sum_{h=1}^{k} N_h S_h \sqrt{C_h}\right)\left(\sum_{h=1}^{k} N_h S_h / \sqrt{C_h}\right)}{N^2 V + \sum_{h=1}^{k} N_h S_h}$$

Donde:

$$V = \frac{e^2}{Z^2 {}_{\alpha/2}}$$

Luego la afijación a los estratos es:

$$n_h = \frac{\left(N_h S_h / \sqrt{C_h}\right)}{\sum_{h=1}^{k}\left(N_h S_h / \sqrt{C_h}\right)} \cdot n$$

Muestreo Aleatorio Estratificado (Desproporcional)

El MAE desproporcionado con afijación no proporcional en los estratos, se utiliza cuando la variable principal que se mide tiene una distribución de probabilidades altamente sesgada o asimétrica. Tienen pocas unidades de muestra con altos valores de la variable y muchas unidades con valores bajos.

Son variables relacionadas al tamaño como las ventas totales en comercios, producción de leche en fincas, número de pacientes por hospital, precio de la casas, número de cargos vacantes en empresas, capital suscrito, entre otras. En este caso la estratificación por tamaño y la distribución desproporcional es más eficiente.

Ejemplo sencillo para mostrar cómo funciona la afijación no proporcional en un MAE, en una población asimétrica, se muestra en la **Tabla 3**:

Tabla 3. Datos necesarios para el cálculo de la afijación en los estratos

Estrato(Número de alumnos por institución	Numero de instituciones o escuelas	Matrícula por estrato	Desviación Estándar por estrato(σ)
< 1000	661	292671	236
1000-3000	205	345302	625
3001-10000	122	672728	2008
>10000	31	573693	10023
	1019		

Fuente: Elaboración propia.

Queremos que el lector perciba en este ejemplo que el número de instituciones es mayor en el estrato cuatro (4) donde hay una matrícula grande y en el estrato uno (1) hay más instituciones (661) pero menos matrícula de alumnos, ello deja ver que es una población asimétrica. Otra cosa es que en poblaciones asimétricas, el estrato de escuelas más grandes tiene mayor Desviación Estándar de la variable de interés Y, en el estrato cuatro (4), la Desviación Estándar es de 10023 en comparación con el estrato uno (1) donde la Desviación Estándar es de 236.

Lo que hacemos es darle mayor tasa muestral a los estratos con mayor matrícula, aun cuando se tengan menos instituciones o escuelas. En la **Tabla 4**, se presenta una ponderación, se multiplica el número de alumnos Nh por la Desviación Estándar encada estrato.

Por otro lado, en este ejemplo vemos como la afijación o reparto de la muestra a los estratos no se hace como siempre, es decir, proporcional al número de elementos en cada estrato de la población, que serían las instituciones, sino que se toma en cuenta la variabilidad en cada estrato y el producto del número de instituciones por la Desviación Estándar en cada estrato.

Tabla 4. Cálculo de la afijación de la muestra en cada estrato.

Estrato	N_h	$N_h \times \sigma_h$	$N_h \times \sigma_h / \sum N_h \times \sigma_h$	Afijación	%
1	661	155.996	0.1857	65	10
2	205	128.125	0.1526	53	26
3	122	244.976	0.2917	101	83
4	31	310.713	0.3700	31	100%
	1019	839.810			

Fuente: Elaboración propia.

En el estrato 1 se van a seleccionar 65 UA, el dos (2) se seleccionan 53, en el trae (3) se seleccionan 101 y en el cuatro(4) se seleccionan 31,

En el estrato 1 el % de afijación se calcula como: 65/661x100 y para los otro se hace de la misma forma.

En el estrato 4 con menos instituciones se toma el 100 % (31) y las restantes (219), porque el tamaño de la muestra era 250, se le restan las 31 del estrato 4 y correspondería a cada estrato el siguiente reparto:

$$219\left\{\frac{0.1857}{0.1857+0.1526+0.2917}\right\}=65$$

$$219\left\{\frac{0.1526}{0.1857+0.1526+0.2917}\right\}=53$$

$$219\left\{\frac{0.2917}{0.1857+0.1526+0.2917}\right\}=101$$

Ejemplo de la aplicación de un MAE

Este ejemplo está en Scheaffer, Mendenhall y Ott, 1987:

Una empresa decide realizar una encuesta por muestreo para estimar el número de horas promedio por semana que se ve televisión en los hogares de un Municipio. El pueblo A circunda una fábrica, el pueblo B es un suburbio exclusivo de una ciudad vecina y C es un área rural. Hay

20 hogares en A, 8 en B y 12 en C.

Se tiene la siguiente información:

Tabla 5. Tiempo en horas que se ve televisión por semana

ESTRATO 1 A				ESTRATO 2 B				ESTRATO C			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	15	41	25	30	14	30	20	1
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								

Fuente: Elaboración propia.

Tabla 6. Resumen del Tiempo en horas que se ve televisión por semana

ESTRATO 1 A	ESTRATO 2 B	ESTRATO 3 C
n1=20	n2=8	n3=12
Y1=33,900	Y2=25,125	Y3=19,000
$S1^2$=35,358	$S2^2$=232,411	$S3^2$=87,636
N1=155	N2=62	N3=93

Fuente: Elaboración propia.

1. Estime el tiempo medio que se ve televisión por semana en todo el Municipio

Un estimador de μ es:

$$\bar{Y}st = \sum_{n=1}^{k} Nh/N.\ \bar{Yh}$$

K=3 (Número de estratos)

$\bar{y}_{st}$ = (155/310) 33,90+ (62/310) 25,125+ (93/310) 19,0= 27,7

Así, estimamos que el número promedio de horas por semana que se ve televisión en los hogares en todo el Municipio es: $\bar{y}_{st}$=27,7 horas.

2. Estime el margen de error o error en la estimación con un 95% de Confianza y construya el intervalo de confianza para la media (μ)

El error en la estimación es:

e= $Z^{\alpha}/_{2}$.x $\sigma_{yst,}$ el producto del valor critico de la distribución Z, para un 95% de confianza, por el Error Estándar del estimador.

La varianza del estimador de la media (μ) es:

$$V\left(\bar{y}_{st}\right)=\frac{1}{N^2}\sum_{h=1}^{k} N_h^2\left(\frac{N_h\, n_h}{Nn}\right)\frac{\sigma_h^2}{n_h}$$

Y un estimador de la varianza del estimador $\bar{y}_{st}$ es:

$$Varianza\ estimada\ de\ \bar{y}_{st}=\hat{V}\left(\bar{y}_{st}\right)=\frac{1}{N^2}\sum_{h=1}^{k} N_h^2\left(\frac{N_h\, n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right)$$

$S^2_{\bar{Y}st}$ =1/(310)²[(155)²(0,871)(35,358)+(62²)(0,871)(232,41)+(93²)(0,871)(87,626)]

$S^2_{\bar{y}st}$ =1,97

.

Recuerde que la estimación es en la muestra y se hace con la Cuasivarianza muestral (S^2).

El error de estimación debe ser menor de e= 1,96x√1,97= 2,7 horas.

Y un intervalo de confianza del 95%, para estimar μ es:

$27,7 \pm Z\alpha/2 \cdot \sqrt{1,97}$

$27,7 \pm 1,96 \cdot \sqrt{1,97}$

Zα/2= 1,96.

Esto es: [24,96; 30,44].

La media del número de horas por semana que se ve televisión en los hogares de ese Municipio está entre 24,96 y 30,44 horas.

Muestreo Sistemático

Se enumeran los elementos de la población de 1 a N, en cualquier orden. Se divide la población en n partes de tamaño k (Intervalo de selección), $K = \frac{N}{n}$. Se elige un número al azar entre 1 y k que se llama i (arranque aleatorio) y de este número en adelante se toman los elementos que ocupen la misma posición en los intervalos k sucesivos: i, i+k, i+2k, 1+3k,. . . . . i + (n-1) k. Este intervalo de selección (1 en K) puede presentarse como tiempo, espacio u orden, una UA puede aparecer cada 10 minutos, cada dos (2) metros o tomarse cada cinco (5) elementos en orden de una lista. Al MS se le llama Muestreo sistemático lineal. La primera unidad o elemento se selecciona al aza r y las siguientes de acuerdo al intervalo de selección K, esa primera unidad, es la que determina la muestra sistemática.

Cada UA tiene una probabilidad de inclusión igual que en MAS, es decir, n/N= π, pero no toda combinación de n UA, tiene la misma probabilidad. Cada muestra sistemática tiene una probabilidad de 1/K. Se puede seleccionar solo muestras que estén separadas por K, nunca dos UA que estén juntas o seguidas podrán formar parte de la muestra sistemática.

Si existe un MM explícito y se conoce N y n, se utiliza K=N/n. Pero el MS no requiere de marco explícito para aplicarse. Si no se conoce N. puede tomarse un valor de K que nos asegure mínimo deseado de tamaño de muestra, y N se estima con el despeje en K=N/n. El valor de K se cuadra para obtener el valor de n. Esto generalmente nos proporciona un n mayor o igual al requerido. Nosotros recomendamos calcular n con la ecuación de MAS para poblaciones infinitas, como aproximación, también como aproximación se puede usar un N que provenga de estudios anteriores, recuerde que el muestreo tiene mucho de arte sin alejarse de la ciencia.

Si se conoce el tamaño de la población N, si K es un número entero, se calcula n. Pero si K no es entero, el tamaño de la muestra puede variar, dependiendo del arranque aleatorio que se utilice; sin embargo, esto no impide la validez del MS. Muchos autores aseveran que en la práctica, es posible utilizar un valor de K sin importar que sea entero o no, si tiene decimales, se aproxima al número inmediatamente superior o siguiente valor, indistintamente de los valores decimales.

Si existe un MM explícito o tangible en físico, y se tiene una información auxiliar clave correlacionada con la variable de interés, entonces los elementos se ordenan en relación a la variable auxiliar y se selecciona un intervalo fijo, para luego, obtener la muestra sistemática.

Entre las características del muestreo sistemático se tienen: fácil llevarlo en el campo. Puede proporcionar mayor información por unidad de coste que MAS. La muestra se dispersa por toda la población.

El MS se puede recomendar cuando:

- ✓ La población se presenta en el marco en forma aleatoria.
- ✓ Las unidades son heterogéneas y grande la población.
- ✓ hay un gran número de estratos en un MAE.
- ✓ hay submuestreo (Generalmente en las últimas etapas de un muestreo Polietápico).
- ✓ La población tiene variación continua.
- ✓ Los lugares son de difícil acceso, como por ejemplo, los barrios de una ciudad.

Con el MS se simplifica el proceso de selección de la muestra en comparación con MAS, la muestra rastrea, se reparte o desparrama por toda la población, no requiere de MM para ser utilizado, tiene también teoría estructurada, está menos expuesto a errores de selección por parte del investigador que hace la entrevista o medición, nos proporciona la obtención de una muestra lo suficientemente representativa, aun no conociendo el tamaño de la población, la población no tiene que ser numerada para su uso, pero los elementos deben tener representación física, es más fácil de usar y menos costoso que MAS, el MS lo que hace es formar estratos de tamaño K, y puede ser tan preciso como MAS con un solo elemento por estrato, en el MS las unidades muestreadas ocurren en la misma posición relativa en cambio en un MAE, la posición la determina la aleatoriedad. El MS tiene mucha facilidad para ser usado en campo, no se usa la tabla de números aleatorios, generalmente proporciona más información por unidad de coste que el MAS.

De igual forma, la muestra sistemática puede verse como una muestra por conglomerados, donde hay K posibles muestras sistemáticas de tamaño n. Seleccionar una muestra sistemática equivale a seleccionar uno de los K Clusters al azar, se selecciona un solo Cluster al azar de la población de Clusters de la población. Así, una muestra sistemática equivale a una muestra aleatoria simple de un solo Cluster, de la población done hay K Clusters. Y la probabilidad de seleccionar un Cluster es de 1/K. Si las muestras son sistemáticas de tamaño n,

tienen la misma probabilidad de ser seleccionadas, pero ello no es cierto, si las muestras no son sistemáticas.

De la misma forma, el MS facilita la localización en lugares de difícil acceso, caminos, terrenos difíciles de entrar, y nos permite ir a elementos que no están definidos en el marco, pero son parte de la población objetivo y además, se utiliza mucho en grandes ciudades donde no se tienen los listados de viviendas u hogares. El MS se utiliza también en situaciones prácticas, donde no se conoce el tamaño de la población, queremos hacer énfasis en esto por su importancia.

El MS se puede utilizar en MM que no estén numeradas la UA o no sea explícito el marco, en estos casos para la selección de la muestra se puede utilizar un instrumento, por ejemplo, una regla u otro, con medida de distancia en una serie de documentos, recibos, entre otros, o hacerse una selección por computadora con un programa específico, como Excel o SPSS.

El MS es preciso si las UA o elementos dentro de la muestra sistemática son heterogéneos e imprecisos, si son homogéneos, ya que, si hay poca variabilidad e la variable de interés o de una correlacionada con ella, se repite la misma información.

Otro aspecto muy importante de hacer notar es que en una muestra sistemática siempre hay elementos muy cercanos que no van a aparecer en la muestra, y no se puede obtener una estimación insesgada de la varianza del estimador en una sola muestra sistemática. Cochran (1976) menciona que no hay método confiable para estimar esta varianza y para superar este problema muchos autores recomiendan el Muestreo Sistemático Replicado.

Este tipo de muestreo es adecuado para las situaciones en donde la población es grande y con alto nivel de varianza.

Entre las áreas más frecuentes donde se utiliza el MS están: En líneas de producción en fabrica; en el Control de Calidad, debido a la rapidez, facilidad y la baja varianza de este tipo de muestreo; en Auditoría, cuando se comprueban

muchas partidas; en la Investigación de Mercados, cuando hay personas en cajas o en movimiento en mercados o lugares donde se presentan los consumidores; en el tráfico de vehículos de todo tipo, en carreteras, caminos o puentes, en bancos, en cajeros automáticos, donde no hay marco lista.

Estimación de una media de la población en muestreo sistemático

El estimador de la media μ es:

$$\mu = \overline{y}_{sy} = \frac{\sum_{i=1}^{h} y_i}{n}$$

Según Scheaffer, Mendenhall y Ott (1987), el subíndice sy significa que pertenece al muestreo sistemático.

El estimador de la varianza del estimador $\bar{y}$ es:

$$\hat{V}\left(\bar{y}_{sy}\right) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$$

Y el límite para el error de estimación es:

$$K\sqrt{\hat{V}(y_{st})} = K\sqrt{\frac{S^2}{n}\left(\frac{N-n}{N}\right)}$$

Si N es desconocido, se elimina $\left(\frac{N-n}{N}\right)$

Se observa que la varianza estimada de $\bar{y}_{sy}$ es similar a la varianza estimada de $\bar{y}$ en el MAS, mientras que la varianza poblacional de $\bar{y}_{st}$ puede expresarse en términos de un coeficiente de correlación entre elementos, y está dada por $V\left(\bar{y}_{sy}\right) = \frac{\sigma^2}{n}\left[1 + (n-1)\,\rho\right]$.

Donde ρ, es la correlación entre los pares de elementos de la muestra sistemática.

Si ρ es alto, hay alta correlación entre los elementos, en relación a la característica de interés y la varianza estimada de $\bar{y}_{sy}$ es mayor que la de la media en MAS para un ρ cerca de cero (0) y N grande, el Muestreo Sistemático es equivalente al MAS.

Scheaffer, Mendenhall y Ott (1987) mencionan que una población aleatoria es aquella donde sus elementos están ordenados al azar, es ordenada si los elementos en la población están ordenados en magnitud, de acuerdo a una característica auxiliar clave correlacionada con la variable de interés u objeto de estudio y es periódica, si los elementos de la población tienen una varianza cíclica.

Una población se dice ser aleatoria, si las unidades de análisis se ordenan al azar, como por ejemplo, una lista de estudiantes ordenada por nombre y apellido, en este caso, el orden de identificación o posición del estudiante no se correlaciona con Y, la variable de interés, por ejemplo, sus calificaciones. Cuando la población es aleatoria resulta indiferente utilizar MAS o MS, siendo. $\rho \approx 0$.

Una población es ordenada cuando sus elementos están ordenados de acuerdo con una característica correlacionada con la variable de interés Y. En este caso $\rho \leq 0$ y se prefiere el MS en lugar del MAS.

Una población es periódica cuando los valores de la variable de interés Y, tienen una variación cíclica y en este caso $\rho \geq 0$, aquí se prefiere el MAS que el MS. Como ejemplo de una población periódica se pueden mencionar las ventas diarias en locales comerciales, muchos autores plantean que para evitar el problema de la periodicidad, se debe cambiar varias veces el arranque aleatorio i, mezclándose los elementos de la población, quienes se comportan como una muestra aleatoria, pudiéndose utilizar las ecuaciones de MAS. Otra forma de eliminar la periodicidad, es utilizar el MS con intervalo variable.

Raj (1980) menciona que el MS funciona mal cuando hay periodicidad en los datos y el intervalo muestral (1 en K) se acomoda a esa periodicidad.

Lo anterior deja claro que el comportamiento del MS (Muestreo Sistemático) depende de la naturaleza de la variable Y, objeto de estudio, si es aleatoria, continua o periódica.

Estimación del total en Muestreo Sistemático

El estimador del total T es:

$\hat{T} = N\bar{y}_{sy}$

Donde se necesita conocer N.

En este caso un estimador de la varianza del estimador es:

$$\hat{V}(\hat{\tau}) = N^2\hat{V}(\bar{y}_{sy}) = N^2\frac{S^2}{n}\left(\frac{N-n}{N}\right)$$

Y un límite para la cota de error es:

$e = K.\ \sigma_{\hat{t}_{sy}}$

$e = K.\ \sqrt{N^2}\ (N\text{-}n/N).\ s^2/n.$

Recuerde que s^2 es la cuasivarianza muestral y con esta cota de error, podemos construir un intervalo de confianza para T.

Estimación de la Proporción Poblacional

Si se utiliza un muestreo sistemático para estimar una proporción poblacional (P), el estimador se denota por $\hat{P}_{sy}$ y sus propiedades son:

$$\hat{P}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Para $y_i = 1$ o 0.

La varianza estimada de $\hat{P}_{sy}$ es:

$$\hat{V}\left(\hat{P}_{sy}\right) = \frac{\hat{P}_{sy}\,\hat{q}_{sy}}{n-1}\left(\frac{N-n}{N}\right)$$

Donde $\hat{q}_{sy} = 1 - \hat{P}_{sy}$ y el límite para el error de estimación es:

$$K\sqrt{\hat{V}\left(\hat{P}_{st}\right)} = K\sqrt{\frac{\hat{P}_{sy}\,\hat{q}_{sy}}{n-1}\left(\frac{N-n}{N}\right)}$$

Si N es muy grande con respecto a n, se puede omitir $\left(\frac{N-n}{N}\right)$ o factor de corrección.

Determinación del tamaño de la muestra sistemática

Para la estimación de la proporción poblacional P, Se utiliza la ecuación $K\sqrt{\sigma_{\bar{y}_{sy}}} = d$ o margen de error en la estimación y K es un múltiplo del margen de error que equivale a $Z_{\alpha/2}$, en una Normal Estándar. Para despejar n se debe conocer σ^2 y P.

Si la población es aleatoria (suposición razonable) se puede usar la ecuación:

$$n = \frac{N\,Z_{\alpha/2}^2\,.pq}{(N-1)\,d^2 + Z^2{}_{\alpha/2}\,.pq}$$

P= Proporción de éxitos

q= Proporción de fallas

Para estimar l total de elementos de una clase dicotómica se utiliza la ecuación:

$n = NPQ/ (N-1)\ e^2/Z^2_{\alpha/2}.N^2 + PQ$

Para estimar la Media y el Total se utilizan las ecuaciones:

$n = N.\sigma^2/ (N-1).e^2/Z^2_{\alpha/2} + \sigma^2$ (Para estimar μ),

$n = N.\ \sigma^2/ (N-1).e^2/Z^2_{\alpha/2}\ .\ N^2 + \sigma^2$ (Para estimar T).

Debe quedar claro que estas ecuaciones solo se utilizan en un MS si la población es aleatoria, donde los resultados de un MS tienen mucha relación con los de MAS.

Ejercicios planteados en contexto donde se aplica el Muestreo Sistemático

1. Suponga que determinada empresa en Madrid requiere conocer el ingreso medio por ventas (en recibos) del mes pasado, donde se registraron N ventas y usted decide seleccionar n recibos para estimar el ingreso medio. Si el número de recibos es alto, no es fácil numerarlos y luego usar una tabla de números aleatorios para seleccionar la muestra de recibos de forma aleatoria. Cree usted entonces que es más fácil e igualmente preciso usar un esquema de Muestreo Sistemático para estimar la media de ingreso con una cota de error y una confianza predeterminada y luego construir un intervalo de confianza.
2. En una empresa de fabricación de botellas, el coordinador de Calidad desea conocer la proporción de botellas con algún defecto en la línea de

producción. Se conoce el número de botellas que se fabrican en un día y elige un número n de botellas a ser muestreadas. Él puede utilizar un Muestreo Sistemático de intervalo k y seleccionar la primera botella con arranque aleatorio y luego cada k consecutivo selecciona las otras, hasta completar el tamaño de muestra seleccionado.

3. Una tienda de descuento grande da bonos al personal de ventas sobre la base de su monto medio en ventas. Puesto que cada vendedor hace cientos de ventas cada mes, el gerente de la tienda decidió basar el importe medio de venta de cada vendedor en una muestra aleatoria de las ventas de la persona. Dado que los registros de ventas se mantienen en libros, el uso de muestreo sistemático aleatorio es conveniente. Supongamos que un vendedor ha hecho 800 ventas al mes, y la gerencia quiere elegir una muestra de 30 ventas para la estimación del importe medio de todas las ventas.
4. Una población se compone de 100 elementos dispuestos en algún orden. Cada estrato de 10 elementos en el orden de la disposición tiende a ser similar en sus valores. Un investigador selecciona "cada 10" muestra sistemática. El primer ítem, elegido al azar, es el 6, y su valor es 20. Los siguientes elementos de la muestra son, por supuesto, el 16, 26, etc. Los valores de todos los ítems de la muestra sistemática son los siguientes: 20, 25, 27, 34, 28, 22, 28, 21, 37, 31.

Muestreo por Conglomerados

Una muestra por conglomerados es una muestra aleatoria en la cual cada unidad de muestreo es una colección o conglomerado de unidades de análisis o elementos. A pesar de que una muestra por conglomerados genera mayor varianza que una aleatoria simple, la ganancia en costo hace favorable su uso. Si el coste de listar todos los elementos o de obtener las observaciones o mediciones

en los elementos se incrementa con la distancia, entonces el muestreo por conglomerados es menos costoso.

En el MPC, si el listado de los elementos de la `población no se tienen porque no los hay, es difícil obtenerlos, imposible o muy costoso, y la población es muy dispersa geográficamente o está formada por delimitaciones naturales, no existe un esquema de muestreo estadístico mejor que el MPC, no se necesita tener los listados de los elementos de la población, sino de los conglomerados que conforman esa población, luego se elige una muestra aleatoria, sistemática aleatoria, de conglomerados, o se eligen con PPT.

En los conglomerados o Clusters elegidos es donde se depura y actualiza el marco de la unidades o elementos, más aun, teniendo la lista de todos los elementos de la población objeto, el MPC se justifica porque reduce el coste, ya que los elementos a observar o medir dentro de los conglomerados seleccionados, se encuentran relativamente cerca, y así, se ahorra en los gastos de traslado y es más fácil la supervisión, coordinación y operatividad. Por otro lado, si es más barato obtener la información en todos los elementos de cada conglomerado, se prefiere usa el MPC en una sola etapa, es decir, se hace enumeración completa en cada conglomerado. Pero si no es barato obtener esa información, se procede a utilizar el MPC Bietápico o Polietápico, estos dos esquemas se utilizan más cuando los conglomerados son homogéneos internamente y además, es barato obtener la información.

Los conglomerados pueden ser de igual o desigual tamaño, sobre esto, Lohr (2010) afirma que la diferencia entre ellos está en que la variación entre totales de la variable de interés en la investigación (Ti) de cada conglomerado más grande, cuando estos son desiguales.

Se puede obtener conglomerados de igual o tamaño parecido de forma real, o planeada, pero en la realidad la mayoría de los conglomerados son desiguales.

Es importante mencionar que los conglomerados pueden ser seleccionados con o sin reposición. Aun cuando se pierde precisión con el uso de conglomerados en

comparación con MAS, esta pérdida de precisión se compensa por la reducción de costes (para recoger los datos, la distancia es más corta y se aminoran los costes de viaje), es decir, el entrevistador puede visitar más unidades en un mismo tiempo y se obtiene mayor información por unidad de coste.

Una estrategia muy utilizada por los muestristas para ganar precisión en un MPC es utilizar mayor número de Clusters, pequeños y dispersos, en la primera etapa y menos unidades secundarias o de segunda etapa.

El muestreo por conglomerados es efectivo cuando:

1. No se encuentra disponible o es muy costoso obtener un marco muestral que liste todos los elementos.
2. El coste de obtener observaciones o mediciones se incrementa con la distancia que separa los elementos de la población, con el muestreo por conglomerados se reduce ese costo, más aun cuando la población se considera grande.

Para realizar un muestreo por conglomerados se debe especificar los conglomerados. Por teoría, deben ser muchos y pequeños ya que generalmente hay correlación entre elementos dentro de un conglomerado.

Una cuadra es un conglomerado de casas, una carnicería es un conglomerado de clientes, un hospital es un conglomerado de enfermos, una escuela en un conglomerado de niños, una planta es un conglomerado de hojas y frutos, Un vuelo en avión, es un conglomerados de pasajeros, una vivienda es un conglomerado de hogares o personas, unas determinadas horas de día es un conglomerado de' programas de televisión, un salón de clase es un conglomerado de estudiantes, un intervalos de minutos es un conglomerado de vehículos en una carretera.

La diferencia entre estratos y conglomerados es que los estratos deben ser homogéneos internamente, pero heterogéneos entre ellos. Los conglomerados deben ser heterogéneos internamente, por ser cada uno una representación de la

población y un conglomerado debe ser muy similar a otro. Se puede usar la técnica de la estratificación y en cada estrato, utilizar el MPC, el MS o el MAS.

Un aspecto importante en MPC se refiere al control de la variabilidad en el tamaño de los conglomerados, debido a que mucha variabilidad en los tamaños va a influir en la varianza de los estimadores de los parámetros de la población, para ello se estila hacer lo siguiente:

1. Tratar de definir o formar los conglomerados de igual tamaño, para esto se puede usar una variable auxiliar clave correlacionada con la variable de interés.
2. Estratificar por tamaño de los conglomerados.
3. Usando un estimador indirecto de razón, con variable auxiliar en el denominador.
4. Utilizando MPC con PPT, muestreando conglomerados con una medida de tamaño correlacionada con la variable de interés o PPT al número de elementos del conglomerado.

Una moderada variación en los tamaños de los conglomerados tiene poco efecto en la varianza de los estimadores como la media, el total, la proporción y la razón.

En la **Tabla 7**, se muestra ejemplos de unidades primarias y secundarias, y variables de interés.

Tabla 7. Unidades primarias y secundarias en Muestreo por Conglomerados (MPC).

Variable de Interés	**Unidades Primarias**	**Unidades Secundarias**
Trabajo, ingreso, drogadicción.	``Manzanas`` o cuadras	Individuos dentro de la manzana o cuadra.
Producción de trigo, Maíz, etc. Carne, leche.	Municipios	Predios

Contenido de calcio, nitrógeno, etc.	Árboles	Hojas o frutos
Gastos, ingresos, periódicos y revistas leídas.	``Manzanas``	Familias
Calificaciones, ingresos de los alumnos, entre otros.	Grupos de alumnos	Alumnos

Fuente: Elaboración propia

¿Cómo se selecciona una muestra por conglomerados?

- Divida la población entera en los conglomerados o Clusters que la componen, bien definidos.
- Los Clusters son las unidades de muestreo, no los elementos de la población.
- Use una selección de una muestra aleatoria simple, generalmente sin reemplazo o sustitución o una muestra sistemática aleatoria.
- Al seleccionar los Clusters en la muestra, concéntrese solo en ellos, revíselos, depúrelos y actualícelos, de ser necesario.
- En estos Clusters seleccionados, haga una enumeración completa, tome todos los elementos en todos los Clusters y registre toda la información necesaria de la variable de interés y otras relacionadas, en cada unidad o elemento.

Tamaño del conglomerado y la correlación intraclase

Hay un fenómeno que ocurre en el MPC llamado correlación intraclase o entre elementos dentro de cada conglomerado, la cual se mide por el coeficiente ρ. Si

ρ=0, significa que la unidades o elementos en cada Cluster están esparcidas aleatoriamente y se dice que no hay correlación intraclase. No obstante, en la práctica este coeficiente es positivo, ubicándose generalmente entre valores de 1,5 hasta 3,0, entre otros. El coeficiente ρ decrementa al aumentar M (tamaño del conglomerado), pero la tasa de decremento es mucho más baja en comparación con el aumento de M.

En el tamaño del conglomerado existe un balance entre ρ y M, por ello se prefieren los conglomerados pequeños, ya que M aumenta más rápido que ρ y la varianza de los estimadores aumenta más al aumentar M. Los conglomerados seleccionados en la muestra, a pesar de ser pequeños, deben atrapar la varianza de la variable Y de interés en toda la población.

Por otro lado, es deseable que el efecto de Diseño (Deff) sea cercano a cero (0), este efecto lo proporciona el término [1+ (M-1) ρ], donde los valores de M y ρ intervienen en el efecto de Diseño de cualquier MPC, Monoetápico, Bietápico o Polietápico.

En la **Tabla 8**, se presenta la notación para Muestreo por Conglomerados de tamaños desiguales, donde se incluye los términos de la población, muestra, definición y ecuación.

Tabla 8. Notación para Muestreo por Conglomerados de tamaños desiguales

Población	Muestra	Definición	Ecuación
M_i		Nº de elementos en el i- ésimo conglomerado	
M	m	Nº totales de elementos	$M = \sum_{i=1}^{N} M\,i \qquad m = \sum_{i=1}^{N} M\,i$
N	n	Nº de conglomerados	
$\bar{M}$	$\bar{m}$	Promedio del Nº de elementos por conglomerados	$\bar{M} = \frac{M}{N} \qquad \bar{m} = \frac{1}{n}\sum_{i=1}^{N} M\ i$

Población	Muestra	Definición	Ecuación
Y_i	Y_i	Total de las observaciones en el i- ésimo conglomerado	
Y_{ij}	Y_{ij}	Valores de la observación j en el conglomerado i	
$\bar{Y}_i$	$\bar{y}_i$	Promedio por conglomerado o del conglomerado i	$\bar{Y}_i = \sum_{i=1}^{Mi} \frac{Y_i}{M_i}$ $\bar{Y}_i = \frac{\sum y_{ij}}{M_i}$
$\bar{y}$	$\bar{y}_n$"	Media por elemento	$\bar{\bar{y}} = \sum_{i=1}^{N} \frac{Y_i}{N}$ $\bar{y}_n'' = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$ $\bar{y}_n'' = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$ „

Fuente: Elaboración propia

Siempre hay N Clusters de tamaño Mi Y M número de elementos en la población.

Estimación de la media μ en MPC (monoetápico) desiguales

Para estimar la media se pueden usar cuatro tipos de estimadores clásicos, sin embargo, el que mayormente se utiliza es:

$$\bar{y}_n'' = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

Este estimador se basa en una estimación de razón, es una media ponderada, donde las ponderaciones son los tamaños e los conglomerados.

La varianza estimada de y_n" es

$$V\left(\bar{y}''_n\right) = \frac{N-n}{Mn\bar{M}^2} \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}''_n Mi\right)^2}{n-1}.$$

Donde:

$$\sum_{i=1}^{n} (yi - \bar{y}_n^{"}\ Mi) = \sum_{i=1}^{n} y_i^2 - 2\bar{y}_n^{"} \sum_{i=1}^{n} y_i\ M_i + \left(\bar{y}_n^{"}\right)^2 \sum_{i=1}^{n} M_i^2$$

Y $\bar{M}$ se estima por $\bar{m} = \frac{\sum_{i=1}^{n} M_i}{n}$

Un límite para el error de estimación es:

$$K\sqrt{\hat{V}\left(y_n^{"}\right)} = K\sqrt{\frac{N-m}{Nn\bar{M}^2} \frac{\sum_{i=1}^{n} \left(y_i - y_n^{"} M_i\right)^2}{n-1}}$$

Cálculo del tamaño de la muestra para estimar μ en MPC (monoetápico) desiguales

Sea la varianza de estimador:

$$V\left(\bar{y}''_{n}\right)=\left(\frac{N-n}{Mn\bar{M}^{2}}\right)\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}''_{n}M_i}{n-1}\right)^{2}$$

Y fijando u error de estimación o cota de margen de error

$$e = \sigma_{\bar{y}n}$$

Se despeja n

$$n=\frac{N\sigma_c^{2}}{N\left(\frac{d^{2}}{k^{2}}\right)\bar{M}^{2}+\sigma_c^{2}},$$

Donde: $s_c^{2}=\frac{\sum_{i=1}^{n}\left(y_i-y''_{n}M_i\right)^{2}}{n-1}$

Es la cuasivarianza muestral, y_i es el total de la variable de interés en el i-ésimo conglomerado.

$\bar{M}$ se estima por $\bar{m}$ en muestra piloto, estudios preliminares o muestra preliminar.

Recuerde que el tamaño de muestra n se refiere al número de conglomerados, no de elementos.

Estimación del total T en MPC (Monoetápico) y desiguales e igual probabilidad de selección

Muchas veces el interés es estimar un total (T) en la población

Uno de los estimadores que se utiliza es:

$\hat{T}$ = M. (∑yi / ∑mi) = M. $\bar{y}_n$ '',

Este estimador se utiliza solo si se conoce M, donde yi es el total de la .variable de interés en el conglomerado i. Si no se conoce M no se puede calcular este estimador. Frecuentemente M no es conocido, entonces se utiliza un estimador que no depende de M, se trabaja es con los totales de los conglomerados muestreados y el estimador es $\hat{T}$= N. $\bar{y}_t$. Lohr (2010) presenta muy bien este tema.

Estimación de la proporción P en MPC (Monoetápico) y desiguales e igual probabilidad se selección

El estimador de $\underline{(P)}$ es:

$$\hat{P} = \frac{\sum_{i=1}^{n} ai}{\sum_{i=1}^{n} Mi}$$, recuerde que es una media,

ai= Nº de elementos con la característica de interés (1), en el conglomerado i.

M_i es el número de elementos en el conglomerado i.

La varianza estimada de $\hat{P}$ es:

$$\hat{V}\left(\hat{P}\right) = \frac{N-n}{Mn\overline{M}^2} \frac{\sum_{i=1}^{n} \left(ai - \hat{P}Mi\right)^2}{n-1}$$

Un límite para el error de estimación será:

$$K\left(\hat{\sigma}_p\right) = K\sqrt{\frac{N-n}{Mn\overline{M}^2} \frac{\sum_{i=1}^{n} \left(ai - \hat{P}Mi\right)^2}{n-1}}$$

Cálculo del tamaño de la muestra para estimar $\underline{P}$ en MPC (monoetápico) desiguales

El tamaño de la muestra para estimar la proporción en la población, viene dado por:

$$n = \frac{N S_c^2}{N \frac{d^2 \overline{M}^2}{k^2} + S_c^2}$$

Donde:

$$S_c^2 = \frac{\sum_{i=1}^{n} \left(a_i - \hat{P} M_i\right)^2}{n-1}$$

Es la cuasivarianza entre conglomerados, la que se calcula por estudio piloto, preliminar, estudios anteriores, y:

$$\frac{\wedge}{M} = \overline{m} \frac{\sum_{i=1}^{n} M_i}{n}$$

Tamaño promedio de conglomerado calculado en la muestra piloto, estudio preliminar o estudios anteriores.

d=e, es el margen de error o cota fijada, y K= $Z^{\alpha}/_{2}$, de la Distribución Normal Estándar.

Ejercicio sobre Muestreo por Conglomerados Desiguales

Este ejercicio fue tomado del escrito del Doctor Jesús mellado, del Departamento de Estadística y Cálculo de la Autónoma de Madrid, donde se presentan a manera de ejemplo 6 conglomerados desiguales, cada uno con su número de elementos o unidades dentro de ellos y cada total de la variable Y de interés. Presentamos este

ejercicio muy ilustrativo por su simplicidad y facilidad de entenderlo, como lo presenta el Profesor Mellado:

Estimación de la media

Una vez seleccionados los conglomerados a muestrear, se obtiene de cada uno su tamaño (m_i) y la suma de la variaable que se desea analizar (y_i). Nótece que es la suma de las variables, no la media.

Después se suma cada una de las columnas y se aplica la siguiente ecuación:

Conglomerado	m_i	y_i
1	32	125
2	28	136
3	25	145
4	27	134
5	26	135
6	30	128
	168	803

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

Como los valores de las sumatorias ya está calculado en la tabla, solamente se sustituyen los valores:

$$\bar{y} = \frac{803}{168} = 4.77$$

Estimación de la varianza de la media

Para el cálculo de la varianza de la media es conveniente agregar dos columnas a la tabla, en la primera se multiplica la media general por el tamaño de cada conglomerado; en la siguiente columna se resta el total de cada conglomerado menos el la columna anterior y se eleva al cuadrado. La columna se suma.

Conglomerado	m_i	y_i	$\bar{y}m_i$	$(y_i - \bar{y}m_i)^2$
1	32	125	152.95	781.336
2	28	136	133.83	4.694
3	25	145	119.49	650.554
4	27	134	129.05	24.467
5	26	135	124.27	115.051
6	30	128	143.39	236.940
	168	803		1813.042

La varianza se calcula con la siguiente ecuación:

$$V(\bar{y}) = \left(\frac{N-n}{Nm\left(\frac{M}{N}\right)^2} \right) \frac{\sum_{i=1}^{n} (y_i - \bar{y}m_i)^2}{n-1}$$

Si N=81 conglomerados y M=2268 elementos en la población. Nótece que se la sumatoria ya está calculada en la tabla anterior.

$$V(\bar{y}) = \left(\frac{81-6}{81(6)\left(\frac{2268}{81}\right)^2} \right) \frac{1813.04}{6-1} = 0.0713$$

Intervalo de confianza de la media

El intervalo de confianza para la media es la siguiente:

$$\bar{y} - 2\sqrt{V(\bar{y})} < \mu < \bar{y} + 2\sqrt{V(\bar{y})}$$

$$4.77 - 2\sqrt{0.071} < \mu < 4.77 + 2\sqrt{0.071}$$

$$4.24 < \mu < 5.31$$

Ejemplo contextualizado de MPC

Se desea determinar si conviene instalar una productora de leche pasteurizada en cierto poblado de España. Para ello, se desea conocer el consumo mensual por persona al mes. Se tiene información municipal de un total de N conglomerados geográficos o conjunto de hogares, de los cuales se puede tomar una muestra aleatoria de n Conglomerados y se ajusta el marco muestral en esos conglomerados, solamente. Se toma la muestra de los n Conglomerados y se entrevista al informante calificado para responder, se visitan todos los hogares, solo de los conglomerados seleccionados. Estimar la media, su intervalo de confianza y el límite del error de estimación.

Aquellos lectores interesados en profundizar sobre el MPC Monoetápico, recomendamos hacerlo con el libro de Muestreo Estadístico de la Profesora Lohr (2010), de la Universidad de Arizona.

Muestreo por Conglomerados en dos etapas

Una muestra por conglomerados en dos etapas se obtiene seleccionando primero una muestra aleatoria de conglomerados y posteriormente una muestra aleatoria de elementos de cada conglomerado muestreado.

Los conglomerados frecuentemente contienen demasiados elementos para obtener una medición de cada uno de ellos. Los elementos en un conglomerado son tan semejantes que la medición de solo algunos de ellos proporciona información completa sobre ese conglomerado.

Una muestra por conglomerados en dos etapas se obtiene seleccionando primero una muestra aleatoria de conglomerados y posteriormente una muestra aleatoria de elementos de cada conglomerado muestreado sea:

N = número de UP en el universo.

M = número de US dentro de cada UP

NM: Número de US n el universo

n: Número de UP en una MAS

m: Número de US en una muestra dentro de cada UP

y = Característica de interés

yij= Valor de la característica en la j-esima US de la i-ésima UP

Parámetros:

$\bar{y}_i$ = media por US en la i-ésima UP de la población

$$\bar{y}_i = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$$

Media por elemento en el universo (parámetro de interés).

Estimadores:

$$\bar{y}_{im} = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$$

Promedio de una muestra de m US en la i-esima UP

$$\bar{y}_{nm} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}$$

Promedio por elemento en una muestra de tamaño nm en la población.

$\bar{y}_{nm}$ estima a μ = $\bar{y}$ en la población

$$\bar{y}_{nm} = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2$$

La varianza de

$Si\ N \to \infty$

$n \to \infty$,

entonces la varianza $\bar{y}_{nm}$ es:

$$V(\bar{y}_{nm}) = \frac{Sb^2}{n} + \frac{\bar{S}_w^2}{nm}$$

Se toma una muestra aleatoria preliminar o piloto y se hace un análisis de la varianza para estimar las cantidades $Sb^2\ y\ \bar{S}_w^2$,

Con los estimadores de $Sb^2\ y\ \bar{S}_w^2$ se puede estimar la $V(\bar{y}_{nm})$ como:

$$\hat{V}(\bar{y}_{nm}) = \frac{\widehat{Sb^2}}{n} + \frac{\widehat{\bar{S}_w^2}}{nm}$$

Otros estimadores de la media y el total en Bietápico

Otro estimador insesgado de la media de la población es:

$$\widehat{\mu} = \frac{N}{Mn}\sum_{i=1}^{n} M_i\overline{y}_i = \frac{1}{\overline{M}}\frac{\sum_{i=1}^{n} M_i\overline{y}_i}{n},$$

Y su varianza estimada es:

$$\widehat{V}(\widehat{\mu}) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\overline{M}^2}\right)\frac{s_b^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

donde:

s^2 b = Cuasivarianza entre Clusters s^2 i = Cuasivarianza dentro de Clusters.

Se puede obtener un estimador insesgado del total de Y es :

$$\widehat{\tau} = \frac{N}{n}\sum_{i=1}^{n} M_i\overline{y}_i = N\frac{\sum_{i=1}^{n} M_i\overline{y}_i}{n},$$

Y un estimador de la varianza del total es:

$$\widehat{V}(\widehat{\tau}) = \left(\frac{N-n}{N}\right)N^2\frac{s_b^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

Tamaño de muestra en un Muestreo Bietápico

Hay varias estrategias que planteamos para el cálculo del tamaño de la muestra en un Muestreo Bietápico:

Recordemos que el tamaño de muestra en un Bietápico está formado por n y m, unidades primarias y secundarias, respectivamente.

1. Para una varianza dada de $\bar{y}_{nm}$, en el computador se juega con los valores n y m. Así {n, m} es el tamaño de muestra que minimiza esa varianza. Esto resulta sencillo con la utilización de cualquier programa de Análisis Estadístico, utilizando el Análisis de la varianza y estimando los componentes de varianza.
2. Una solución práctica:
    - Se define el conglomerado o unidad primaria.
    - Se seleccionan los conglomerados con igual probabilidad (si los conglomerados son parecidos en tamaño), si no son parecidos en tamaño se utiliza PPT.
    - Si se usa PPT en las unidades primarias, se selecciona igual número de unidades secundarias en la segunda etapa. Si son similares los conglomerados en tamaño, se puede usar una misma fracción de muestreo.
    - Se calcula el tamaño de la muestra n usando la ecuación de MAS y se incrementa n usando la tasa de respuesta esperada y el efecto de diseño deff. Esto se hace de la siguiente manera: $n_1 = n_0 / 1 + n_0 / N$, luego se incrementa este tamaño de muestra considerando el % de tasa de no respuesta esperada, entonces $n_2 = n_1 /1 -$ % NR. Luego, n_2 se incrementa incluyéndole el efecto de diseño deff, $n = n_2$. Deff, donde n es el tamaño de muestra definitivo.
    - ¿Cuántas unidades primarias se utilizan?, depende de los factores de coste, tamaño, operatividad y de la homogeneidad o heterogeneidad dentro de los conglomerados. El número de unidades primarias lo elige el equipo investigador.
    - Se divide el valor de n entre el número de unidades primarias seleccionadas y ello nos proporciona el valor de las m unidades

secundarias en cada unidad primaria, aun cuando se haya seleccionado las unidades primarias con PPT.

3. Otra forma es utilizar la ecuación :

$$n = \frac{p \times (1-p) \times D \times z^2}{e^2 \times b}$$

donde:

n El número de Clusters seleccionado.

p Proporción de éxito esperada.

D El efecto del diseño deff.

z El valor critico punto de corte de la distribución Normal Estándar

e El margen de error aceptable o admisible

b Número de unidades secundarias

Ejercicio de Muestreo Bietápico

Un fabricante de prendas de vestir tiene N = 90 plantas localizadas a lo largo de los Estados Unidos y quiere estimar el promedio de horas que las máquinas de coser estaban para reparaciones en los últimos meses (Sin funcionar). Debido a que las plantas están muy dispersas, decide usar el muestreo de conglomerados, especificando cada planta como un grupo de máquinas. Cada planta contiene muchas máquinas y la comprobación del registro de reparación para cada máquina requeriría mucho tiempo. Por lo tanto, utiliza el muestreo en racimo o conglomerados en dos etapas. Se dispone de suficiente tiempo y dinero para probar n = 10 plantas y aproximadamente 20% de las máquinas en cada planta. Los datos resultantes se dan en la tabla siguiente:

Planta	M_i	m_i	Tempo sin funcionar (en horas)	$\bar{y}_i$	s_i^2
1	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5	5.40	11.38
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5	4.00	10.67
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4	5.67	16.75
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10	4.80	13.29
5	52	10	11, 4, 3, 1, 0., 2, 8, 6, 5, 3	4.30	11.12
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4	3.83	14.88
7	42	8	3, 7, 6, 7, 8, 4, 3, 2	5.00	5.14
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3	3.85	4.31
9	40	8	6, 4, 7, 3, 9, 1, 4, 5	4.88	6.13
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4	5.00	11.80

Queremos estimar el tiempo medio de inactividad por máquina, y sabemos que el número total de máquinas en todas las plantas es de M = 4500.

Estimación insesgada:
Un estimador insesgado de la media de la población es:

$$\hat{\bar{y}}_{unb} = \frac{N}{nK}\sum_{i=1}^{10} M_i \bar{y}_i = \frac{90}{(10)(4500)}(2400.59) = 4.80118 \text{ horas}$$

El estimador de la varianza es: $\hat{V}(\hat{\bar{y}}_{unb}) = \left(\frac{N}{K}\right)^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{K^2 n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$

Los dos términos pueden ser calculados separadamente. Tenemos que $\frac{N}{K^2 n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$

$$= \left(\frac{90}{(4500)^2 10}\right)\left[\left(1-\frac{10}{50}\right)(2500)\left(\frac{11.38}{10}\right) + \left(1-\frac{13}{65}\right)(4225)\left(\frac{10.67}{13}\right) + \ldots + \left(1-\frac{11}{56}\right)(3136)\left(\frac{11.80}{11}\right)\right] = 0.0097722267 ,$$

En este ejercicio K=M

El tamaño medio del conglomerado se estima por: $\overline{M}_s = 52.2$, tal que

$$s_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{y}_i - \overline{M}_s\hat{\bar{y}}_{unb}\right)^2 = \frac{1}{9}\left[((50)(5.40)-(52.2)(4.80118))^2 + \ldots((50)(5.40)-(52.2)(4.80118))^2\right]$$

$= 892.3468481$, y $\left(\frac{N}{K}\right)^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} = \left(\frac{90}{4500}\right)^2\left(1-\frac{10}{90}\right)\left(\frac{892.3468481}{10}\right) = 0.0317278879$.

Entonces la varianza estimada es $\hat{V}(\hat{\bar{y}}_{unb}) = 0.0415001146$, y el error estándar es: $SE(\hat{\bar{y}}_{unb}) = 0.2037157692$.
Un intervalo del 95% de confianza del tiempo medio sin funcionar es:
$\hat{\bar{y}}_{unb} \pm 1.96SE(\hat{\bar{y}}_{unb}) = (4.4019\ hrs.,\ 5.2005\ hrs.)$.

Entonces la media del tiempo sin funcionar las maquinas está entre 4.4019 y 5.2005 horas, con una confianza de un 95%.

Estimación en MPC (monoetápico) con probabilidades proporcionales al tamaño (PPT)

$$\text{Sea } \pi_i = \frac{M_i}{M}$$

Donde:

π_i= Probabilidad de que la i-esima Unidad de Muestreo (Conglomerados) sea seleccionada.

Estimación en MPC (monoetápico) de μ con PPT

Un estimador para la varianza de $\bar{y}_{ppt}$ será

$$\bar{y}_{ppt} = \frac{1}{M}\hat{T}_{ppt} = \frac{M}{Mn}\sum_{i=1}^{n}\bar{y}_i = \frac{\sum_{i=1}^{n}\bar{y}_i}{n}$$

$$\hat{V}(\bar{y}_{ppt}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}_{ppt}\right)^2$$

El límite de error de estimación será:

$$K\sqrt{\hat{V}\left(\bar{y}_{ppt}\right)} = k\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_{ppt})^2}$$

Estimación en MPC (monoetápico) del total con PPT

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

La varianza estimada de $\hat{T}_{ppt}$ es:

$$\hat{V}\left(\hat{T}_{ppt}\right) = \frac{M^2}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_{ppt})^2$$

El límite para el error del estimador es:

$$K\sqrt{\hat{V}\left(\hat{T}_{ppt}\right)} = k\sqrt{\frac{M^2}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_{ppt})^2}$$

El trabajar con PPT, es un aspecto muy interesante y el de mayor uso en las muestras complejas, es decir, el uso de la estratificación acompañada con el uso de PPT en la selección de los conglomerados desiguales en una sola etapa y sin reemplazo.

Cuando seleccionamos una muestra probabilística, no todos los elementos o unidades tienen la misma probabilidad de inclusión. Lo que si se requiere es conocer la probabilidad de cada elemento, que no necesariamente es la misma.

Si la variable de interés o investigación Y es aproximadamente proporcional a una variable auxiliar clave X, se prefiere utilizar el muestreo con probabilidades desiguales de inclusión. Se conoce el valor de la variable X en todos los individuos de la población y generalmente X se refiere a una medida de tamaño del elemento o unidad. La variable X debe estar correlacionada linealmente con Y.

Sin embargo, hay muchas formas de muestrear sin reemplazo y con probabilidades desiguales de selección. Un esquema muy utilizado es el PPT o probabilidad proporcional al tamaño, en nuestro caso se refiere al tamaño del

conglomerado (Número de elementos dentro del conglomerado), este esquema tiene la ventaja que la varianza de los estimadores se reduce y sus errores estándar, además no necesitamos estratificar con esa variable X, ya que ganamos más usando PPT.

En el caso concreto de MPC Monoetápico se justifica el uso de PPT si hay diferencia entre los tamaños de los conglomerados y sus totales. Una forma de controlar la variabilidad en el tamaño de los conglomerados es utilizar PPT. Debe quedar claro que el tamaño no se refiere necesariamente al número de elementos, sino a alguna característica del elemento o unidad vinculada a la variable de interés o de investigación Y. En el caso específico del MPC se utiliza mucho PPT en casi todos los estudios, es decir, que ls conglomerados que tengan mayor número de elementos, tienen mayor probabilidad de inclusión en la muestra.

El MPC con PPT al tamaño del conglomerado, reduce el límite del error de estimación, más aun, cuando el total de conglomerado Y_i está altamente correlacionado con el número de elementos en el conglomerado, se evidencia mucho más la reducción.

En la actualidad, cuando se utiliza el MPC en cualquier etapa, con PPT y sin reemplazo, uno de los estimadores para el total y la media, que produce la mejor estimación lineal insesgada, es el estimador de Horvitz-Thompson.

Generalidades del Muestreo Polietápico

Este esquema de muestreo es muy utilizado en encuestas complejas de hogares, sin embargo, puede ser aplicado siempre y cuando se establezcan jerarquías, por ejemplo: arboles; ramas dentro de árboles, frutos dentro de ramas y finalmente se realizan las determinaciones. Es decir que se deben definir las unidades de primera etapa o unidades primarias, unidades de segunda etapa o secundarias, unidades de tercera etapa o terciarias y así, sucesivamente, pero no es conveniente en la práctica, tener más de tres etapas, ya que aumenta la varianza

de los estimadores. Generalmente va acompañado con estratificación. En el muestreo de hogares es uno de los esquemas más utilizado.

Es menos eficiente que MAS, pero facilita la operatividad y se utiliza cuando no hay listado de los elementos de la población, o aun existiendo el listado, la disminución de los costes y las labores de encuesta lo favorecen. En cada etapa la unidad de muestreo cambia, y en la última etapa es donde se tiene la unidad de análisis.

En relación al tamaño de la muestra cuando se utiliza un esquema Polietápico, para determinar el tamaño de la muestra sugerimos utilizar la ecuación de MAS, y tomar en cuenta el % de no respuesta esperado y el efecto del diseño (Deff), para incluirlos en el cálculo de n.

Para elegir el número de unidades finales en la última etapa de un Polietápico, hay que tener en cuenta la variabilidad de la variable Y de interés. Si los conglomerados o unidades primarias son diferentes, pero hay mucha homogeneidad dentro de ellas, se seleccionan muchas unidades primarias, pero si se consideran semejantes las unidades primarias y hay mucha heterogeneidad dentro de ellas, entonces se eligen pocos conglomerados**.**

A medida que utilizamos más etapas en un muestreo hay más error, se aumenta el error por cada etapa.

Estimadores de Razón, Regresión y Diferencia en Muestreo Aleatorio simple

En los métodos comunes o clásicos de muestreo se utilizan estimadores simples de los parámetros poblacionales, tales como: μ, T y P. Existen métodos de estimación donde utilizan información auxiliar o complementaria para estimar los parámetros μ, T y P. Estos métodos requieren del uso de una variable auxiliar clave X junto con la variable de interés Y, donde X está linealmente correlacionada con Y. Si se mide Y, y una o más variables auxiliares X, se pueden estimar

parámetros con mayor precisión. Si se tiene información auxiliar de X, sea la media o el total de la población, esta información se utiliza para estimar la media o el total de la variable de interés Y. Con la información auxiliar X, no se necesita conocer el tamaño de la población N, sino el total de la variable X, si se desea estimar el total de Y.

Los estimadores de razón resultan muy apropiados cuando se presenta una relación aproximada de proporcionalidad directa entre la variable de interés y variables auxiliares clave X.

Estimación de Razón

El estimador de razón se puede utilizar para:

1. Estimar la proporción de la variable Y (Variable de interés) respecto a X (Variable auxiliar)
2. Estimar la media de la población.
3. Estimar el total de la población

   Es decir, se estiman los parámetros R, μ y T, con un estimador de razón.

El uso del estimador de razón es más efectivo cuando la relación entre la respuesta Y, y una variable auxiliar X es lineal a través del origen y la varianza de Y es proporcional a X, como se observa en la **Figura 4**.

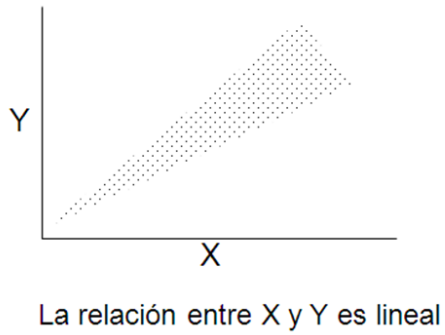

Figura 4. Relación entre la variable de interés Y, y la variable auxiliar X.

Fuente: Elaboración propia

Se calcula el coeficiente de correlación lineal entre X y Y, para un coeficiente mayor de 0,50 el estimador de razón proporciona una estimación más precisa de μ_y o T_y o P_y que la que darían los estimadores clásicos.

Si la relación entre X y Y es lineal a través del origen, los estimadores de razón son aproximadamente insesgados.

El objetivo de usar un estimador de razón para estimar la media o el total de la población es utilizar la ventaja de la correlación lineal entre X y Y.

Como resumen en relación al estimador de razón podemos decir que se utiliza mayormente:

- Cuando nos centramos en estimar un total de la población, queda entendido que también se puede estimar una media al usar un estimador de la razón.
- Cuando no se conoce el tamaño de la población.
- Cuando X es fácil de obtener y Y es más difícil de obtener y costosa, su obtención, o se tiene que destruir la unidad al medirla.
- Si lo que se desea es obtener un estimador de la razón R.
- Cuando hay una correlación lineal entre la variable de investigación (Y) y la variable auxiliar (X).

La razón en la población es:

$$R = \frac{\mu_y}{\mu_x} = \frac{\sum_{1}^{N} y}{\sum_{1}^{N} x} = \frac{Y}{X}$$

Y un estimador de la razón en la población es:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{1}^{n} y}{\sum_{1}^{n} x}$$

En algunos textos se le llama r, porque se calcula en la muestra.

Como mencionamos anteriormente, la razón es el número medio de unidades de Y por cada unidad de X.

Estimador de razón del total poblacional T y:

$$\hat{T}y = \frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi} (Tx)$$

T_X es el total de X en la población.

Varianza estimada de *T y:*

$$\hat{V}(\hat{T}y) = (Tx)^2 V(r) = T^2x \,.\, (N-n/nN)\,(1/\mu^2x) \frac{\sum_{i=1}^{n} (yi - rx_i)^2}{n-1}$$

Donde μx y Tx, son la media y el total poblacionales, respectivamente, para la variable aleatoria X, y Y es la variable de interés o de investigación.

$$2\sqrt{\hat{V}}\,(\hat{T}y) = 2\sqrt{T^2}\,X(N-n/nN)\,(1/\mu^2X)\sum_{I=1}^{n}(Yi-rx_i)^2/n-1$$

Un estimador de la media de la población es:

$$\hat{\mu}_y = \bar{y}_R = \hat{R}\,\mu_x$$

Donde μ_x es la media de la variable auxiliar en la población.

Ejemplo sencillo del uso del estimador de razón:

Se desea estimar el número total de habitantes en 120 ciudades y se conoce el total de habitantes T_X en un estudio anterior. Se tomó una muestra de n= 49 ciudades. Los resultados en la muestra fueron:

$\sum y = 626$ $\sum x = 5054$

Además, se conoce el total de habitantes por el estudio (X) = $T_X = 22919$

Entonces, el estimador del total de habitantes es:

$\sum y/\sum x.\ T_X = 626/5054.\ (29919) = 28397$ habitantes.

Estimación de Regresión

Si la relación entre Y y X (variable auxiliar) es lineal, pero no necesariamente pasa por el origen, entonces X puede ser tomada para construir un estimador de regresión.

Si se quiere estimar la media de Y o sea μ_y, debe conocerse μ_x antes de que se utilice el estimador.

Se considera que X ya ha sido observada y Y es una variable aleatoria que será observada.

Un estimador de regresión de una media (μ) viene dado por:

$$\overline{y}_{lr} = \overline{y} + b(\overline{X} - \overline{x})$$, donde:

$\overline{y}$ = media de la muestra

$\overline{X}$ = media de la población, de la variable auxiliar

$\overline{x}$ = media de la muestra de la variable auxiliar

b = coeficiente de regresión lineal

y el estimador de la varianza de $\bar{y}_{lr}$ es:

$$\hat{v}(\bar{y}_{lr}) = \left(\frac{N-n}{Nn}\right)\left(\frac{1}{n-2}\right)\left[\sum_{i-1}^{n}(y_i-\bar{y})^2 - b^2.\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \quad (1)$$

y un limite para el error de estimación es:

$$K\sqrt{\hat{v}(y_{lr})} = K\sqrt{\hat{\sigma}_{\bar{y}_{lr}}{}^2} \quad (2)$$

Donde:

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

b se llama coeficiente de regresión lineal de **y** en **x** en una población finita.

Aunque **b** se calcula de la muestra, algunas veces se fija el valor de **b** por estudios anteriores o encuestas repetidas

Los siguientes dos (2) ejercicio fueron tomados de Scheaffer, Mendenhall y Ott (1987).

Ejercicio 1

Antes de ingresar a un colegio se hizo un examen de conocimientos en matemáticas a 486 estudiantes. Se seleccionó una muestra aleatoria simple de n=10 y se anotó el puntaje en el examen de conocimientos (x) y el puntaje del examen final de calculo (y) Los resultados aparecen en la siguiente tabla:

ESTUDIANTE	PUNTAJE DE CONOCIMIENTOS (X)	PUNTAJES EN CALCULO (y)
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75

Estime la media μ_y para la población de estudiantes (N=486)

Solución:

$\bar{y} = 76 \quad \bar{x} = 46 \quad$ b=0,766 $\quad \bar{X} = 52$

Entonces:

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x})$$

$$\bar{y}_{lr} = 76 + (0,766)(52 - 46) = 80$$

Entonces 80 es el puntaje medio estimado de cálculo para los 486 estudiantes de la población

Para el cálculo de la varianza del estimador $\bar{y}_{lr}$ se tiene: (usando la ecuación (1)

$$v(\hat{\bar{y}}_{lr}) = \left(\frac{486 - 10}{486(10)}\right)\left(\frac{1}{8}\right)\left[2056 - (0,766)^2(2,474)\right]$$

$$v(\hat{\bar{y}}_{lr}) = 7,397$$

Y un límite para el error de estimación con un 95% de confianza es:

$$1,96\sqrt{7,397} = 5,4$$

Estimación por Diferencia

El estimador de diferencia $\bar{y} + (\mu_x - \bar{x})$ se utiliza cuando la figura (gráfica) de Y contra X muestra que los puntos caen a lo largo de una línea recta con pendiente igual a 1, como se observa en la **Figura 5**.

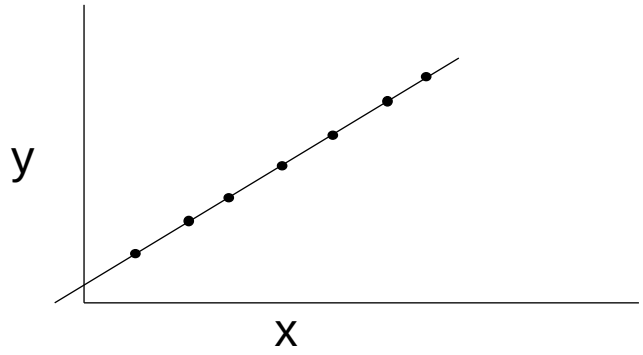

Figura 5. Línea recta con pendiente uno.
Fuente: Sheaffer, Mendenhall. y Ott (1987).

Para estimar la media o total poblacional, es similar al de regresión, pero no se calcula el coeficiente b. Este se fija igual a 1.

Estimador de diferencia de la media poblacional (μ)

$$\bar{y}_D = \bar{X} + (\bar{y} - \bar{x}) \qquad (1)$$

Donde:
$\bar{X}$ = Media de la variable auxiliar en la población
$\bar{x}$ = Media de la variable auxiliar en la muestra
$\bar{y}$ =Media de la variable de interés en la muestra

Un estimador de la varianza de $\bar{y}_D$ es:

$$\hat{v}(\bar{y}_D) = \left(\frac{N-n}{Nn}\right)\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \qquad (2)$$

Donde: $\overline{d} = \overline{y} - \overline{x}$ y $d_i = y_i - x_i$

Un límite para el error de estimación viene dado por:

$$K \sqrt{V\left(\hat{\overline{y}}_D\right)} \qquad (3)$$

Ejercicio 2

Suponga que una población contiene 180 artículos inventariados con un valor establecido en el libro de $ 13,320. Denote por x_i el valor en el libro y por y_i el valor intervenido del i-ésimo artículo.

Se seleccionó una muestra aleatoria irrestricta de tamaño n=10 artículos, la cual produjo los siguientes resultados:

MUESTRA	VALOR INTERVENIDO y_i	VALOR EN EL LIBRO x_i	d_i
1	9	10	+1
2	14	12	+2
3	7	8	-1
4	29	26	+3
5	45	47	-2
6	109	112	-3
7	40	36	+4
8	238	240	-2
9	60	59	+1
10	170	167	+3

Se aclara que los auditores frecuentemente están interesados en comparar el valor intervenido de los artículos con el valor asentado en el libro. Los valores en el libro son conocidos para cada artículo en la población y los valores intervenidos se obtienen con una muestra.

Se utilizan los valores en el libro para obtener una estimación del total o la media del valor intervenido en la población

Solución:

$$\overline{y} = 72{,}1 \qquad \overline{x} = 71{,}7 \quad \text{y} \quad \overline{X} = 74{,}0$$

$$\overline{y_D} = 74 + (72{,}1 - 71{,}7) = 74{,}0$$

Aplicando luego la ecuación (2) se tiene:

$$V(\hat{\overline{y}}_D) = \left[\frac{180-10}{(180)10}\right](6{,}27) = 0{,}59$$

Y el límite para el error de estimación es:

Aplicando la ecuación (3), para un 95% de confianza:

$$1{,}96\sqrt{0{,}59} = 1{,}50$$

Muestreo No Probabilístico (MNP)

Se puede decir que es el más usado en las Ciencias Sociales, ya que el muestreo probabilístico requiere de una infraestructura Estadística que muchas veces no es accesible a los Investigadores. En el MNP la mayoría de las veces no se encuentra el Marco de Muestreo de interés y es costoso y difícil de elaborar, y el objetivo del investigador no es la Generazibilidad de los resultados a la población. Además, se desea seleccionar sujetos con determinadas características en un contexto social.

Los investigadores en Ciencias Sociales saben de la existencia y ventajas del MP, pero aducen que son demasiados costosos y difíciles de abordar y utilizan frecuentemente los MNP, por su facilidad y menor coste, pero en ellos hay mucha carga subjetiva a la hora de Diseñar una Muestra. Una excepción de este escenario, en las Ciencias Sociales, es el Muestreo de Hogares, donde se ha hecho norma la utilización de combinaciones o Diseños Mixtos denominados Muestras Complejas.

El MNP es insustituible en Investigación Cualitativa, donde más que preocuparse por el número de sujetos de la muestra, se preocupa por recoger información con riqueza y profunda de los significados de conceptos, el discurso, desde la vivencia de los propios entrevistados para generar teoría, necesitándose una gran capacidad del investigador para observar e interpretar la información aportada por los casos. Es muy utilizado en Investigaciones exploratorias.

El muestreo no probabilístico se utiliza en Marketing en pruebas de concepto, pruebas de empaque, pruebas de nombre y pruebas de texto, donde por lo general no es necesario hacer extrapolaciones a la población. El interés se centra en la proporción de la muestra que da varias respuestas o manifiesta diversas actitudes. Las muestras para estos estudios se obtienen usando, por ejemplo, el muestreo por cuotas en centros comerciales o combinaciones de ésta con otros esquemas No Probabilísticos. Si no hay un Marco Muestral con estructura probabilística, lo más seguro es que se utilice un MNP.

Hay una inmensa variedad de aplicaciones del MNP, haciendo énfasis en el muestreo de propósito, de conveniencia y sin norma. Además, es posible también hacer combinaciones de esquemas como de Muestreo por cuotas con muestreo de Bola de Nieve en un mismo estudio o de Cuotas con un muestreo de Conveniencia.

Existe una tipología de esquemas de muestreo, una de ellas se presenta en la **Figura 6**, donde se observa la clásica división entre los muestreos probabilísticos y no probabilísticos. Hay una amplia clase de muestreo en estos dos tipos (el probabilístico y el no probabilístico).

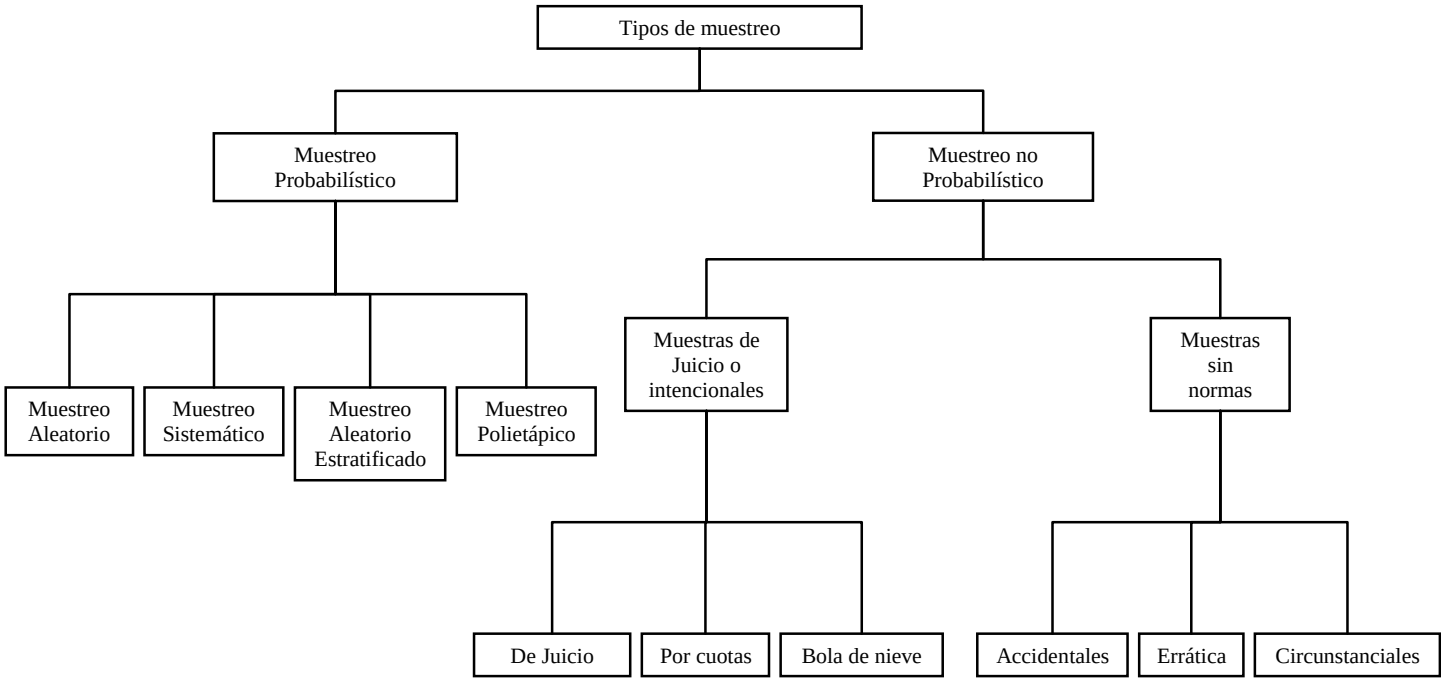

***Figura 6*:** Tipos de Muestreo.
Fuente: Elaboración propia

Entre los muestreos probabilísticos están: el Muestreo Aleatorio Simple(MAS), el Muestreo Aleatorio Estratificado (MAE), el Muestreo Sistemático (MS), el Muestreo por Conglomerados(MPC), el Bietápico y el Polietápico, mientras que en los no probabilísticos se tiene: muestreo de juicio, por cuotas, bola de nieve y los

muestreos sin norma: accidentales, circunstanciales, por conveniencia o comodidad.

En el muestreo no probabilístico a pesar de reducirse los costos en general, y de tener el problema de no contar con un marco muestral, los resultados pueden contener sesgos y no existe la forma de estimar la probabilidad de que cada elemento de la población tenga que ser incluido en la muestra, por consiguiente, no se aplica la teoría del muestreo que conocemos.

Seguidamente hacemos una comparación entre los esquemas de muestreo probabilístico y no probabilístico:

MUESTREO PROBABILÍSTICO	MUESTREO NO PROBABILÍSTICO
Se le llama científico, aleatorio, cuantitativo.	Se le llama de propósito, cualitativo
Su objetivo es la medición.	Su objetivo es la comprensión de un fenómeno social.
Generaliza los resultados a una población estadística.	Se produce la transferibilidad a un contexto.
Se enfoca en el perfil estadístico de la población.	Se enfoca en información profunda de los casos.
Generalmente se utiliza >50 casos.	Generalmente se utiliza <30 casos.
La muestra se selecciona antes del estudio.	La muestra se selecciona en el proceso o estudio
Se selecciona la muestra por un proceso aleatorio.	Se selecciona la muestra por el juicio del investigador o sin norma.
El tamaño de la muestra se define mayormente por una ecuación.	El tamaño de la muestra se define por saturación.
Se tiene Marco Muestral.	No se tiene marco Muestral.

Seijas (2006) clasifica las muestras no probabilísticas en intencional u opinática y las llamadas sin norma. Las opináticas se clasifican en muestras de juicio, por cuotas, y bola de nieve o cadena, mientras que las muestras sin normas se clasifican en accidentales, circunstanciales, por conveniencia.

Muestras intencionales, opináticas o de propósito

Las intencionales son aquellas en donde la ecuación personal del investigador está presente en la selección de la muestra, y **las muestras sin norma** son aquellas en donde hay una selección circunstancial o errática.

El objetivo de las muestras de propósito no es la medición u obtener el perfil estadístico, sino la comparación de los procesos sociales y humanos en toda su complejidad, se eligen los casos con el fin de tener un conocimiento profundo del fenómeno.

La muestra de juicio es aquella en la que se seleccionan los casos que son típicos de la población, según el criterio del investigador.

El muestreo por cuotas es una combinación de la estratificación y el muestreo por opinático, donde se fija a priori los porcentajes o cuotas de elementos o individuos que reúnen ciertas características de la población. Se seleccionan las unidades de muestreo de forma tal que se cumpla las cuotas fijadas en la muestra. Se corre el riesgo de que la muestra sea representativa de algunas características de la población, pero no de la variable de interés.

La muestra está formada por subgrupos definidos a priori en la población. Se usan los porcientos de la población.

Las características Demográficas que generalmente se emplean son: La edad, sexo, ocupación, nivel económico, raza, Localización Geográfica.

La población se divide en estratos, donde la ponderación se obtiene del censo o encuesta anterior, con el tamaño de la muestra y los porcentajes, en cada estrato se asignan las cuotas a cada investigador, quien utiliza su juicio para seleccionar la unidad de análisis.

El investigador queda en libertad de seleccionar su muestra siempre que satisfaga la cuota fijada y ello, puede sesgar la selección hacia individuos más accesibles o atractivos para el investigador que selecciona la muestra.

Este esquema trata garantizar que la muestra se parezca a la población, pero no quiere decir que la haga una muestra representativa, esto podría ocurrir solo si hay una alta correlación entre las variables de control y la característica de interés que se investiga.

Lo que apoya al muestreo por cuotas es el tiempo y el costo. Se utiliza mucho en Muestreo Polietápico

Muestreo de Bola de Nieve

En el muestreo de bola de nieve se seleccionan casos raros o poco comunes, se le pide a cada entrevistado que identifique uno o más casos con la característica de interés. Se desea medir una característica en la población, donde no hay marco y los sujetos son muy difíciles de encontrar como: Drogadictos, homosexuales, corruptos, delincuentes, coleccionistas, población de buzos de mar profundo, niños con labio leporino, personas con coches mu viejos, familias con trillizos. Cada individuo localizado puede nominar a otros de la población.

Para que la muestra tenga algún grado de aleatoriedad, el primer grupo en la etapa cero (0) se selecciona aleatoriamente. La cadena se detiene cuando no se consigue más nominaciones o los sujetos se niegan a contestar

Lo importante es que el individuo desee cooperar con la investigación, además se va a crear un marco de muestreo y hay mayor seguridad de los entrevistados pertenezcan a la población.

El muestreo Bola de Nieve produce estimaciones con sesgo y hay poca representatividad estadística, no se debe extrapolar los resultados

Este esquema de muestreo se usa mucho combinado con otros esquemas no probabilísticos, de propósito o intencionales

Muestras Sin Norma

En las muestras sin norma se toman los casos que llegan a la mano o se tiene más facilidad de su selección, como por ejemplo, los que se toman en la calle, los que pasan por una esquina, en la radio, la televisión, los que se convocan a una reunión y se toman los que asistan, se toma cualquier animal de un grupo, en algunas condiciones se toman accidentalmente los casos disponibles en el momento, sin convocatoria. Dentro de las muestras sin normas es importante considerar las muestras de conveniencia y las muestras casuales, accidentales o fortuitas, muestra de voluntarios.

Muestra de conveniencia

Los sujetos se seleccionan bajo la conveniente accesibilidad y proximidad con relación al investigador. Le resulta más sencilla la selección en base a su proximidad geográfica o relación de amistad. Son accesibles al investigador o desean participar en el estudio como voluntarios. Se selecciona la muestra con mucha facilidad.

Cuando se utilizan estudiantes voluntarios, personas cautivas, el uso de sujetos en una clínica, un profesor selecciona estudiantes de su clase, familiares, correos

personales. Suele utilizarse en estudios exploratorios, para luego ver posibilidad de plantear hipótesis. Útil para depurar un instrumento.

Muestras casuales, fortuitas, accidentales

Los individuos se eligen de manera casual, sin ningún juicio. Se elige un lugar o un medio y seleccionan sujetos que accidentalmente se encuentren a su disposición (el que pase). Sus resultados son orientadores. Cuando se hacen entrevistas en la calle por parte de una emisora de radio o televisión, cuando se aborda a un grupo de gente que sale de un cine, en un centro comercial a una hora determinada se entrevistan los sujetos, en la calle se entrevista los sujetos sobre su opinión en relación a un producto o tema. Los heridos que llegan a un hospital.

Muestra de voluntarios

Todos los elementos de la población tienen la oportunidad de participar. El voluntario generalmente posee características diferentes. El sujeto se selecciona el mismo en la muestra. Es informado mediante un anuncio y se interesa en contactar al investigador. Por ejemplo, cuando se inserta un cuestionario en un periódico, pagina, WEB, blog, entre otros y se pide la opinión de la gente. El sujeto tiene interés en el tópico y no es representativo de la gente que no tiene interés.

Este esquema de voluntarios es barato, rápido y permite obtener gran cantidad de datos con poco esfuerzo y tiempo limitado.

Muestreo virtual online, por redes sociales

Se utilizan las redes sociales (Network) para acceder a elementos que no se hubiesen detectado por otros medios como registros administrativos, censos,

investigaciones, entre otros. Muy usado en las llamadas poblaciones ocultas. Es utilizado en esquemas de muestreo mixto.

Las personas interesadas en obtener más información sobre el muestreo no probabilístico pueden consultar a:

Mixed Methods Sampling

A Typology With Examples

Charles Teddlie
Fen Yu
Louisiana State University, Baton Rouge

Aspectos teórico-metodológicos clave del Muestreo Estadístico en las Investigaciones

Existen aspectos clave que deben comprenderse antes de realizar cualquier investigación por muestreo (Cochran, 1976; Kish, 1972; Scheaffer, Mendenhall y Ott, 1987; Sukhatme y Sukhatme, 1970; Lohr, 2010).

Según nuestra opinión, los términos universos y población estadística son diferentes, el universo es el conjunto completo finito o infinito de elementos o unidades de análisis que tiene al menos una característica o especificación en común y población estadística es el conjunto de medidas de una variable aleatoria que se realiza en cada unidad de análisis; es decir, que, en un universo, puede haber varias poblaciones estadísticas.

El MAS se utiliza cuando se tiene un marco muestral restringido con todas las unidades de análisis del universo y no existen variables que se relacionen con la variable de interés, donde se desea estimar el valor del parámetro, siendo, además, la población relativamente homogénea, en la que no se observa un patrón de variación. El MAS tiene la desventaja de que se necesita listar todas las

unidades de análisis de la población y puede aumentar el costo, producto de la dispersión de la muestra, por eso se aplica a poblaciones pequeñas.

Los métodos estadísticos asumen que los datos fueron obtenidos usando muestras aleatorias en forma directa o indirecta. Los otros diseños se comparan con el MAS en lo que se conoce como efecto diseño, donde se compara la variabilidad de cualquier diseño con la variabilidad del MAS. También se aplica a condiciones experimentales, donde se seleccionan al azar las unidades experimentales a las que se les aplica determinado tratamiento. Los puntos muestrales pueden no estar uniformemente repartidos en la población, sobre todo si la muestra es grande y al utilizarlo, se puede estar ignorando cualquier información auxiliar que pudiese ser de interés para reducir el error de estimación.

El MAS con probabilidad proporcional al tamaño (PPT), se utiliza cuando la población estadística es asimétrica; esto es; pocos elementos o individuos afectan grandemente la respuesta y muchos la afectan poco, ejemplo de variables asimétricas: capital suscrito, total de ventas en tiendas, precio de las casas, número de cargos vacantes en empresas y total de producción de leche por finca. En este esquema se mejora el diseño de la muestra, dándole a las empresas grandes, mayor probabilidad de pertenecer a la muestra, incluso hay empresas que tienen probabilidad uno (1) de formar parte de la muestra.

El MAE se utiliza cuando la población tiene una estructura que justifica la formación de estratos. Existen variables correlacionadas con la variable de interés u objeto de estudio, se cuenta con un marco de muestreo amplio y se conoce el tamaño de la población (N), de cada estrato y la sumatoria de las N_i de cada estrato. En el MAE se obtiene información global sobre el parámetro poblacional e información del valor del parámetro en cada estrato. Hay mayor precisión en los estimadores, estos tienen menor error de muestreo y menor tamaño de muestra a igual error que en MAS, además favorece la administración del muestreo, ya que las labores se realizan por estrato, esto trae una reducción en el costo por unidad de observación o medición. Se puede, incluso, formar estratos donde cada elemento tiene probabilidad uno (1) de formar parte de la muestra, esto se conoce

como estrato autorepresentado, lo que le proporciona mayor representatividad a la muestra.

No se aconseja utilizar un alto número de estratos (Kish, 1972) debido a que estratos pequeños contribuyen poco a la ganancia con la estratificación y puede aparecer estratos vacíos o con muy pocos elementos. El autor ha conseguido mayor precisión en los estimadores usando pocos estratos, pero los estratos deben ser muy diferentes entre ellos y debe existir homogeneidad dentro de estrato. En el MAE se utiliza la información auxiliar para formar estratos. Los estratos son homogéneos internamente con respecto a la variable de interés o una variable correlacionada con esta. Si se utiliza afijación proporcional u óptima, las estimaciones del parámetro tienen precisión tan o mejor que MAS. El MAE asegura mayor representatividad de la muestra, seleccionando una muestra aleatoria en cada estrato. Es útil si se hace necesario tener diferentes esquemas de muestreo en cada estrato, que pudiese reducir el costo total. Cuando los estratos se forman con variables correlacionadas con la variable de interés, se produce un aumento en la precisión de las estimaciones, comparado con MAS, sobre todo si el tamaño de la muestra es grande.

En el MAE se necesita un marco de muestreo para definir los estratos. La ganancia en precisión está relacionada con la correlación entre la variable objeto de estudio y la/s variable/s auxiliar/es. Una vez formados los estratos, se toma la muestra dentro de cada a estrato usando MAS, MAE, MS o MPC.

Algunas veces se desea formar estratos, pero no se puede ubicar las unidades de muestreo en los estratos hasta después de seleccionar la muestra; es decir, no se conoce a priori el tamaño de los estratos, en este caso se puede utilizar la postestratificación, la cual consiste en estratificar una muestra, no la población. Para postestratificar se toma una muestra aleatoria simple o sistemática y se forman estratos con esa muestra. La efectividad de la postestratificación, después de consultar teóricos (Cochran, 1976; Ras, 1980; Pérez, 2000) se justifica solo si la muestra aleatoria es grande, esto coincide con los resultados de un esquema

MAE con afijación proporcional, donde coinciden prácticamente las varianzas de los estimadores de ambos procedimientos de muestreo.

Para la aplicación del MS, es preferible, pero no necesario, contar con un listado de elementos de la población en estudio. En él se toma en cuenta el coeficiente de elevación (K=N/n) y se selecciona un punto o arranque aleatorio entre 1 y K para iniciar la toma de la muestra. Se toman muestras en espacio y tiempo en un patrón específico, en intervalos de espacio o tiempo. Asegura que la población objeto sea representada uniformemente. Combina la aleatoriedad con la cobertura de población. Si el coeficiente de elevación (k) está asociado con la variable de interés, se producen estimaciones sesgadas, como si se tomara muestras en el mismo punto. En campo es conveniente desde el punto de vista práctico y operativo, aun en área geográficas grandes. Tiene mucha utilidad en el muestreo de hogares, tanto en áreas rurales como urbanas, y tiene un área de cobertura más completa que MAS. Es común su uso en las últimas etapas de un esquema Polietápico, ya que ahorra costo por unidad de análisis. En el MS se corre el riesgo de introducir homogeneidad en la muestra si existe periodicidad en la población, si se eligen casos donde la periodicidad es constante.

En relación al MPC, Sudman (citado por Díaz, 2001) menciona aspectos que deben tenerse en cuenta al realizar una selección de conglomerados: Los conglomerados deben delimitarse perfectamente, sin traslapes o duplicaciones, se debe conocer (aproximadamente) el número de conglomerados, la conglomeración aumenta el error de estimación, comparado con MAS. Los conglomerados deben seleccionarse aleatoriamente o con PPT. El autor, de acuerdo a la teoría, considera cada conglomerado como una reproducción a escala de la población diana, ello conlleva a seleccionar muchos conglomerados de tamaño pequeño y dispersos, no coincide con Sudman, en este aspecto. El autor considera que aun cuando se tenga el listado, se prefiere el MPC, por la reducción en el costo del muestreo y solo se necesita el listado de las unidades de muestreo primarias o de primera etapa.

Si los conglomerados son grandes, se justifica un muestreo en dos o más etapas, Bietápico o Polietápico, ambos son casos especiales del MPC. El submuestreo influye en la precisión de los estimadores, en ellos cambia la unidad de muestreo en cada etapa y según Díaz (2001), se debe utilizar el muestreó Bietápico o Polietápico cuando sea difícil o no se tenga una lista exhaustiva de los elementos o unidades de análisis de la población, que imposibilita utilizar MAS, MAE y MS. También se recomienda el submuestreo cuando se realizan estudios de ámbito nacional e internacional, en los cuales se evidencia una alta dispersión de los elementos de la población. Es importante acotar que el error de muestreo en MPC es mayor que en MAS y MAE, pero la reducción de costos sin sacrificar información vale la pena.

El muestreo no probabilístico o de juicio se utiliza cuando las características de interés de la población se miden en pequeña escala como la toma de muestras microbiológicas, muestras de tejidos, muestras de suelos, agua y aire, muestras en plantas, hojas, frutos y toda clase de líquidos, y cuando en las condiciones sobre las que se va a muestrear no es posible utilizar los métodos probabilísticos. Así mismo, el muestreo no probabilístico es de mucha utilidad en los esquemas de muestreo en varias etapas o Polietápico, debido a la facilidad de la toma de muestra en las últimas etapas En los esquemas de muestreo no probabilísticos o de juicio como se les llama, no aplica la teoría del muestreo estadístico y no envuelve la selección aleatoria de las unidades de análisis de la población.

La selección de unidades se hace con la participación del investigador con muestras de opinión o sin norma (Seijas, 2006). Este tipo de muestreo es menos costoso, se realiza en menos tiempo, pero depende del juicio personal para la selección de las unidades. Igualmente, no puede aplicarse el análisis estadístico inferencial para obtener conclusiones respecto a la población objeto. Las conclusiones se basan en el juicio personal y no se conoce el margen de error en la estimación, ni el nivel de confianza para generalizar los resultados de la muestra hacia la población, las inferencias se basan en el juicio de personas, no en la teoría estadística.

REFERENCIAS BIBLIOGRÁFICAS

Arias, F. (2006). *El proyecto de investigación científica*. Introducción a la metodología científica. (5ª ed.). Caracas: Episteme.

Cochran, W. (1976). *Técnicas de muestreo*. (2ª edición). (Trad. E. Casas. Díaz). México: COMPAÑÍA EDITORIAL CONTINENTAL, S.A.

Díaz, V, (2001). *Organización y Gestión de los trabajos de campo con encuestas personales y telefónicas*. (1ª edición). Barcelona: Ariel.

Gómez, Á. y A. Higuera. (1986). *Bases para el manejo integral de plagas: revisión crítica de la investigación entomológica*. Fondo Nacional de Investigaciones Agropecuarias. Serie D, Nº 1-21. Maracaibo: Editorial Maracaibo, S.R.L.

Kish, L. (1972). *Muestreo de encuestas*. México: Trillas.

Lohr, S.L. (2010). *Sampling: Design and Analysis. Second Edition. Arizona State University*. Brooks/Cole. CENGAGE Learning. Boston.

Pérez, C. (2000). *Técnicas de muestreo estadístico.* Teoría, práctica y aplicaciones informáticas. México: Alfaomega Grupo Editor, S.A. de C.V.

Ras, D. (1979). *La estructura de las encuestas por muestreo.* (Trad. E. Suárez.). México: FONDO DE CULTURA ECONÓMICA.

Ras, D. (1980). *Teoría del muestreo*. (1ª edición). México: Fondo de cultura económica.

Scheaffer, R.; Mendenhall, W. y Ott L. (1987). *Elementos de muestreo*. (Trad. G. Rendón y J. Gómez). México: Grupo Editorial Ibero América, S.A. de C.V.

Seijas, F. (2006). *Investigación por muestreo*. Universidad Central de Venezuela. Caracas: Ediciones de la Biblioteca de la Universidad Central de Venezuela.

Statistical Analysis System. (2005). SAS/STAT. *User's Guide (Version 8,2).* Cary, NC.

Sudman, S. (1976). *Applied Sampling*. New York: Academy Press.

Sukhatme, P. y B. Sukhatme. (1970). *Sampling theory of survey with applications.* USA: International Standard Book.

Walpole, R.; Myers, R. y Myers, S. (1999). *Probabilidad y estadística para ingenieros.* (6ª edición). México: Prentice Hall Hispanoamericana, S.A.