

Implementación de un diseño muestral en dos etapas

Angela Gissel López Rodríguez

Jhon Alejandro Ramirez Daza

December 4, 2023



Departamento de Estadística
Universidad Nacional de Colombia
Bogotá, Colombia

1 Introducción

La Gran Encuesta Integrada de Hogares (GEIH) constituye un estudio exhaustivo y representativo llevado a cabo en Colombia con el propósito de recopilar datos detallados sobre diversos aspectos socioeconómicos de los hogares y la fuerza laboral del país. Este relevamiento, a cargo del Departamento Administrativo Nacional de Estadística (DANE), es una fuente primordial para analizar y monitorear los indicadores socioeconómicos en Colombia.

Desarrollada de manera continua a lo largo del año, la GEIH abarca una muestra considerable de hogares en todo el territorio colombiano. Su diseño posibilita obtener estimaciones representativas a nivel nacional, departamental y regional, proporcionando así una visión integral de la realidad social y laboral en Colombia.

La encuesta recopila información detallada sobre diversas variables socioeconómicas, como empleo, ingresos, educación, vivienda, salud y satisfacción laboral, entre otras. Estos datos son cruciales para analizar y seguir las condiciones de vida de la población colombiana.

La GEIH emplea una metodología de muestreo probabilístico que garantiza que cada hogar e individuo tengan la misma probabilidad de ser seleccionados, asegurando así la confiabilidad estadística y generalización de los resultados a toda la población.

La relevancia de la GEIH radica en su capacidad para ofrecer información actualizada y confiable sobre diversos aspectos socioeconómicos en Colombia. Los datos recopilados son utilizados por investigadores, formuladores de políticas, instituciones académicas y otros interesados en la toma de decisiones basada en evidencia.

Este trabajo se centra en utilizar los datos de la GEIH para analizar la realidad laboral en Colombia. Se seleccionarán variables clave, como departamento de residencia, tiempo de desplazamiento al trabajo, satisfacción laboral, tipo de contrato, ingresos mensuales y antigüedad laboral, con el objetivo de obtener una visión integral de los aspectos relacionados con el empleo y la calidad de vida de los trabajadores colombianos.

El análisis de estos datos permitirá identificar patrones, tendencias y desafíos en el mercado laboral, así como comprender las desigualdades socioeconómicas y proponer recomendaciones para mejorar las condiciones laborales y promover el bienestar de los trabajadores en Colombia.

2 Resumen descriptivo de los datos.

- Departamento de Trabajo (P388): Esta variable indica el departamento donde el individuo de estudio desempeña principalmente su actividad laboral.
- Tipo de Contrato Laboral (P6460): Esta variable identifica si el contrato de trabajo de cada individuo es de término indefinido, a término fijo, o si no cuenta con esa información.
- Ingresos Mensuales Antes de Descuentos (P6500): Se refiere al monto total que el individuo ganó el mes pasado en el empleo actual, antes de cualquier deducción.
- Duración Continua en el Empleo Actual (P6426): Indica el tiempo que el individuo lleva trabajando de manera ininterrumpida en la actual empresa, negocio, industria, oficina, firma o finca.
- Número de Meses Trabajados en los Últimos 12 Meses (P6790): Esta variable cuantifica la cantidad de meses que el individuo ha estado empleado durante el período de los últimos 12 meses.
- Licencia por Enfermedad Pagada (P6424S5): Esta variable indica si el individuo de estudio ha gozado de licencia por enfermedad remunerada.

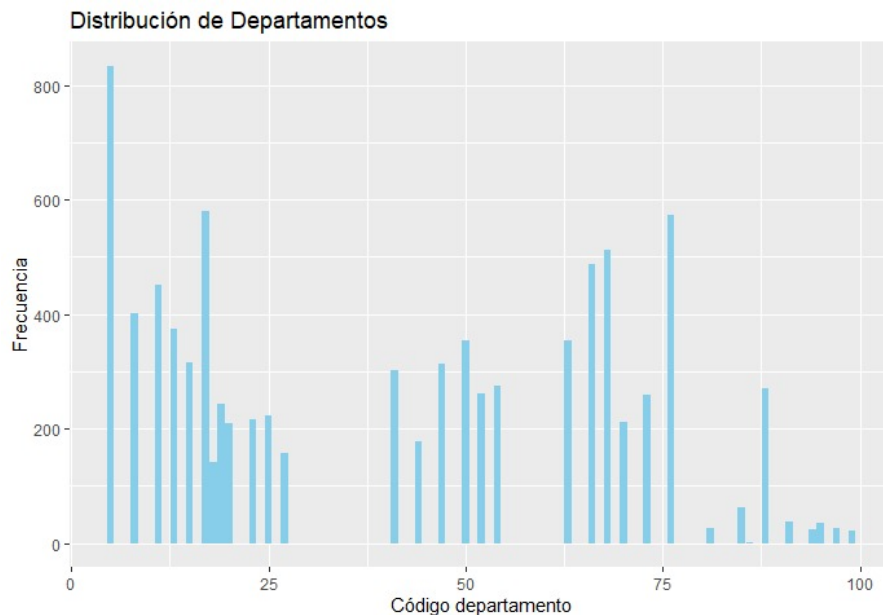


Figure 1: Distribucion de los departamentos

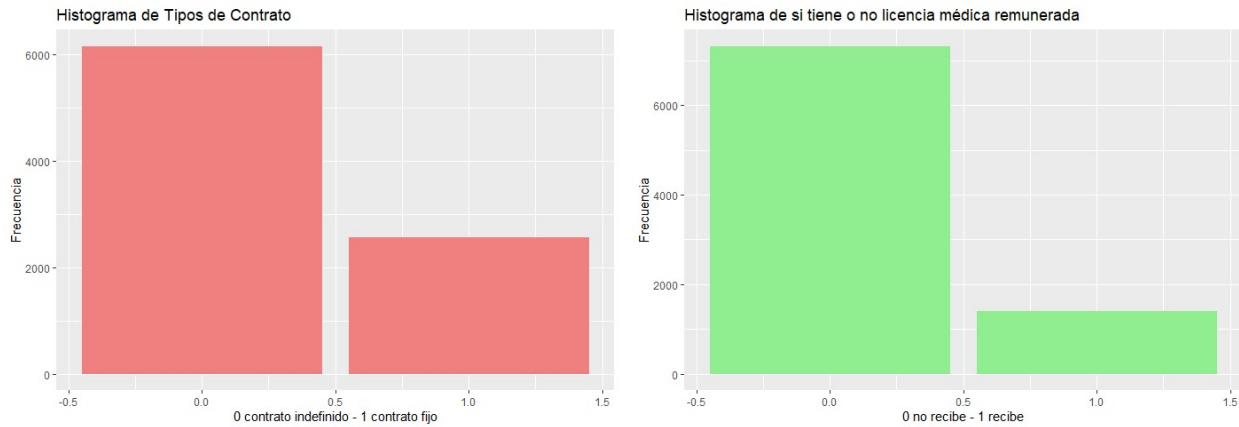


Figure 2: Histogramas de las variables "Tipo de contrato" y "Licencia médica remunerada", donde en el primer histograma la primera barra hace referencia a las personas que poseen contrato indefinido y la segunda a las que tienen contrato fijo, y en el segundo histograma la primera barra haciendo referencia a las personas que no reciben licencia médica remunerada y la segunda a las que sí.

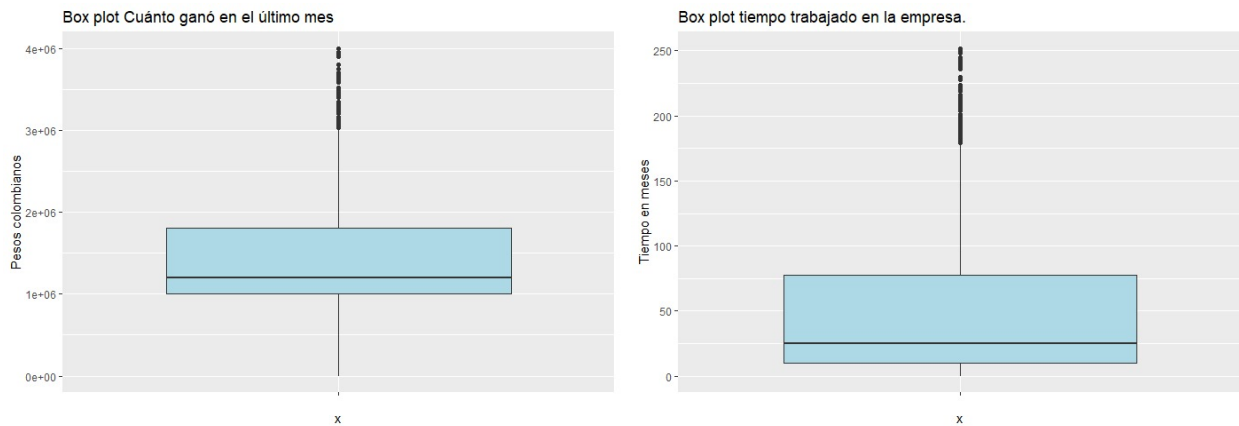


Figure 3: Box plot de las variables "Cuánto ganó en el último mes" y "Tiempo trabajado en la empresa"

Algunas características de las variables son:

- La variable "Licencia enfermedad" es binaria, con valores 0 y 1.
- La variable "Tipo de contrato" es binaria, con valores 0 y 1. La media (0.2944) sugiere que, en promedio, alrededor del 29.44% de los individuos tienen un contrato distinto al de término indefinido.
- La variable "Ingresos Mensuales" muestra valores en millones de pesos colombianos. La

media (1,533,554) y la mediana (1,200,000) indican una distribución sesgada hacia la derecha, con algunos individuos que ganan salarios significativamente más altos.

- La variable "Trabajo por meses en el año" muestra la duración promedio de empleo en meses, con valores que van desde 1 hasta 12. La media (11.14) sugiere que, en promedio, los individuos han trabajado alrededor de 11 meses.
- La variable "Duración en el empleo" muestra el tiempo total de trabajo en la empresa, que varía desde 0 hasta 252 meses (21 años). La media (54.65) y la mediana (25.00) indican una distribución sesgada hacia la derecha, con algunos individuos que tienen un tiempo total de trabajo considerablemente más largo.

3 Justificación teórica e implementación del diseño muestral en dos etapas MAS-MAS

En el desarrollo de este estudio, se implementó un diseño muestral en dos etapas con el objetivo primordial de asegurar la representatividad de la muestra y obtener conclusiones robustas para la población objetivo.

En la primera etapa del muestreo, se optó por un diseño de muestreo aleatorio simple (MAS), utilizando Unidades Primarias de Muestreo (UPM's) representadas por los departamentos. Esta elección se basó en la premisa de que cada departamento constituye una partición y a su vez es una entidad significativa en términos de las variables de interés para el estudio. Al aplicar un diseño MAS en esta etapa, se garantiza que cada departamento tenga igual probabilidad de ser seleccionado, lo que proporciona una base imparcial para la representación de todo el país.

En la segunda etapa del muestreo, se implementó un diseño de muestreo aleatorio simple a nivel de hogares. Esta elección se justifica por la necesidad de obtener información detallada a nivel de unidades familiares, considerando que las características a analizar podrían variar dentro de un mismo departamento. Al aplicar un diseño MAS a nivel de hogares, se asegura que cada hogar tenga la misma probabilidad de ser seleccionado, permitiendo así una representación equitativa de las viviendas en la muestra.

La combinación de estos dos diseños muestrales en dos etapas, busca lograr una muestra que capture tanto la variabilidad a nivel departamental como las particularidades a nivel de hogar. Este enfoque proporciona una base sólida para la validez de los resultados obtenidos, per-

mitiendo realizar inferencias y generalizaciones significativas sobre la población objetivo del estudio.

Este diseño estratégico busca maximizar la representatividad y la validez de las conclusiones, proporcionando una estructura sólida para el análisis detallado de las variables de interés a nivel departamental y hogar en el contexto de la investigación.

4 Cálculo del tamaño de muestra apropiado indicando la fórmula a utilizar y realizar las correspondientes justificaciones teóricas correspondientes.

En la primera etapa de muestreo se llevo a cabo un diseño MAS tomando como UPMs los departamentos, para calcular el tamaño de muestra se realizó el calculo con la formula para el tamaño de muestra para el total, la media y la proporción poblacional, que están definidas de la siguiente manera:

El tamaño de muestra para el total poblacional está definido por:

$$n_{m_I} = \frac{z_{1-\frac{\alpha}{2}}^2 N_I^2 S_{YU_I}^2}{EM^2 + z_{1-\frac{\alpha}{2}}^2 N_I S_{YU_I}^2} = \frac{n_0}{1 + \frac{n_0}{N}}$$

donde $n_0 = \frac{z_{1-\frac{\alpha}{2}}^2 N_I^2 S_{YU_I}^2}{EM^2}$

El tamaño de muestra para la media poblacional está definido por:

$$n_{m_I} = \frac{z_{1-\frac{\alpha}{2}}^2 S_{YU_I}^2}{EM^2 + \frac{z_{1-\frac{\alpha}{2}}^2 S_{YU_I}^2}{N_I}}$$

El tamaño de muestra para la proporción poblacional está definido por:

$$n_{m_I} = \frac{z_{1-\frac{\alpha}{2}}^2 \frac{N_I}{N_I-1} p(1-p)}{EM^2 + z_{1-\frac{\alpha}{2}}^2 \frac{1}{N_I} \frac{N_I}{N_I-1} p(1-p)}$$

Una vez conseguidos los tamaños de muestra se evaluo cual era el más grande y se siguio a la siguiente etapa. Se uso como S la desviación estandar de una muestra piloto de 100 obcer-

vaciones aleatorias de la población, de esta misma se obtuvo la proporción para las variables categóricas.

En la segunda etapa una vez teniendo el tamaño de muestra se volvió a aplicar un diseño MAS sobre la muestra seleccionada, donde de igual forma se realizó el cálculo del tamaño con las tres formulas para el total, la media y la proporción para cada UPM es decir para cada departamento en la muestra m_I y de igual forma se seleccionó el tamaño de muestra mayor.

5 Estimación de los parámetros asociados con las variables de tipo continuo o discreto como también de las variables categóricas.

Para la estimación de los parámetros de proporciones y promedios se tiene en cuenta el estimador Hurvitz-Thompson como sigue

$$\hat{t}_\pi = \sum_{h=1}^H \hat{t}_{\pi h} = \sum_{h=1}^H \frac{N_{IH}}{n_{Ih}} \hat{t}_{i\pi h}$$

Luego, la proporción estimada está dada por:

$$\hat{p} = \frac{\hat{t}_\pi}{N}$$

La varianza para los estimadores de la proporción se estima de la siguiente manera

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^H \frac{N^2_{Ih}}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{t_{S_{Ih}}} + \frac{N_{Ih}}{n_{Ih}} \sum_{S_{Ih}} \frac{N^2_{ih}}{n_{ih}} \left(1 - \frac{n_{ih}}{N_{ih}}\right) S^2_{Y_{S_{ih}}}$$

$$\hat{V}(\hat{p}) = \frac{\hat{V}(\hat{t}_\pi)}{N^2}$$

Además el promedio estimado está dado por

$$\hat{\mu} = \frac{\hat{t}_\pi}{N}$$

La varianza para los estimadores del promedio se estima de la siguiente manera

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^H \frac{N^2_{Ih}}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{t_{SIh}} + \frac{N_{Ih}}{n_{Ih}} \sum_{S_{Ih}} \frac{N^2_{ih}}{n_{ih}} \left(1 - \frac{n_{ih}}{N_{ih}}\right) S^2_{Y_{s_{ih}}}$$

$$\hat{V}(\hat{\mu}) = \frac{\hat{V}(\hat{t}_\pi)}{N^2}$$

Se estima así mismo el coeficiente de variación para el total

$$CV(\hat{t}_\pi) = \frac{\sqrt{\hat{V}(\hat{t}_\pi)}}{\hat{t}_\pi}$$

Puede observarse en la siguiente tabla los resultados para las estimaciones del total

Variable	CV%
Licencia por enfermedad	0.1538
Tipo de contrato	0.1015
Ingreso mes	0.0334
Meses trabajo año	0.0127
Tiempo trabajando	0.0750

Table 1: Estimaciones de los parametros respecto a las variables

6 Resultados

Estabilidad de las Variables

Las variables se analizaron con respecto a las totales, que son las más grandes. Los coeficientes de varias variables son bajitos, alrededor de 0.0, lo que sugiere que son estables y muestran consistencia.

Proporciones de Variables Categóricas

- Licencia por Enfermedad: Aproximadamente el 10.8% de la población tiene licencias por enfermedad remuneradas.
- Tipo de Contrato: Cerca del 18% de la población tiene contrato de término fijo.

Promedios de Variables Numéricas

- Trabajo en el Año: En promedio, la gente ha trabajado alrededor de 9.7 meses en el último año.
- Ingresos Mensuales: El promedio de ingresos mensuales es de 978 mil pesos colombianos.
- Tiempo de Trabajo Actual: En promedio, los individuos llevan trabajando alrededor de tres meses en el trabajo actual.

7 Conclusiones

Las variables analizadas muestran coeficientes bajos, indicando estabilidad y consistencia en los datos.

En términos de beneficios laborales, como licencias por enfermedad remuneradas, solo el 10.8% de la población tiene acceso a este beneficio.

En cuanto a la naturaleza contractual, aproximadamente el 18% de la población tiene contratos de término fijo.

El tiempo promedio de trabajo en el último año es de 9.7 meses, sugiriendo cierta variabilidad en la duración del empleo.

El ingreso mensual promedio es de 978 mil pesos colombianos, proporcionando una visión de la distribución de los salarios en la muestra.

Los individuos, en promedio, llevan trabajando alrededor de tres meses en el trabajo actual, lo que destaca la relativa novedad de las posiciones actuales.

8 Bibliografía.

- DANE. (2023). Microdatos - Gran Encuesta Integrada de Hogares - GEIH - 2023. Recuperado de <https://microdatos.dane.gov.co/index.php/catalog/782/study-description>
- Trujillo, L., Polo, M. L., & González, L. M. (s.f.). Notas de Clase (Apuntes de clase). Recuperado de Notas_de_Clase.pdf.