KIBERNUM
IT Academy

Módulo 4 – Inferencia Estadística

Muestra y Muestreo

Ciencia de Datos

Objetivos de Aprendizaje

- Utiliza los conceptos básicos de estadística Inferencial.
- Reconocer técnicas de muestreo.
- Realizar cálculos de probabilidad utilizando la distribución muestral para resolver un problema.

KIBERNUM

Contenido:

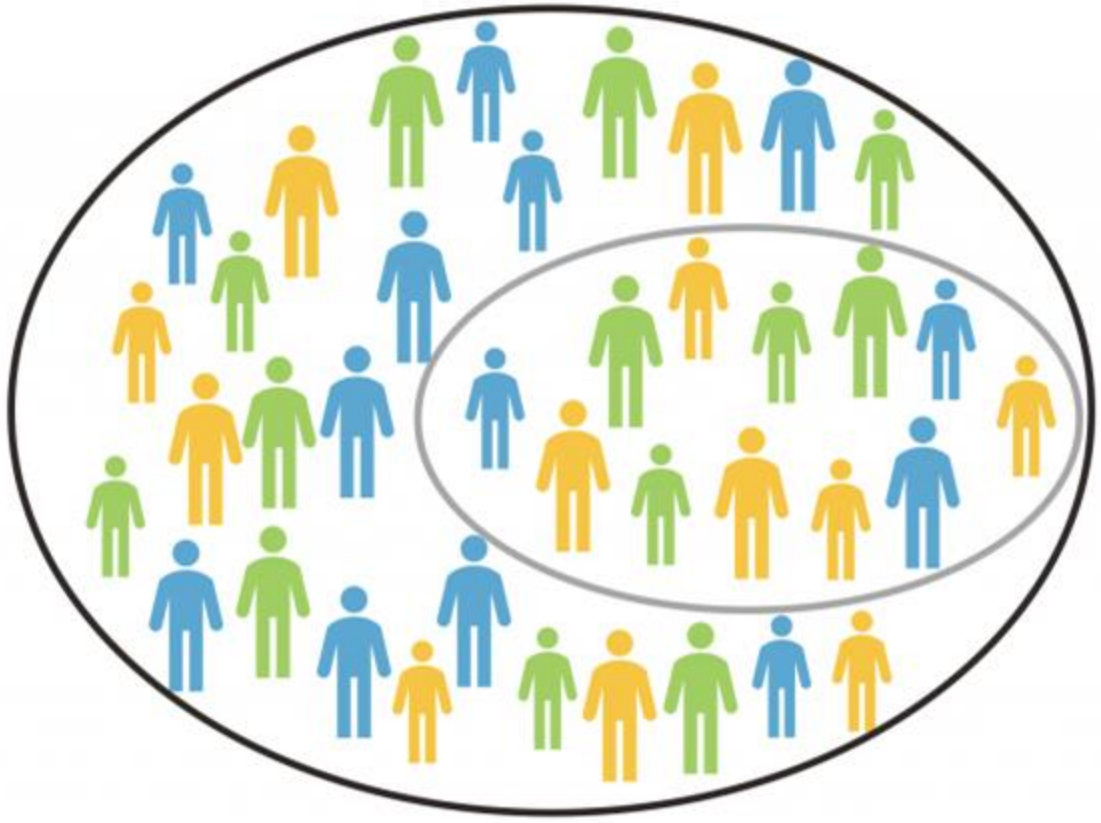
1. Muestreo.
2. Tamaño Muestral.
3. Tipos de Muestreo.

KIBERNUM

Muestreo

KIBERNUM

Muestra Poblacional



Población

Muestra

Población

Los términos **población** (o universo) y **muestra** son términos **relativos**.

La **población** es el **conjunto de elementos (sujetos, objetos o indicadores) que presentan determinada característica o propiedad en común**, que el investigador quiere analizar al realizar la investigación y que satisfacen un conjunto predeterminado de criterios establecidos (definidos) por el analista. Es decir, son **los "casos" investigados**, que pueden ser personas, animales, registros de cualquier tipo, muestras de laboratorios, etc., pero que son siempre **elementos que comparten una determinada característica predefinida por el investigador**, en base a la cual se agrupan en una determinada población.

Muestra Poblacional

El analista debe definir precisamente los criterios que permitan decidir, ante cada caso o elemento, si pertenece o no a la población investigada, es decir, debe determinar estrictamente el marco muestral o los límites de la población.

Cuando el tamaño de la población es **muy grande**, la investigación no se realiza en toda su extensión , sino en un **subconjunto** o parte de ella, denominada muestra, y después **se generalizan los hallazgos** obtenidos a toda la población.

La muestra debe ser representativa de la población.

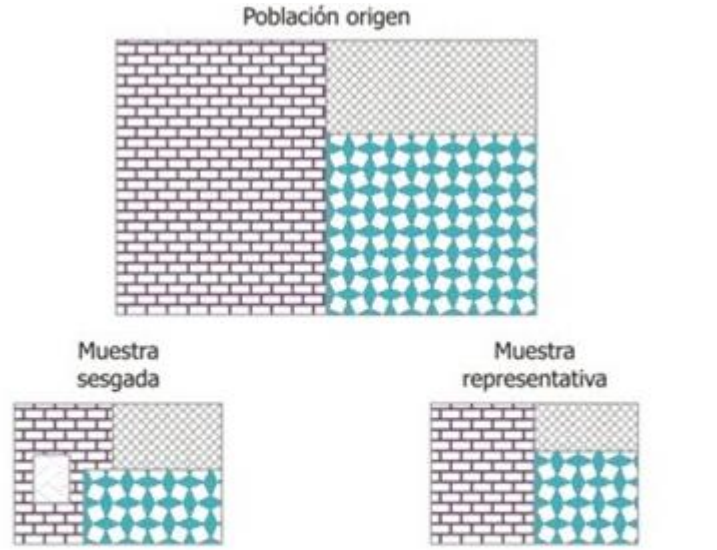


Figura 5.5: Población, muestra representativa y muestra sesgada.
Fuente: Pantoja, 2009.

KIBERNUM

Muestra Poblacional

La muestra es el **subconjunto** de la población donde se efectúa o lleva a cabo la investigación con la finalidad de **generalizar** posteriormente los resultados a toda la población.

Para que dicha generalización sea lícita, la muestra debe poseer **las mismas (o muy similares) características básicas (relevantes) de la población investigada**, es decir, debe ser **representativa** de la población.

KIBERNUM

Muestra Poblacional

- Si **N** es el tamaño de la población y **n** el tamaño de la muestra, siendo **N** suficientemente **grande**, pueden extraerse un cierto número de muestras **distintas** de tamaño **n**.

- Si en cambio **N** es un número **pequeño** (por ejemplo, 30 o 40 casos), convendrá **analizar directamente a toda la población**, es decir, no extraer una muestra o subconjunto.

KIBERNUM

Unidad de Observación y de Muestreo

- La **unidad de observación** es cada uno de los elementos (sujetos, objetos o indicadores) que integran la población, y en los que se analizarán las variables investigadas.

- La **unidad de muestreo o de análisis** es el elemento utilizado para seleccionar la muestra, es decir, cada uno de los elementos que integran la muestra.

- Por lo general, la unidad de observación (poblacional) y la unidad de análisis (muestral) son la misma, pero hay casos en que no: si se desea investigar el maltrato familiar de los menores, y no hay modo de seleccionar directamente las unidades de observación (los menores maltratados), se seleccionan las unidades de análisis (los hogares o casas donde habitan los menores maltratados) para poder llegar a ellos.

Ventajas del Muestreo

Permite profundizar más el análisis de las variables involucradas en el fenómeno investigado permite mayor control de dichas variables.

¿Por qué calcular el tamaño de la muestra?

- Una muestra puede estudiarse con mayor rapidez que una población.
- El estudio de una muestra es menos costoso.
- Toma menos tiempo el estudio a realizar.
- Los resultados son más precisos.

¿Cuándo calcular el tamaño de la muestra?

- Cuando no se puede estudiar toda la población.
- Cuando se quieren estudiar dos o más grupos y establecer diferencias.
- Cuando se quieren estimar parámetros, prevalencia, promedio, porcentaje y tasas.

Tamaño Muestral

KIBERNUM

Determinando el Tamaño de la Muestra

El tamaño muestral se determina en función de varios factores, como el objetivo del estudio, la variabilidad de los datos, el nivel de confianza deseado y el margen de error permitido.

A continuación, se describen los pasos generales para determinar el tamaño muestral:

Definir el objetivo del estudio: El objetivo del estudio debe ser claro y específico. Es importante saber qué se quiere medir o evaluar y qué resultados se esperan obtener.

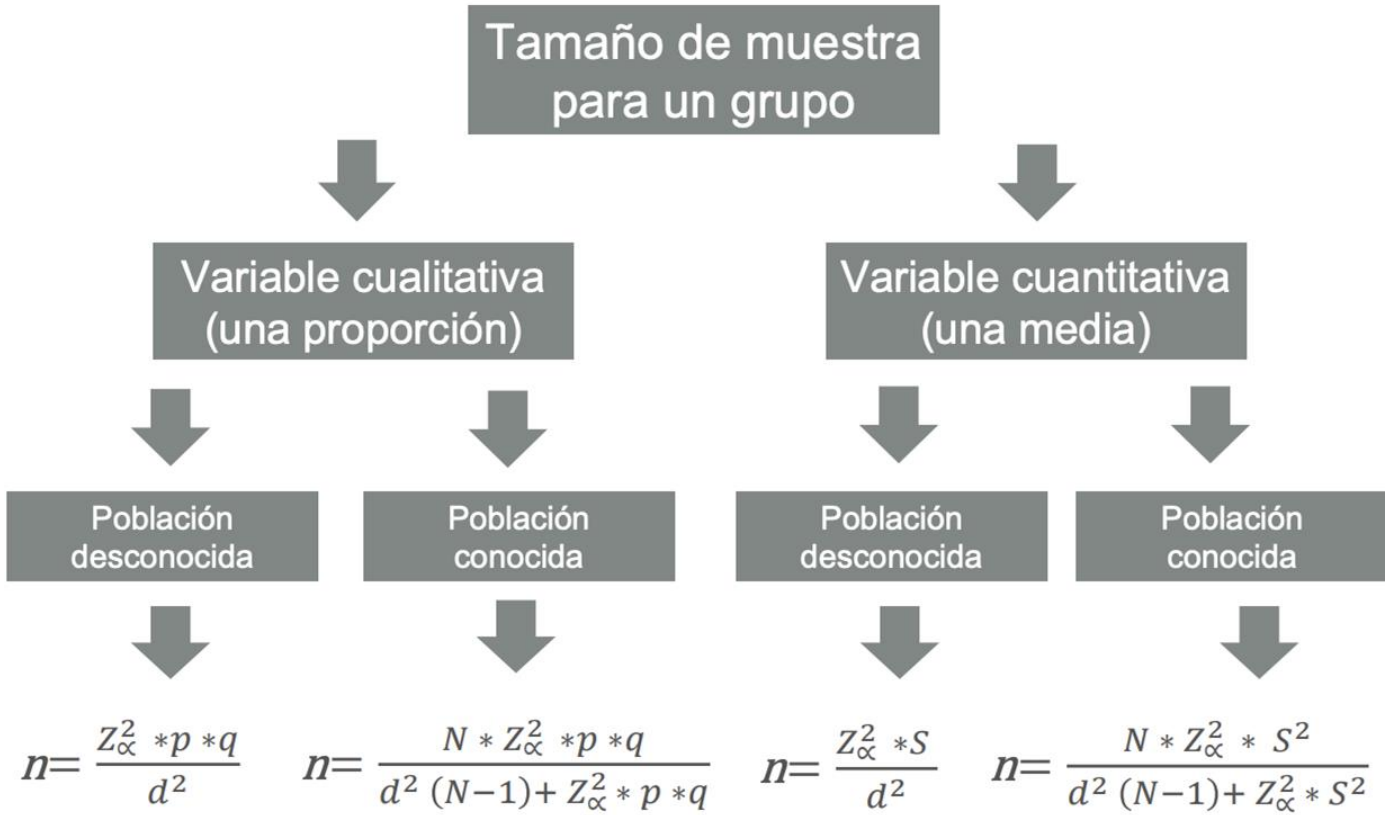
Identificar la población: Es importante definir claramente la población a la que se quiere generalizar los resultados. La población puede ser un grupo de personas, objetos o eventos que comparten características similares.

Determinar el nivel de confianza: El nivel de confianza es la probabilidad de que los resultados obtenidos en la muestra sean representativos de la población. El nivel de confianza se expresa en porcentaje y se suele fijar en un 95% o un 99%

Establecer el margen de error: El margen de error es la cantidad de variabilidad que se permite en los resultados. Se expresa en porcentaje y suele estar entre el 2% y el 5%

Determinando el Tamaño de la Muestra

Una vez que se han determinado el nivel de confianza y el margen de error, se puede utilizar una fórmula estadística para calcular el tamaño muestral necesario. La fórmula varía según el tipo de variable y el conocimiento de parámetros de la población.



Determinando el Tamaño de una Muestra

Para determinar el tamaño de muestra necesario para **estimar una proporción con población desconocida**, se puede utilizar la siguiente fórmula:

$$n = \frac{Z^2 * p * (1 - p)}{e^2}$$

Donde:

- **n** = tamaño muestral.
- **Z** = valor crítico de la distribución normal estándar para el nivel de confianza seleccionado.
- **p** = proporción de la población con la característica que se está estudiando. (Si no se conoce, se utiliza un valor estimado)
- **e** = margen de error.

La proporción **(p)** se puede estimar a partir de estudios previos o mediante una estimación conservadora basada en el conocimiento experto del tema. El margen de error **(e)** se define como la máxima cantidad de error que se puede tolerar en la estimación de la proporción.

KIBERNUM

Determinando el Tamaño de la Muestra

Por ejemplo, si se quiere estimar la proporción de personas que usan un determinado servicio en una ciudad, con un nivel de confianza del 95% y un margen de error del 5%, y se estima que la proporción es del 50%, entonces la fórmula para calcular el tamaño de muestra sería:

$$n = \frac{Z^2 * p * (1 - p)}{e^2}$$

$$n = \frac{(1.96)^2 * 0.5 * (1 - 0.5)}{0.05^2}$$

$$n = 384.16$$

Por lo tanto, se necesitaría una muestra de al menos 385 personas para estimar la proporción con un nivel de confianza del 95% y un margen de error del 5%. Es importante tener en cuenta que, si la proporción estimada es menor o mayor que el 50%, el tamaño de muestra necesario podría ser mayor o menor que el ejemplo mencionado.

Determinando el Tamaño de la Muestra

Para determinar el tamaño de muestra necesario para **estimar una proporción con población finita (conocida)**, se utiliza la siguiente fórmula:

$$n = \frac{N * Z^2 * p * (1 - p)}{e^2 * (N - 1) + Z^2 * p * (1 - p)}$$

Donde:

n = tamaño muestral.
N = tamaño población o universo.
Z = valor crítico de la distribución normal estándar para el nivel de confianza seleccionado.
p = proporción de la población con la característica que se está estudiando. (Si no se conoce, se utiliza un valor estimado).
e = margen de error.

Determinando el Tamaño de la Muestra

Por ejemplo, si se quiere estimar la media de la edad de los estudiantes universitarios en una ciudad con un nivel de confianza del 95% y un margen de error de 2 años, y se estima que la desviación estándar es de 5 años, entonces la fórmula para calcular el tamaño de muestra sería:

$$n = \left(\frac{1.96 \cdot 5}{2}\right)^2 = 24.01$$

Por lo tanto, se necesitaría una muestra de al menos 25 estudiantes universitarios para estimar la media de la edad con un nivel de confianza del 95% y un margen de error de 2 años. Si no se conoce la desviación estándar de la población, se puede utilizar una estimación conservadora (por ejemplo, s = 10), lo que aumentaría el tamaño de muestra necesario a alrededor de 97.

Determinando el Tamaño de la Muestra

Para determinar el tamaño de muestra necesario para **estimar una media con población desconocida**, se utiliza la siguiente fórmula:

$$n = \frac{Z^2 * s^2}{e^2}$$

Donde:

n = tamaño muestral
Z = valor crítico de la distribución normal estándar para el nivel de confianza seleccionado
s = desviación estándar de la muestra (o una estimación de la desviación estándar de la población)
e = margen de error

Si la desviación estándar de la población (σ) es conocida, se puede utilizar en lugar de la desviación estándar de la muestra (s).

Determinando el Tamaño de la Muestra

Como se puede apreciar, el nivel de confianza corresponde a una puntuación Z.
Este es un valor constante necesario para la ecuación. Conviene conocer algunas puntuaciones Z para los niveles de confianza más comunes:

90%	→	Z = 1.645
95%	→	Z = 1.96
99%	→	Z = 2.576

KIBERNUM

Tipos de Muestreo

KIBERNUM

Tipos de Muestreo

La representatividad de la muestra tiene que ver, entonces, con que ésta posea aproximadamente las mismas características básicas que posee la población. Y esto, a su vez, tiene que ver con la manera de seleccionar u obtener la muestra (es decir, con los procedimientos de extracción de la muestra) y con el tamaño de la muestra.

Distintos procedimientos de obtención de muestras definen distintos tipos de muestreo.

En general, los muestreos se califican en **probabilísticos** y **no probabilísticos**.

Tipos de Muestreo

Todos y cada uno de los elementos que integran la población tienen la misma probabilidad conocida de ser seleccionados.

Probabilístico (Aleatorio)

- Aleatorio Simple
- Sistemático
- Estratificado
- Por Conglomerado

Cuando **no todos** tienen la misma posibilidad de ser elegidos, o esta, probabilidad **no se conoce.**

No Probabilístico

- Por cuotas
- Accidental
- Por conveniencia
- Bola de nieve

Muestreo Probabilístico

Pese a que nunca hay garantías absolutas de representatividad, en los muestreos probabilísticos el error de muestreo, es decir, el margen de error o riesgo de equivocarse al generalizar los resultados obtenidos en la muestra a toda la población, puede calcularse. Este margen de error **(“desviación estándar”)** se define de antemano.

En los muestreos no probabilísticos el margen de error se desconoce, y por ende, no puede calcularse.

Muestreo Aleatorio Simple

Cada uno de los elementos o unidades de la población tiene aquí **la misma probabilidad conocida de ser seleccionado**, y esto se logra mediante la **selección al azar** de dichos elementos.

Se confecciona primero un **listado numerando correlativamente todas las unidades de la población** (denominada "**marco muestral**"), para lo cual es necesario, previamente, haber **definido correctamente la población** (es decir, haberla delimitado de un modo estricto y concreto).

KIBERNUM

Muestreo Aleatorio Simple

Ventajas

- La simpleza de su procedimiento y el bajo costo.
- No es necesario dividir la población en subgrupos ni tomar ningún otro paso adicional antes de seleccionar miembros de la población al azar.

Desventajas

- No puede usarse cuando la población es demasiada grande, o potencialmente infinita.
- No es posible confeccionar el listado numerado de todas las unidades (es decir, el marco o estructura muestral).
- Además, dependiendo del tamaño de la población, puede tornarse un método muy lento.

Muestreo Sistemático

El **muestreo sistemático** es un método de **muestreo probabilístico** en el que los elementos de una población se ordenan de manera sistemática y se seleccionan periódicamente para formar una muestra representativa.

En el muestreo sistemático, se selecciona el primer elemento de la muestra aleatoriamente, y luego se seleccionan los elementos restantes de manera sistemática, utilizando un intervalo de muestreo que se determina dividiendo el tamaño de la población por el tamaño de la muestra deseada. Por ejemplo, si se desea una muestra de 100 elementos de una población de 1000, se seleccionaría cada décimo elemento después del primer elemento.

Este método es útil cuando la población es grande y está ordenada, ya que permite ahorrar tiempo y recursos al no tener que revisar cada elemento individualmente. Sin embargo, es importante tener en cuenta que el muestreo sistemático puede estar sujeto a sesgos si hay patrones sistemáticos en la población que no se tienen en cuenta al seleccionar los elementos de la muestra.

Muestreo Sistemático

Ventajas

- Es un método relativamente sencillo y rápido de implementar.
- Puede reducir la variabilidad en la muestra y, por lo tanto, mejorar la precisión de las estimaciones.
- Es menos costoso que otros métodos de muestreo probabilístico, como el muestreo aleatorio simple.

Desventajas

- Si hay patrones sistemáticos en la población, como ciclos o estacionalidad, el muestreo sistemático puede introducir sesgos en la muestra.
- Si la población no está ordenada, el muestreo sistemático puede ser difícil de aplicar.
- La elección del punto de partida puede tener un impacto significativo en los resultados del muestreo sistemático, y puede ser difícil determinar cuál es el mejor punto de partida.

Caso de uso

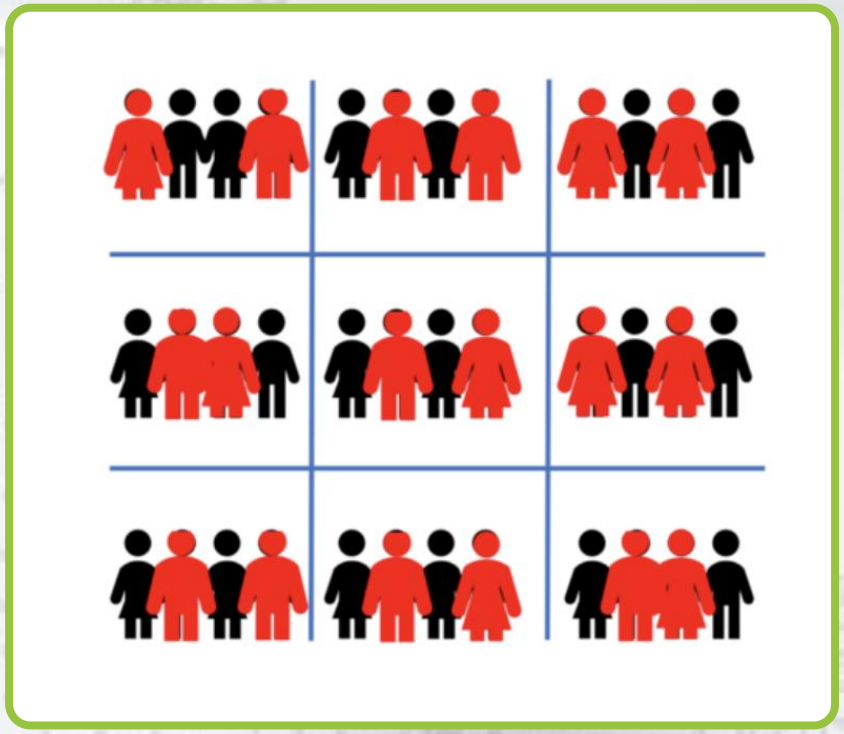
- Se usa cuando los datos relevantes no exhiben patrones.

Muestreo Estratificado

El muestreo estratificado es un **método de muestreo probabilístico** que implica la **división de la población en subgrupos mutuamente exclusivos y homogéneos** llamados estratos. A continuación, se selecciona una muestra aleatoria de cada estrato de acuerdo con un plan de muestreo específico.

En el muestreo estratificado, **la población se divide en estratos en función de características importantes que se cree que afectan la variable de interés**. Por ejemplo, si se está llevando a cabo una encuesta sobre la satisfacción del cliente de una empresa, los estratos podrían ser los diferentes productos o servicios ofrecidos por la empresa.

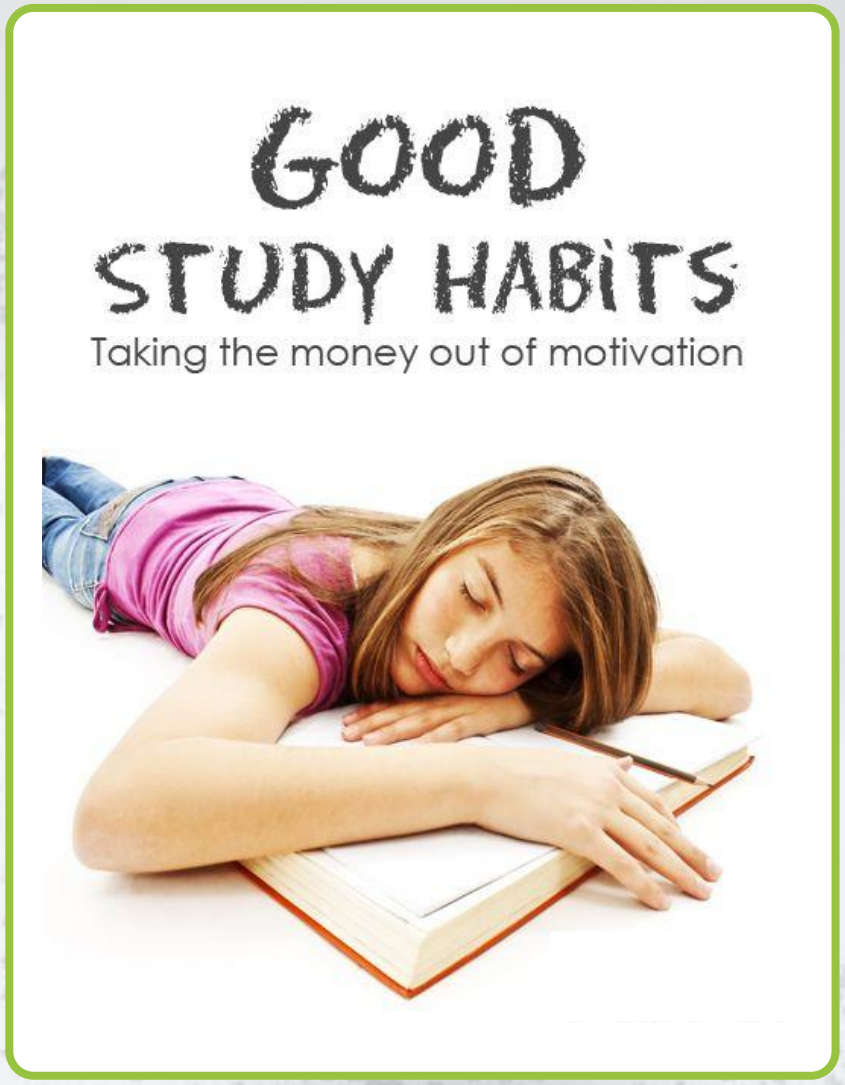
Una vez que se han identificado los estratos**, se selecciona una muestra aleatoria de cada estrato**. Esto asegura que cada subgrupo de la población esté representado en la muestra, lo que aumenta la precisión de las estimaciones y reduce la variabilidad.



KIBERNUM

Muestreo Estratificado - Ejemplo

Por ejemplo, un investigador está analizando la relación entre las variables hábitos de estudio y nivel de aprendizaje logrado. Si el investigador sospecha que la variable hábitos de estudio se comporta de manera diferente respecto de la variable nivel socioeconómico, porque supone, acertadamente o no, que los alumnos de bajo nivel socioeconómico tienen menos hábitos de estudio que los alumnos de elevado nivel socioeconómico , podría formar tres estratos o subgrupos en la población total: alumnos de bajo nivel socioeconómico, alumnos de nivel medio, y alumnos de alto nivel.



KIBERNUM

Muestreo Estratificado

Ventajas

- El muestreo estratificado puede proporcionar estimaciones más precisas y confiables que otros métodos de muestreo probabilístico, especialmente cuando la población es heterogénea.
- Permite la comparación de subgrupos dentro de la población, lo que puede ser útil en estudios que buscan identificar diferencias en la variable de interés entre diferentes grupos.
- Es más eficiente que otros métodos de muestreo probabilístico en términos de costos, ya que permite que la muestra sea más pequeña y aun así representativa.

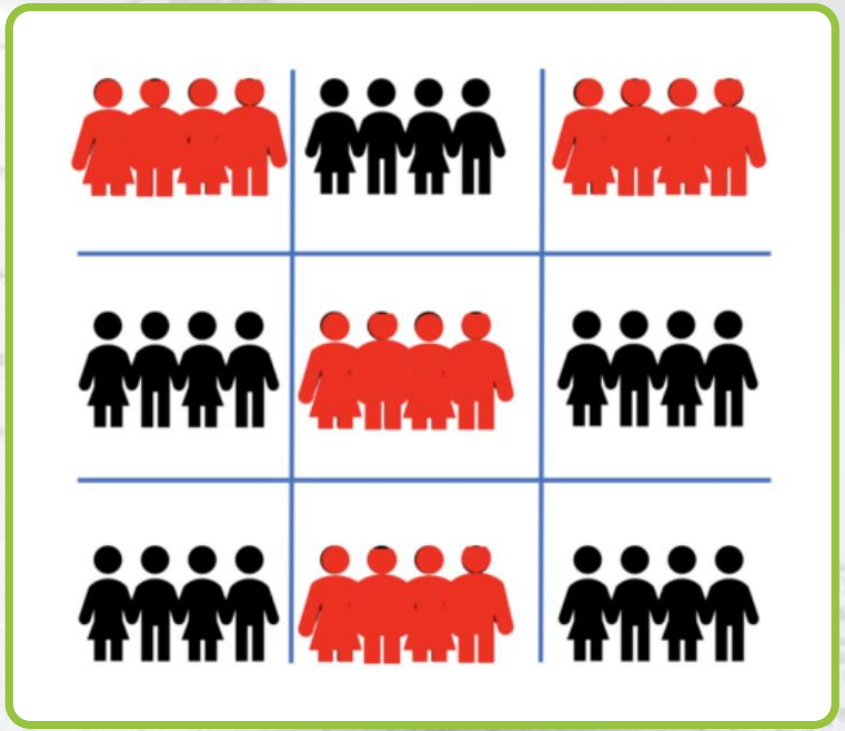
Desventajas

- El muestreo estratificado puede ser más complejo y costoso de implementar que otros métodos de muestreo.
- La identificación y selección de los estratos puede ser un desafío y puede requerir información previa sobre la población.
- Si la estratificación se basa en variables incorrectas o mal elegidas, el muestreo estratificado puede introducir sesgos en la muestra.

Muestreo por Conglomerado (clustering)

Es un método de muestreo probabilístico que se utiliza para poblaciones **grandes y dispersas**, en lugar de individuos, se seleccionan conglomerados que están agrupados de forma natural, por ejemplo: casas, cuadras, manzanas, colonias, etc. La selección de la muestra de cada conglomerado es aleatoria.

Se selecciona en primer lugar el conglomerado más alto y a partir de este se selecciona un subgrupo. Y así sucesivamente hasta legar a las unidades de análisis. También, se denomina muestreo por etapas múltiples.



KIBERNUM

Muestreo por Conglomerado (clustering)

Ventajas

- El muestreo por conglomerados puede ser más eficiente que otros métodos de muestreo probabilístico, especialmente cuando la población es grande y dispersa.
- Puede ser más fácil y más práctico implementar el muestreo por conglomerados en comparación con otros métodos de muestreo probabilístico, especialmente cuando se trata de poblaciones muy grandes.
- Si los conglomerados se eligen cuidadosamente, el muestreo por conglomerados puede proporcionar estimaciones precisas y confiables.

Desventajas

- El muestreo por conglomerados puede ser menos preciso que otros métodos de muestreo probabilístico, especialmente si los conglomerados no son heterogéneos.
- Si se eligen conglomerados inadecuados, el muestreo por conglomerados puede introducir sesgos en la muestra.
- La estimación de los errores de muestreo en el muestreo por conglomerados puede ser más complicada que en otros métodos de muestreo probabilístico.

Caso de uso

- Se utiliza cuando todos los individuos de cada grupo pueden ser representativos de las poblaciones.

Dudas y consultas

Fin Presentación

KIBERNUM