Universidad de San Carlos de Guatemala
Escuela de Ciencias Físicas y Matemáticas
Departamento de Matemática

ELEMENTOS DE TEORÍA DE MUESTREO EN ENCUESTAS DE HOGARES

Sergio Alexander Alay Arellano
Asesorado por Cristian José Alvarez Bran

Guatemala, septiembre de 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

ELEMENTOS DE TEORÍA DE MUESTREO EN ENCUESTAS DE HOGARES

TRABAJO DE GRADUACIÓN
PRESENTADO A LA JEFATURA DEL
DEPARTAMENTO DE MATEMÁTICA
POR

SERGIO ALEXANDER ALAY ARELLANO
ASESORADO POR CRISTIAN JOSÉ ALVAREZ BRAN

AL CONFERÍRSELE EL TÍTULO DE
LICENCIADO EN MATEMÁTICA APLICADA

GUATEMALA, SEPTIEMBRE DE 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

CONSEJO DIRECTIVO INTERINO

Director	M.Sc. Jorge Marcelo Ixquiac Cabrera
Representante Docente	Arqta. Ana Verónica Carrera Vela
Representante Docente	M.A. Pedro Peláez Reyes
Representante de Egresados	Lic. Urías Amitaí Guzmán García
Representante de Estudiantes	Elvis Enrique Ramírez Mérida
Representante de Estudiantes	Oscar Eduardo García Orantes
Secretario	M.Sc. Freddy Estuardo Rodríguez Quezada

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

Director	Director
Examinador	Examinador 1
Examinador	Examinador 2
Examinador	Examinador 2
Secretario	Secretario

AGRADECIMIENTOS

DEDICATORIA

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
$:=$	es definido por
$\cong$	es isomorfo a
$\Leftrightarrow$	si y sólo si
$\varnothing$	conjunto vacío
E^c	complemento de E
$\subsetneq$	estrictamente contenido
$E \setminus F$	diferencia entre E y F
$E \Delta F$	diferencia simétrica entre E y F
$\mathcal{P}(X)$	conjunto potencia de X
χ_E	función característica de E
$E_n \uparrow$	E_n es una sucesión creciente
$\mathfrak{L}$	σ-álgebra de los conjuntos Lebesgue-medibles
$\mathscr{S}$	espacio muestral
$\mathfrak{A}$	σ-álgebra de eventos
$(\mathscr{S}, \mathfrak{A}, P)$	espacio de probabilidad
$\mathscr{D}$	espacio de las funciones de prueba
$\mathscr{D}'$	espacio de las distribuciones
δ_0	medida de Dirac, función δ de Dirac o δ-función
$\Phi^\times$	espacio antidual de Φ
$\Phi \subset \mathcal{H} \subset \Phi^\times$	espacio de Hilbert equipado o tripleta de Gel'fand
$\lvert\psi\rangle$	vector *ket*
$\langle\psi\rvert$	funcional *bra*
$\langle\varphi\rvert\psi\rangle$	*braket*

OBJETIVOS

General

Escriba el objetivo general.

Específicos

Enumere los objetivos específicos.

1.

2.

INTRODUCCIÓN

1. PRELIMINARES

1.1. Teoría elemental de probabilidad

1.1.1. ¿Qué es un espacio de probabilidad?

1.1.2. Variables aleatorias

1.1.2.1. Esperanza

1.1.2.2. Varianza

1.1.2.3. Función densidad

1.1.2.4. Función distribución

1.1.3. Distribuciones de probabilidad

1.1.3.1. Binomial

1.1.3.2. Normal

1.1.3.3. Teorema del Límite Central

1.1.4. Estimadores

1.1.5. Intervalos de confianza

¿Cómo se definen? y en la práctica, ¿cómo se interpretan?

2. ¿Qué son las encuestas?

Consideraremos una encuesta como una herramienta de investigación que se utiliza para recolectar datos e información de un grupo específico de personas, conocido como muestra, con el objetivo de inferir características de la población general de donde se tomó la muestra. Una encuesta se caracteriza por utilizar instrumentos tales como cuestionarios o entrevistas, en los que se plantean preguntas estandarizadas que permiten obtener respuestas comparables y cuantificables.

A lo largo de este capítulo se discutirán las generalidades, terminología e importancia de las encuestas.

2.1. ¿Qué elementos tiene una encuesta?

Con el propósito de asegurar la calidad y la efectividad de una encuesta, resulta fundamental considerar una serie de elementos clave que nos ayudarán para realizar su diseño, implementación y análisis. Estos elementos, además de garantizar la precisión de los datos recolectados, facilitan la interpretación y la aplicación práctica de los resultados y, principalmente, permiten generalizar a la población. A continuación se presentan los componentes que conforman una encuesta.

- **Unidad de observación.** Es el objeto sobre el que se realiza una medición. En el estudio de las poblaciones humanas, las unidades de observación suelen ser personas.

- **Población objetivo.** Es la colección completa de las unidades de observación que deseamos estudiar. Usualmente, definir la población objetivo es una parte difícil del estudio[1] y la elección de esta afectará profundamente a las estadísticas resultantes.

[1]Por ejemplo, en un estudio de eficacia de un nuevo programa educativo ¿la población objetivo debe incluir solo a los estudiantes de un cierto grado escolar?, ¿deberíamos limitar la población objetivo a estudiantes de una región específica o, incluir estudiantes de todo el país? intentá resumir esto.Ya.

- **Muestra.** Subconjunto de la población que se utiliza para inferir o sacar conclusiones sobre toda la población.

- **Unidad de muestreo.** Es una unidad que puede ser seleccionada para formar parte de la muestra. Se puede dar la situación de que queramos estudiar personas, pero no dispongamos de una lista de todas las personas de la población objetivo. En su lugar, los hogares sirven como unidades de muestreo y las unidades de observación son las personas que viven en los hogares.

- **Marco de muestreo.** Es una lista, un mapa u otra especificación de las unidades de muestreo de la población a partir de las cuales puede seleccionarse una muestra. Es importante que este marco cuente con información que permita ubicar de manera pertinente[2] a cada una de las unidades que lo conforman.

  Esto me gusta, pero no sé si aquí, dejemoslo por el momento y evaluemos si de repente hay un mejor lugar donde ponerlo Todos los marcos de muestreo presentan algún nivel de desactualización con respecto a la población objetivo. Por ejemplo, un marco de muestreo de áreas basado en cartografía, puede estar desactualizado debido a que es posible:

  - entrevistar a la misma persona en varias ocasiones,
  - nunca realizar la entrevista a una persona que no tiene un hogar fijo de residencia.

Esto también me gusta pero, de nuevo, no sé si aquí Es esencial saber que en las encuestas de personas, la población total suele ser menor que la población objetivo, es decir, no todas las peronas de la población objetivo están incluidas en el marco de muestreo, y varias personas no responderán a la encuesta.

Veamos cómo identificar algunos de estos elementos en dos de las encuestas que se han realizado en Guatemala.

2.1 Ejemplo. El Ministerio de Eduación -MINEDUC- realizó el estudio "Situación nutricional de los escolares inscritos en los centros educativos públicos que son beneficiarios del Programa de Alimentación Escolar, 2023" mediante la medición de peso y longitud/talla en una muestra estadísticamente significativa. En este estudio, las unidades de observación son los estudiantes matriculados en escuelas públicas,

[2]Para una encuesta de hogares con entrevistas *in situ*, las direcciones de los hogares de las personas a entrevistar. En el caso de una encuesta telefónica, los teléfonos de las personas a entrevistar.

ya que son los "objetos" sobre los cuales se realizaron las mediciones de peso y longitud. La población objetivo del estudio son todos los estudiantes matriculados en escuelas públicas durante el año 2023. Esta es la colección completa de unidades de observación que se desea estudiar.

2.2 Ejemplo. La Encuesta de Evaluación de la Calidad de los Servicios Públicos Básicos -ENCASBA- 2019, brinda información acerca de los servicios públicos que proporciona el Organismo Ejecutivo, la Municipalidad de Guatemala y otras instituciones públicas, la misma se sustenta en la evaluación que hizo la población informante del hogar, la cual podría ser el jefe o jefa del hogar o una persona de 18 años o más de edad. La población objetivo está compuesta por las viviendas y los hogares del municipio de Guatemala. Dentro de este contexto, las unidades de muestreo son los hogares, ya que cada hogar puede ser seleccionado para formar parte de la muestra. La muestra es un subconjunto representativo de estos hogares seleccionados para inferir conclusiones sobre la percepción de la calidad de los servicios en toda la población objetivo del municipio. El marco de muestreo consiste en un listado detallado de todas las unidades primarias de muestreo (UPM) del municipio, permitiendo identificar de manera precisa las unidades de muestreo elegibles para la encuesta.

Existen dos enfoques principales en el diseño de encuestas que se utilizan para recopilar datos sobre poblaciones:

- **Encuestas transversales.** Recopilan datos en un solo punto en el tiempo. Se utilizan para describir las características de una población en un momento específico y son útiles para realizar análisis de prevalencia o describir condiciones actuales. Como ejemplo de este enfoque podemos mencionar las *encuestas repetidas* que, además de ser una serie de encuesta aplicadas en diferentes momentos, utilizan el mismo diseño metodológico. Para este tipo de encuestas, la selección de hogares se realiza de forma independiente para cada aplicación.

- **Encuestas longitudinales.** Siguen las mismas unidades de observación (como hogares o individuos) a lo largo del tiempo. Esto permite analizar cambios y desarrollar comprensiones causales sobre cómo y por qué ocurren los cambios.

2.2. Relevancia de las encuestas de hogares

Las encuestas de hogares son utensilios esenciales en la investigación social, económica y demográfica. Proveen datos en diferentes tópicos:

1. *Salud y educación.* Proveen datos importantes sobre el acceso y calidad de servicios de salud y educación, lo que permite identificar desigualdades y áreas donde se requiere intervención. Esto es crucial para mejorar la cobertura y calidad de estos servicios (Incluir referencia: UNICEF - MICS).

2. *Investigación económica.* Datos sobre ingresos, gastos, empleo y otras variables económicas que son fundamentales para el análisis macro y microeconómico (Incluir referencia: OECD - Household Wealth).

3. *Calidad de Vida y Bienestar.* Recogen información sobre la calidad de vida y el bienestar subjetivo de las personas, incluyendo aspectos como la satisfacción con la vida, condiciones de vivienda, y acceso a servicios básicos. Ayudando así a construir un panorama integral del bienestar de la población (Incluir referencia: OECD - How's Life?).

4. *Desarrollo Sostenible.* En el contexto de los Objetivos de Desarrollo Sostenible -ODS-, las encuestas de hogares son una fuente de datos para monitorear el progreso hacia dichos objetivos (Incluir referencia: UN - SDG Indicators).

El diseño de la encuesta dependerá del objetivo de la medición, lo que se pretende al diseñar una encuesta de hogares es que esta sea un instrumento confiable, que brinde estimaciones exactas y precisas, ya que, de lo contrario, no se podrían monitorear de manera consistente las Políticas Públicas y los indicadores de interés.

2.3. Errores muestrales y no muestrales

En este apartado se trata de describir el paradigma de los errores que se cometen en una encuesta y cómo medirlos de manera acertada y acotarlos sobre la base del *principio de representatividad.* Para esto empezamos definiendo dicho principio.

Principio de representatividad en encuestas. Afirma que cada elemento incluido en una muestra se representa a sí mismo y a un grupo de unidades que no pertenecen a la muestra seleccionada, cuyas características son cercanas a las del elemento incluido en la muestra

Este principio es crucial para garantizar que los resultados de la encuesta sean generalizables y reflejen con precisión las opiniones, comportamientos o características de toda la población objetivo. Existen dos fuentes principales de error cuando se realiza una encuesta:

1. **Error muestral.** Este tipo de error ocurre porque no se incluyó a todas las personas de la población y se seleccionó una muestra. La magnitud del error muestral depende del tamaño de la muestra y del método de muestreo utilizado. En general, un tamaño de muestra mayor tiende a reducir el error muestral.

2. **Error no muestral.** Este tipo de error abarca todos los demás errores que pueden ocurrir en una encuesta que no están relacionados con el hecho de que se ha utilizado una muestra en lugar de un censo. Estos errores pueden surgir en cualquier etapa del proceso de la encuesta, desde el diseño del cuestionario hasta la recolección y el análisis de los datos. El error no muestral puede ser sistemático o aleatorio. Entre los diferentes tipos de errores no muestrales están:

    - *Error de cobertura.* Ocurre cuando algunos miembros de la población objetivo no tienen la oportunidad de ser seleccionados en la muestra.
    - *Error de respuesta.* Surge cuando los encuestados no proporcionan respuestas precisas, ya sea intencionalmente o por confusión; malentendidos del cuestionario o respuestas socialmente deseables.
    - *Error de no respuesta.* Se produce cuando ciertas personas seleccionadas para la muestra no responden a la encuesta, y estas personas pueden tener características u opiniones diferentes a las de quienes sí responden.
    - *Errores de procesamiento.* Incluyen errores en la entrada de datos, codificación o análisis de los datos de la encuesta.

Dentro de los errores no muestrales existe un tipo de error denominado **sesgo** que afecta la validez de los resultados de la encuesta; es una tendencia o inclinación que distorsiona la verdad de la representatividad de los datos, llevándolos a ser no objetivos o inexactos. Los sesgos pueden ocurrir en cualquier etapa del proceso de la encuesta: desde el diseño hasta la recolección y análisis de los datos.

Según [11] las dos fuentes más importantes de sesgo se resumen de la siguiente manera:

1. **Sesgo de selección.** Ocurre cuando parte de la población objetivo no está en el marco de muestreo o cuando el marco está incompleto y presenta deficiencias. En [1] se establece que este tipo de sesgo se presenta si:

- La selección de la muestra depende de cierta característica asociada a las propiedades de interés. Por ejemplo, la frecuencia con la que los adolescentes hablan con los padres acerca del SIDA.
- La muestra se realiza mediante elección deliberada o mediante un juicio subjetivo. Por ejemplo, si el parámetro de interés es la cantidad promedo de gastos en compras en un centro comercial y el encuestador elige a las personas que salen con muchos paquetes, entonces la información estaría sesgada puesto que no está reflejado el comportamiento verdadero de las compras.
- Existen errores en la especificación de la población objetivo. Por ejemplo, en encuestas electorales, cuando la población objetivo contiene a presonas que no están registrafas como votantes ante la organización electoral de su país.
- Existe sustitución deliberada de unidades no disponibles en la muestra. Si, por alguna razón, no fue posible obtener la medición y consecuente observación de la característica de interés para algún individuo en la población, la sustitución de este elemento debe hacerse bajo estrictos procedimientos estadísticos y no debe ser subjetiva en ningún modo.
- Existe ausencia de respuesta. Este fenómeno puede causar distorsión de los resultados cuando los que no responden a la encuesta difieren críticamente de los que sí respondieron.
- La muestra está compuesta por respondientes voluntarios. Los foros radiales, las encuestas de televisión y los estudios de portales de internet no proporcionan información confiable.

2. **Sesgo de medición.** Se da cuando el instrumento con el que se realiza la medición tiene una tendencia a diferir del valor verdadero que se desea averiguar. Éste sesgo debe ser considerado y minimizado en la etapa de diseño de la encuesta. Nótese que ningún análisis estadístico puede revelar que una pesa añadió a cada persona dos kilos de más en un estudio de salud. De igual manera, en [1] se encuentran algunas de las situaciones en las que se presenta el sesgo de medición:

    - Cuando el respondiente miente. Esta situación se presenta a menudo en encuestas que se pregunta acerca del ingreso salarial, alcoholismo y drogadicción, nivel socioeconómico e incluso edad.

- Difícil comprensión de las preguntas. Por ejemplo: ¿No cree que no esté en buen momento para invertir? La doble negación en la pregunta es muy confusa para el respondiente.
- Las personas tienden a olvidar. Es bien sabido que las malas experiencias suelen ser olvidadas; esta sutiación debe acotarse si se está trabajando en una encuesta de criminalidad.
- Distintas respuestas a distintos entrevistadores. En algunas regiones es muy probable que la raza, edad o género del encuestador afecte directamente la respuesta del entrevistado.
- Leer mal las preguntas o polemizar con el respondiente. El encuestador puedeinfluir notablemente en las respuestas. Por lo anterior, es muy importante que el proceso de entrenamiento del entrevistador sea riguroso y completo.

El siguiente ejemplo ilustra cómo es que los sesgos afectan a una encuesta.

2.3 Ejemplo. Durante las elecciones generales de 2015 en Guatemala, se llevaron a cabo varias encuestas preelectorales para predecir los resultados de la primera vuelta presidencial. Sin embargo, los resultados de las encuestas no coincidieron con los resultados finales, generando críticas y cuestionamientos sobre la metodología empleada. Entre los problemas que se identificaron, tenemos:

1. **Muestra no representativa**
    - *Subrepresentación*[3] *de votantes rurales:* Era recurrente que las encuestas no incluían adecuadamente a los votantes en áreas rurales, que representan una parte significativa del electorado en Guatemala. Esto llevó a un sesgo en los resultados hacia las preferencias de los votantes urbanos.
    - *Dificultades logísticas:* La geografía y el acceso limitado a algunas regiones hicieron que las encuestas no pudieran captar completamente las opiniones de todas las áreas del país.
2. **Tasa de respuesta baja**
    - *Desconfianza en las encuestas:* Muchos cuidadanos mostraron desconfianza hacia las encuestas, lo que resultó en tasas de respuesta bajas y potencialmente sesgadas.

[3]explicar qué significa esto

- *Acceso a tecnología:* En algunas encuestas que utilizaron métodos en línea o telefónicos, hubo subrepresentación de personas sin acceso a tecnología, como teléfonos o internet.

3. **Errores en la proyección de participación**
   Participación inesperada: Las encuestas no lograron capturar cambios en la participación electoral, como el aumento en el número de votantes jóvenes o de primer voto, que influyeron en los resultados finales.

4. **Formulación de preguntas**
   Influencias en las respuestas: Las preguntas utilizadas en algunas encuestas podrían haber sido percibidas como sesgadas o influyentes, lo que podría afectar la sinceridad de las respuestas.

2.4. Encuesta Multi-etápica

Las encuestas multi-etápicas son un tipo de diseño muestral utilizado en la investigación estadística y de encuestas. En este enfoque, la selección de la muestra se realiza en varias etapas, en lugar de seleccionar directamente a los individuos de la población objetivo. Este método es particularmente útil cuando se trata de poblaciones grandes y dispersas, donde sería costoso o logísticamente difícil realizar un muestreo directo. En general. en América Latina son muy comunes los diseños de selección en dos etapas:

- **Primera etapa.** Se seleccionan unidades grandes, llamadas Unidades Primarias de Muestreo (UPM), que pueden ser regiones geográficas, bloques o conglomerados.

- **Segunda etapa.** Dentro de cada UPM se seleccionan unidades más pequeñas como viviendas u hogares.

También es posible encontrar en algunos países diseños divididos en más de dos etapas. Por ejemplo, en una primera etapa se seleccionan municipios, en una segunda etapa se seleccionan UPM dentro de los municipios seleccionados y, en la tercera, se selecciona una muestra de hogares en aquellas UPM seleccionadas en la segunda etapa pertenecientes a los municipios seleccionados en la primera etapa de muestreo.

2.4.1. Características de las Encuestas Multi-etápicas

Según [10], el muestreo multi-etápico presenta dos características fundamentales que lo hacen estadísticamente robusto y eficiente en la planificación logística del proceso de recopilación de datos

- **La independencia,** que implica que no hay ninguna correlación en el diseño de muestreo de las UPM. Esto quiere decir que en cada UPM se puede ejecutar con independencia cualquier estrategia de muestreo que se considere apropiada para seleccionar la submuestra de hogares.

- **La varianza,** que implica que sin importar qué diseño de muestreo se haya ejecutado en la primera etapa para seleccionar las UPM, la segunda etapa de selección podrá ejecutarse de manera independiente a la primera. Es decir, el submuestreo de los hogares es independiente del muestreo de las UPM.

3. Selección de la muestra

El cálculo del tamaño de muestra es un elemento fundamental en la planificación y ejecución de encuestas, ya que determina la precisión y la representatividad de los resultados obtenidos. En este capítulo se aboradará el proceso de determinar el tamaño de muestra óptimo basándose en la Teoría de Probabilidad. Este enfoque permite asegurar que la muestra seleccionada sea suficientemente grande para reflejar las características de la población, pero también lo suficientemente pequeña para ser práctica y costo-efectiva.

3.1. Cálculo del tamaño de muestra

Vamos a tratar de abordar el problema de *determinar el valor de alguna cantidad o indicador de la población objetivo (dichos valores pueden ser: ingreso total, cantidad total de personas, proporción de personas con cierta característica, índice total de alguna característica de las personas, etc)*. El camino seguro pero exhaustivo sería el de recabar información de toda la población para poder conocer el indicador deseado, pero, desde ya, esto tiene demasiadas desventajas: económicas, logísticas, tiempo, errores de cobertura, calidad de los datos. Por lo tanto, lo ideal es realizar la recabación de información en solamente un pequeño porcentaje de la población para lograr determinar el valor del indicador en cuestión.

Pretender usar una cantidad limitada de información para determinar con precisión un valor exacto a nivel poblacional es poco realista. Pero es razonable utilizar esa información para obtener una idea general de la situación. Por esto, en lugar de buscar un valor exacto, nos enfocamos en determinar un rango que contenga el valor del indicador, algo como "La cantidad total de personas en el país está entre 15 y 20 millones". Idealmente, quisiéramos que este rango sea lo más estrecho posible para tener una estimación más concreta del valor real del indicador.

Sin embargo, para hacerlo correctamente debemos superar obstáculos clave como asegurar que la cantidad de información disponible sea suficiente, que esta

información refleje fielmente la realidad de la población, y encontrar una manera de evaluar la calidad de la estrategia empleada. Por esta razón utilizamos la Teoría de la Probabilidad como marco metodológico, ya que nos proporciona las herramientas necesarias para realizar este ejercicio de manera fundamentada, considerando toda la lógica mencionada.

3.1.1. ¿Cómo se define el tamaño de muestra de una encuesta a partir de la Teoría de probabilidad?

Sea s una muestra de $n \in \mathbb{N}$ individuos que pertenecen a una población $X = \{x_1, x_2, \ldots, x_N\}$, y sea $\mathcal{S}$ la colección de todas esas posibles muestras s. Cada individuo x_i tiene asociado un número real y_i que captura alguna cantidad del parámetro poblacional θ, por ejemplo, el ingreso del individuo que contribuye al ingreso promedio de la población

$$\frac{1}{N}\sum_{i=1}^{N} y_i. \tag{3.1}$$

Consideremos una función $\hat{\theta}\colon \mathcal{S} \to \mathbb{R}$ que asigna a cada posible muestra $s \in \mathcal{S}$ un valor real, que representará la información pertinente para el objetivo de estimar al parámetro poblacional y que se puede obtener de la medición de la muestra seleccionada. Por ejemplo, si el objetivo es ubicar el ingreso promedio de X, una función natural a considerar como punto de partida es el ingreso promedio de la muestra:

$$\hat{\theta}(s) = \frac{1}{n}\sum_{x_i \in s} y_i \tag{3.2}$$

donde y_i corresponde al ingreso del individuo x_i.

La función es "natural" en el sentido de que emula en la muestra lo que se desea saber de la población, pero a parte de ello no tiene por qué ser necesariamente una función útil para ubicar a (3.1), puede que la muestra s seleccionada tenga características muy singulares respecto al resto de la población y $\hat{\theta}(s)$ no esté ni cerca del valor poblacional. Sin embargo, si consideramos esta función en el contexto más general de todas las posibles muestras s que se pueden obtener, al promediar

todos los resultados de calcular $\hat{\theta}(s)$ se llega a que

$$\begin{aligned}\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\hat{\theta}(s) &= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{x_i\in S} y_i \\ &= \frac{1}{|\mathcal{S}|}\sum_{i=1}^{N}\frac{y_i}{n}\cdot|\{s\in\mathcal{S}\colon x_i\in s\}| \\ &= \frac{1}{\binom{N}{n}}\sum_{i=1}^{N}\frac{y_i}{n}\binom{N-1}{n-1} \\ &= \frac{1}{N}\sum_{i=1}^{N} y_i.\end{aligned} \tag{3.3}$$

Es decir, el promedio de los $\hat{\theta}(s)$ es exactamente igual al parámetro deseado. A primera vista esto quizás pueda parecer, como mucho, un resultado bonito e interesante pero que carece de utilidad práctica. Después de todo, disponer de $\hat{\theta}(s)$ para toda posible muestra es equivalente a tener la información de toda la población, así que hemos vuelto al punto de partida. Sin embargo, intuitivamente el promedio de ciertas cantidades se interpreta como esa medida "resumen" o "punto medio" del contexto general de esas cantidades y esta noción se potencia en el contexto de la Teoría de Probabilidad. Si $\mathcal{S}$ es un espacio de probabilidad, es decir, cada $s\in\mathcal{S}$ tiene asignada una probabilidad $p(s)$ de ser la muestra seleccionada, entonces $\hat{\theta}$ es una variable aleatoria. Si además, cada s tiene la misma probabilidad de ser la muestra

$$p(s) = \frac{1}{|\mathcal{S}|}$$

entonces el cálculo que acabamos de hacer nos dice que f es un estimador insesgado del parámetro poblacional, es decir

$$\begin{aligned}\mathrm{E}[\hat{\theta}] &= \sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{S}|}f(s) \\ &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \theta.\end{aligned}$$

Lo interesante de esto es que es posible ubicar la esperanza de una variable aleatoria en cierto rango con cierta "confianza" asociada a esta estructura probabilística que existe en $\mathcal{S}$, gracias primordialmente al famoso Teorema del Límite Central. En

resumen, *por medio de un estimador insesgado $\hat{\theta}$ del parámetro poblacional θ, existe una buena forma de ubicar a θ y todas las tareas alrededor de la muestra estarán íntimamente ligadas a esto.*

En las encuestas se considera un tamaño fijo de elementos de una población para pertenecer a una muestra s sobre la cual se recopilará información. Una de las primeras tareas al diseñar una encuesta consiste en definir ese tamaño de elementos de la población, esto se hace al **ubicar el parámetro de interés** definido por el objetivo principal de la encuesta.

Ubicar el parámetro θ consiste en construir un intervalo de confianza para θ con un nivel de significancia α y una amplitud A. Tanto α como A son cantidades que pueden depender del criterio de las personas intersadas en los resultados de la encuesta puesto que son quienes tienen, o deberían tener, una noción adecuada de las dimensiones de las cantidades recopiladas, además de las aplicaciones que se desprenden de los resultados de la encuesta, por ejemplo: la elaboración de Políticas Públicas o los análisis socioeconómicos (Manual de Encuestas Sobre Hogares).

Por el Teorema del Límite Central tenemos que para un tamaño de muestra $|s| = n$ suficientemente grande en una población grande y un estimador insesgado $\hat{\theta}$ de θ se cumple que

$$\mathrm{P}\left(\left|\hat{\theta} - \theta\right| \leq c\sqrt{\mathrm{V}(\hat{\theta})}\right) \approx \phi(c), \quad c \in \mathbb{R}$$

donde ϕ es la función de distribución de una variable aleatoria normal estándar (en otras palabras, $\phi(c)$ es un valor conocido). Esto implica que el intervalo

$$I = \left[\hat{\theta} - c\sqrt{\mathrm{V}(\hat{\theta})}, \hat{\theta} + c\sqrt{\mathrm{V}(\hat{\theta})}\right] \tag{3.4}$$

contiene al parámetro θ con probabilidad $\phi(c)$. Como podemos ver en (3.4), la amplitud A de I depende de c, que está ligado al nivel de significancia que usualmente se establece en un 95 %, y de la varianza $\mathrm{V}(\hat{\theta})$ del estimador.

Supongamos que la varianza del estimador es una función decreciente del tamaño de muestra n, es decir $\mathrm{V}(\hat{\theta}) = h(n)$ con $n\colon \mathbb{N} \to \mathbb{R}$ tal que $a < b \implies h(a) > h(b)$.

Si se busca que la amplitud de I no supere un umbral A_0, entonces

$$\begin{aligned} 2c\sqrt{\mathrm{V}(\hat{\theta})} \leq A_0 &\iff \\ h(n) \leq \left(\frac{A_0}{2c}\right)^2 &\iff \\ n \geq h^{-1}&\left(\left(\frac{A_0}{2c}\right)^2\right). \end{aligned}$$

Es decir, podemos hallar el tamaño "óptimo" de muestra tomando el n más pequeño que satisface la desigualdad anterior.

La forma que toma $\mathrm{V}(\hat{\theta})$ dependerá de la estructura probabilística de $\mathcal{S}$, así como del estimador $\hat{\theta}$.

Anteriormente vimos que un diseño de muestreo consiste en una medida de probabilidad definida en $\mathcal{S}$, que consiste en asignar a cada $s \in \mathcal{S}$ un número $p(s) \in (0, 1)$ que captura "qué tan probable" es que s sea *la* muestra.

3.2. Diseño muestral

En el ámbito de las encuestas de hogares, el diseño muestral ocupa un lugar central como fundamento de los procesos inferenciales que permiten obtener información precisa y confiable sobre la población de un país. La premisa básica de este enfoque radica en que una muestra represenativa puede reflejar las características más importantes de una población extensa. Esta idea, aunque ambiciosa, se respalda en la implementación de los procedimientos robustos y metodologías estadísticas rigurosas que garantizan estimaciones válidas y exactas, incluso a partir del estudio de una fracción relaticamente pequeña de la población objetivo.

El diseño muestral, por tanto, no solo facilita el levantamiento de datos con eficiencia operativa y presupuestaria, sino que también constituye la base para generar indicadores sociales y económicos fundamentales. Estos indicadores permiten evaluar Pólíticas Públicas, monitorear el desarrollo y guiar la toma de decisiones infomradas para mejorar las condiciones de vida de la sociedad.

3.2.1. ¿Qué es un diseño muestral?

Un diseño muestral es un conjunto de procedimientos y estrategias que definen la forma en que se seleccionará una muestra a partir de una población objetivo. En términos prácticos, es el proceso que garantiza que cada unidad de la población tenga

una probabilidad conocida y no nula de ser incluida en la muestra. Esto permite que las estimaciones derivadas sean represenativas y precisas, evitando sesgos y proporcionando bases sólidas para realizar inferencias estadísticas.

El diseño muestral se construye a partir de principios fundamentales como la aleatorización y la inclusión:

- **Aleatorización:** Las unidades que componen la muestra se seleccionan de forma probabilística, asegurando que cualquier combinación plausible de hogares o individuos pueda ser parte de la muestra.

- **Inclusión:** Todas las unidades de la población tienen una probabilidad no nula de ser seleccionadas, garantizando que ninguna unidad quede excluida del proceso.

3.2.2. Muestreo Aleatorio Simple

El diseño muestral más sencillo de todos, sobre el cual toda l a teoría parte, es el Muestreo Aleatorio Simple. La definición consiste en considerar cada una de las $s \in \mathcal{S}$ como igual de probables, es decir:

$$p(s) = \frac{1}{|\mathcal{S}|}. \tag{3.5}$$

Bajo este diseño, el estimador natural de un promedio poblacional resulta tener varianza

$$\mathrm{V}_{\mathrm{MAS}}(\hat{\theta}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

donde S^2 y N son constantes de la población, y n es el tamaño de la muestra. Se puede verificar que esta es una función decreciente en n, así que por lo descrito anteriormente existe una forma sistemática en la cual uno define el tamaño de muestra para una encuesta bajo este diseño:

$$n = \frac{S^2}{\left(\frac{A_0}{2}\right)^2 \cdot \frac{1}{c^2} + \frac{S^2}{N}} \tag{3.6}$$

Al hacer $A_0/2 = e$, $c = z_{\alpha/2}$ y suponiendo que N es mucho más grande que S se llega a la fórmula que se encuentra en los libros de texto:

$$n = \left(\frac{z_{\alpha/2} S}{e}\right)^2.$$

3.2.3. Muestreo estratificado

3.3. ¿Qué es el DEFF?

3.3.1. ¿Cómo se calcula el tamaño de muestra en encuestas de hogares?

3.4. Algoritmos de selección

4. Factores de expansión

CONCLUSIONES

1. Conclusión 1.
2. Conclusión 2.
3. Conclusión 3.

RECOMENDACIONES

1. Recomendación 1.
2. Recomendación 2.
3. Recomendación 3.

BIBLIOGRAFÍA

[1] S. Lohr. *Sampling: Design And Analysis.* 2.ª ed. CRC press/Chapman & Hall/Taylor & Francis Group, Nueva York, 2019.

[2] R. De la Madrid. The rigged Hilbert space of the free hamiltonian. Consultado en marzo de 2005 en `http://arxiv.org/abs/quant-ph/0210167`.

[3] J. Escamilla-Castillo. *Topología.* 2.ª ed. s.e., Guatemala, 1992.

[4] N. Haaser y J. Sullivan. *Análisis real.* Tr. Ricardo Vinós. Trillas, México, 1978.

[5] P. Halmos. *Teoría intuitiva de los conjuntos.* 8.ª ed. Tr. Antonio Martín. Compañía Editorial Continental, S.A., México, 1973.

[6] F. Kronz. Quantum theory: von Neumann versus Dirac. Consultado en marzo de 2005 en `http://plato.stanford.edu/entries/qt-nvd/`.

[7] K. Liu, X. Sun, and S.-T. Yau. Goodness of canonical metrics on the moduli space of Riemann surfaces. *Pure Appl. Math. Q.*, **10**(2):223–243, 2014

[8] E. Leader and C. Lorcé, The angular momentum controversy: What's it all about and does it matter?, *Phys. Rept.* **541**, 163 (2014).

[9] S. Sternberg. Theory of functions of a real variable. Consultado en abril de 2005 en `http://www.math.harvard.edu/~shlomo`.

[10] Comisión Económica para América Latina y el Caribe (CEPAL), *Diseño y análisis estadístico de las encuestas de hogares de América Latina*, Metodologías de la CEPAL, N° 5 (LC/PUB.2023/14-P), Santiago, 2023.

[11] A. Gutiérrez, *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*, Ediciónes de la U, Bogotá, 2016.