UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

TRABAJO DE FIN DE GRADO

ESTRATEGIAS DE MUESTREO ESTADÍSTICO PARA LA ESTIMACIÓN DE PARÁMETROS

Autora: JELENA FÁTIMA ALLPACCA MEDINA

Tutora: MARÍA TERESA CABERO MORÁN

Grado en Estadística

Salamanca, 2024

UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

TRABAJO DE FIN DE GRADO

ESTRATEGIAS DE MUESTREO ESTADÍSTICO PARA LA ESTIMACIÓN DE PARÁMETROS

Firmado por autora:

Firmado por tutora:

Firmado por CABERO MORAN MARIA TERESA - ***6717** el día 01/07/2024 con un certificado emitido por AC FNMT Usuarios

Grado en Estadística

Salamanca, 2024

Certificado de la tutora de TFG

D./Dña. María Teresa Cabero Morán, profesora del Departamento de Estadística de la Universidad de Salamanca,

HACEN CONSTAR:

Que el trabajo titulado "Estrategias de muestreo estadístico para la estimación de parámetros", que se presenta, ha sido realizado por Jelena Fátima Allpacca Medina, con DNI ****1060 y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 1 de julio de 2024

Firmado por CABERO MORAN MARIA TERESA - ***6717** el día 01/07/2024 con un certificado emitido por AC FNMT Usuarios

Fdo.: María Teresa Cabero Morán

Índice

Introducción

El concepto de estadística ha experimentado una profunda evolución desde sus orígenes hasta el presente. Inicialmente, se limitaba a la simple enumeración o conteo de personas, objetos, eventos o fenómenos, pero ha cambiado hacia la propuesta de métodos complejos para analizar, predecir y asistir en la toma de decisiones relacionadas con el comportamiento de grupos. La estadística se enfoca en el estudio de conjuntos de datos que exhiben variabilidad, extendiendo sus conclusiones a los grupos o poblaciones de los cuales provienen los datos.

Como disciplina matemática, la estadística demuestra su utilidad en la recolección, organización y análisis de datos, adaptándose a las necesidades y objetivos específicos de cada contexto. Su importancia se refleja en diversos campos del conocimiento, al proporcionar a los investigadores herramientas para poner a prueba hipótesis, validar o refutar teorías y llegar a conclusiones sólidas. Este ámbito de estudio, reconocido como **la ciencia del análisis de datos**, se establece con el propósito principal de facilitar la comprensión de fenómenos y eventos en el entorno, utilizando información derivada tanto de experimentos como de investigaciones empíricas. Según el INE, la estadística se puede entender como dos significados:

El primero: "*A alguien elaborando estadísticas, que consiste en recopilar información sobre algo.*"

Y el segundo: *"Estadística cuando se tratan los datos y se obtienen conclusiones, y oímos cosas como a la vista de los resultados de los experimentos se puede concluir que no hay evidencia empírica para afirmar que tal producto es perjudicial para la salud."*

Con este tipo de reflexiones, se comienza a comprender la utilidad de la estadística en la vida cotidiana, aunque no constituye una definición técnica, refleja los pensamientos que puede tener una persona en su día a día al escuchar la palabra estadística.

La relevancia de la estadística se evidencia en diversas situaciones, como en el proceso de realizar un análisis conjunto de revisiones de la literatura antes de iniciar una investigación. En este caso, se lleva a cabo una exhaustiva búsqueda bibliográfica para recopilar toda la información publicada sobre el tema específico en cuestión. Sin embargo, la diversidad de metodologías, revistas de difusión y duración de los estudios pueden dificultar la obtención de conclusiones definitivas. En tales circunstancias, el análisis estadístico se convierte en una herramienta invaluable para identificar elementos compartidos entre los estudios, descubrir patrones subyacentes o relaciones ocultas.

Otra situación donde la estadística juega un papel crucial son los ensayos clínicos, especialmente en la evaluación de nuevos medicamentos o tratamientos. Antes de que un fármaco pueda considerarse seguro y efectivo, debe someterse a pruebas rigurosas en grupos

de personas seleccionadas específicamente. Estos ensayos implican una selección cuidadosa de la población y la muestra, junto con la definición de un período de tiempo adecuado y la planificación de varias fases experimentales. Además, la estadística se emplea para establecer parámetros que faciliten la evaluación de la eficacia del nuevo tratamiento, así como para compararlo con otros medicamentos ya disponibles en el mercado. En este contexto, la estadística ofrece un marco objetivo y riguroso para la interpretación de los resultados y la toma de decisiones informadas en el ámbito clínico.

Estos son solo dos ejemplos de cómo la estadística puede ser de ayuda en investigaciones o estudios, entre muchos otros.

Por otra parte, estas ayudan conllevan en algunos casos un mal uso de la estadística por ciertos medios de comunicación, como pueden ser los gráficos 3D que resultan ser engañosos y ofrecen una perspectiva ilusoria de la realidad, alterar las escalas cortando los ejes por un valor previamente meditado en vez del valor 0 o el valor adecuado, gráficos que no dicen nada con datos que no tienen coherencia entre ellos y no están bien contextualizados, mezcla de tipo de datos como por ejemplo los datos acumulados con no acumulados y compararlos entre sí, ausencia de datos en los gráficos y modificación de proporciones o áreas en los gráficos. Es por eso que es importante conocer qué es realmente la estadística y como funciona para así evitar cometer este tipo de errores o dejarse influenciar al verlos en los medios de comunicación.

En su uso más común, la estadística se entiende como una simple colección de datos, listas, tablas y otros métodos de presentación de información. Sin embargo, considero que esta visión es bastante limitada. La estadística es una disciplina mucho más rica y compleja que no solo se encarga de recopilar datos, sino que también los analiza y los interpreta para extraer conclusiones significativas.

Para entender mejor su alcance, es importante reconocer que la estadística se divide en dos ramas principales: la estadística descriptiva y la estadística inferencial.

La **estadística descriptiva** abarca un conjunto de procedimientos y técnicas que se utilizan para recoger datos, clasificarlos, ordenarlos, representarlos y simplificarlos, sin que ello implique la obtención de conclusiones a la población general. Su objetivo principal es la capacidad de resumir y representar de manera precisa y detallada la información contenida en conjuntos de datos voluminosos, sin aventurarse a realizar inferencias o generalizaciones sobre la totalidad de la población en cuestión. Este enfoque analítico, que se centra en la presentación objetiva y sistemática de los datos, constituye un paso fundamental en el proceso de análisis estadístico, al proporcionar una vista global y detallada de la distribución, tendencias y características fundamentales de los datos observados.

La **estadística inferencial**, conocida también como analítica, deductiva o matemática, se conoce como el conjunto de métodos y técnicas que se aplican para obtener conclusiones de un colectivo a partir de la información obtenida de una parte de él. Este campo de estudio se caracteriza por su enfoque en la toma de decisiones fundamentadas en la aceptación o el rechazo de hipótesis específicas, considerando siempre un margen de error que permite evaluar la incertidumbre asociada a los resultados obtenidos.

Ilustración 1:Esquema de los tipos de estadística

Por lo tanto, la estadística descriptiva tiene como objetivo la descripción del colectivo estudiado, y este puede ser toda la población (censo) o una parte de la misma (muestra). Cuando en una población dada se selecciona adecuadamente un grupo, la descripción de estos (estadística descriptiva) se puede ampliar a toda la población; en este caso se infiere el conocimiento de todos a partir de unos pocos (estadística inferencial).

En la estadística descriptiva, no se corren riesgos, se describe lo que hay, pero en el caso de la estadística inferencial se atribuyen cualidades de esa pequeña parte al colectivo y hay un riesgo de error que puede aumentarse por medio del cálculo de probabilidades.

Dentro de este marco conceptual, el presente trabajo se enfoca en explorar y analizar el muestreo estadístico como una herramienta fundamental de la estadística inferencial. Este enfoque particular implica una investigación detallada sobre los métodos y técnicas de muestreo utilizados para seleccionar muestras representativas de una población más amplia.

En el mundo actual, la obtención precisa y efectiva de información sobre el comportamiento, necesidades y particularidades de un universo, ya sea un conjunto amplio de individuos, objetos o entidades, se ha convertido en una necesidad inevitable. Para abordar esta tarea, se recurre a las técnicas de muestreo, las cuales permiten extraer muestras representativas, que es una muestra que incluye todos los aspectos relevantes de la población de la cual se extrae, permitiendo hacer inferencias válidas y generalizables sobre la población total a partir del análisis de la muestra. Estos métodos están guiados por la teoría del muestreo que están sustentado en el cálculo de probabilidades, conocimientos de álgebra y cálculo infinitesimal, y son fundamentales para garantizar la fiabilidad y la eficiencia en la recopilación de datos, asegurando, así, la validez de los resultados obtenidos.

En ocasiones, se piensa que estudiar la población completa es lo mejor, es decir, realizar un censo, pero no llega a ser viable por los costes, dificultad o por la destrucción de elementos de la población durante el estudio, entonces resulta más practico seleccionar un subconjunto de elementos a estudiar qué debe ser lo más representativo posible de la población, esto hace que el riesgo de error sea mínimo. Para la obtención de ese subconjunto, usamos el muestreo, sus diferentes técnicas ayudan a los distintos tipos de situaciones a realizar la mejor obtención de la muestra.

Entre algunas definiciones de muestreo:

"*El muestreo es el proceso mediante el cual se selecciona un grupo de observaciones que pertenecen a una población. Esto, con el fin de realizar un estudio estadístico.*" (Westreicher, 2021)

"*Muestreo: Acción de escoger muestras representativas de la calidad o condiciones medias de un todo.*" (RAE, 2023)

"El muestreo es una manera de obtener información deseada de una población basada en ciertos criterios." (Nicolás, 2011)

Se puede observar que el muestreo es una herramienta de investigación científica que tiene como función determinar qué parte de la población debe examinarse con la finalidad de hacer inferencia sobre la población teniendo en cuenta los errores o sesgos que conlleva el uso del muestreo. El subconjunto obtenido se llama **muestra**, esta debe ser representativa de la población que se va estudiar teniendo algunas características básicas de la misma, las cuales pueden ser: tamaño, distribución de variables demográficas, geografía, ocupación o sector, etc.

Algunos artículos nombran al muestreo como un arte, ya que se requiere pericia por parte del investigador en el diseño de la muestra a la hora de determinar la técnica, la selección de los estimadores, la elección de un tamaño adecuado de la muestra con un nivel de confianza aceptable y el uso de marcos muestrales actualizados.

En la actualidad, la estadística es muy necesaria en el día a día, podemos ver que ayuda en muchos sectores de la vida diaria y que llega a verse muy usada en los medios de comunicación, no por ello se puede decir que se hace bien. En las investigaciones donde se quiere estudiar poblaciones o universos muy extensos, se usan la estadística inferencial ya que se desea obtener conclusiones del conjunto sin tener que estudiar a cada individuo, por ello se usa el muestro.

El principal objetivo de este trabajo reside en identificar y seleccionar el método óptimo de muestreo estadístico para la estimación de bases de datos. La importancia de esta tarea radica en la necesidad de garantizar la precisión y la fiabilidad de los datos recopilados, ya que estos forman la base sobre la cual se toma decisiones críticas en diversos campos, desde la investigación científica hasta la planificación estratégica en empresas y organizaciones.

Con el propósito de lograr este objetivo, se seguirá un enfoque estructurado que abarcará diversos aspectos.

En primer lugar, se realizará una exposición exhaustiva de los fundamentos teóricos relacionados con la estadística, el muestreo y los principales tipos de muestreos aleatorios, con una breve mención al muestreo no aleatorio y sus tipos.

Posteriormente, se procederá a la aplicación práctica de estos conceptos sobre una base de datos específica, previamente sometida a los procesos necesarios para su adecuado manejo. Acto seguido, se llevarán a cabo pruebas piloto para cada tipo de muestreo, seguidas de los cálculos pertinentes destinados a obtener la estimación requerida.

Finalmente, se compararán los resultados obtenidos a fin de derivar conclusiones pertinentes para el estudio.

Es importante destacar que, a lo largo de este proyecto, se priorizará el uso de métodos de muestreo aleatorios o al azar, excluyendo deliberadamente los muestreos no aleatorios con el fin de minimizar posibles fuentes de error. En el caso práctico, se demostrará la ejecución del muestreo aleatorio simple, el muestreo aleatorio estratificado, así como el muestreo por conglomerados en una o varias etapas, y una combinación de muestreo estratificado y por conglomerados.

Capítulo 1

Muestreo Estadístico

El muestreo es un conjunto de técnicas estadísticas que implican el análisis y la obtención de conclusiones acerca de un subconjunto pequeño de elementos (muestra) para inferirlas a todo el conjunto de elementos (población), permitiendo generalizar las conclusiones a todo el conjunto de individuos, de los que se han dado valores a distintas variables.

Antes de continuar, se deben definir varios términos que se usaran a lo largo del trabajo para no llegar a confusiones.

- El **universo** es un conjunto de individuos, objetos o entes diferentes que se puede someter a estudio.
- La **población** es el conjunto de valores o categorías que toma una variable (aleatoria o estadística) en el universo.
- La **muestra** es un subconjunto representativo de la población que se estudiara para la obtención de conclusiones globales.
- Un **estadístico** es cualquier función real medible de la muestra de una variable aleatoria o estadística.
- Un **parámetro** es cualquier función real medible de la población de una variable aleatoria o estadística.

Para adquirir las ventajas que el muestreo permite (Briones, 1990) se necesita manejar los recursos teóricos y prácticos que permitan decidir diseños muestrales adecuados. En este caso, el muestreo está basado en su teoría que resulta en cuatro principios: dos axiomas, un teorema y una ley.

Axioma 1: Todo trabajo con muestras requiere un conocimiento previo del universo del cual se extraerán dichas muestras.

Se indica que, antes de realizar el muestreo para extraer la muestra, es necesario conocer el universo que se va a estudiar. Aunque parezca obvio, se debe tener sumo cuidado en esta etapa, ya que, si no se conocen las características básicas del universo, ¿cómo se obtendrá una muestra representativa de dicho universo? Además, ¿qué certeza se tendrá de que las conclusiones obtenidas sean fiables y válidas?

Según Azorín y Crespo (1986, p.69): *"El diseño óptimo de la muestra, en particular la determinación previa de su tamaño óptimo, sólo podría conseguirse a partir del conocimiento de la población."*.

Esto tiene sentido; en algunas poblaciones, al ser tan extensas, no se pueden apreciar las características básicas y necesarias para realizar un estudio. Muchas características que se dan

por hechas pueden no ser necesarias para el estudio y no deben seguir considerándose para la extracción de una muestra.

Axioma 2: La similitud entre los componentes del universo no son propiedades dadas sino distinciones hechas por el observador.

Al igual que existe un alto número de personas en el mundo, también hay variedad de puntos de vista, por lo cual, se puede decir que las características de los elementos son vistas por el observador, dependiendo de lo que se quiera estudiar o a lo que se quiere dar mayor importancia. Por eso, dependiendo de lo que se quiere estudiar, el investigador tiene que tener fijado el objetivo de su estudio y así dejar en claro las características de la población necesarias para el estudio.

Teorema: Teorema Central del Límite

Una sencilla definición es (Daniel; 1985, p.108): *"Sin tener en cuenta la forma funcional de la población de donde se extrae la muestra, la distribución de las medias muestrales, calculadas con muestras de tamaño* $\boldsymbol{n}$ *extraídas de una población con media* $\boldsymbol{\mu}$ *y varianza finita* $\boldsymbol{\sigma^2}$*, se aproxima a una distribución normal con media μ y varianza* $\boldsymbol{\sigma^2/\, n}$*, cuando* $\boldsymbol{n}$ *aumenta. Si* $\boldsymbol{n}$ *es grande, la distribución de medias muestrales puede aproximarse mucho a una distribución normal".*

Esto se puede entender que da igual el origen de la muestra, la distribución de su media muestral tendrá una distribución aproximadamente normal, cuanto más aumente su tamaño.

Este teorema es uno de los conceptos fundamentales en estadística y probabilidad, desempeña un papel crucial en la inferencia estadística y tiene aplicaciones en muchos campos. Ayuda mucho cuando se hace inferencia en poblaciones excesivamente grandes, siendo la distribución normal la más sencilla y conocida, facilitando cálculos y toma de decisiones.

Ley: La Ley de los Grandes Números

Según Robinson (1981, p.16): *"Si un experimento es repetido más y más veces, entonces la frecuencia relativa del evento tiende a acercarse a la probabilidad del evento".*

Esta ley dice que cuantas más veces se repita un experimento, por ejemplo, el lanzar una moneda, la frecuencia con la que se repetirá un determinado suceso, el que salga cara o cruz, se acercara a una constante, esa constante llegará a ser la probabilidad de que ocurra ese suceso.

Con el fundamento teórico abordado, se debe tener en cuenta algunos factores para encontrar la muestra necesaria en la investigación o estudio que se quiera realizar, como, por ejemplo, que la muestra sea representativa de la población.

La muestra debe ser **representativa**, es decir, debe tener un tamaño suficiente, presentar características iguales o similares a las de la población y las observaciones deben extraerse con el método correcto de muestreo para la población y el estudio. Estas características recaen sobre el investigador y debe tener sumo cuidado a la elección de las mismas.

El tamaño de la muestra debe ser suficiente, es decir, el número de observaciones debe ser óptimo para el estudio. A veces se piensa que, si el número de observaciones que tenemos es mayor, es mejor y se cae en un error muy común, en los estudios no son necesarias muestras de grandes tamaños, al contrario, puede ser que se necesiten muestras más pequeñas.

Las muestras no tienen que ser representativas en todas las características de la población, solo tiene que ser en las importantes para el estudio.

La elección del método de muestreo debe ser óptima, es decir, la técnica usada debe ser la correcta para ese caso específico. Un mal uso de las técnicas puede darnos un fallo en la muestra y eso conlleva a más errores ya que no resultara representativa de la población.

Al realizar un muestreo se debe tener en cuenta ciertos conceptos que son clave para hallar una muestra.

- **Tamaño de la población:** es el número total de variables, objetos o entes en un grupo específico que se está estudiando o investigando. Cada uno de los elementos que lo componen se denominan **unidad de muestreo**. Este número debe ser adecuado para cada tipo de estudio.
- **Margen de error**: indica la cantidad de variabilidad que está asociada con una muestra utilizada para estimar un parámetro de interés en una población más amplia. En palabras más sencillas, indica lo mucho que refleja los resultados las opiniones de la población en general.
- **Nivel de confianza:** es el porcentaje que indica la certeza del intervalo de confianza calculado a partir de una muestra representa la probabilidad de que dicho intervalo contenga el verdadero valor del parámetro poblacional de interés. En otras palabras, este porcentaje refleja cuán confiables son los límites aleatorios del intervalo al estimar el parámetro real de la población.

Teniendo eso en cuenta, se piensa que ya se puede obtener una muestra rápidamente, pero también se debe pensar en las posibles fuentes de error que se pueden cometer.

- **Errores no muestrales o sesgos**. Este tipo de errores son también llamados errores humanos ya que incluyen las buenas o malas decisiones de los investigadores. Existen tres tipos más comunes: **sesgo de selección** (relacionado con las alteraciones en los resultados derivados de los métodos utilizados para la selección de los individuos), **de información** (cuando se observa diferencias sistemáticas en el modo en que los datos son obtenidos por parte de los investigadores) y **de confusión** (cuando los resultados obtenidos están influenciados indirectamente por el efecto de variables relacionadas con el perfil de la persona encuestada).
- **Errores muestrales.** Es la diferencia entre el verdadero valor del parámetro y el estimado. Este error se puede minimizar usando la aleatorización, es decir, usando técnicas de muestreo aleatorias o al azar en las que la elección de los elementos se debe al azar.

Como se mencioná anteriormente, la elección de la técnica de muestreo debe ser la correcta y es una de las decisiones más importantes que los investigadores deben tener en cuenta, para ello se debe conocer algunas características de la población objeto de estudio. Se puede distinguir el muestreo probabilístico y el no probabilístico.

Muestreo aleatorio o probabilístico: los elementos de la población son elegidos al azar, ya sea con la misma probabilidad (con reposición) o con distinta probabilidad (sin reposición).

Los más conocidos son: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo sistemático, muestreo por conglomerados y el muestreo bietápico o polietápico.

De todas estas técnicas, se hablará en profundidad más adelante.

Muestreo no probabilístico: los elementos se eligen por criterios que no están basados en el azar, normalmente estos criterios son a juicio del investigador. Se usan cuando no es posible la realización de los otros debido a restricciones principales de coste o cuando se está realizando una exploración inicial y se quiere obtener cierto grado de representatividad. Los más conocidos son: el muestreo por rutas, intencional, por cuotas, cadena, bola de nieve y conveniencia; de

estas técnicas se comentará brevemente a continuación, ya que este estudio no optará por ese tipo de muestreo.

El muestreo por rutas: se eligen los elementos estableciendo una ruta o camino a través de la población objetivo y se obtiene información según los criterios establecidos, por ejemplo, se planea un camino del punto A al B, en ese camino solo se quiere información de las personas del género femenino que fumen.
Resulta de utilidad cuando la población es excesivamente grande y se encuentra muy dispersa.
El muestreo intencional, discrecional u opinático: se seleccionan específicamente los elementos de la muestra con el propósito de representar lo que el investigador crea necesario para el estudio.
Este método se usa para representar ciertas características o grupos específicos dentro de la población, por ejemplo, si se quiere realizar un estudio sobre los hábitos alimenticios de los estudiantes universitarios, optaremos por seleccionar estudiantes de diferentes facultades, edades, géneros y niveles socioeconómicos de manera intencional.
El muestreo por cuotas: se eligen los elementos estableciendo ciertas características representativas del estudio y se selecciona un número específico de elementos que satisfacen dichas características.
Por ejemplo, se quiere hacer un estudio sobre preferencias de sabor de helado, se desea tener una muestra que represente la diversidad de edades en la población y, para ello, se establecen cuotas para diferentes grupos de edad (30% menores de 18, 50% de 18 a 30 años y 20% de 31 en adelante), luego se selecciona los participantes de cada grupo de edad de manera que se cumpla esas cuotas.
El muestreo en cadena: se eligen ciertos elementos de la población que están vinculados entre sí, y estos se van agregando a la muestra gradualmente hasta alcanzar el tamaño deseado, esta técnica depende de las conexiones sociales o profesionales entre individuos, te permite aprovechar las redes sociales para reclutar participantes de una manera eficiente.
El muestreo de bola de nieve: se utiliza típicamente cuando resulta difícil ubicar a los miembros de la población. Aquí, el investigador solicita a los individuos en la muestra que proporcionen datos de contacto de otros individuos que puedan ser adecuados para el estudio, y así sucesivamente hasta llegar al tamaño deseado de participantes.
Un ejemplo sencillo podría ser un estudio sobre la experiencia de los estudiantes internaciones en una universidad, se comenzaría entrevistando a un estudiante internacional conocido, durante la entrevista se mencionará a otros estudiantes internacionales que conocen o podrían estar interesados en el estudio. Se podría en contacto con esos estudiantes y se pedirá información de otros estudiantes, así sucesivamente hasta tener el número de entrevistas necesarias.
El muestreo por conveniencia: involucra la inclusión de sujetos disponibles y accesibles, lo que puede comprometer la representatividad de la muestra y la inclusión de sesgos, por lo tanto, se utiliza generalmente en las etapas iniciales del estudio.
Por ejemplo, se quiere realizar un estudio sobre el uso de dispositivos móviles en una cafetería universitaria, simplemente se va a la cafetería objeto de estudio y se entrevista a los individuos que estén cerca, esto es conveniente para el investigador, pero se compromete la representatividad ya que solo se pregunta a los estudiantes que lleguen a la hora en la que se está en la cafetería.

Tanto los métodos de muestreo probabilístico como los no probabilísticos tienen ventajas e inconvenientes, por eso se debe saber cuál elegir dependiendo del objetivo de la investigación, recursos y limitaciones. Por ejemplo, el muestreo probabilístico requiere más tiempo, dinero y esfuerzo al contrario de los no probabilísticos, pero estos últimos tiene más limitaciones y sesgos.

En conclusión, el principal objetivo del muestreo estadístico es establecer un número óptimo de observaciones, suficiente y representativo que participaran en un estudio como fuente de información para que los resultados puedan ser extrapolados a la población o poblaciones semejantes. Es natural cometer errores en la elección de la muestra, por ello, se debe tener cuidado al elegir qué tipo de muestreo se usara en el estudio, es de las primeras decisiones que se debe hacer con sumo cuidado ya que eso conlleva muchos errores.

Capítulo 2

Principales Muestreos Aleatorios

A continuación, se hablará de los distintos tipos de muestreo aleatorio, estas técnicas se usarán en el caso práctico del trabajo, aunque antes se explicarán algunos conceptos que se deben tener en cuenta en el uso del muestreo.

Los parámetros que se usan dependen del tipo de variable con la que se esté trabajando, puede ser de dos tipos:

Variable cuantitativa: describe una característica de un elemento con números.

Total poblacional: $\tau = x_1 + \cdots + x_N$

Media poblacional: $\mu = \frac{\tau}{N}$

Varianza poblacional: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$

Variable cualitativa: describe una característica de un elemento sin la necesidad de números. Se llamara a a_i: si un elemento tiene o no la característica tomando 0 y 1 como valores, siendo 0 que no tenga la característica y 1 si la tiene.

Total poblacional con una característica: $\tau_A = A = a_1 + \cdots + a_N$

Proporción poblacional con una característica: $p_A = \frac{\tau_A}{N}$

Varianza poblacional: $\sigma_A^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - p_A)^2 = p_A \cdot (1 - p_A)$

Después al obtener la muestra se usan los siguientes estadísticos para estimar las características de la muestra hallada, también depende de la variable que se esté usando.

En las **variables cuantitativas** se utiliza:

Media muestral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Cuasivarianza muestral: $s_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Varianza muestral: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n} \cdot s_c^2$

Cuasivarianza muestral corregida: $s_{C_N}^2 = \left(1 - \frac{1}{N}\right) \cdot s_c^2$

Si N es grande, la cuasivarianza muestral corregida se aproxima a la cuasivarianza muestral.

En las **variables cualitativas** se usa:

Proporción: $\hat{p}_A = \frac{1}{n}\sum_{i=1}^{n} A_i$

Error de muestreo

Se puede escribir el intervalo de confianza para el parámetro θ, este será el parámetro que queremos estimar, como:

$$\hat{\theta} \pm k \cdot \sqrt{Var(\hat{\theta})} = \hat{\theta} \pm k \cdot \sigma(\hat{\theta})$$

A la cantidad $k \cdot \sqrt{Var(\hat{\theta})} = e$ se le llama **error de muestreo**.

Si la distribución es normal:

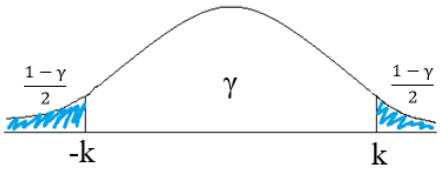


Ilustración 2. Distribución normal

k depende del nivel de confianza γ, si k = 1’96 su nivel de confianza es γ = 95%, esto se puede ver dando valores a k y el nivel de confianza de cada uno.

En el muestreo estadístico, es frecuente tomar el valor k = 2 ya que se asegura una confianza del 75% y del 95’45% si hay normalidad. Así, $e = 2 \cdot \sigma(\hat{\theta})$ se le llama **límite de error**.

Error absoluto y error relativo

El error relativo se usa para comparar distintas medidas y no tiene unidades.

$$e_r = \frac{e}{|\hat{\theta}|}$$

El error absoluto es la diferencia entre el valor real y el valor medido, tiene las mismas unidades que el estadístico y su estimador.

$$e = |\theta - \hat{\theta}|$$

Se relacionan por: $e = e_r \cdot |\hat{\theta}|$

Si el error viene dado en porcentaje, en proporciones será un error absoluto a menos que se diga lo contrario, y será relativo en medias y totales, salvo que sean medias de porcentajes.

Estimadores indirectos

En el muestreo, es usual utilizar la información de alguna variable auxiliar, Y, relacionada con la variable objeto de estudio, X, trata de mejorar la precisión de un estimador simple (muestreo aleatorio simple o estratificado) utilizando la información de la variable auxiliar supuestamente relacionada con la variable en estudio, aunque sea el método de recoger observaciones se usan los estimadores indirectos.

Se quiere estimar X (variable cuantitativa) y se utiliza la información de Y (variable cuantitativa), estas deben estar correlacionadas. Los métodos indirectos más usados son:

- **Razón**, basado en la razón entre X e Y.
- **Regresión**, basado en la regresión entre X e Y.
- **Diferencia**, la pendiente es próxima a 1.

Estimador de razón

Es un estimador no lineal porque es un cociente.

Si X e Y tienen una fuerte correlación positiva, la estimación de razón da mejores resultados que el muestreo aleatorio simple.

Para que el estimador de razón sea más eficiente, se ha de cumplir que:

$$Var_{RAZÓN}(\hat{\mu}) < Var_{MAS}(\hat{\mu})$$

Para que ocurra lo anterior, $\rho > \frac{1}{2} \cdot \frac{CV(Y)}{CV(X)}$.

No se puede calcular el coeficiente de variación (CV) ya que no tenemos los datos de la población; tampoco ρ, pero sí podemos estimarlo para una muestra:

$$r > \frac{1}{2} \cdot \frac{\widehat{CV(Y)}}{\widehat{CV(X)}} = \frac{1}{2} \cdot \frac{s_y \cdot \bar{X}}{s_x \cdot \bar{Y}}$$

Consideramos la recta de regresión $x = a + b \cdot y$, $b = \frac{s_{xy}}{s_y^2}$ y $a = \bar{X} - b \cdot \bar{Y}$.

Si la recta de regresión pasa por el origen de ordenadas, el estimador de razón es insesgado. Esto ocurre cuando a es nulo, a = 0 o se aproxima a 0.

Para estimar los parámetros se usa $R = \frac{\mu_x}{\mu_y}$ o su estimación, $\hat{R} = r = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\bar{X}}{\bar{Y}}$.

Para la estimar la **media**: $\hat{\mu}_x = r \cdot \mu_y$ **,** siendo μ_y: la estimación poblacional de los datos auxiliares, y para estimar el **total**: $\hat{\tau}_x = N \cdot \mu_x$.

La varianza teórica de la media es: $Var(\hat{\mu}_x) = \frac{N-n}{N-1} \cdot \frac{\sigma_R^2}{n}$, siendo la varianza de la razón:

$\sigma_R^2 = \frac{\sum_{i=1}^{N} \left(x_i - R_{y_i}\right)^2}{n}$.

La varianza teórica de la razón es: $Var(r) = \frac{1}{\mu_y^2} \cdot \frac{N-n}{N-1} \cdot \frac{\sigma_R^2}{n}$.

La varianza estimada de la media es: $V\hat{a}r(\hat{\mu}_x) = \left(1 - \frac{n}{N}\right) \cdot \frac{s_R^2}{n}$, siendo la cuasivarianza de la razón:

$s_R^2 = \frac{\sum_{i=1}^{n} \left(X_i - r_{Y_i}\right)^2}{n-1}$.

La varianza teórica estimada de la razón es: $V\hat{a}r(r) = \frac{1}{\mu_y^2} \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_R^2}{n}$.

Para el cálculo del tamaño de la muestra, n, dependiendo del parámetro, la n_∞ cambia, aunque se tiene en cuenta que siempre se debe usar la varianza de la razón.

Para la media: $n_\infty = \frac{k^2 \cdot \sigma_R^2}{e^2}$, para la razón: $n_\infty = \frac{k^2 \cdot \sigma_R^2}{e^2 \cdot \mu_y^2}$ y para el total: $n_\infty = \frac{N^2 \cdot k^2 \cdot \sigma_R^2}{e^2}$, sin redondear ese resultado se sustituye en la fórmula de n, $n = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$.

Estimador de regresión

Se toma X como la variable dependiente, la cual se quiere estimar, e Y la variable independiente. A partir de la Y estimamos la X y consideramos la recta de regresión:

$x = a + b \cdot y$, siendo: $b = \frac{s_{xy}}{s_y^2}$, $r = \frac{s_{xy}}{s_x \cdot S_y}$ y $a = \bar{X} - b \cdot \bar{Y}$.

Si la recta de regresión no pasa por el origen de ordenadas, la estimación de regresión es mejor que la de razón y esto ocurre cuando a no es nulo, $a \neq 0$.

Para estimar la **media**: $\hat{\mu}_{XL} = \bar{X} + b \cdot (\mu_y - \bar{Y})$, siendo: $b = \frac{s_{xy}}{s_y^2}$.

La varianza teórica de la media: $Var(\hat{\mu}_{XL}) = \frac{N-n}{N-1} \cdot \frac{\sigma_L^2}{n}$, siendo: $\sigma_L^2 = \sigma_X^2 - \beta^2 \cdot \sigma_Y^2$ y β es el coeficiente de la recta de regresión poblacional, X = α + βY.

La estimación de la varianza teórica de la media: $V\hat{a}r(\hat{\mu}_{XL}) = \left(1 - \frac{n}{N}\right) \cdot \frac{s_L^2}{n}$, siendo:

$s_L^2 = \frac{1}{n-2} \cdot [\sum_{i=1}^{n}(X_i - \bar{X})^2 - b^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2] = \frac{n}{n-2} \cdot \left[s_x^2 - b^2 \cdot s_y^2\right] = \frac{n}{n-2} \cdot [s_x^2 \cdot (1 - r^2)]$.

Para estimar el **total**: $\hat{\tau}_{XL} = N \cdot \hat{\mu}_{XL}$.

Para el cálculo del tamaño de la muestra en este caso, n, se usa la varianza de regresión (σ_L^2) en la fórmula de : n_∞. El procedimiento es el mismo que en los demás muestreos.

Como se puede observar en la explicación de los estimadores, hay semejanzas y diferencias. A continuación, se comentará brevemente.

> Si la dispersión de los puntos es mayor a medida que los valores de X e Y aumenten, el estimador de razón es mejor.
>
> Si X e Y tienen una fuerte correlacion positiva, el de razón tiene mejor resultado que el muestreo aleatorio simple.
>
> Si los puntos están en una línea recta con varianza constante que no pasa por el origen y con pendiente muy diferente de la unidad, la estimación de regresión resulta mejor.

Muestreo Aleatorio Simple

Si una muestra de tamaño n se extrae de una población P de tamaño N de tal manera que tiene la misma probabilidad de ser extraída que otra que tiene el mismo tamaño, se dice que la muestra es aleatoria simple (m a s) y el procedimiento estadístico para extraerla se llama muestreo aleatorio simple (M A S).

Se pueden extraer muestras **con reposición** y **sin reposición**, la diferencia es que en las muestras con reposición cualquier unidad extraída vuelve de nuevo a la población, por ello puede volver a ser elegida; al contrario que la muestra sin reposición que la unidad extraída no vuelve a la población.

En este caso, se debe buscar tanto estimadores para el muestreo con reposición y sin reposición, al igual que para variables cualitativas y cuantitativas.

Siendo θ el parámetro que se va a estudiar, entonces se tiene:

Para estimar el total en variables cuantitativas: $\tau \rightarrow \hat{\tau} = N \cdot \hat{\mu}$, y cualitativas: $\tau_A \rightarrow \hat{\tau}_A = N \cdot \hat{p}_A$.

Para estimar la media: $\mu \rightarrow \hat{\mu} = \bar{X}$,y la proporción: $p_A \rightarrow \hat{p}_A = \hat{p}_A$.

Junto con el estimador, con objeto de medir el error, se debe calcular la varianza teórica del estimador: $Var(\hat{\theta})$. Las varianzas de cada parámetro en el muestreo con reposición son:

$$\tau \rightarrow Var(\hat{\tau}) = N^2 \cdot \frac{\sigma^2}{n} \qquad \tau_A \rightarrow Var(\hat{\tau}_A) = N^2 \cdot \frac{p_A \cdot (1-p_A)}{n}$$

$$\mu \rightarrow Var(\hat{\mu}) = \frac{\sigma^2}{n} \qquad p_A \rightarrow Var(\hat{p}_A) = \frac{p_A \cdot (1-p_A)}{n}$$

Y para el muestreo sin reposición las varianzas teóricas de cada parámetro son:

$$\tau \rightarrow Var(\hat{\tau}) = \frac{N-n}{N-1} \cdot N^2 \cdot \frac{\sigma^2}{n} \qquad \tau_A \rightarrow Var(\hat{\tau}_A) = \frac{N-n}{N-1} \cdot N^2 \cdot \frac{p_A \cdot (1-p_A)}{n}$$

$$\mu \rightarrow Var(\hat{\mu}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \qquad p_A \rightarrow Var(\hat{p}_A) = \frac{N-n}{N-1} \cdot \frac{p_A \cdot (1-p_A)}{n}$$

A $\frac{N-n}{N-1}$ se le llama **fracción de corrección por poblaciones finitas**.

Como se puede observar estas son varianzas teóricas de parámetros estimados, es decir, todas dependen de la varianza poblacional. En algunos casos, no se tendrá información de la varianza poblacional y se tendrán que estimar las varianzas teóricas de los parámetros; eso lo hacemos con la cuasivarianza muestral.

En el muestreo con reposición, las varianzas estimadas son:

$$\hat{Var}(\hat{\tau}) = N^2 \cdot \frac{s_c^2}{n} \qquad \hat{Var}(\hat{\tau}_A) = N^2 \cdot \frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n-1}$$

$$\hat{Var}(\hat{\mu}) = \frac{s_c^2}{n} \qquad \hat{Var}(\hat{p}_A) = \frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n-1}$$

En el muestreo sin reposición las varianzas estimadas son:

$$\hat{Var}(\hat{\tau}) = N^2 \cdot (1-f) \cdot \frac{s_c^2}{n} \qquad \hat{Var}(\hat{\tau}_A) = N^2 \cdot (1-f) \cdot \frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n-1}$$

$$\hat{Var}(\hat{\mu}) = (1-f) \cdot \frac{s_c^2}{n} \qquad \hat{Var}(\hat{p}_A) = (1-f) \cdot \frac{\hat{p}_A \cdot (1-\hat{p}_A)}{n-1}$$

Se puede observar la expresión $1-f$, esta expresión es igual a $1-\frac{n}{N}$.

La fracción de muestreo $\left(\frac{n}{N}\right)$ sirve para describir la proporción de la población total que se incluye en la muestra, es decir, indica qué tan grande es la muestra en relación con la población total.

Se puede observar en las fórmulas que tanto en el muestreo con reposición como sin reposición la varianza total se puede calcular a partir de la varianza de la media o proporción, simplemente multiplicando por el tamaño de la población al cuadrado.

$$Var(\hat{\tau}) = N^2 \cdot Var(\hat{\mu}) \qquad Var(\hat{\tau}_A) = N^2 \cdot Var(\hat{p}_A)$$

$$V\hat{a}r(\hat{\tau}) = N^2 \cdot V\hat{a}r(\hat{\mu}) \qquad V\hat{a}r(\hat{\tau}_A) = N^2 \cdot V\hat{a}r(\hat{p}_A)$$

Por último, se explicará el cálculo de tamaño muestral para el MAS. Como se ha ido mencionando, se deben tener en cuenta muchos factores para hacer un buen muestreo, los siguientes pasos para obtener una buena estimación es saber qué tamaño de muestra es el óptimo para tener resultados fiables y válidos. Para ello, teniendo un error, se busca el tamaño muestral.

Fijando un error, e, y el parámetro que se quiere estimar, en este caso será la media.

Para el muestreo con reposición, se calcula directamente el tamaño muestral $n = \frac{k^2 \cdot \sigma^2}{e^2}$, este resultado se redondea por exceso.

Para el muestro sin reposición, se calcula primero una n auxiliar, $n_\infty = \frac{K^2 \cdot \sigma^2}{e^2}$, sustituyendo en la fórmula del tamaño muestral, $n = \frac{n_\infty}{1+\frac{n_\infty - 1}{N}}$, sin redondear y este resultado se redondea por exceso. Si A es cualitativa, la n auxiliar toma la siguiente fórmula: $n_\infty = \frac{K^2 \cdot p \cdot (1-p)}{e^2}$.

Muestreo Aleatorio Estratificado

En algunos casos la población está dividida en subgrupos claramente identificables por alguna característica, a esos subgrupos se le llaman **estratos**, los cuales son homogéneo entre ellos, es decir, los elementos de un mismo estrato tienen características comunes y heterogéneo con el resto. Esto significa que no son similares a los otros elementos fuera de ese estrato. Llamamos L al número de estratos identificados en la población, en cada uno de ellos se realiza un muestreo aleatorio simple.

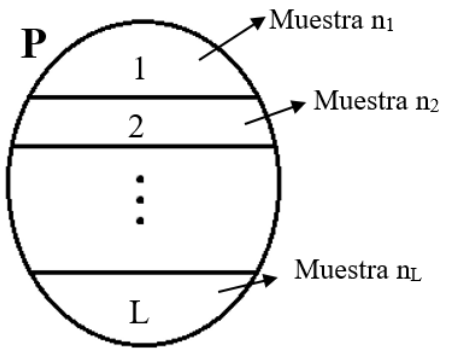


Ilustración 3. Población P dividida en L estratos.

$$N_1 = tamaño\ del\ estrato\ 1$$
$$n_1 = tamaño\ de\ la\ muestra\ 1$$
$$N_2 = tamaño\ del\ estrato\ 2$$
$$n_2 = tamaño\ de\ la\ muestra\ 2$$
$$\dots$$
$$N_L = tamaño\ del\ estrato\ L$$
$$n_L = tamaño\ de\ la\ muestra\ L$$

Se verifica que: $N = \sum_{i=1}^{L} N_i = N_1 + \cdots + N_L$ y $n = \sum_{i=1}^{L} n_i = n_1 + \cdots + n_L$.

Este muestreo tiene ciertas ventajas frente al anterior como, por ejemplo, se puede llegar a toda la población, observaciones con mejor calidad, se necesita menor tamaño de muestra y disminuye la varianza (se hacen grupos más homogéneos).

A continuación, se pueden ver las estimaciones de cada parámetro.

Para el **total:** $\hat{\tau} = \sum_{i=1}^{L} N_i \cdot \hat{\mu}_i$ siendo $\hat{\mu}_i$ la media de cada estrato.

La varianza teórica del total en el muestreo sin reposición: $Var(\hat{\tau}) = \sum_{i=1}^{L} N_i^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$

La varianza teórica del total en el muestreo con reposición: $Var(\hat{\tau}) = \sum_{i=1}^{L} N_i^2 \cdot \frac{\sigma_i^2}{n_i}$

La varianza estimada del total en el muestreo sin reposición: $V\hat{a}r(\hat{\tau}) = \sum_{i=1}^{L} N_i^2 \cdot \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s_{c_i}^2}{n_i}$

La varianza estimada del total en el muestreo con reposición: $V\hat{a}r(\hat{\tau}) = \sum_{i=1}^{L} N_i^2 \cdot \frac{s_{c_i}^2}{n_i}$

Para la **media:** $\hat{\mu} = \sum_{i=1}^{L} W_i \cdot \hat{\mu}_i$ llamamos $W_i = \frac{N_i}{N}$ a la fracción que corresponde al tamaño del estrato con respecto al tamaño de la población.

Se le llama **afijación** a la fracción de muestra del estrato con respecto al tamaño entero de la muestra, $w_i = \frac{n_i}{n}$.

La varianza teórica de la media en el muestreo sin reposición: $Var(\hat{\mu}) = \sum_{i=1}^{L} W_i^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$

La varianza teórica de la media en el muestreo con reposición: $Var(\hat{\mu}) = \frac{1}{N^2} \cdot \sum_{i=1}^{L} N_i^2 \cdot \frac{\sigma_i^2}{n_i}$

La varianza estimada de la media en el muestreo sin reposición:

$$V\hat{a}r(\hat{\mu}) = \frac{1}{N^2} \cdot \sum_{i=1}^{L} N_i^2 \cdot \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s_{c_i}^2}{n_i}$$

La varianza estimada de la media en el muestreo con reposición: $V\hat{a}r(\hat{\mu}) = \frac{1}{N^2} \cdot \sum_{i=1}^{L} N_i^2 \cdot \frac{s_{c_i}^2}{n_i}$

En este caso se puede poner al error como: $e = \sqrt{\sum_{i=1}^{L} W_i^2 \cdot e_i^2}$

Para la **proporción:** $\hat{p} = \frac{1}{N} \cdot \sum_{i=1}^{L} N_i \cdot \hat{p}_i$.

La varianza teórica de la proporción en el muestreo sin reposición:

$$Var(\hat{p}_i) = \frac{1}{N^2} \cdot \sum_{i=1}^{L} N_i^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{p_i \cdot (1 - p_i)}{n_i}$$

La varianza teórica de la proporción en el muestreo con reposición:

$$Var(\hat{p}_i) = \frac{1}{N^2} \cdot \sum_{i=1}^{L} N_i^2 \cdot \frac{p_i \cdot (1 - p_i)}{n_i}$$

La varianza estimada de la proporción en el muestreo sin reposición:

$$\hat{Var}(\hat{p}_i) = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \cdot \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{\hat{p}_i \cdot (1 - \hat{p}_i)}{n_i - 1}$$

La varianza estimada de la proporción en el muestreo con reposición:

$$\hat{Var}(\hat{p}_i) = \sum_{i=1}^{L} \left(\frac{N_i}{N}\right)^2 \cdot \frac{\hat{p}_i \cdot (1 - \hat{p}_i)}{n_i - 1}$$

Para calcular el tamaño de la muestra se debe tener en cuenta que se calculan la n y $n_1, \dots, n_L$. Conociendo que $n = n_1 + \cdots + n_L$ de forma que se halla primero $n_1, \dots, n_L$ y luego se suman para obtener n.

Se usará la fórmula de la afijación para el cálculo de las n teniendo en cuenta que $0 \leq w_i \leq 1$ y $\sum_{i=1}^{L} w_i = 1$.

En la expresión de la afijación se despeja n, quedando: $n_i = w_i \cdot n$.

Entonces, dependiendo de la afijación que se tenga, se podrán hallar las n_i. Para ello, se comentarán los distintos tipos de afijación que existen.

- **Afijación constante:** cuando la muestra está repartida de forma uniforme en los estratos, es decir, todas las n_i son iguales. Entonces queda la fórmula de afijación como:

$$w_i = \frac{n_i}{n} => w_1 = \cdots = w_L = \frac{1}{L}$$

- **Afijación proporcional:** guarda la misma proporción la muestra y la población, un claro ejemplo es cuando la población está dividida en hombres y mujeres, y guarda una proporción 40% ~ 60%, entonces las muestras deben guardar la misma proporción.

$$\frac{N_i}{N} = \frac{n_i}{n};\ W_i = w_i\ \forall\ i = 1, \dots, L$$

- **Afijación óptima:** a cada estrato se le añade un coste para la extracción de una observación y a ese coste se le llamara c_i. Habrá tantos como estratos existan.
  Hay dos tipos, fijando un coste con el error mínimo y fijando un error con el coste mínimo.

  <u>Coste fijo y error mínimo</u>
  Se llamara C al coste total, $C = \sum_{i=1}^{L} n_i \cdot c_i$ y c_i al coste de extraer una observación en el estrato i, los cuales se conocen.
  Para el tamaño muestral de cada estrato se calcula, $n_i = n \cdot w_i => C = n \cdot \sum_{i=1}^{L} w_i \cdot c_i$.
  Para el tamaño muestral total se calcula, $n = \frac{C}{\sum_{i=1}^{L} w_i \cdot c_i}$ siendo $w_i = \frac{N_i \cdot \sigma_i / \sqrt{c_i}}{\sum_{i=1}^{L} N_i \cdot \sigma_i / \sqrt{c_i}}$.
  Y se calculan los distintos tamaños muestrales, $n_i = n \cdot w_i$, teniendo en cuenta que la n calculada no se redondea y n_i se redondea sin pasarse del coste total y con el mínimo error.

  <u>Error fijo y coste mínimo</u>
  En este caso tenemos un error conocido, el coste se calcula: $C = n \cdot \sum_{i=1}^{L} w_i \cdot c_i$ y $n_i = n \cdot w_i$. Se halla el tamaño muestral con la siguiente expresión: $n = \frac{\sum_{i=1}^{L} N_i^2 \cdot \sigma_i^2 / w_i}{\frac{e^2 \cdot N^2}{K^2} + \sum_{i=1}^{L} N_i \cdot \sigma_i^2}$.

Se puede observar mejor estos procedimientos en el siguiente esquema:

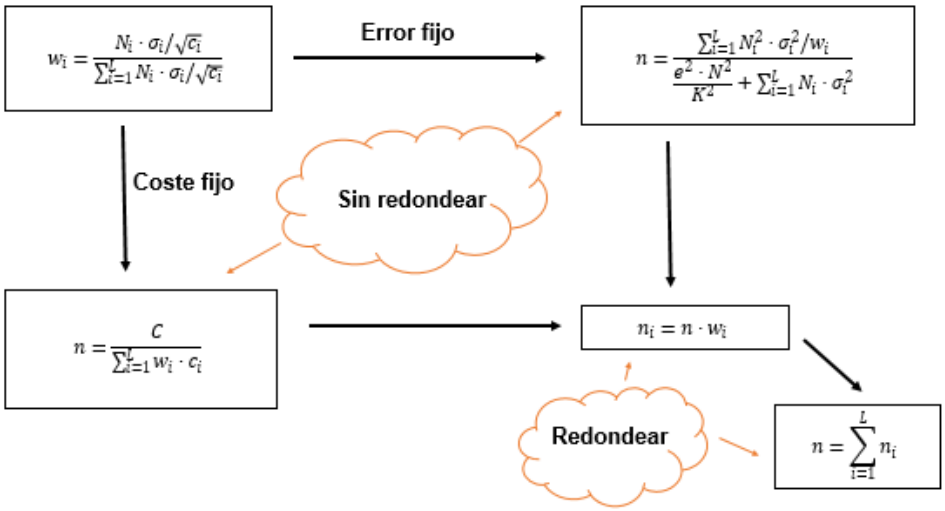


Ilustración 4:Esquema para el cálculo del tamaño muestral

- **Afijación de Neyman:** cuando todos los costes son constantes, por ejemplo, las encuestas realizadas por teléfono en España.
  Se calcula la afijación de cada estrato con la siguiente fórmula: $w_i = \frac{N_i \cdot \sigma_i}{\sum_{i=1}^{L} N_i \cdot \sigma_i}$.

Estimadores indirectos combinados con estratificado

Se considera la población dividida en L trozos por alguna característica concreta, cada trozo se le llama estrato que resulta ser homogéneo dentro de sí y heterogéneo con los demás, y en cada uno de los estratos, se realiza un estimador indirecto, con X e Y.

Se usa la siguiente notación:

N: tamaño de la población, $N_1 + \cdots + N_L = N$.

N_i: tamaño del estrato i, i = 1, …, L.

n: tamaño de la muestra, $n_1 + \cdots + n_L = n$.

n_i: tamaño de la muestra del estrato i, i = 1, …, L.

$\boldsymbol{\mu_{yi}}$: media de la población de la variable Y en el estrato i, i = 1, …, L.

$\boldsymbol{\tau_{yi}}$: total, poblacional de la variable Y en el estrato i, i = 1, …, L.

$X_{i1}, \dots, X_{in_i}$ es la muestra en el estrato i = 1, …, L para la población definida por la variable X.

$Y_{i1}, \dots, Y_{in_i}$ es la muestra en el estrato i = 1, …, L para la población definida por la variable Y.

$\boldsymbol{f_i}$ es la fracción de muestreo en el estrato i =1, …, L. $f_i = n_i/N_i$

$\boldsymbol{W_i}$ es la proporción de estrato en el total poblacional en el estrato i =1, …, L. $W_i = N_i/N$

Estimador de razón combinado con estratificado.

Para un **estimador separado** (RS), para la estimación de la media se usa la estimación de la r en cada estrato, $r_i = \frac{\bar{X}_i}{\bar{Y}_i}$ y la r del estimado separado: $r_{RS} = \sum_{i=1}^{L} \frac{N_i}{N} \cdot r_i = \sum_{i=1}^{L} W_i \cdot r_i$.

$$\hat{\mu}_{X_{RS}} = \sum_{i=1}^{L} \frac{N_i}{N} \cdot \hat{\mu}_{X_i} = \sum_{i=1}^{L} \frac{N_i}{N} \cdot r_i \cdot \mu_{y_i}$$

Para la varianza teórica y su estimación:

$Var(\hat{\mu}_{X_{RS}}) = \sum_{i=1}^{L} W_i^2 \cdot Var(\hat{\mu}_{X_i})$, siendo :$Var(\hat{\mu}_{X_i}) = \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma^2_{R_{RSi}}}{n_i}$

$V\hat{a}r(\hat{\mu}_{X_{RS}}) = \sum_{i=1}^{L} W_i^2 \cdot V\hat{a}r(\hat{\mu}_{X_i})$, siendo: $V\hat{a}r(\hat{\mu}_{X_i}) = \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s^2_{r_{RSi}}}{n_i}$ y

cuasivarianza: $s^2_{r_{RSi}} = \frac{\sum_{i=1}^{n_i} (x_{ij} - r_i \cdot y_{ij})^2}{n_i - 1}$.

Para un **estimador combinado** (RC), la estimación para la media:

$\hat{\mu}_{X_{RC}} = r_{RC} \cdot \mu_y$, siendo: $r_{RC} = \frac{\bar{X}}{\bar{Y}} = \frac{\sum_{i=1}^{L} W_i \cdot \bar{X}_i}{\sum_{i=1}^{L} W_i \cdot \bar{Y}_i} = \frac{\sum_{i=1}^{L} N_i \cdot \bar{X}_i}{\sum_{i=1}^{L} N_i \cdot \bar{Y}_i}$ y $\mu_y = \frac{\sum_{i=1}^{L} N_i \cdot \mu_{yi}}{N}$.

Para la varianza teórica y su estimación:

$Var(\hat{\mu}_{X_{RC}}) = \sum_{i=1}^{L} W_i^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma^2_{R_{RCi}}}{n_i}$

$V\hat{a}r(\hat{\mu}_{X_{RC}}) = \sum_{i=1}^{L} W_i^2 \cdot (1 - \frac{n_i}{N_i}) \cdot \frac{s^2_{r_{RCi}}}{n_i}$, siendo: $s^2_{r_{RCi}} = \frac{\sum_{i=1}^{n_i} [(x_{ij} - \bar{X}_i) - r_{RC}(y_{ij} - \bar{Y}_i)]^2}{n_i - 1}$.

Estimador de regresión combinado con estratificado.

Para un **estimador separado** (RS), la media estimada:

$\hat{\mu}_{x_{LRS}} = \sum_{i=1}^{L} \frac{N_i}{N} \cdot \hat{\mu}_{x_{Li}} = \sum_{i=1}^{L} W_i^2 \cdot (\bar{X}_i - b_i \cdot (\mu_{y_i} - \bar{Y}_i))$, siendo: $b_i = \frac{s_{x_i y_i}}{s^2_{y_i}}$.

Para la varianza teórica y su estimación:

$V\hat{a}r(\hat{\mu}_{X_{LRS}}) = \sum_{i=1}^{L} W_i^2 \cdot V\hat{a}r(\hat{\mu}_{X_{Li}})$

$V\hat{a}r(\hat{\mu}_{X_{Li}}) = \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s^2_{L_{RSi}}}{n_i}$, siendo: $s^2_{L_{RSi}} = \frac{1}{n_i - 2} \cdot \left(\sum_{i=1}^{n_i - 1} \left((x_{ij} - \bar{X}_i)^2 - b_i^2 \cdot \sum_{i=1}^{n_i - 1} (y_{ij} - \bar{Y}_i)^2\right)\right)$

$$= \frac{n_i}{n_i - 2} \cdot \left(s^2_{x_i} - b_i^2 \cdot s^2_{y_i}\right) = \frac{n_i}{n_i - 2} \cdot \left(s^2_{x_i}(1 - r_i^2)\right)$$

Para un **estimador combinado** (RC), la estimación de la media: $\hat{\mu}_{x_{LRC}} = \bar{X} + b \cdot (\mu_y - \bar{Y})$,

siendo: $\bar{X} = \sum_{i=1}^{L} W_i \cdot \bar{X}_i$, $\bar{Y} = \sum_{i=1}^{L} W_i \cdot \bar{Y}_i$, $\mu_y = \frac{\sum_{i=1}^{L} N_i \cdot \mu_{y_i}}{N}$ y $b = \frac{\sum_{i=1}^{L} W_i^2 \cdot (1 - f_i) \cdot \frac{s^2_{y_i}}{n_i} \cdot b_i}{\sum_{i=1}^{L} W_i^2 \cdot (1 - f_i) \cdot \frac{s^2_{y_i}}{n_i}}$.

Para la varianza teórica: $Var(\hat{\mu}_{Lx_{RC}}) = \sum_{i=1}^{L} W_i^2 \cdot \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s^2_{L_{RCi}}}{n_i}$, siendo:

$$s^2_{L_{RCi}} = s^2_{x_i} - 2bs_{x_i y_i} + b^2 s^2_{y_i}$$

Para el cálculo del tamaño de la muestra se debe tener en cuenta la afijación dependiendo del estimador usado para usar la varianza correcta.

$$\omega_i = \frac{N_i \cdot \sqrt{varianza}/\sqrt{c_i}}{\sum_{i=1}^{L} N_i \cdot \sqrt{varianza}/\sqrt{c_i}}$$

Según el muestreo usado:

Para el **coste fijo** y minimizar el error.

$$n = \frac{C}{\sum_{i=1}^{L} \omega_i c_i} \qquad\qquad C = \sum_{i=1}^{L} \omega_i c_i$$

Para el **error fijo** y mínimo coste.

$$n = \frac{\sum_{i=1}^{L} N_i^2 \cdot \frac{varianza}{\omega_i}}{N^2 \frac{e^2}{k^2} + \sum_{i=1}^{L} N_i \cdot varianza}$$

Y luego, con la n resultante se calcula las distintas, $n_i = n \cdot \omega_i$ y se redondean según coste o error.

Muestreo Sistemático

En este tipo de muestreo, se selecciona un elemento al azar, la semilla y se representará con x_0, y para escoger el resto de la muestra, se utiliza intervalos regulares basados en un periodo k.

Las ventajas que tiene este muestreo es que es un método fácil de realizar, elimina la posibilidad de autocorrelación y evita los costes.

Se utiliza a menudo en investigaciones sociales, económicas y de mercado, en el ámbito de estadística se usan en teoría de colas o control de calidad.

El proceso de este muestreo es el siguiente:

Se elabora una lista ordenada de los N elementos de la población, se divide la población en n fragmentos. El tamaño de estos fragmentos será: $k = \frac{N}{n}$, donde k se le llama **coeficiente de elevación**, obtenemos la semilla x_0 aleatoriamente y este será el primer elemento de la muestra, los siguientes elementos de la muestras serán los que ocupen la posición k en los fragmentos divididos, es decir, seguirá esta sucesión aritmética: $x_0, x_0 + k, x_0 + 2k, \dots, x_0 + (n-1)k$.

Los estimadores de este muestreo son:

Si $N = n \cdot k$, el estimador de la media es: $\hat{\mu} = \bar{X}$.

Si $N \neq n \cdot k$ entonces $N > n \cdot k$ y generalizando queda: $k \leq \frac{N}{n}$.

Los estimadores en el muestreo sistemático son los siguientes:

Para la media: $\hat{\mu} = \bar{X}$ Para la proporción: $\hat{p} = \frac{X}{n}$

Para el total para las dos variables se multiplica por N.

Para la varianza teórica de la media: $Var(\hat{\mu}) = \frac{\sigma^2}{n} \cdot (1 + (n-1) \cdot p)$

Como se ha visto, las expresiones para el muestreo sistemático tienen cierta similitud con el muestreo aleatorio simple, se puede observar ciertas relaciones.

Las varianzas del estimador de la media son: $Var_{MS}(\hat{\mu}) = \frac{\sigma^2}{n} \cdot (1 + (n-1) \cdot p)$ y $Var_{MAS}(\hat{\mu}) = \frac{\sigma^2}{n}$.

Recordando el coeficiente de correlación es: $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ siendo $-1 \leq \rho \leq 1$, entonces con esto se puede explicar la relación entre estos dos muestreos.

- Ordenadas al azar: $\rho = 0$

$$Var_{MS}(\hat{\mu}) = Var_{MAS}(\hat{\mu}) => e_{MS} = e_{MAS}$$

- Unidades ordenadas: $\rho < 0$

$$Var_{MS}(\hat{\mu}) < Var_{MAS}(\hat{\mu}) => e_{MS} < e_{MAS}$$

- Ordenadas de forma periódica: $\rho > 0$

$$Var_{MS}(\hat{\mu}) > Var_{MAS}(\hat{\mu}) => e_{MS} > e_{MAS}$$

En el caso práctico no se hará este muestreo ya que la población tiene cierto patrón por lo cual se verá comprometida la aleatoriedad y por eso los resultados pueden estar sesgados.

Muestreo por Conglomerados

La población se divide en trozos que se llamaran **conglomerados**, estos poseen características semejantes a la población y son heterogéneos dentro de sí y homogéneos con los demás. Se extrae una muestrea aleatoria simple de n de los N conglomerados y una vez extraído se estudia todos los individuos pertenecientes al conglomerado.

En este caso, se explicará la notación que se usa en esta técnica, porque cambian algunos conceptos de los anteriores muestreos y para las estimaciones se debe dejar claro.

N: número de conglomerados en la población.

n: número de conglomerados seleccionados en la muestra.

M: número de individuos en población, $M_1 + \ldots + M_N = M$.

M_i: número de individuos en conglomerado i, $i = 1, \ldots, N$.

$\bar{M} = \frac{M}{N}$: tamaño promedio del conglomerado en la población.

m: número individuos en la muestra.

$\bar{m} = \frac{\sum_i^{n=1} M_i}{n}$: número medio de individuos por conglomerado por muestra.

En caso de no saber el tamaño promedio del conglomerado en la población, este se puede estimar como el número medio de individuos por conglomerado por muestra, $\hat{\bar{M}} = \bar{m}$.

Para la estimación de la **media**: $\hat{\mu} = \frac{\sum_{i=1}^{n} \tau_i}{\sum_{i=1}^{n} M_i}$, siendo τ_i: el total en conglomerado i para X.

Se usara estimador de una razón: $\begin{cases} X_i \sim \tau_i \\ Y_i \sim M_i \end{cases}$ para calcular la varianza teórica y su estimación.

$Var(\hat{\mu}) = \frac{N-n}{N-1} \cdot \frac{1}{\bar{M}^2} \cdot \frac{\sigma_\tau^2}{n}$, siendo: $\sigma_\tau^2 = \frac{\sum_{i=1}^{N}(\tau_i - \mu \cdot M_i)^2}{n}$.

$V\hat{a}r(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{\bar{M}^2} \cdot \frac{S_\tau^2}{n}$, siendo: $S_\tau^2 = \frac{\sum_{i=1}^{n}(\tau_i - \hat{\mu} \cdot M_i)^2}{n-1}$.

Si $\bar{M}$ es desconocido se estima por $\bar{m}$, como se ha comentado antes.

La estimación del **total**, su varianza teórica y su estimación.

$$\hat{\tau} = M \cdot \hat{\mu}$$

$$Var(\hat{\tau}) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma_\tau^2}{n} \qquad V\hat{a}r(\hat{\tau}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{S_\tau^2}{n}$$

Para calcular el estimador de la **proporción,** se debe tener en cuenta que: $\boldsymbol{a_i}$: es el número total de individuos que cumplen una condición, y, de esa forma, se tiene: $\hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i}$ y su varianza estimada: $V\hat{a}r(\hat{p}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{\bar{M}^2} \cdot \frac{S_a^2}{n}$,siendo: $s_a^2 = \frac{\sum_{i=1}^{n}(a_i - \hat{p} \cdot M)^2}{n-1}$.

Para calcular el tamaño de muestra n, en el caso de querer estimar la media, se realizará de la misma forma que en el muestreo aleatorio simple cambiando en el denominador de la n infinito añadiendo el tamaño promedio del conglomerado en la población.

Se calcula primero la n auxiliar, $n_\infty = \frac{k^2 \cdot \sigma_\tau^2}{\bar{M}^2 \cdot e^2}$, sustituyendo en la fórmula del tamaño muestral, $n = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$, sin redondear y este resultado se redondea por exceso.

Muestreo Polietápico

Los muestreos polietápicos son procesos que involucran la selección de muestras en **varias fases o etapas** hasta alcanzar los elementos que conformarán la muestra final. Cada fase implica la aplicación de métodos de muestreo como aleatorio simple, estratificado, por conglomerados, cuotas, rutas, entre otros. Las unidades de muestreo cambian en cada extracción subsiguiente, definiendo así cada fase del diseño muestral. Usualmente, la primera fase implica la formación de conglomerados.

Se hacen dos tipos de mezcla de muestreos:

- Si se tiene una lista, la selección es al azar.
- Si no se tiene la lista, se puede combinar: Aleatorio + No aleatorio.

No se deben ignorar las correlaciones intraclases para que los resultados tengan fiabilidad y que hay variabilidad interclase. Llamamos **intra** a la variabilidad dentro de los conglomerados y la variabilidad entre los conglomerados a la que se le llamará **inter**.

Las ventajas de esta técnica serán los costes y tiempos más efectivos, eficacia para recolectar datos de una población geográficamente dispersa, mayor representatividad de la muestra de la población, flexibilidad y mayor eficacia cuando hay un elevado número de participantes y se quiere hacer una selección.

Aunque tiene ciertas desventajas, por ejemplo, para una misma n, es más preciso el muestreo aleatorio simple, se requiere metainformación y el cálculo de estimaciones es complejo.

La selección de muestras se realiza en cada etapa según normas de extracción del muestreo utilizado. Las estimaciones serán las del muestreo elegido en cada etapa y las varianzas tendrán en cuenta todas las etapas, añadiendo un número de sumandos igual al de fases.

Muestreo por Conglomerados en dos etapas o bietápico

Se realizan en dos etapas aleatorias: la **primera etapa** es la muestra de conglomerados, es decir, se hallará un muestreo aleatorio simple de los conglomerados en los que está dividida la población previamente, y la **segunda etapa** será la muestra de elementos de cada conglomerado. Se realizará un muestreo aleatorio simple de los elementos.

En este tipo de muestreo, se pueden diferenciar dos unidades de muestreo: las **unidades primarias** son los conglomerados y las **unidades secundarias** son los elementos del conglomerado.

Un conglomerado frecuentemente contiene demasiados elementos para obtener una medición de cada uno de ellos. La medición de una parte de ellos proporciona información del conglomerado completo, con la buena elección de la muestra.

Esta técnica se usa solo cuando está justificado económicamente, cuando la reducción de los costos supere las pérdidas de precisión. Esto ocurre en las siguientes situaciones:

- Resulta difícil, costoso o imposible construir una lista completa de elementos de la población.
- Cuando existen grandes poblaciones, ya que implementar cualquier otro método de muestreo sería muy costoso.
- La población se concentra en conglomerados naturales (ciudades, escuelas, hospitales, etc.).

Las **ventajas** de este muestreo son:

- No es necesario utilizar todas las unidades de los conglomerados seleccionados.
- Es más fácil conseguir una lista de conglomerados que de elementos.
- Se necesitan menos recursos y el coste es menor (menos gastos de viaje y gastos de administración).
- Disminuir los costos al aumentar la eficacia del muestreo.

Las **desventajas** que se pueden encontrar en este muestreo son:

- Menor precisión.
- Difiere del muestreo estratificado, donde el motivo es incrementar la exactitud.
- Se complican las operaciones algebraicas a la hora de calcular varianzas, pues aumenta las fuentes de variación: debida a la selección de conglomerados y debida al submuestreo en la unidad primaria.

Para la selección de muestras aleatorias por conglomerados en dos etapas, tenemos un primer problema que es la selección de conglomerados apropiados. Las condiciones deseables para la selección son: proximidad geográfica de los elementos dentro de un conglomerado y tamaños de conglomerados convenientes para su manejo.

Puede haber dos tipos de elección, basados en el coste:

- **Elección 1**: Pocos conglomerados (conglomerados grandes con elementos heterogéneos) y muchos elementos en cada uno, se requiere muestra grande para tener estimaciones precisas.
- **Elección 2**: Muchos conglomerados (conglomerados pequeños tienen elementos homogéneos) y pocos elementos en cada uno, se requiere una muestra pequeña y se obtiene información precisa.

En este caso, la notación es la misma que en el muestreo por conglomerados, simplemente se añaden ciertos puntos a tener en cuenta.

x_{ij} : j-ésima observación en la muestra del i-ésimo conglomerado.

$\bar{X}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{m_i}$: media muestral para el i-ésimo conglomerado.

Las fracciones de muestreo son: $f_1 = \frac{n}{N}$ y $f_2 = \frac{m_i}{M_i}$.

En este muestreo, para hallar las estimaciones se debe tener en cuenta si es conocida o no la M y la variabilidad inter e intra.

Para una M conocida (**Estimador Insesgado**):

El estimador de la **media** resulta ser: $\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} M_i \bar{X}_i}{n}$

$\hat{Var}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{1}{\bar{M}^2} \frac{s_b^2}{n} + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{c_i}^2}{m_i}$, siendo: $s_b^2 = \frac{\sum_{i=1}^{n} (M_i \bar{X}_i - \bar{M}\hat{\mu})^2}{n-1}$

INTER **INTRA**

El estimador del **total** es: $\hat{\tau} = M \cdot \hat{\mu} = N \cdot \frac{\sum_{i=1}^{n} M_i \bar{X}_i}{n}$

$$\hat{Var}(\hat{\tau}) = M^2 \cdot \hat{Var}\,(\hat{\mu}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_b^2}{n} + \frac{1}{n} \sum_{i=1}^{n} M_i^2 \cdot \left(1 - \frac{m_i}{M_i}\right) \cdot \frac{s_{c_i}^2}{m_i}$$

Para una M desconocida (**Estimador de Razón**):

La estimación de la **media** es: $\hat{\mu}_r = \frac{\sum_{i=1}^{n} M_i \bar{X}_i}{\sum_{i=1}^{n} M_i}$.

$\hat{Var}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{1}{\bar{M}^2} \frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{c_i}^2}{m_i}$,siendo: $s_r^2 = \frac{\sum_{i=1}^{n} M_i^2 (\bar{X}_i - \hat{\mu}_r)^2}{n-1}$

Se recuerda que: $\widehat{\bar{M}} = \frac{\sum_{i=1}^{n} M_i}{n}$ y $\widehat{M} = \frac{\sum_{i=1}^{n} M_i}{n} \cdot N$.

Para M conocida (**Estimador de Proporción Insesgado**):

La estimación de la **proporción** es: $\hat{p} = \frac{N}{M} \cdot \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{n}$

$\hat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{1}{\bar{M}^2} \frac{s_b^2}{n} + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\hat{p}_i \hat{q}_i}{m_i - 1}$, siendo: $s_b^2 = \frac{\sum_{i=1}^{n} (M_i \hat{p}_i - \bar{M}\hat{p})^2}{n-1}$

Para M desconocida (**Estimador de Proporción Sesgado**):

La estimación de la **proporción** es: $\hat{p}_r = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}$

$\hat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{1}{\bar{M}^2} \frac{s_r^2}{n} + \frac{1}{nN\bar{M}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\hat{p}_i \hat{q}_i}{m_i - 1}$, siendo: $s_r^2 = \frac{\sum_{i=1}^{n} M_i^2 (\hat{p}_i - \hat{p})^2}{n-1}$

Algo a tener en cuenta es, si s_b^2 es pequeña (inter) en comparación con la parte de $s_{c_i}^2$ o $\hat{p}_i \hat{q}_i$ (intra), se pueden seleccionar pocos conglomerados y muchos elementos dentro de cada conglomerado.

El cálculo del tamaño de la muestra, es mucho más complicado que en el muestreo por conglomerados ya que se debe encontrar n y m_i, es decir, el número de conglomerados y los individuos dentro de cada conglomerado.

Depende de dos fuentes de variación, entre conglomerados y dentro de los conglomerados.

A mayor variabilidad entre conglomerados y menor variabilidad dentro de los conglomerados, n es grande y m_i es pequeña.

A menor variabilidad entre conglomerados y mayor variabilidad dentro de los conglomerados, n es pequeña y m_i es grande.

Para explicar el cálculo del tamaño de la muestra se utiliza una situación simplificada:

Todos los conglomerados tienen $\bar{M}$ elementos, es decir, $M_1 = M_2 = \cdots = M_N = \bar{M}$.

Se extraen $\bar{m}$ elementos de cada conglomerado, $m_1 = m_2 = \cdots = m_n = \bar{m}$.

Se usa una situación simplificada para facilitar los cálculos y compresión (los conceptos básicos pueden ser entendidos sin la distracción de las variaciones y complejidades adicionales), para enfocar en los conceptos fundamentales, reducción de complejidad inicial y aplicabilidad general (se pueden adaptar y extender a situaciones más complejas).

La optimización del tamaño de muestra dependerá de las fracciones de muestreo y se debe tener en cuenta que si σ_1^2, $\bar{m}$ disminuye y que si σ_w^2 aumenta, $\bar{m}$ aumenta.

Las fracciones de corrección por poblaciones finitas de muestreo son despreciables**,** es decir, menores que el 1%.

$$f_1 = \frac{n}{N} < 0{,}01, f_2 = \frac{\bar{m}}{\bar{M}} < 0{,}01$$

Las fracciones de corrección, tienden a 0 y quedaría resumido: $Var(\hat{\mu}) = \frac{\sigma_1^2}{n} + \frac{\sigma_w^2}{n\bar{m}}$

Las fracciones de corrección son no despreciables, es decir, mayores que el 1%.

$$f_1 = \frac{n}{N} \neq 0, f_2 = \frac{\bar{m}}{\bar{M}} \neq 0$$

Quedaría resumido: $Var(\hat{\mu}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{\sigma_1^2}{n} + \frac{\bar{M}-\bar{m}}{\bar{M}-1} \cdot \frac{\sigma_w^2}{n\bar{m}}$

En el caso de media se tiene lo siguiente:

La estimación de la **media** será: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \bar{X}_i$

La varianza teórica es: $Var(\hat{\mu}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{1}{\bar{M}^2} \cdot \frac{\sigma_b^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2 \cdot \frac{M_i - m_i}{M_i - 1} \cdot \frac{\sigma_i^2}{m_i}$, siendo:

$\sigma_b^2 = \frac{\bar{M}^2 \cdot \sum_{i=1}^{N}(\mu_i - \mu)^2}{N}$ y $\sigma_i^2 = \frac{\sum_{j=1}^{M_i}(x_{ij} - \mu_i)^2}{M_i}$, sus estimaciones: $\hat{\sigma}_b^2 = S_b^2 = \frac{\bar{M}^2 \cdot \sum_{i=1}^{n}(\bar{X}_i - \hat{\mu})^2}{n-1}$ y

$\hat{\sigma}_i^2 = S_{c_i}^2 = \frac{\sum_{j=1}^{m_i}(X_{ij} - \bar{X}_i)^2}{m_i - 1}$.

Se llamará a $\sigma_1^2 = \frac{\sum_{i=1}^{N}(\mu_i - \mu)^2}{N} \rightarrow \hat{\sigma}_1^2 = s_1^2 = \frac{\sum_{i=1}^{n}(\bar{X}_i - \hat{\mu})^2}{n-1}$ y $\sigma_w^2 = \frac{\sum_{i=1}^{N}\sigma_1^2}{N} \rightarrow \hat{\sigma}_w^2 = s_w^2 = \frac{\sum_{i=1}^{n} s_{c_i}^2}{n}$.

El mejor ajuste para la estimación de la varianza es: $\hat{\sigma}_1^2 = s_1^2 - \frac{s_w^2}{\bar{m}}$

La $\bar{m}$ es la del estudio previo, donde se tiene calculadas las varianzas. A veces, en la práctica no es posible, puede salir negativa y se realiza la primera.

De esta forma se tiene: $Var(\hat{\mu}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{\sigma_1^2}{n} + \frac{\bar{M}-\bar{m}}{\bar{M}-1} \cdot \frac{\sigma_w^2}{n\bar{m}}$

- **Fracciones de muestreo despreciables**
  $c_1 =$ coste de extraer un conglomerado y $c_2 =$ coste de extraer un elemento del conglomerado.

  Con un coste fijo se usa:

  $C = n \cdot c_1 + n \cdot \bar{m} \cdot c_2$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{\sigma_w^2 \cdot c_1}{\sigma_1^2 \cdot c_2}}$

  Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

  Con un error fijo se usa:

  $e = k \cdot \sqrt{Var(\hat{\mu})}$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{\sigma_w^2 \cdot c_1}{\sigma_1^2 \cdot c_2}}$

  Se despeja el tamaño muestral (n) de la fórmula del error (e).

- **Fracciones de muestreo no despreciables**
  Con un coste fijo se usa:

  $C = n \cdot c_1 + n \cdot \bar{m} \cdot c_2$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1}{\frac{\sigma_1^2}{\sigma_w^2} - \frac{1}{\bar{M}}}}$

  Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

  Con un error fijo se usa:

  $e = k \cdot \sqrt{Var(\hat{\mu})}$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1}{\frac{\sigma_1^2}{\sigma_w^2} - \frac{1}{\bar{M}}}}$

  Se despeja el tamaño muestral (n) de la fórmula del error (e).

En el caso de la proporción, se tiene que la estimación de la proporción será: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i$ y la varianza teoría es: $Var(\hat{p}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{1}{\bar{M}^2} \cdot \frac{\sigma_b^2}{n} + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2 \cdot \frac{M_i - m_i}{M_i - 1} \cdot \frac{p_i \cdot q_i}{m_i}$.

Siendo: $\sigma_b^2 = \frac{\bar{M}^2 \cdot \sum_{i=1}^{N}(p_i - p)^2}{N}$ y $\sigma_i^2 = \frac{\sum_{j=1}^{M_i}(a_{ij} - p_i)^2}{M_i}$, sus estimaciones: $\hat{\sigma}_b^2 = s_b^2 = \frac{\bar{M}^2 \cdot \sum_{i=1}^{n}(\hat{p}_i - \hat{p})^2}{n-1}$ y $\hat{\sigma}_i^2 = s_{c_i}^2 = \frac{\sum_{j=1}^{m_i}(A_{ij} - \hat{p}_i)^2}{m_i - 1}$.

Se llamará a $\sigma_1^2 = \frac{\sum_{i=1}^{N}(p_i - p)^2}{N} \rightarrow \hat{\sigma}_1^2 = s_1^2 = \frac{\sum_{i=1}^{n}(\hat{p}_i - \hat{p})^2}{n-1}$ y $\sigma_w^2 = \frac{\sum_{i=1}^{N} p_i \cdot q_i}{N} \rightarrow \hat{\sigma}_w^2 = s_w^2 = \frac{\sum_{i=1}^{n} \hat{p}_i \cdot \hat{q}_i}{n}$.

El mejor ajuste para la estimación de la varianza es: $\hat{\sigma}_1^2 = s_1^2 - \frac{s_w^2}{\bar{m}} = s_1^2 - \frac{\frac{\sum_{i=1}^{n} s_{c_i}^2}{n}}{\bar{m}}$

De esta forma: $Var(\hat{\mu}) = \left(\frac{N-n}{N-1}\right) \cdot \frac{\sigma_1^2}{n} + \frac{\bar{M}-\bar{m}}{\bar{M}-1} \cdot \frac{\sigma_w^2}{n\bar{m}}$

- **Fracciones de muestreo despreciables**
  Con un coste fijo se usa:

  $C = n \cdot c_1 + n \cdot \bar{m} \cdot c_2$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{\sigma_w^2 \cdot c_1}{\sigma_1^2 \cdot c_2}}$

Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

Con un error fijo se usa:

$e = k\sqrt{Var(\hat{p})}$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{\sigma_W^2 \cdot c_1}{\sigma_1^2 \cdot c_2}}$

Se despeja el tamaño muestral (n) de la fórmula del error (e).

- **Fracciones de muestreo no despreciables**
  Con un coste fijo

$C = nc_1 + n\bar{m}c_2$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1}{\frac{\sigma_1^2}{\sigma_W^2} - \frac{1}{\bar{M}}}}$

Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

Con un error fijo

$e = k\sqrt{Var(\hat{p})}$ $\qquad\qquad$ $\bar{m} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1}{\frac{\sigma_1^2}{\sigma_W^2} - \frac{1}{\bar{M}}}}$

Se despeja el tamaño muestral (n) de la fórmula del error (e).

Muestreo por Conglomerados en tres etapas

Los muestreos en una o dos etapas se pueden generalizar a cualquier número de etapas, esto implica mayor complejidad en el manejo de expresiones.

Se puede generalizar en que un muestreo en r etapas es un muestreo en etapas (r > 2) y englobando las r-1 etapas en una sola. Las expresiones se obtienen aplicando de forma recursiva las ya vistas.

La notación usada para tres etapas es:

N: número de conglomerados

n: número de conglomerados de la muestra

$\bar{M}$: número medio de conglomerados en cada conglomerado de 1º etapa

$\bar{m}$: número medio de conglomerados en muestra extraídos de 1º etapa

$\bar{T}$: tamaño conglomerado promedio en población en 2º etapa

$\bar{t}$: número de elementos promedio en los conglomerados seleccionados en 2º etapa

i: contador en 1º etapa (i = 1, …, n)

j: contador en 2º etapa (j =1, …,$\bar{m}$)

k: contador en 3º etapa (k= 1, …,$\bar{t}$)

Las estimaciones nos resultan de la siguiente forma:

X_{ijk} = k-ésimo elemento del j-ésimo conglomerado de la 2º etapa del i-ésimo conglomerado

$\bar{X}_{ij} = \frac{\sum_{k=1}^{\bar{t}} X_{ijk}}{\bar{t}}$: media de los elementos del conglomerado en 2º etapa

$\bar{X}_i = \frac{\sum_{j=1}^{\bar{m}} \bar{X}_{ij}}{\bar{m}}$: media de las medias en cada conglomerado de 1º etapa

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \bar{X}_i}{n} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \frac{\bar{X}_{ij}}{\bar{m}}}{n} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \frac{\sum_{k=1}^{\bar{t}} X_{ijk}}{\bar{t}}}{n\bar{m}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \sum_{k=1}^{\bar{t}} x_{ijk}}{n\bar{m}\bar{t}}$$

La estimación de la varianza es:

INTER **INTRA**

$$\hat{Var}(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + \left(1 - \frac{\bar{m}}{\bar{M}}\right)\frac{s_w^2}{n\bar{m}} + \left(1 - \frac{\bar{t}}{\bar{T}}\right)\frac{s_u^2}{n\bar{m}\bar{t}}$$

S_b^2 = varianza interconglomerados de 1º etapa.

$$s_b^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (\bar{X}_i - \hat{\mu})^2$$

S_w^2 = varianza interconglomerados de 2º etapa.

$$s_w^2 = \frac{1}{n(\bar{m}-1)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \left(\bar{X}_{ij} - \bar{X}_i\right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_{c_i}^2$$

S_u^2 = varianza intraconglomerados de 3º etapa.

$$s_u^2 = \frac{1}{n\bar{m}(\bar{t}-1)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \sum_{k=1}^{\bar{t}} \left(X_{ijk} - X_{ij}\right)^2 = \frac{1}{n\bar{m}} \cdot \sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} s_{c_{ij}}^2$$

Las cuasivarianzas aumentarán cuantas más etapas se hagan, en la varianza teórica se tendrá que tener tantos términos como etapas se realicen.

El cálculo del tamaño de la muestra en forma simplificada, considerando que el número de conglomerados es N, $\left(f_1 = \frac{n}{N}\right)$, el tamaño del conglomerado promedio en 2º etapa es constante $\left(f_2 = \frac{\bar{m}}{\bar{M}}\right)$ y el tamaño del conglomerado promedio en 3º etapa es constante $\left(f_3 = \frac{\bar{t}}{\bar{T}}\right)$.

- **Fracciones de muestreo despreciables: $f_1 = f_2 = f_3 = 0$**

  La varianza teórica es: $Var(\hat{\mu}) = \frac{\sigma_b^2}{n} + \frac{\sigma_w^2}{n\bar{m}} + \frac{\sigma_u^2}{n\bar{m}\bar{t}}$.

  El coste total es: $C = c_1 \cdot n + c_2 \cdot n \cdot \bar{m} + c_3 \cdot n \cdot \bar{m} \cdot \bar{t}$.

  c_1 = coste de extraer un conglomerado en 1º etapa, c_2 = coste de extraer un conglomerado en 2º etapa y c_3 = coste de extraer un elemento en 3º etapa.

  $$\bar{t} = \sqrt{\frac{c_2 \cdot \sigma_u^2}{c_3 \cdot \sigma_w^2}} \qquad \bar{m} = \sqrt{\frac{c_1 \cdot \sigma_w^2}{c_2 \cdot \sigma_b^2}}$$

  Para **coste fijo** – Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

  $$C = c_1 \cdot n + c_2 \cdot n \cdot \bar{m} + c_3 \cdot n \cdot \bar{m} \cdot \bar{t}$$

  Para **error fijo** – Se despeja el tamaño muestral (n) de la fórmula del error (e).

$$e = k \cdot \sqrt{Var(\hat{\mu})}$$

- **Fracciones de muestreo no despreciables: $f_1 \neq 0, f_2 \neq 0, f_3 \neq 0$**

  La varianza teórica es: $Var(\hat{\mu}) = \frac{N-n}{N-1} \cdot \frac{\sigma_b^2}{n} + \frac{\bar{M}-\bar{m}}{\bar{M}-1} \cdot \frac{\sigma_w^2}{n\bar{m}} + \frac{\bar{T}-\bar{t}}{\bar{T}-1} \cdot \frac{\sigma_w^2}{n\bar{m}\bar{t}}$.

  El coste total es: $C = c_1 \cdot n + c_2 \cdot n \cdot \bar{m} + c_3 \cdot n \cdot \bar{m} \cdot \bar{t}$.

  $$\bar{t} = \sqrt{\frac{c_2}{c_3} \cdot \frac{1}{\frac{\sigma_w^2}{\sigma_u^2} - \frac{1}{\bar{T}}}} \qquad \bar{m} = \sqrt{\frac{c_1}{2} \cdot \frac{1}{\frac{\sigma_b^2}{\sigma_w^2} - \frac{1}{\bar{M}}}}$$

  Para **coste fijo** – Se despeja el tamaño muestral (n) de la fórmula del coste total (C).

  $$C = c_1 \cdot n + c_2 \cdot n \cdot \bar{m} + c_3 \cdot n \cdot \bar{m} \cdot \bar{t}$$

  Para **error fijo** – Se despeja el tamaño muestral (n) de la fórmula del error (e).

  $$e = k \cdot \sqrt{Var(\hat{\mu})}$$

Muestreo por Conglomerados combinado con Estratificado

La población está dividida en L estratos que a su vez está dividida en conglomerados y se extrae una muestra la cual se estudia todos sus elementos.

Dependiendo del tipo de estimador se tiene:

- **Estimador separado:** con lleva mayor sesgo y un tamaño de muestra para cada estrato mayor a 20.

  La media estimada es: $\hat{\mu} = \frac{1}{M} \cdot \sum_{i=1}^{L} M_i \cdot \hat{\mu}_i$, siendo: $\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} \tau_{ji}}{\sum_{j=1}^{n_i} M_{ji}}$.

  La varianza estimada es: $\hat{Var}(\hat{\mu}) = \frac{1}{M^2} \cdot \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s_{\tau_i}^2}{n_i}$, siendo:
  $s_{\tau_i}^2 = \frac{\sum_{j=1}^{n_i} \left(\tau_{ij} - \hat{\mu}_i \cdot M_{ij}\right)^2}{n_i - 1}$.

- **Estimador combinado:** tiene mayor varianza y un tamaño de muestra menor o igual a 20 o razones entre estratos iguales.

  La media estimada es: $\hat{\mu} = \frac{\bar{\tau}}{\bar{M}} = \frac{\sum_{i=1}^{L} W_i \cdot \bar{\tau}_i}{\sum_{i=1}^{L} W_i \cdot \bar{M}_i} = \frac{\sum_{i=1}^{L} N_i \cdot \bar{\tau}_i}{\sum_{i=1}^{L} N_i \cdot \bar{M}_i} = \frac{\sum_{i=1}^{L} N_i \cdot \bar{\tau}_i}{\sum_{i=1}^{L} M_i}$.

  La varianza estimada es: $\hat{Var}(\hat{\mu}) = \frac{1}{M^2} \sum_{i=1}^{L} N_i^2 \cdot \left(1 - \frac{n_i}{N_i}\right) \cdot \frac{s_{\tau_{ci}}^2}{n_i}$, siendo:
  $s_{\tau_{ci}}^2 = \frac{\sum_{j=1}^{n_i} \left[(\tau_{ij} - \bar{\tau}_i) - \hat{\mu} \cdot (M_{ij} - \bar{M}_i)\right]^2}{n_i - 1}$.

Para el cálculo del tamaño de la muestra se tiene: $\omega_i = \frac{\frac{N_i \cdot \sqrt{varianza}}{\sqrt{c_i}}}{\sum_{i=1}^{L} \frac{N_i \cdot \sqrt{varianza}}{\sqrt{c_i}}}$ y dependiendo del muestreo.

Para el **coste fijo**:

$$n = \frac{C}{\sum_{i=1}^{L} \omega_i c_i}$$

$$C = \sum_{i=1}^{L} \omega_i c_i$$

Para un **error fijo**:

$$n = \frac{\sum_{i=1}^{L} N_i^2 \cdot \frac{varianza_i}{\omega_i}}{M^2 \frac{e^2}{k^2} + \sum_{i=1}^{L} N_i \cdot varianza_i}$$

Se redondean n_i según coste o error: $n_i = n \cdot \omega_i$

Muestreo por Conglomerados en dos etapas combinado con Estratificado

La población se divide en L estratos, cada estrato se divide en N_i conglomerados. Este muestreo, como su nombre indica, se divide en dos etapas: en la primera etapa, se realiza una muestra de n_i conglomerados y, en la segunda etapa se realiza una muestra de m_i elementos.

La notación que se usa en esta técnica es la siguiente:

h: contador de estratos, $h = 1, \dots, L$.

i: contador de conglomerados en primera etapa, $i = 1, \dots, n_h$.

j: contador de elementos en segunda etapa, $j = 1, \dots, m_i$.

N_h: número de conglomerados en estrato.

$N = \sum_{h=1}^{L} N_h$: número de conglomerados totales.

n_h: número de conglomerados en la muestra en estrato.

$n = \sum_{h=1}^{L} n_h$: número de elementos en la muestra totales.

M_h: número de individuos en el estrato.

$M = \sum_{h=1}^{L} M_h$: número de individuos en la población.

m_h: número de individuos en la muestra en el estrato.

$m = \sum_{h=1}^{L} m_h$: número de individuos en la muestra.

M_{hi}: número de individuos en el estrato h del conglomerado i. ($M = \sum_{i=1}^{n_h} M_{hi}$)

m_{hi} = número de individuos en la muestra en el estrato h del conglomerado i. ($m = \sum_{h=1}^{n_h} m_{hi}$)

$f_{h_1} = \frac{n_h}{N_h}$: fracción de muestreo en primera etapa en el estrato h.

$f_{h_2} = \frac{\bar{m}_h}{\bar{M}_h}$: fracción de muestreo en segunda etapa en el estrato h.

$\bar{M}_h$: tamaño del conglomerado promedio en estrato h.

$\bar{m}_h$: tamaño de muestra promedio en el estrato h.

$W_h = \frac{M_h}{M} = \frac{M_h}{\sum_{h=1}^{L} M_h} = \frac{N_h \cdot \bar{M}_h}{\sum_{h=1}^{L} N_h \cdot \bar{M}_h} = \frac{N_h \cdot \bar{M}_h}{N \cdot \bar{M}}$: proporción del tamaño del estrato h con respecto al total. Se puede estimar la $\bar{M}_h$de la siguientes formas: $\hat{\bar{M}}_h = \frac{\sum_{i=1}^{n_h} M_{hi}}{n_h}$ y $\bar{M}_h = \hat{\bar{M}}_h \cdot N_h$.

Para estimar los parámetros, se debe tener en cuenta si se conoce o no la M y M_h, así será un estimador insesgado o no.

- **Estimador insesgado conociendo M y M_h**

  La media estimada: $\hat{\mu} = \sum_{h=1}^{L} W_h \cdot \hat{\mu}_h$, siendo: $\hat{\mu}_h = \frac{N_h}{M_h} \cdot \frac{\sum_{i=1}^{n_h} M_{hi}\bar{X}_{hi}}{n_h}$.

  La varianza estimada: $\hat{Var}(\hat{\mu}_h) = \left(1 - \frac{n_h}{N_h}\right)\frac{1}{\bar{M}_h^2}\frac{S_{b_h}^2}{n_h} + \frac{1}{n_h N_h \bar{M}_h^2}\sum_{i=1}^{n_h} M_{hi}^2\left(1 - \frac{m_{hi}}{M_{hi}}\right)\frac{s_{c_{hi}}^2}{m_{hi}}$,

  siendo: $s_{b_h}^2 = \frac{\sum_{i=1}^{n_h}(M_{hi}\bar{X}_{hi} - \bar{M}_h\hat{\mu}_h)^2}{n_h - 1}$.

- **Estimador de razón desconociendo M y M_h**

  La media estimada: $\hat{\mu}_r = \sum_{h=1}^{L} W_h \cdot \hat{\mu}_{r_h}$, siendo: $\hat{\mu}_{r_h} = \frac{\sum_{i=1}^{n_h} M_{hi}\bar{X}_{hi}}{\sum_{i=1}^{n_h} M_{hi}}$.

  La varianza estimada: $\hat{Var}(\hat{\mu}_{r_h}) = \sum_{h=1}^{L} W_h^2 \cdot \hat{Var}(\hat{\mu}_{r_h})$, siendo:

$$\hat{Var}(\hat{\mu}_{r_h}) = \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{1}{\overline{M_h}^2}\frac{s_{r_h}^2}{n_h} + \frac{1}{n_h N_h \overline{M_h}^2} \cdot \sum_{i=1}^{n_h} M_{hi}^2 \cdot \left(1 - \frac{m_{hi}}{M_{hi}}\right) \cdot \frac{s_{c_{hi}}^2}{m_{hi}}$$

$$s_{r_h}^2 = \frac{\sum_{i=1}^{n} M_{hi}^2\left(\bar{X}_{hi} - \hat{\mu}_{r_h}\right)^2}{n_h - 1}$$

Capítulo 3

Material y Métodos

Se procedió a buscar una base de datos óptima para la aplicación de todos los muestreos requeridos. En este contexto, el tema de la base de datos se consideró secundario, ya que el objetivo principal del trabajo era comparar distintos métodos de muestreo probabilístico. Por lo tanto, se buscó una base de datos con una amplia gama de variables que pudieran utilizarse para formar estratos, conglomerados, etc., incluyendo variables cuantitativas que permitieran calcular estimaciones y comparar resultados.

Una vez identificadas las características básicas requeridas para la base de datos, se realizó una búsqueda en diversas plataformas para encontrar una extensa, es decir, con el mayor número posible de individuos. Se seleccionaron tres opciones encontradas en la página del Instituto Nacional de Estadística (INE).

Con relación a los temas más destacados, se consideró la inserción laboral de titulados universitarios y las migraciones exteriores. Para el primer tema, se evaluaron dos bases: "Graduados universitarios por titulación y sexo" y "Tasas de actividad, empleo y desempleo de los graduados universitarios por comunidad autónoma de su universidad y área de estudio". Para el segundo tema, se evaluó "Flujo de inmigración procedente del extranjero por comunidad autónoma, año, sexo, grupo de edad y país de nacimiento (España/Extranjero)".

La primera base de datos no fue seleccionada debido a que, aunque contaba con un gran número de individuos, disponía de pocas variables para llevar a cabo los muestreos.

La segunda base de datos se consideró como una de las opciones finales, ya que cumplía con los requisitos necesarios: podía ser estratificada (tasas de actividad, empleo y desempleo) y conglomerada (área de estudio o comunidad autónoma de la universidad), además de contar con un gran número de individuos.

La última base de datos también cumplía con los requisitos necesarios y, como un valor adicional, incluía datos desde 2008 hasta 2021, lo que la convertía en la opción óptima entre las tres bases de datos.

Continuando con la descripción de la base de datos seleccionada, se identificaron las siguientes variables:

- **Comunidad y ciudades autónomas**: La comunidad o la ciudad autónoma, a la que pertenece el individuo, se usan 19 opciones, las 17 comunidades autónomas y las 2 ciudades autónomas.
- **Sexo**: género del individuo, con dos opciones: hombre y mujer.

- **País de nacimiento**: país de nacimiento del individuo, con dos opciones: España y extranjero.
- **Grupo quinquenal de edad**: edad del individuo, dividido en 19 intervalos de edad, desde 0 a 4 años hasta 90 años o más.
- **Periodo**: año de los datos, desde 2008 hasta 2021.
- **Flujo migratorio**: cantidad de personas que se trasladan de un área geográfica a otra en un período de tiempo específico, ya sea de manera temporal o permanente. La variable cuantitativa necesaria para la estimación de parámetros.

Como se puede observar, se dispone de suficientes variables para la formación de estratos y conglomerados.

Se procedió con la limpieza y ordenación de los datos, uniéndolos desde el año 2011 hasta 2021, lo que resultó en un total de 15884 individuos. Se añadió un identificador, y las variables se organizaron en el siguiente orden: periodo, país, intervalo de edad, sexo, comunidad autónoma y flujo. Todo esto se llevó a cabo utilizando el programa Microsoft Excel.

En principio, se planificó realizar los siguientes muestreos probabilísticos: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo sistemático, muestreo por conglomerados y muestreos polietápicos, incluyendo muestreo por conglomerados en dos etapas, además del muestreo por conglomerados combinado con estratificado. Tras analizar más detenidamente la base de datos, se identificó la posibilidad de ampliar los muestreos añadiendo el muestreo por conglomerados en tres etapas y el muestreo por conglomerados en dos etapas combinado con el estratificado. También se observó que el muestreo sistemático no resultaba factible, por lo que se eliminó de la lista de técnicas.

Una vez determinados los muestreos que se van a realizar, se procedió a considerar las variables que se utilizarían en cada tipo de muestreo, surgieron las siguientes ideas al respecto:

En el **muestreo aleatorio simple,** se estima la media del flujo de migración.

En el **muestreo aleatorio estratificado**, se utiliza las categorías de la variable país como estratos para estimar la media del flujo de migración.

Para el **muestreo por conglomerados**, se emplea los 19 conglomerados definidos por la variable intervalo de edad para estimar la media del flujo de migración.

En el **muestreo por conglomerados en dos etapas**, se forma conglomerados utilizando las variables edad y comunidad autónoma. En la primera etapa, se agrupa por intervalos de edad y en la segunda etapa por comunidad autónoma, para luego estimar la media del flujo de migración.

En el **muestreo por conglomerados en tres etapas**, se forma conglomerados utilizando las variables edad y comunidad autónoma. Se realiza conglomerados primero por intervalo de edad y luego por comunidad autónoma. Se selecciona una muestra aleatoria simple dentro de cada comunidad autónoma para estimar la media del flujo de migración.

El **muestreo por conglomerados combinado con estratificado** emplea las variables país y edad. Los países se utilizan como estratos, y dentro de cada estrato se seleccionan conglomerados basados en el intervalo de edad. Se analiza toda la muestra para estimar la media del flujo de migración.

En el **muestreo por conglomerados en dos etapas combinado con estratificado**, se emplea las variables país, edad y comunidad autónoma. Los países se utilizan como estratos, y dentro de cada estrato se selecciona conglomerados basados en el intervalo de edad. Se utiliza también los conglomerados de la comunidad autónoma para estimar la media del flujo de migración.

Se realizará el cálculo de la media y varianza poblacional, así como una prueba piloto en cada muestreo para determinar el tamaño muestral, utilizando un error relativo del 15%. Todos estos cálculos se llevarán a cabo utilizando el programa Microsoft Excel. Además, se utiliza el programa R-studio para ayudar en la selección del tamaño muestral de manera aleatoria, ya que en algunos casos puede superar los 1000 individuos.

Entre los recursos que resultaron de utilidad para el desarrollo del trabajo, destaca la página web donde se encontró la base de datos, el INE.

El Instituto Nacional de Estadística es una institución gubernamental que se encarga de recolectar, procesar, examinar y difundir datos estadísticos oficiales en el país. Su principal misión es suministrar información precisa y fidedigna sobre diversos aspectos de la sociedad, como la demografía, la actividad económica, el mercado laboral, la educación, la salud, entre otros.

Obtienen datos a través de censos, encuestas y otras fuentes de información, y emplean métodos estadísticos para analizar y presentar estos datos de forma clara y útil para diversos usuarios, como funcionarios públicos, investigadores, empresas y el público en general.

La información proporcionada por el INE es esencial para la elaboración de políticas públicas, la toma de decisiones empresariales, la investigación académica y la evaluación de indicadores clave de progreso social y económico en el país.

En cuanto a los cálculos realizados en la parte práctica, se hicieron con Microsoft Excel y R-Studio.

Excel es una aplicación de hoja de cálculo desarrollada por Microsoft, es una de las herramientas más utilizadas para el análisis de datos, la creación de informes, la realización de cálculos, la elaboración de gráficos y la gestión de datos en forma de hojas de cálculo. Excel es conocido por su interfaz intuitiva basada en celdas, lo que lo hace accesible incluso para aquellos que no son expertos en programación o análisis de datos.

Algunas de las características clave de Microsoft Excel incluyen: cálculos y fórmulas, gráficos y visualizaciones, análisis de datos, automatización y programación (se pueden automatizar tareas repetitivas utilizando macros y programación en VBA) e integración con otras aplicaciones de Microsoft Office, como Word y PowerPoint.

R-Studio es una plataforma integral diseñada para facilitar el desarrollo de proyectos en el ámbito del lenguaje de programación R, reconocido por su amplia aplicación en el campo de la estadística y el análisis de datos. Gracias a su extenso conjunto de paquetes y su capacidad para manipular y representar datos de manera eficiente, R se ha convertido en una herramienta imprescindible en diversos ámbitos profesionales.

Este entorno de desarrollo proporciona una interfaz amigable que simplifica la tarea de escribir, depurar y ejecutar código en R, al mismo tiempo que permite la visualización de resultados a través de gráficos y tablas dinámicas. Además, R-Studio ofrece una serie de herramientas integradas para la gestión de proyectos, lo que facilita la organización y seguimiento de cada etapa del proceso de desarrollo.

Una de las características más destacadas de R-Studio es su capacidad para integrarse con otros lenguajes de programación, lo que permite a los usuarios combinar diferentes herramientas y recursos según las necesidades específicas de cada proyecto. Esta versatilidad, junto con su funcionalidad para facilitar la colaboración en equipo, convierte a R-Studio en una opción preferida por científicos de datos, analistas y estadísticos que buscan optimizar su flujo de trabajo y obtener resultados precisos y confiables. Además, R-Studio es libre y gratuito.

Capítulo 4

Caso Práctico

A continuación, se explica cada caso práctico realizado, el cual se compone de una prueba piloto para estimar el tamaño muestral óptimo y el muestreo final para estimar la media y error de cada uno, esto se hace para obtener comparaciones y conclusiones.

Cabe destacar que en cada muestreo se usa un error relativo del 15% y se quiere estimar la media en cada método, en algunos casos se usaron costes para hallar el tamaño muestral.

En el archivo de Excel llamado: “Base de datos”, en la primera hoja está la población de estudio del trabajo, ya limpia y codificada para mayor facilidad, en la siguiente imagen podemos ver las distintas variables, número de individuos en cada uno y totales.

PERIODO	TOTAL
2021	1444
2020	1444
2019	1444
2018	1444
2017	1444
2016	1444
2015	1444
2014	1444
2013	1444
2012	1444
2011	1444

PAIS	TOTAL
España	7942
Extranjero	7942

SEXO	TOTAL
Hombre	7942
Mujer	7942

INTERVALO DE EDAD	TOTAL
De 0 a 4 años	836
De 5 a 9 años	836
De 10 a 14 años	836
De 15 a 19 años	836
De 20 a 24 años	836
De 25 a 29 años	836
De 30 a 34 años	836
De 35 a 39 años	836
De 40 a 44 años	836
De 45 a 49 años	836
De 50 a 54 años	836
De 55 a 59 años	836
De 60 a 64 años	836
De 65 a 69 años	836
De 70 a 74 años	836
De 75 a 79 años	836
De 80 a 84 años	836
De 85 a 89 años	836
90 y más años	836

CCAA	TOTAL
Andalucía	836
Aragón	836
Principado de Asturias	836
Islas Baleares	836
Canarias	836
Cantabria	836
Castilla y León	836
Castilla - La Mancha	836
Cataluña	836
Comunitat Valenciana	836
Extremadura	836
Galicia	836
Comunidad de Madrid	836
Región de Murcia	836
Comunidad Foral de Navarra	836
País Vasco	836
La Rioja	836
Ceuta	836
Melilla	836

Ilustración 5: Distintas variables de la base de datos

Se puede observar claramente que la base tiene simetría en cada variable, esto se debe a que no está precisamente hecha para realizar muestreos en ella. En la segunda hoja, se pueden ver los cálculos necesarios para la media y varianza poblacional, con esto se tiene:

N	15884
μ	312,0181
σ^2	791832,8804

Ilustración 6: Número de individuos, media y varianza poblacional de la base de datos

Todo esto se puede observar mejor en el anexo 1. A continuación, se explicará al detalle cómo se realizó los distintos tipos de muestreo y los resultados obtenidos.

Muestreo Aleatorio Simple

En el archivo: "Muestreo Aleatorio Simple", en la primera hoja está la prueba piloto y en la segunda, el muestreo aleatorio simple definitivo con los cálculos necesarios para la estimación de la media.

Para llevar a cabo la prueba piloto, se determinó un tamaño muestral inicial aleatorio de 60 individuos. Este tamaño se seleccionó para obtener una muestra preliminar que permitiera calcular estimaciones iniciales necesarias para determinar el tamaño muestral óptimo para estudios posteriores.

Utilizando la función "ALEATORIO.ENTRE" en la hoja de cálculo, se seleccionaron aleatoriamente 60 individuos de la base de datos.

Una vez seleccionada la muestra aleatoria, se procedió a calcular dos estadísticos fundamentes: la media y cuasivarianza de la variable de interés en la muestra, en este caso el flujo migratorio. Para ello se usó la función "PROMEDIO" y "VAR.S. Estos cálculos se ven reflejados en el anexo 2.

N	15884
n	60
$\hat{\mu} = \bar{X}$	298,240476
s_c^2	475679,082
e_r	0,15
n_∞	950,732541
n	897,093716
n	898

$$e = e_r \cdot |\hat{\mu}|$$

e	44,7361

$$n_\infty = \frac{k^2 \cdot \sigma^2}{e^2}$$

$$n = \frac{n_\infty}{1 + \frac{n_\infty - 1}{N}}$$

Ilustración 7:Estimación del tamaño muestral del MAS

Con los cálculos hechos en la prueba piloto, se obtuvo el tamaño muestral de 898. En el muestreo final se realizó la extracción de los individuos con la misma función, con esto se calculó la media estimada y por consiguiente su error.

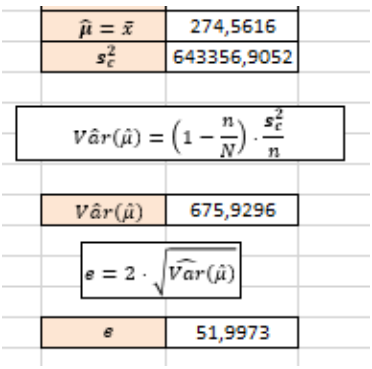


Ilustración 8: Estimación de la media y error en el MAS

Como se observa, se obtiene una media estimada de 274’5616 individuos por año y un error absoluto de 51’9973 individuos por año (un error relativo de casi el 19%).

En este caso, el error es superior al fijado inicialmente. Esto se puede deber a que la muestra seleccionada es muy diversa y este tipo de muestreo no capta toda la variabilidad presente, también se puede deber a la alta varianza de la variable de estudio.

Todo esto se observa mejor en el anexo 3.

Muestreo Aleatorio Estratificado

En el archivo: “Muestreo Aleatorio Estratificado”, en la primera hoja esta la prueba piloto y en la segunda se encuentra el muestreo final con la estimación de la media y error.

En este caso, como estrato, se considera la variable “País”, la cual se divide en dos tipos: “España” y “Extranjero”, también se considera la afijación óptima con error fijo y coste constante.

En la prueba piloto, se eligieron al azar los tamaños muestrales de cada estrato. Se puede observar que son iguales, pero es por coincidencia. Con ello, se halló el tamaño muestral.

Todos estos cálculos se pueden ver reflejados en el anexo 4.

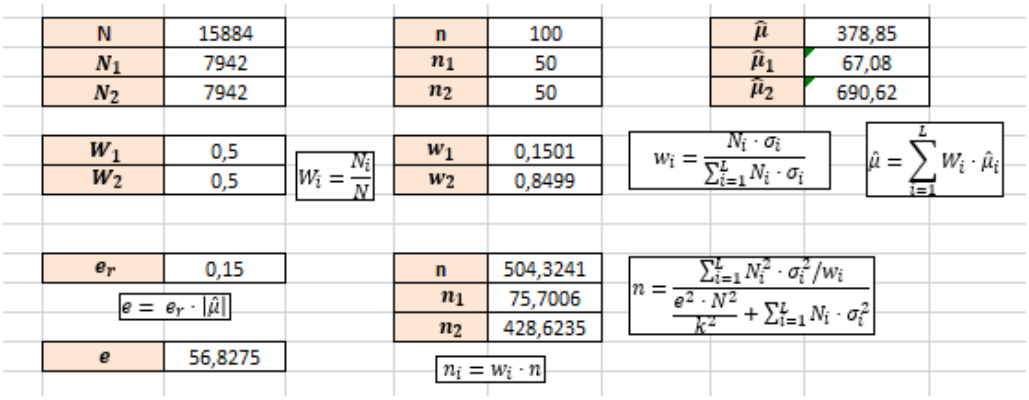


Ilustración 9: Estimación del tamaño muestral del MAE

Con todo calculado se obtiene que el tamaño muestral es 504’3241 individuos, en el primer estrato 75’7006 y en el segundo 428’6235, se redondea al mayor buscando el menor error, para ella se hacen varios casos.

CASO	ERROR
n = 503, n_1 = 75, n_2 = 428	e = 56’9062
n = 505, n_1 = 76, n_2 = 429	e = 56’7876
n = 506, n_1 = 76, n_2 = 430	e = 56’7286

Se puede ver que el error disminuye cuanto más individuos se coge del primer estrato, con esto visto se elige el tercer caso que tiene el menor error. A continuación, se usó un código de R para agilizar la recogida de individuos de la muestra (anexo 5).

En el muestreo final, se obtiene los siguientes resultados:

N	15884		n	505		$\hat{\mu}$	346,3659
N_1	7942		n_1	76		$\hat{\mu}_1$	54,0526
N_2	7942		n_2	430		$\hat{\mu}_2$	638,6791
W_1	0,5	$W_i = \frac{N_i}{N}$	w_1	0,1505	$w_i = \frac{n_i}{n}$	$\hat{\mu} = \sum_{i=1}^{L} W_i \cdot \hat{\mu}_i$	
W_2	0,5		w_2	0,8515			
$\hat{Var}(\hat{\mu})$	822,7586	$Var(\hat{\mu}) = \sum_{i=1}^{L} W_i^2 \cdot \frac{N_i - n_i}{N_i - 1} \cdot \frac{\sigma_i^2}{n_i}$					
e	57,3675	$e = 2 \cdot \sqrt{\widehat{Var}(\hat{\mu})}$					

Ilustración 10: Estimación de la media y error en el MAE

Se observa que se estudia una población de 505 individuos, obteniendo una media de 346’3659 individuos por año y un error absoluto de 57’3675 individuos por año (un error relativo casi del 17%).

Vuelve a ser un error relativo mayor, se puede deber a que hay un número insuficiente de estratos, la mala proporción de la muestra entre. Todos los cálculos se pueden ver en el anexo 6.

Muestreo por Conglomerados

En el archivo: “Muestreo por Conglomerados”, en la primera hoja esta la prueba piloto y en la segunda, se encuentra el muestreo definitivo con los cálculos requeridos para la estimación.

Para este caso, se usó la variable “Intervalo de edad” como la variable con conglomerados, teniendo 19 intervalos de edad, es decir, 19 conglomerados.

En la prueba piloto, se optó por elegir 4 conglomerados al azar y seleccionar todos los individuos de esos conglomerados. Se estudió una población de 3344 individuos. Esto se ve reflejado en el anexo 7.

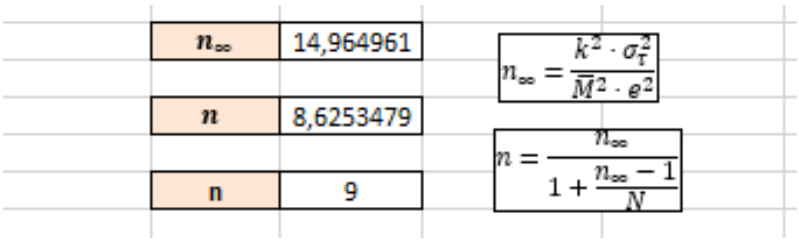


Ilustración 11: Estimación del tamaño muestral del MpC

Al terminar los cálculos, en el muestreo final, se eligen 9 conglomerados aleatorios estudiando 7524 individuos.

Se obtiene la media es 274’8993 individuos por año y un error absoluto de 110’7161 individuos por año (un error relativo de casi 40%).

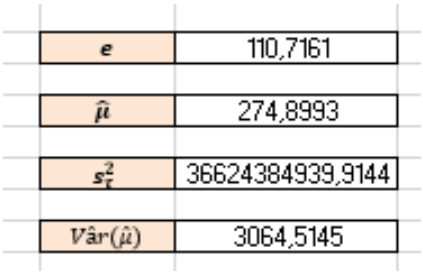


Ilustración 12: Estimación de media y error del MpC

Se puede observar que tenemos un error relativo mucho mayor que en los otros casos, esto puede deberse a la homogeneidad entre conglomerados. Si son demasiados similares entre sí la variabilidad será baja y puede no capturarla adecuadamente de la población total.

Los cálculos se reflejan en el anexo 8.

Muestreo por Conglomerados en dos etapas

En el archivo: “Muestreo por Conglomerado en 2 etapas”, en la primera hoja esta la prueba piloto y en la segunda se encuentra el muestreo final con la estimación de la media y error.

En este caso, también se seleccionarán como conglomerados los rangos de la variable 'Intervalo de edad'. En este muestreo existen dos etapas. En la primera, se eligen los conglomerados al azar que se van a estudiar y, en la segunda etapa, en cada conglomerado se coge una muestra.

En esta instancia se eligieron 4 conglomerados y 230 individuos aleatorios en cada conglomerado. Se estudian 920 individuos. También se usó el programa R con un código para la elección aleatoria de los individuos de cada muestra (anexo 8).

Para determinar el tamaño muestral en este caso, se tomó en consideración el costo asociado a la extracción de los datos. Se establecieron, dos tipos de costos específicos: el costo de extracción de un conglomerado, extraer datos de un conglomerado completo tiene un costo fijo de 100000€, y el costo de extracción de un individuo, extraer datos de cada individuo dentro de ese conglomerado tiene un costo adicional de 20€ por individuo. Se observa que el coste de extracción de un conglomerado es mayor al coste de la extracción del individuo, esto se debe a que con los conglomerados es más caro ya que son varios individuos y están muy dispersos, en cambio, al tener ya el conglomerado a la búsqueda de los individuos es más rápida y sencilla.

Se observa que las fracciones de muestreo son no despreciables, esto significa que el tamaño de la muestra seleccionada representa una proporción significativa en la población total, con una M conocida, por lo tanto, se usa los cálculos para un estimador insesgado. Como en los otros casos se usa un código de R para agilizar y facilitar las recogidas de individuos (anexo 9).

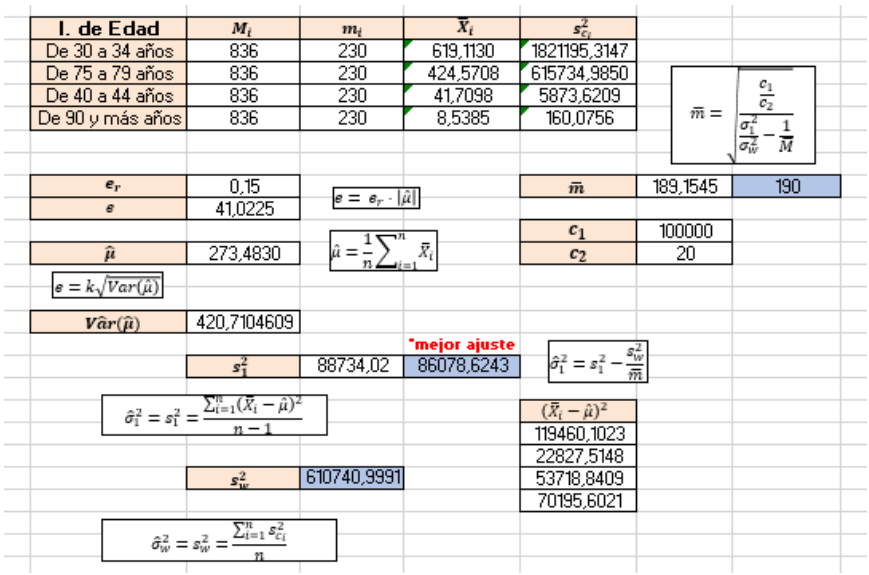
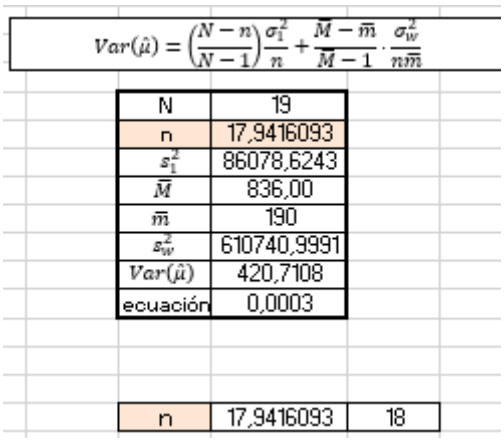


Ilustración 13: Estimación de la media de individuos por conglomerados

Con la función “Solver” de Excel se calcula el tamaño muestral, en este caso el número de conglomerados que se elijan.



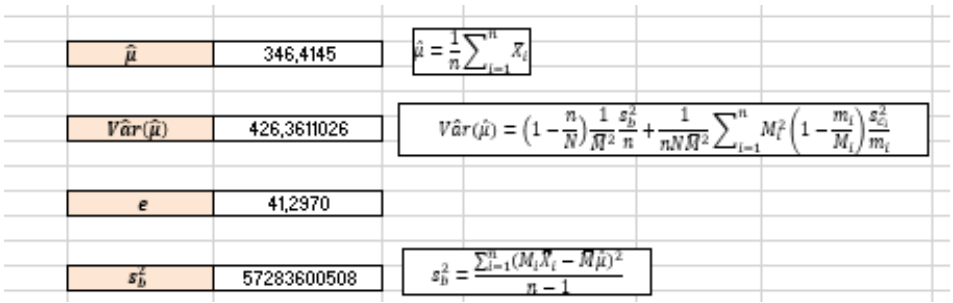
$$Var(\hat{\mu}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma_1^2}{n} + \frac{\overline{M}-\overline{m}}{\overline{M}-1}\cdot\frac{\sigma_w^2}{n\overline{m}}$$

N	19	
n	17,9416093	
s_1^2	86078,6243	
$\overline{M}$	836,00	
$\overline{m}$	190	
s_w^2	610740,9991	
$Var(\hat{\mu})$	420,7108	
ecuación	0,0003	
n	17,9416093	18

Ilustración 14: Estimación del tamaño muestral del MpC en 2 etapas

Se obtiene un tamaño de 18, es decir, estudiamos 18 conglomerados al azar y una muestra de 190 en cada conglomerado. Será una población de 3420 individuos. Esto se refleja en el anexo 10.

En el muestreo final, se observa:



$\hat{\mu}$	346,4145	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i$
$\hat{Var}(\hat{\mu})$	426,3611026	$\hat{Var}(\hat{\mu}) = \left(1-\frac{n}{N}\right)\frac{1}{\overline{M}^2}\frac{s_b^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{ci}^2}{m_i}$
e	41,2970	
s_b^2	57283600508	$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\overline{x}_i - \overline{M}\hat{\mu})^2}{n-1}$

Ilustración 15:Estimación de la media y error del MpC en 2 etapas

La media estimada es de 246’4145 individuos por año y su error absoluto es 41’2970 (un error relativo de casi el 12%). Se ve una gran disminución del error relativo. Se puede deber a que se estudiaron la mayoría de conglomerados.

Todos los cálculos y tablas se ven reflejadas en el anexo 11.

Muestreo por Conglomerados en tres etapas

En el archivo: “Muestreo por Conglomerado en 3 etapas”, en la primera hoja esta la prueba piloto y en la segunda se encuentra el muestreo final con las estimaciones.

Este tipo de muestreo tiene tres etapas en las cuales se necesitan dos variables divididas en conglomerados. En la primera etapa, se eligen al azar los conglomerados de la variable “Intervalo de edad”, en la segunda, dentro de cada conglomerado, se eligen al azar de la variable “Comunidad y ciudad autónoma” y, en la tercera, se elige una muestra aleatoria de elementos.

En la prueba piloto, se optó por coger 9 intervalos de edad, dentro de cada uno se eligieron 2 CCAA y en cada uno de ellos 21 individuos. Se estudia una población de 378 individuos.

Para la recogida de individuos, se usó un código de R (anexo 12).

Se optó por incluir costos específicos al calcular el tamaño muestral, los cuales se detallan de la siguiente manera: el costo de extracción de un conglomerado asciende a 100000 €, el costo de extracción de conglomerados en la segunda etapa se establece en 1000 € y el costo de extracción de un individuo está fijado en 20 €. Se puede ver una disminución del coste en cuanto a las etapas. Es normal ya que resulta más sencillo y barato si se tiene ya los conglomerados extraídos.

Se observa que las fracciones de muestreo son no despreciables con una M conocida, por lo tanto, se usan las expresiones para un estimador insesgado. Como los cálculos y tablas son más complejos y grandes, en el anexo 13 se pueden observar mejor.

Se puede ver que se tienen que seleccionar 18 conglomerados al azar en cada uno se cogen 10 conglomerados aleatorios y dentro de cada uno 14 individuos. Se estudia una población de 2520 individuos.

$$Var(\hat{\mu}) = \frac{N-n}{N-1}\cdot\frac{\sigma_b^2}{n} + \frac{\overline{M}-\overline{m}}{\overline{M}-1}\cdot\frac{\sigma_w^2}{n\overline{m}} + \frac{\overline{T}-\overline{t}}{\overline{T}-1}\cdot\frac{\sigma_u^2}{n\overline{m}\overline{t}}$$

n	17,5500
N	19
σ_b^2	4798,2592
$\overline{m}$	10
$\overline{M}$	19
σ_w^2	4294,0043
σ_u^2	13967,8138
$\overline{t}$	14
$\overline{T}$	44
$Var(\hat{\mu})$	38,2238
ecuacion	0,00
n	18

Ilustración 16: Estimación del tamaño muestral del MpC en 3 etapas

En el muestreo final, se puede ver que la media estimada es de 323,4158 individuos por año y su error absoluto es de 66,8921 individuos por año (error relativo de casi 21%).

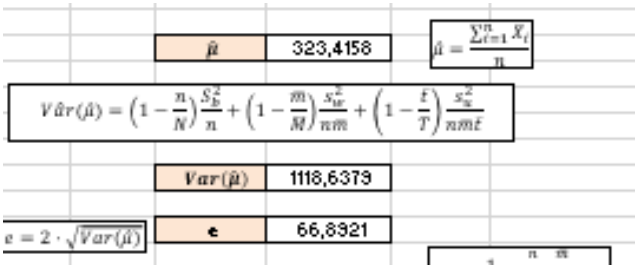


Ilustración 17: Estimación de la media y error del MpC en 3 etapas.

En este caso, se aumentó el error relativo. Se puede deber a alguna mala elección de conglomerados en alguna etapa. También en que este método no sea el adecuado para esta base de datos.

Todos las tablas y cálculos del muestreo final se ven en el anexo 14.

Muestreo por Conglomerados combinado con Estratificado

En el archivo: "Muestreo por Conglomerado combinado con Estratificado", en la primera hoja, está la prueba piloto y, en la segunda, se encuentra el muestreo final con las estimaciones.

En este caso, primero la población está divida en dos estratos, "España" y "Extranjero"; dentro de cada estrato hay 19 conglomerados, referente a la variable "Intervalo de Edad", de los cuales elegiremos unos al azar y estudiaremos todos los individuos de cada uno.

En la prueba piloto, se seleccionó del primer estrato 6 y del segundo 3 conglomerados. En total se estudian 3762 individuos.

Se tomó la decisión de considerar costes especificados al calcular el tamaño muestral para garantizar precisión de las estimaciones con un margen de error mínimo. Los costes establecidos son: coste de extracción de un conglomerado se fijó en 800 € y el coste de extracción de un individuo se estableció en 600 €. En este caso, los costes varían más que en los anteriores muestreos. Esto se debe a que se ha dividido la población en estratos y resulta más económico la extracción de conglomerados en esta forma. Aunque la recolección de individuos resulta casi igual de costosa ya que están más dispersos.

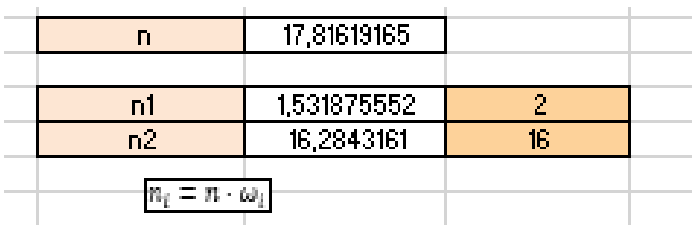


Ilustración 18: Estimación del tamaño muestral del MpC combinado con Estratificado

Se observa que el tamaño muestral es de 17'8162; en el primer estrato, se cogen 1'5319 y, en el segundo, 16'2843 conglomerados. Con esto hallado se busca el caso con el menor error posible. Para ello, se calcula, para cada caso, su error y se selecciona el mínimo.

CASO	ERROR
$n = 18, n_1 = 2, n_2 = 16$	e = 28,4905
$n = 19, n_1 = 3, n_2 = 16$	e = 26,5768

El caso optimo es coger 3 conglomerados en el primero y 16 en el segundo estrato. Se tiene una población de 7942 individuos. Se puede ver que aumentando el número de conglomerados en el estrato 1, se disminuye el error. Estos cálculos se reflejan en el anexo 15.

En el muestreo final, se tiene las siguientes estimaciones:

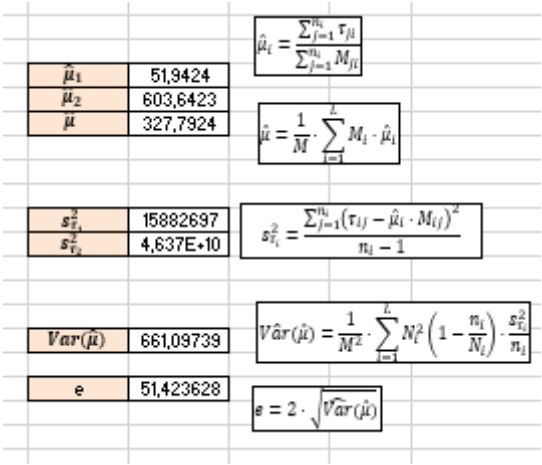


Ilustración 19:Estimación de la media y error en MpC combinado con Estratificado

Se ve que la media total estimada es de 327’7924 individuos por año y el error absoluto es de 51’4236 individuos por año (error relativo de casi 16%). También se observa una media estimada en el estrato “España” de 51’9424 individuos por año y en el estrato “Extranjero” de 603’6423 individuos por año.

En este caso, se puede ver, en comparación a los anteriores, que el error relativo ha disminuido, aunque respecto al error inicial tiene cierto aumento.

Todos estos cálculos y tablas se ven en el anexo 16.

Muestreo por Conglomerados en dos etapas combinado con Estratificado

En el archivo: “Muestreo por Conglomerado en 2 etapas combinado con Estratificado”, en la primera hoja, está la prueba piloto y, en la segunda, se encuentra el muestreo final con las estimaciones.

En este caso, se cogieron los dos estratos de la variable “País”, dentro de cada estrato se realizó un muestreo por conglomerado en 2 etapas; es decir, se eligieron al azar los conglomerados a estudio en cada estrato y luego una muestra en cada uno.

En la prueba piloto, en el primer estrato se eligieron 4 y en el segundo 5. Dentro de cada uno, se cogió una muestra de 169 y 65, teniendo así una población de 1001 individuos. Para facilitar la recogida de individuos de las muestras, se usó un código de R (anexo 17).

Se pueden ver que los cálculos de fracciones muestrales resultan despreciables y se obtiene un tamaño muestral total de 35. Todo esto se refleja en el anexo 18.

	ESPAÑA	EXTRANJERO
n	18,9387109	15,95718236
$\hat{\sigma}^2_{1_h}$	33394,3873	62084,7864
$s^2_{w_h}$	10645,3583	289988,2437
$\overline{m}$	40	154
$\overline{M}_h$	7942	7942
N_h	19	19
Var	19,98727026	773,4391077
ecuacion	-0,02	-0,03
n	19	16

Ilustración 20:Estimación del tamaño muestral en MpC en 2 etapas combinado con Estratificado

Se observa que, en el primer estrato, se deben estudiar todos los conglomerados y en el segundo 16. También se eligen 40 y 154 individuos, la población de estudio es de 3224 individuos.

En el muestreo final, se obtiene una media total de 279’8904 individuos por año y un error absoluto de 49’7098 individuos por año (error relativo de casi 18%).

Se observa que el error relativo es mayor que el inicial, pero en comparación con otros casos es menor.

Todos los cálculos se ven reflejados en el anexo 19.

Los resultados de todos los muestreos se ven en la siguiente tabla:

MUESTREO	MEDIA	ERROR RELATIVO
Muestreo Aleatorio Simple	274’5616	18’9383%
Muestreo Aleatorio Estratificado	346’3659	16’5627%
Muestreo por Conglomerados	274’8993	40’2751%
Muestreo por Conglomerados en 2 etapas	246’4145	11’9213%
Muestreo por Conglomerados en 3 etapas	323’4158	20’6830%
Muestreo por Conglomerados combinado con Estratificado	327’7924	15’6878%
Muestreo por Conglomerados en 2 etapas combinado con Estratificado	279’8904	17’7605%

Como se puede ver, el menor error reside en el muestreo por conglomerados en dos etapas, esto se puede deber a que se estudia una muestra de casi todos los conglomerados.

Se pueden agrupar los muestreos por sus errores relativos. Se tiene un error menor al 15%: el muestreo por conglomerado en dos etapas; entre un 15% y 20%: el muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerado combinado con estratificado y muestreo por conglomerados en dos etapas combinado con estratificado; y, por último, mayores al 20%: muestreo por conglomerados y muestreo por conglomerados en tres etapas.

Conclusiones

Durante este trabajo, se han evaluado los diferentes tipos de muestreo aleatorio, destacando tanto sus ventajas como sus desventajas. A través de varios casos prácticos, se ha buscado confirmar estas características de manera práctica.

Se concluye que no existe un método óptimo definitivo para esta tarea debido a la presencia inevitable de ciertos sesgos. Estos pueden surgir por la elección de la población de estudio, el tipo de muestreo utilizado o las decisiones tomadas por el investigador.

Los casos prácticos realizados han subrayado la importancia crítica de seleccionar el método de muestreo más adecuado para cada base de datos específica, especialmente en aquellos casos donde el error aumenta considerablemente.

Durante el trabajo, se han comentado los posibles sesgos en cada método:

1. Muestreo aleatorio simple: El error es superior al fijado inicialmente, lo que puede deberse a que la muestra seleccionada es muy diversa y este tipo de muestreo no capta toda la variabilidad presente. Este error se podría disminuir aumentando el tamaño muestral o buscando las fuentes de variabilidad.

2. Muestreo aleatorio estratificado y sus derivados: El error puede residir en un número insuficiente de estratos o en una mala proporción de la muestra entre estratos.

3. Muestreo por conglomerado: Tiene el mayor error de todos, puede deberse al tamaño de los conglomerados elegidos sea demasiado grandes, puede haber una falta de representación de la variabilidad intraconglomerado.

4. Muestreo por conglomerados en varias etapas: Los errores pueden deberse a una mala elección de conglomerados en alguna etapa.

Algunos sesgos pueden originarse por la manipulación involuntaria de los datos o las muestras por parte de los investigadores. A veces, la falta o excesiva atención puede eliminar la aleatoriedad de las muestras.

Se ha observado que el método más adecuado para esta base de datos y la estimación de la media es el muestreo por conglomerados en dos etapas. Los métodos que presentan errores entre el 15% y el 20% podrían ser ajustados para mejorar la precisión, mientras que aquellos con errores superiores al 20% deberían descartarse para esta base de datos específica.

Bibliografía

Artículos

Ramos, D. D. P. (s. f.). D[a] IRENE HERNANDO SAMIT.

Bustamante, F. M. (s. f.). Teoría del muestreo: Particularidades del diseño muestral en estudios de la conducta social.

Azorín, F. y Sánchez-Crespo, J.L. (1986). Métodos y Aplicaciones del Muestreo. Madrid: Alianza Editorial.

Daniel, W.W. (1985). Estadística aplicada a las Ciencias Sociales y la Educación. México: McGraw-Hill.

Robinson, E.A.(1981). Statistical Reasoning and Decision Making. Houston, TX: Goose Pond Press.

Nicolás, S. (2011). *(Introductory Notions of Statistical Sampling)*. 6(1) 89-105.

Páginas web

Estadística básica: ¿Qué es la estadística? (s. f.). GCFGlobal.org. https://edu.gcfglobal.org/es/estadistica-basica/que-es-la-estadistica/1/

Historia estadistica. (s. f.). www.ine.es. https://www.ine.es/explica/docs/historia_estadistica.pdf

Sriram, R. (2023, 11 abril). 6 Essential Application of Statistics. El Blog de Kolabtree. https://www.kolabtree.com/blog/es/6-aplicaciones-esenciales-del-analisis-estadistico/

Westreicher, G. (2022, 24 noviembre). Muestreo. Economipedia. https://economipedia.com/definiciones/muestreo.html#google_vignette

muestreo. (2023). RAE. https://dle.rae.es/muestreo

Muestreo. (s. f.). www.chospab.es. https://www.chospab.es/calidad/archivos/Metodos/Muestreo.pdf

Cgb, O. M. C. o. I. (2023, 24 septiembre). El teorema del límite central: una joya de la estadística. https://es.linkedin.com/pulse/el-teorema-del-l%C3%ADmite-central-una-joya-de-la-omar-miguel

Roldán, P. N. (2021, 16 febrero). Ley de los grandes números. Economipedia. https://economipedia.com/definiciones/ley-los-grandes-numeros.html

Unir, V. (2022, 18 abril). Tipos de muestreo: los principales y sus características. UNIR. https://www.unir.net/ingenieria/revista/tipos-de-muestreo/

Universidad Europea. (2022, 28 junio). Tipos de muestreo: qué es y cuáles son. https://universidadeuropea.com/blog/tipos-de-muestreo/