UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Julio 2016

COMPARACIÓN ENTRE DOS ESTRATEGIAS BIETÁPICAS DE MUESTREO PARA LA ESTIMACIÓN DE UN TOTAL EN DIFERENTES ESCENARIOS SIMULADOS

COMPARISON BETWEEN TWO STRATEGIES OF SAMPLING IN TWO STAGES FOR A TOTAL ESTIMATED IN DIFFERENT SIMULATED CONTEXT

Rodrigo Antonio Gómez Hernández.[a]
rodrigogomez@usantotomas.edu.co

Resumen

En este trabajo se muestra la comparación entre dos estrategias de muestreo bietápico a través de los coeficientes de variación estimados (CVE). La primera estrategia es un muestreo proporcional a una variable de información auxiliar, acompañado de un muestreo aleatorio simple denominado πPTMAS; la segunda estrategia es un muestreo aleatorio estratificado acompañado de un muestreo aleatorio simple denominado ESTMAS-MAS, ambos con el estimador de Horvitz y Thompson. Esta comparación se hizo utilizando una variable de respuesta simulada en varios escenarios, donde se varió por un lado el coeficiente de correlación entre la variable de interés y la variable de información auxiliar (total personas por área geográfica de la base de datos de viviendas, hogares y personas - VIHOPE del DANE) y por otro lado se varió la razón entre el intercepto del modelo lineal que las asocia y el promedio de la variable simulada, esto con el fin de determinar el comportamiento de cada estrategia en cada escenario simulado y determinar la de mejor comportamiento según el coeficiente de variación estimado (CVE) arrojado por cada estrategia.

***Palabras clave*:** Estrategias de muestreo bietápicas, Simulación, CVE.

Abstract

This work is about the comparison between two strategies of two stages sampling through estimated coefficients of variation (ECV) produced by each strategy. The first strategy is a sampling proportional to a variable of auxiliary information, in addition to a simple random sampling named πPS-SRS; the second strategy is a stratified random sampling named STSI-SRS. Both strategies use the Horvitz - Thompson estimator. This comparison was made by using a variable of simulated answer in several scenarios, where it changed the correlation coefficient between the variable of interest and the variable of auxiliary information (total of people per geographical area of the database of Houses, Homes and People - VIHOPE of DANE) and it changed too the rate between the intercept of linear model which links them and the average of the simulated variable, in order to determine the performance of each strategy in each simulated scenario and decide which one has the best performance, according to the estimated coefficient of variation (ECV) produced by each strategy.

***Keywords*:** Two stages sampling strategies, Simulation, ECV .

[a]Estudiante de estadística Universidad Santo Tomás Bogotá

Índice

1. OBJETIVOS **IV**

1.1. Objetivo General IV

1.2. Objetivos Específicos IV

2. INTRODUCCIÓN **5**

3. MARCO TEÓRICO **6**

3.1. Definiciones Básicas 6

3.1.1. Marco de muestreo 6

3.1.2. Muestra Probabilística 6

3.1.3. Soporte de muestreo 6

3.1.4. Diseño de muestreo 7

3.1.5. Muestreo aleatorio sin reemplazo 7

3.1.6. Probabilidad de inclusión 7

3.1.7. Estimadores 8

3.1.8. Estimador de Horvitz y Thompson 9

3.1.9. Coeficiente de Variación Estimado (CVE) 10

3.2. Estrategias de Muestreo 10

3.2.1. Muestreo Aleatorio Simple (MAS) 11

3.2.2. Tamaño de Muestra 12

3.2.3. Muestreo Proporcional al Tamaño de Información Auxiliar πPT 13

3.2.4. Muestreo Estratificado 14

3.2.5. Muestreo aleatorio estratificado 16

3.2.6. Muestreos Bietápicos 18

3.2.7. Muestreo bietápico πPT-MAS 20

3.2.8. Muestreo bietápico ESTMAS-MAS 21

4. CONSTRUCCIÓN DE LOS ESCENARIOS **22**

4.0.1. Depuración de la base de datos y descripción de las variables 22

4.0.2. Creación de los Escenarios 23

5. MUESTREO BIETÁPICO ESTMAS-MAS **26**

6. MUESTREO BIETÁPICO πPT-MAS **28**

7. COMPARACIÓN DE LOS RESULTADOS **30**

7.1. Correlación (ρ) $> 90\,\%$ 30

7.2. Correlación (ρ) < 90 % 32

8. CONCLUSIONES **36**

9. ANEXOS **37**

9.1. Anexo A.- Descripción de la Base de Datos 37

9.2. Anexo B.- Tablas de Resultados Para los Muestreos Realizados 38

9.3. Anexo C.- Códigos en el Sofware Estadístico R requeridos en este trabajo 40

9.3.1. Código para la creación de los escenarios 40

9.3.2. Código para el muestreo ESTMAS-MAS 42

9.3.3. Código para el Muestreo Bietápico πPT-MAS 45

9.4. Anexo D.- Gráficos de distribución de las variables creadas 48

10.REFERENCIAS BIBLIOGRÁFICAS **50**

1. OBJETIVOS

1.1. Objetivo General

En este trabajo se propuso como objetivo general:

Determinar cuáles son los escenarios óptimos para los diseño de muestreo πPT-MAS y ESTMAS-MAS, a través de la comparación de los coeficientes de variación estimada (CVE) arrojados por cada diseño en cada escenario.

1.2. Objetivos Específicos

En este trabajo se propuso como objetivos específicos:

- Simular para cada área geográfica y para cada municipio las 49 variables respuesta que variarán por un lado la correlación entre variable creada y la variable auxiliar; por el otro lado la razón entre el intercepto del modelo lineal que la relaciona estas variables y la media de la variable respuesta.
- Crear para cada diseño de muestreo un algoritmo computacional que permita calcular un tamaño de muestra para primera y segunda etapa de cada estrategia además de obtener las estimaciones para el total y para la varianza del total estimado a través de 5000 simulaciones realizadas para cada una de las 49 variables creadas.
- Determinar cual estrategia de muestreo presenta mejores resultados en cada uno de los escenarios planteados a través de la comparación de los CVE obtenidos en las simulaciones.

2. INTRODUCCIÓN

Al contar dentro de un marco de muestreo con una variable de información auxiliar de tipo continuo para todos y cada uno de los elementos de dicho marco, se obtiene una ventaja enorme para mejorar la precisión de las estimaciones a realizar, pero antes de plantear la que a ojo cerrado sería la mejor estrategia de muestreo con información auxiliar, es menester que el analista o estadístico que tiene a cargo el estudio, determine si la variable de información auxiliar y la variable de interés se correlacionan en un buen grado y además; si el intercepto del modelo lineal que las relaciona es pequeñísimo con respecto a las unidades numéricas de observación.[1].

Este trabajo presenta la comparación de dos estrategias de muestreo bietápicas en 49 escenarios diferentes; la primera estrategia es un muestreo proporcional a una variable de información auxiliar acompañada de un muestreo aleatorio simple πPT-MAS; esta primera estrategia consta en su primera etapa con la que se cataloga por muchos académicos como la mejor estrategia de muestreo referente a información auxiliar πPT-MAS[2] , ya que contar con información auxiliar y tener probabilidades proporcionales a esta información aumenta la precisión y más aún el muestreo πPT es una estrategia de muestreo sin reemplazo y con un tamaño de muestra fijo; lo que potencia este muestreo con relación a otros que también utilizan información auxiliar. Pero ¿Qué pasa cuando la información de tipo auxiliar no está muy bien correlacionada con la variable de interés? o ¿Qué pasa si la razón entre el intercepto del modelo lineal que relaciona a la variable de interés y la variable de información auxiliar con respecto a la media de la variable de interés es alto?, se compara en este trabajo la estrategia de muestreo πPT-MAS con respecto a una estrategia de muestreo que también utiliza información auxiliar pero no para obtener probabilidades de inclusión sino para dividir la población en grupos homogéneos y escoger de cada grupo formado una muestra aleatoria; este tipo de muestreo se conoce como el muestreo aleatorio estratificado que ira acompañado en una segunda etapa por un muestreo aleatorio simple ESTMAS-MAS. En ambas estrategias los tamaños de muestra para la primera etapa son iguales, en la segunda etapa la diferencia de las unidades muestrales seleccionadas no supera en promedio el 5 %.

En la primera parte de este trabajo se definen algunos conceptos básicos y se realiza un marco teórico no extenso sobre el cual se sientan las bases para la realización de este trabajo y dentro de esta parte se presentan formalmente los datos a través de una breve descripción de estos mismos.

La segunda parte de este trabajo, después de mostrar las características de la base de datos utilizada, se describe la forma en que se construyen para cada escenario la variable respuesta con relación a la variable Total de personas por unidad de área de la base viviendas, hogares y personas (VIHOPE 2005), donde tanto el coeficiente de correlación entre la variable respuesta y la variable auxiliar fue graduado al 99 %, 97 %, 95 %, 93 %, 90 %, 70 %, 50 %, 30 %, 20 % y 10 % y dentro de cada uno de estos 5 posibles escenarios se simuló el modelo lineal ajustándolo de tal forma que la razón entre el intercepto de dicho modelo y la media de la variable respuesta variara al 90 %, 70 %, 50 %, 30 %, 20 % y 10 %.

En la tercera parte se muestran los resultados de las estrategias realizadas sobre cada una de los 49 escenarios simulados, los cuales, a través de un proceso de Montecarlo, se repitieron 5.000 veces para cada estrategia de muestreo y así obtener un estimativo del total de la característica de interés, de la estimación de la varianza de dicha estimación y el CVE de la estimación.

Por último se redactan las conclusiones del trabajo y se evalua el comportamiento de cada estrategia sobre los escenarios creados.

[1] La mejor forma de determinar si un intercepto es pequeño o no es la razón entre el intercepto del modelo que relaciona las variables y la media de la variable respuesta $\frac{\beta_o}{\bar{y}} \to 0$

[2] Gutiérrez (2009) propone este método cuando exista información aproximadamente proporcional a la característica de interés ya que aumenta junto con el estimador de Horvitz y Thopson de forma dramática la eficiencia de la estratégia

3. MARCO TEÓRICO

Dentro del enfoque matemático y estadístico, todo desarrollo investigativo tiene que ir respaldado por un marco teórico que apunte al problema de investigación. Este trabajo no pretende hacer un nuevo enfoque teórico hacia algún problema investigativo, más bien es un ejercicio esbozado para demostrar lo que teóricamente se plantea en lo que a la comparación de estrategias de muestreo se refiere.

Dentro de esta sección se enunciarán algunos conceptos básicos pero importantes en lo que a la teoría del muestreo se refiere, como marco muestral, soporte, el significado de una muestra probabilística, la probabilidad de inclusión, entre otros, para terminar definiendo formalmente lo que es un diseño de muestreo. Paso seguido se enfatizará en la definición de una estrategia de muestreo y se hará enfasis en las estrategias que hacen parte de este trabajo, entre ellas: Muestreo Aleatorio Simple (MAS), Muestreo Proporcional al Tamaño de Información (πPT) y el Muestreo Aleatorio Estratificado. por último se planteará el marco teórico para las estrategias de muestreo bietápicas del πPT-MAS y el ESTMAS-MAS.

3.1. Definiciones Básicas

3.1.1. Marco de muestreo

La gente entiende al muestreo como una subconjunto de individuos que pertenecen a cierta población de interés y estos individuos por lo general se escogen sin tener un marco muestral, Särndal et al. (2003) define un marco muestral como cualquier material o dispositivo utilizado para obtener acceso a una población finita de interés; puede ser bien una lista física o en medio electrónico que contienen todos los miembros de la población, esta lista permite el reconocimiento de cada individuo y su ubicación, más aún, el encargado de algún estudio por muestreo debe verificar la calidad de este marco ya que puede ser obsoleto, tener información repetida o puede faltarle información.

3.1.2. Muestra Probabilística

El término probabilístico es confundido muchas veces con el término azar, escoger una muestra al azar no es propiamente hacer un muestreo probabilístico; Gutiérrez (2009) define una muestra probabilística como aquella que cumple dos importantes condiciones: 1. se puede construir (al menos en teoría) el listado de todas las muestras posibles y 2. las probabilidades de selección de cada una de esas nuestras son conocidas u otorgadas de antemano. Solo los muestreos probabilísticos permiten inferir estimaciones realizadas en la muestra para la población y así tomar decisiones para toda la población que esté bajo estudio. Una muestra es probabilística cuando se tiene un marco muestral, teóricamente se puede construir la colección de todas las posibles muestras y dar una probabilidad mayor o igual a cero a cada muestra, de tal forma que la suma de todas las probabilidades otorgadas sea 1.

3.1.3. Soporte de muestreo

Un soporte representado por la letra Q es la colección de todas las posibles muestras que se pueden realizar del total de unidades muestrales de la población y se dice simétrico si para cualquier muestra $s \in Q$ todas las posibles permutaciones de s están en Q. Cuando otorgamos a la muestra la característica de ser de tamaño fijo, el número de posibles muestras está dado por $\binom{N}{n}$, de las cuales en la vida práctica por lo general solo se puede escoger una sola realización. Para el desarrollo de este trabajo se tomará un tamaño de muestra fijo.

En este orden de ideas se denota con la letra S a las muestras que pueden generarse de un marco y s a una realización de esta misma.

3.1.4. Diseño de muestreo

Un diseño de muestreo es una función que le otorga a cada posible muestra del soporte la probabilidad de ser seleccionada, se define de la siguiente manera:

$$Pr(S = s) = p(s) \qquad \text{para todo } s \in Q. \tag{1}$$

Esta función debe cumplir dos condiciones importantes:

1. $p(s) \geqslant 0$ para todo $s \in Q$
2. $\sum_{s \in Q} p(s) = 1$

3.1.5. Muestreo aleatorio sin reemplazo

Es importante resaltar que una muestra de tamaño n puede ser tomada de la población de los individuos que no hayan sido tomados en anteriores pasos; este tipo de muestreo se denomina muestreo sin reemplazo y Gutiérrez (2009) define este tipo de muestreo como la realización de un vector de tamaño N (Tamaño de la población) donde la componente k tomará el valor 1 si el k-ésimo individuo pertenece a la muestra o 0 en cualquier otro caso.

$$s = (I_1, I_2, ..., I_N)' \in \{0, 1\} \qquad con\ I_k(s) = \begin{cases} 1 & si \quad k \quad \in \quad s \\ 0 & si \quad k \quad \notin \quad s \end{cases}$$

$I_k(s)$ se conoce como la variable indicadora y como se dijo solo toma dos posibles valores $0, 1$ definidos según el k-ésimo individuo pertenezca o no a la muestra.

3.1.6. Probabilidad de inclusión

De manera intuitiva se puede pensar que si la muestra es en particular (y no en todos los casos) menor a la población entonces algún elemento k de la población pueden pertenecer o no a una muestra seleccionada. A cada elemento de la población se le asigna en un diseño de muestreo una medida de probabilidad de inclusión, esta medida es la probabilidad de que el elemento pertenezca a la muestra. Esta medida se conoce como probabilidad de inclusión π_k y se denota como sigue:

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s). \tag{2}$$

Dentro del conjunto de todas las posibles muestras, se puede definir la probabilidad de inclusión como el producto interior entre el vector de variables indicadoras, por el vector de probabilidades otorgada a cada posible muestra, matemáticamente se define:

$$\pi_k = E(I_k(S)) = \sum_{s \in Q} I_k(s) p(s) \tag{3}$$

De manera similar se tiene una probabilidad de segundo orden π_{kl} que no es más que la probabilidad de que los elementos k y l estén conjuntamente en la muestra y se expresa como:

$$\pi_{kl} = Pr(k \in S \text{ y } l \in S) = Pr(I_k I_l = 1) = \sum_{s \ni k \text{ y } l} p(s). \tag{4}$$

3.1.7. Estimadores

Dentro del inferencia estadística y aun dentro del muestreo, existen funciones que se aplican a cada elemento perteneciente a la muestra y toman valores dentro de los reales, estas funciones reciben el nombre de *Estadísticas*, más aún, si esta función se utiliza para estimar algún parámetro poblacional recibe el nombre de *Estimador* y a su realización dentro de la muestra se le llama *Estimación*. Wackerly et al. (2010) define una estimación como una regla, a menudo expresada como fórmula, que indica cómo calcular el valor de una estimación con base en las mediciones realizadas sobre la muestra.

Al ser una estadística una variable aleatoria, se define para dicha variable el valor esperado y varianza y dichos valores dependerán de los valores que asuma la muestra realizada s. Sea G una estadística, su valor esperado y su varianza quedan expresadas de la siguiente manera:

$$E(G) = \sum_{s \in Q} p(s)G(s). \tag{5}$$

$$Var(G) = E[G - E(G)]^2 = \sum_{s \in Q} p(s)[G(s) - E(G)]^2. \tag{6}$$

Como ya se mencionado anteriormente para los objetivos de este trabajo se necesitará un estimador para el total poblacional con un muestreo probabilístico sin reemplazo y de tamaño fijo, el estimador que de define más adelante cumple con los requisitos de ser para estimar el total poblacional además de ser insesgado para el total y que es un estimador para muestreo sin reemplazo.

Dentro de la teoría de los estimadores Wackerly et al. (2010) meciona las propiedades que debe tener un estimador puntual y entre estas propiedades las más importante son:

- **La eficiencia Relativa** que mide la razón entre las varianzas de dos estimadores para cierto parámetro θ, es decir, suponga que $\hat{\theta_1}$ y $\hat{\theta_2}$ son estimadores para θ, entonces si $\hat{\theta_1}$ es mas eficiente que $\hat{\theta_2}$ implica que la razón e $\frac{Var(\hat{\theta_1}}{Var(\hat{\theta_2}} < |1|$

- **Consistencia** Se dice que el estimador $_n$ es un estimador consistente de θ si, para cualquier número positivo ς

  $$\lim_{n \to \infty} P(|_n - \theta| > \epsilon) = 0$$

  Esto indica que entre mayor sea el tamaño de muestra, la probabilidad de que el parámetro y el estimador disten una distancia mayor a ϵ es cero, entre más grade sea el tamaño de muestra mejor será la aproximación del estimador al parámetro.

- **Suficiencia** Sea Y_1, Y_2, ... , Y_n, una muestra aleatoria de distribución de probabilidad con parámetro desconocido θ. Entonces se dice que el estadístico $U = g(Y_1, Y_2, ..., Y_n)$ es suficiente para θ si la distribución condicional de Y_1, Y_2, ... , Y_n dada U no depende de θ.

  Estas propiedades unida a la estimación insesgada de la varianza mínima por medio del teorema de Rao-Blackwell, demuestran a un estimador como un UMVUE, es decir de varianza mínima, insesgado y suficiente.

- **Sesgo y Error Cuadrático Medio** Cuando se realiza una estimación puntual, difícilmente una estimación obtenida de una sola muestra podrá reproducir identicamente el parámetro buscado, pero se busca que el valor esperado del estimador reproduzca en promedio al parámetro, Lohr (2009) define el sesgo de un estimador $\hat{t}$ como:

  $$B(\hat{T}) = E(\hat{T}) - T \tag{7}$$

Si $B\left[\hat{T}\right] = 0$ decimos que el estimador es insesgado para T. Se define a continuación el error cuadrático del error como el valor esperado del sesgo al cuadrado:

$$ECM\left[\hat{T}\right] = E\left[\hat{T} - t\right]^2 \tag{8}$$

A través de un sencillo paso matemático se puede concluir que

$$ECM\left[\hat{T}\right] = Var(\hat{T}) + B^2(\hat{T}) \tag{9}$$

En caso de que el estimador sea insesgado entonces el error cuadrático medio será la varianza del estimador.

3.1.8. Estimador de Horvitz y Thompson

Conocido también como el estimador π, fue publicado por primera vez en 1952 en la revista JASA (Journal of the American Statistical Association) y controversialmente fue publicado un año antes por Narain en una revista de baja circulación, este estimador actúa bajo el principio de representatividad, es decir, cada elemento de la muestra y_k se representa así mismo y a elementos del universo que no están en la muestra pero que presentan características similares al del elemento seleccionado. El estimador del total poblacional para Horvitz y Thompson está descrito por la selección:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} (10)$$

Donde y_k es el valor tomado para el elemento k, el término $\frac{1}{\pi_k}$ se conoce como el factor de expansión y es el inverso multiplicativo de la probabilidad de inclusión del elemento k-ésimo. Cuando las probabilidades de inclusión de son mayores a cero entonces el estimador de Horvitz y Thompson es insesgado para el total poblacional, es decir:

$$E(\hat{t}_{y,\pi}) = t_y$$

esto es sencillamente demostrable ya que por medio de la función indicadora y recurriendo al estimador de Horvitz y Thompson tenemos

$$E(\hat{t}_{y,\pi}) = E\left(\sum_U \frac{y_k}{\pi_k} I_k(S)\right) = \sum_U \frac{y_k}{\pi_k} E\left(I_k(S)\right) = \sum_U \frac{y_k}{\pi_k}\pi_k = \sum_U y_k = t_y$$

El subúndice U indica que la suma se hace sobre el universo. En cuanto a la varianza de este estimador se puede construir a partir de la misma definición del estimador con ayuda de la función indicadora y teniendo en cuenta las funciones indicadores de primer y segundo nivel:

$$Var_1\left(\hat{t}_{y,\pi}\right) = \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{11}$$

El subíndice 1 hace acotación a que no es la única función para hallar la varianza del estimador, cuando la muestra es de tamaño fijo, surge un segunda función para calcular la varianza del estimador y esta dada por la expresión:

$$Var_2\left(\hat{t}_{y,\pi}\right) = \sum\sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{12}$$

Se observa que para las funciones descritas anteriormente es necesario tener el total de la población para su cálculo, se muestran entonces a continuación para cada función de la varianza del estimador, una estimación de la varianza del estimador del total:

$$\widehat{Var}_1\left(\hat{t}_{y,\pi}\right) = \sum\sum_{S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{13}$$

Este primer estimador es para cuando no se tiene un tamaño de muestra fijo, cuando tenemos un muestreo con tamaño de muestra fija el estimador toma la forma:

$$\widehat{Var}_2\left(\hat{t}_{y,\pi}\right) = \sum\sum_{U} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{14}$$

Ambos estimadores son insesgados para la varianza del estimador, fácilmente demostrables al aplicar la función indicadora $I_k(S)I_l(S)$ ya que $E(I_k(S)I_l(S)) = \pi_{kl}$. Para concluir este importante punto, queda por hablar de la estimación pero no puntual sino por intervalo, [3] con lo cual se puede construir un intervalo de confianza dado en la siguiente expresión con nivel de confianza $(1-\alpha)$ para el total de la población :

$$C\left(1-\alpha\right) = \left[\hat{t}_{y,\pi} - z_{1-\alpha 2}\sqrt{Var\left(\hat{t}_{y,\pi}\right)}, \hat{t}_{y,\pi} + z_{1-\alpha 2}\sqrt{Var\left(\hat{t}_{y,\pi}\right)}\right] \tag{15}$$

Como se mencionó anteriormente se necesita tener el estudio de toda la población para obtener la varianza del estimador, más aun al reemplazar la varianza del estimador por su estimador se tiene:

$$C\left(1-\alpha\right) = \left[\hat{t}_{y,\pi} - z_{1-\alpha 2}\sqrt{\widehat{Var}\left(\hat{t}_{y,\pi}\right)}, \hat{t}_{y,\pi} + z_{1-\alpha 2}\sqrt{\widehat{Var}\left(\hat{t}_{y,\pi}\right)}\right] \tag{16}$$

Entonces al realizar un muestro escogiendo solo una muestra del conjunto de posibles muestras, se podrá estimar un intervalo de confianza en los cuales estará la estimación calculada.

3.1.9. Coeficiente de Variación Estimado (CVE)

El que un estimador sea o no preciso depende de la forma en que se planee toda la estrategia de muestreo, pero un buen indicador para esta precisión es el Coeficiente de Variación Estimado que es parecido en forma y contexto al coeficiente de variación estadística que se conoce comúnmente. Este coeficiente también estima la razón entre una variabilidad medida de los datos al calcular un parámetro con respecto a la estimación de dicho parámetro. De forma más exacta se pude calcular el Coeficiente de variación de la siguiente forma:

$$cve\left(\hat{t}\right) = \frac{\sqrt{\widehat{Var}(\hat{T})}}{\hat{T}} \tag{17}$$

3.2. Estrategias de Muestreo

Gutiérrez (2009) define una estrategia de muestreo como la dupla formada por un diseño de muestreo $p(\cdot)$ sobre un soporte Q y un estimador $\hat{T}$ de un parámetro T. Cabe destacar que cada diseño de muestreo trae implícita tanto la forma de distribución de la masa de probabilidad sobre todo el soporte como

[3] Gutiérrez (2009) propone que esto se cumple cuando el tamaño de muestra es suficientemente grande (y dependiendo del tamaño del comportamiento de la población puede bastar algunas docenas de individuos).

la forma propia que toma el estimador, se describirán en en esta parte los diseños de muestreo que se utilizarán acompañados del estimador de Horvitz-Thompson y la forma que toma bajo cada diseño.

3.2.1. Muestreo Aleatorio Simple (MAS)

Se distingue como uno de los diseños de muestreo básicos ya que supone homogeneidad en el valor de la características de interés de los elementos poblacionales, esto indica que cada una de las muestras pertenecientes a un soporte Q tendrá la misma probabilidad de selección. Además de ser un diseño básico es útil en dos aspectos, el primero es que se utiliza para comparar la eficiencia relativa con otros diseños de muestreo, la segunda ventaja que brinda el diseño MAS es su utilidad para la selección de unidades primarias en un muestreo por conglomerados.

En particular, cuando el diseño es de tamaño fijo las probabilidades de inclusión están dadas por:

$$p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#s = n \\ 0 & \text{en otro caso} \end{cases} \tag{18}$$

Como se vio anteriormente cuando de un universo de tamaño N y se desea escoger una muestra de tamaño fijo n, existen $\binom{N}{n}$ posibles muestras, sobre las cuales se distribuye toda la masa de probabilidad de tal forma que cada una de las muestras tenga una probabilidad $p(s)_k > 0$ para todo k perteneciente a la muestra y $\sum_{s \in Q} p(s) = 1$

- **Estimador Horvitz y Thompson para el MAS**
  Como se mencionó antes este estimador es exclusivo para diseños de muestreo sin reemplazo y depende de las probabilidades de inclusión de primer y segundo orden para su construcción, en este caso se define la probabilidad de inclusión de primer orden como:

$$\pi_k = Pr(I_k(S) = 1) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \tag{19}$$

  Para la probabilidad de inclusión de segundo orden tenemos bajo una analogía similar:

$$\pi_{kl} = Pr(k \in S \text{ y } l \in s) = \frac{n(n-1)}{N(N-1)} \tag{20}$$

  Por último tenemos la covarianza entre variables indicadoras definidas como:

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l = -\frac{n}{N^2}\frac{(N-n)}{(N-1)} & \text{para } k \neq l \\ \pi_k(1-\pi_k) = \frac{n(N-n)}{N^2} & \text{para } k = l \end{cases} \tag{21}$$

  Con estos resultados se obtiene el estimador de Horvitz y Thompson para el total, la varianza del estimador del total y el estimador de la varianza del total:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} = \sum_S \frac{y_k}{\left(\frac{n}{N}\right)} = \frac{N}{n}\sum_s y_k \tag{22}$$

$$Var_{MAS}\left(\hat{t}_{y,\pi}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yU} \tag{23}$$

$$\widehat{Var}_{MAS}\left(\hat{t}_{y,\pi}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yS} \tag{24}$$

con

$$S_{yU}^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - \overline{y}_U)^2 \quad y \quad S_{yS}^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \overline{y}_S)^2$$

La varianza poblacional y la varianza muestral respectictivamente

- **Efecto de Diseño** $Deff$
  Sobre la estimación del total poblacional, es importante verificar si el diseño de muestreo escogido con su respectivo estimador es eficiente con respecto a un diseño ya establecido; esto con el fin de comparar la pérdida o ganancia de eficiencia. El Diseño de Muestreo Aleatorio Simple acompañado del estimador Horvitz y Thompson para el total poblacional son el referente para comprobar dicha eficacia, el $Deff$ se define a continuación:

  Siendo $\left(\hat{T}, p(\cdot)\right)$ una estrategia de muestreo, acompañada de la estrategia $\left(\hat{T}_\pi, MAS\right)$ ambas utilizadas para la estimación del parámetro T se define el efecto de diseño:

$$Deff = \frac{Var_p(\hat{T})}{Var_{MAS}\hat{T}_\pi}. \tag{25}$$

  En el caso de este trabajo donde las estimaciones son para el total poblacional, el $Deff$ para el diseño de muestreo acompañado por el estimador HT tomará la forma:

$$Deff = \frac{Var_p(\hat{t}_{y,\pi})}{\frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{yU}^2}. \tag{26}$$

3.2.2. Tamaño de Muestra

Lohr (2009) especifica que el tamaño de muestra tiene una relación directa con el error de muestreo que el investigador desee asumir; y para esta relación se debe encontrar un modelo matemático que le permita relacionar el tamaño de muestra con dicho error de muestreo, para esto (y cuando no se tienen información valida de estudios anteriores) debe realizar un estudio piloto el cual no necesita tanta precisión y servirá para medir tanto la efectividad de la encuesta, como para determinar el tamaño de muestra necesario. En caso de estrategias de muestreo con probabilidades simples para media poblacional, la precisión deseada se tiene en términos de error absoluto:

$$P\left(|\overline{y} - \overline{y}_U| \leqslant e\right) = 1 - \alpha$$

Donde e es llamado margen de error en la mayoría de encuestas y es el investigador quien debe dar valores razonables dependiendo del presupuesto y los objetivos para α y e. La forma más sencilla de relacionar la precisión y el tamaño de la muestra es con ayuda del intervalo de confianza y buscando un valor de n que satisfaga la expresión:

$$e = z_{\alpha 2}\sqrt{\left(1-\tfrac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}}$$

Ecuación que al resolver para n quedará expresada como:

$$n \geqslant \frac{n_o}{1+\frac{n_o}{N}} \tag{27}$$

con $n_o = \left(\frac{z_{\alpha/2}S_{yU}}{e}\right)^2$. La función ***ss4m*** del paquete ***samplesize4surveys*** del software estadístico R, ayuda a calcular el tamaño de la muestra mínimo con errores predefinidos. En el desarrollo de este trabajo se especificará más sobre esta función.

3.2.3. Muestreo Proporcional al Tamaño de Información Auxiliar πPT

Catalogado por muchos como el mejor diseño de muestreo cuando existe información auxiliar de tipo continuo que esté altamente correlacionada con la característica de interés en el estudio, este muestreo asigna probailidades de inclusión proporcionales al tamaño de la característica de información auxiliar, es decir:

$$\pi_k = \frac{nx_k}{t_x} \tag{28}$$

Gutiérrez (2009, pág 135) indica que este diseño de muestreo es una generalización de diseños de muestreo sin reemplazo ya que si la característica de información auxiliar es constante para todos y cada uno de los elementos de la población se tendría que:

$$\begin{aligned}\pi_k &= \frac{nx_k}{t_x} \\ &= \frac{nC}{NC} = \frac{n}{N}\end{aligned}$$

Con lo que se tiene un diseño de muestreo con probabilidades simples. Por otro lado tenemos que la razón $\frac{nx_k}{t_x}$ puede ser mayor a uno para algunos elemento, a dichos elelmentos se les conoce como **elementos de inclusión forzosa**, lo que indica que se debe excluir a dichos elementos del listado de elementos e iniciar nuevamente el cálculo de las probabilidades de selección de la siguiente manera:

$$\pi_k = \frac{(n-n^*)x_k}{\sum_{k\in U*} x_k} \qquad 0 < \pi_k \leqslant 1; \quad k \in U^* \tag{29}$$

Donde n^* corresponde al número de elementos de inclusión forzosa y U^* a los elementos de la población excluyendo a los de inclusión forzasa. A los elementos de inclusión forzosa se les asigna un $pi_k = 1$ de tal forma que se tenga al final $\sum_{k\in U} \pi_k = n$.

- **Estimador Horvitz y Thompson para el πPT**
  Como se mencionó antes el diseño de muestreo πPT es un diseño generalizado de los muestreos sin reemplazo, por tanto, el estimador para el total $\hat{t}$, la varianza de este estimador del total $Var_{\hat{t}_{y,\pi}}$ y el estimador de la varianza $\widehat{Var}_{\hat{t}_{y,\pi}}$ están dados por las siguientes expresiones:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \tag{30}$$

$$Var_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{31}$$

$$\widehat{Var}_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{32}$$

  Gutiérrez (2009, pág 144) afirma que las probabilidades de inclusión de segundo orden, aunque servirán teóricamente para calcular y estimar la varianza del estimador de Horvitz-Thompson, son ineficientes pues cuando el tamaño de muestra crece, su cálculo se vuelve un trabajo tedioso y complicado, en muchos casos imposible de acabar, autores como Tillé proponen estimaciones de las varianzas que no requieren el cálculo de las probabilidades de inclusión de segundo orden

para la familia de diseños exponenciales, la aproximación de la varianza del estimador de Horvitz-Thompson está dada por:

$$Var(\hat{t}_{y,\pi}) = \sum_{k \in U} \frac{b_k}{\pi_k^2}(y_k - y_k^*)^2 \tag{33}$$

donde

$$y_k^* = \pi_k \frac{\sum_{l \in U} b_l y_l/\pi_l}{\sum_{l \in U} b_l} \tag{34}$$

Para escojer el b_k se utiliza el estimador propuesto por Hájek, Gutiérrez (2009, Pagina 145):

$$b_k = \frac{N\pi_k(1-\pi_k)}{(N-1)} \tag{35}$$

Un estimador de la anterior aproximación de la varianza está dada por

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{k \in S} \frac{c_k}{\pi_k^2}(y_k - \hat{y}_k^*)^2 \tag{36}$$

donde

$$\hat{y}_k^* = \pi_k \frac{\sum_{l \in S} c_l y_l/\pi_k}{\sum_{l \in S} c_l} \tag{37}$$

En este caso Gutiérrez (2009) utiliza el estimador porpuesto por Deville (1993) para escogenr a c_k

$$c_k = (1-\pi_k)\frac{n}{(n-1)} \tag{38}$$

El paquete ***TeachingSampling*** del software estadístico R construido por el Ph.D. Andrés Gutiérrez implementa este cálculo para la estimación de la varianza.

La librería ***survey*** del Software estadístico R con una función ***svydesign*** que computa un diseño a dos etapas y obtiene la estimación de totales con la función ***svytotal***, utiliza el método de Brewer, Lumley (2012) en su escrito en "Describing PPS desingns to R", define un tipo de aproximación de la varianza para diseños de muestreo con probabilidades desiguales donde se debe presentar una aproximación para π_{ij} para el cálculo de la expresión Δ_{ij} en el cálculo de la varianza estimada se puede realizar la siguiente estimación truncada de Hartley-Rao:

$$\Delta_{ij} = 1 - \frac{n - \pi_i - \pi_j + \sum_{k=1}^{N} \frac{\pi_k^2}{n}}{n-1} \tag{39}$$

Como se ve en la anterior ecuación lo único que se necesita para conocer el valor estimado de Δ_{ij} es el valor de los π_i. El cálculo poblacional puede estimarse a partir de la muestra con la siguiente estimación:

$$\Delta_{ij} = 1 - \frac{n - \pi_i - \pi_j + \sum_{k=1}^{n} \frac{\pi_k}{n}}{n-1} \tag{40}$$

3.2.4. Muestreo Estratificado

Cuando se utiliza información de tipo auxiliar, no para calcular las probabilidades de inclusión de cada individuo como se realizaba en la estrategia de muestreo πPT, sino para clasificar a los individuos en grupos, se habla de una estrategia de muestreo estratificado. Esta clasificación se hace con la idea de agurpar a los individuos de tal forma estos tengan características semejantes dentro del grupo (varianza pequeña al interior del grupo), pero que difieran entre grupos (varianzas grandes entre los grupos formados).

Los grupos formados reciben el nombre de **estratos** y cumplen dos condiciones que los definen; la primera es que la unión de los estratos debe formar el Universo de estudio (todos los indiviuos deben pertenecer a algún estrato), y la segunda es los estratos sean mutuamente excluyentes (ningún individuo puede pertenecer a dos o más estratos a la vez). Dicho de otra manera:

- $\bigcup_{h=1}^{H} U_h = U$
- $U_h \bigcap U_i = \varnothing \qquad h \neq i$

El tamaño de cada estrato U_h es N_h, entonces:

$$\sum_{h=1}^{H} N_h = N \tag{41}$$

El total población es la suma de los totales poblacionales dentro de cada estrato, es decir:

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} t_{yh} \tag{42}$$

donde $t_{yh} = \sum_{k \in U_h} y_k$

Gutiérrez (2009) comenta que dependiendo de la naturaleza de cada estrato se puden aplicar diferentes estrategias de muestreo en diferentes estratos y aprovechar las ventajas de la información auxiliar dentro de cada uno de ellos he incluso si es necesario realizar censo dentro de algún estrato. De manera independiente se pueden seleccionar las muestras dentro de cada estrato, así la muestra aleatoria S quedará definida como la unión entre las posibles muestras S_h de cada estrato:

$$S = \bigcup_{h=1}^{H} S_h. \tag{43}$$

De manera obvia la unión de las realizaciones de la muestra s_h formará la realización s

$$s = \bigcup_{h=1}^{H} s_h. \tag{44}$$

En este orden de ideas, si el tamaño de muestra en cada estrato es n_h, entonces el tamaño de la muestra seleccionada será:

$$n = \sum_{h=1}^{H} n_h. \tag{45}$$

Gutiérrez (2009) indica que el soporte Q dentro de una estrategia de muestreo estratificado es el producto cartesiano de los soportes dentro de cada estrato:

$$Q^H = \bigtimes_{h=1}^{H} Q_h. \tag{46}$$

La cantidad total de posibles nuestras (cardinalidad de Q), será entonces el producto de la cantidad de posibles muestras dentro de cada estrato:

$$\#Q^H = \prod_{h=1}^{H} \#Q_h. \tag{47}$$

- **Estimador Horvitz y Thompson para el Muestreo Estratificado**
  Uno de los objetivos del muestreo estratificado es obtener mejores estimaciones sobre alguna variable de interés, en dicho caso, si dentro de cada estrato $\hat{t}_{yh}$ es un estimador insesgado para la característica de interés t_{yh}, dentro del estrato h; entonces el estimador para el total de la población es la suma del estimado dentro de cada estrato y la varianza de dicho estimador también será la suma de las varianzas de cada estrato, es decir:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} \qquad Var(\hat{t}_y) = \sum_{h=1}^{H} Var(\hat{t}_{yh}) \tag{48}$$

  De esta forma el estimador de Horvitz y Thompson para el total, su varianza y la estimación de la varianza para el muestreo estratificado quedarán expresados como:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} \tag{49}$$

$$Var_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} Var_{p_h}(\hat{t}_{yh,\pi}) \tag{50}$$

$$\widehat{Var}_{EST}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \widehat{Var}_{p_h}(\hat{t}_{yh,\pi}) \tag{51}$$

  donde

$$\hat{t}_{yh,\pi} = \sum_{k \in S_h} \frac{y_k}{\pi_k} \tag{52}$$

  con $Var_{p_e}(\hat{t}_{yh,\pi})$ como la varianza de $\hat{t}_{yh,\pi}$ en el h-ésimo estrato y $\widehat{Var}_{p_h}(\hat{t}_{yh,\pi})$ la estimación de $Var_{p_h}(\hat{t}_{yh,\pi})$ en el h-ésimo estrato.

3.2.5. Muestreo aleatorio estratificado

Conocido como el muestreo ESTMAS, este diseño, dentro de cada estrato da probabilidades de inclusión iguales a cada uno de los individuos dentro de cada estrato, es decir, escoge dentro de cada estrato una muestra asletoria simple de tamaño n_h, puesto que dentro de cada estrato el comportamiento de la característica de interés es homogéneo, no es necesario recoger toda la información dentro del estrato, sino una catidad de individuos que informen sobre el comportamiento de la variable de interés dentro del estrato.

Dentro del muesteo aleatorio estratificado simple sin reemplazo con tamaños de muestra por estrato $n_1, n_2, ..., n_H$ fijos, la probailidad de seleccionar una muestra de tamaño n está dada por:

$$p(s) = \begin{cases} \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}}, & \text{si } \sum_{h=1}^{H} n_h = n \\ 0, & \text{en otro caso} \end{cases} \tag{53}$$

cabe denotar que $\sum_{s \in Q^H} p(s) = 1$ porque $\#Q^H = \prod_{h=1}^{H} \binom{N_h}{n_h}$.

El estimador de Horvitz y Thompson para un muestreo aleatorio estratificado
Para desarrollar esta parte es necesario detallar la construcción de las probabilidades de inclusión de primer y segundo orden. Las probabilidades de inclusión de primer orden están dadas por:

$$\pi_k = \frac{n_h}{N_h} \quad \text{si } k \in U_h \tag{54}$$

las probabilidades de inclusión de segundo orden están dadas por:

$$\pi_{kl} = \begin{cases} \dfrac{n_h}{N_h}, & \text{si } k = l, k \in U_h, \\ \dfrac{n_h}{N_h}\dfrac{n_h - 1}{N_h - 1}, & \text{si } k, l \in U_h, \\ \dfrac{n_h}{N_h}\dfrac{n_i}{N_i}, & \text{si } k \in U_h, l \in U_i, i \neq h. \end{cases} \tag{55}$$

respectivamente. La covarianza de las variables indicadoras está dada por

$$\Delta_{kl} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{N_h - n_h}{N_h}, & \text{si } k = l, k \in U_h, \\ -\dfrac{n_h}{N_h^2}\dfrac{(N_h - n_h)}{(N_h - 1)}, & \text{si } k, l \in U_h, \\ 0, & \text{si } k \in U_h, l \in U_i, i \neq h. \end{cases}$$

(56)

Con estas probabilidades de inclusión se puede definir el estimador de Horvitz y Thompson para el muestreo aleatorio estratificado, la varianza de dicho estimador y la estimación de la varianza de la siguiente forma:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} y_k \tag{57}$$

$$Var_{MAE}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{yU_h}^2 \tag{58}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) S_{ys_h}^2 \tag{59}$$

Método de estratificación Lavalle-Hidiriglou

Cuando tenemos una población que cuenta con información auxiliar que permite dividir dicha población en grupos mutuamente excluyentes, el implantar una estrategia de muestreo estratificada es ideal ya que permitirá aumentar la precisión de las estimaciones, el problema es como dividir la población de forma tal que la estrategia sea más eficiente y aun más cuando dicha población presenta sesgo en su distribución. Raj (1980) indica que el número estratos a crear dentro de una población depende de la intuición que tenga el analista o estadístico sobre la estructura de población, es decir, no hay un algoritmo que indique según la característica poblacional el número de estratos a crear, pero se sabe que entre mas estratos se creen la precisión de la estimación aumentará pero lo hace a la par que los costos de levante de la infromación ya que se requerirá más unidades muestrales encuestadas.

Una vez son determinados por el analista o estadístico el núemro de estratos L, cada uno contiene N_h elementos de la población, de cada estrato se desea escoger una muestra n_h de elementos, con $h = (1, 2, ..., L)$. Gunning & Horgan (2004) indica que el método de estratificación de Lavallee-Hidiroglou es un método iterativo especializado para poblaciones sesgadas que consiste en un "$Take - all$" en todos los estratos acompañado de un número de "$Take - some$" en algunos estratos de tal manera que el

tamaño de la muestra se reduce al mínimo para un determinado nivel de fiabilidad. De esta forma se desea según Lavalleé & Hidiroglou (1988) se desea encontrar un algoritmo iterativo de cuatro pasos límites para cada estrato, tomando el primer paso como arbitrario, se resume en la siguiete secuencia:

- PASO 1: Empezar con algunos límites arbitrarios $b^{'}_{(1)} < ... < b^{'}_{(L-1)}$.
- PASO 2: Calcular las proporciones $W^{'}_{h}$ con medias u_h y varianza S^2_h basados en dichos límites con $h = 1, ..., L-1$.
- PASO 3: Reemplace el conjunto inicial de límites por $b^{''}_{(1)} < ... < b^{''}_{(L-1)}$ donde:

$$b^{''}_{(1)} = \frac{-B^{'}_{h} + \sqrt{B^{'}_{h} - 4a^{'}_{h}Y^{'}_{h}}}{2a^{'}_{h}}$$

- PASO 4: Repita los pasos 2 y 3 hasta que dos conjuntos consecutivos de límites sean iguales o diferenciable por cifras insignificantes.

Para esto basta con identificar a $W_h = \frac{N_h}{N}$, $B = \sum_{h=1}^{L-1}(W_h S_h)^2 (W_h u_h)^2$, los términos para a y Y están comprendidas dentro del cálculo de los límites y las ponderaciones que se estimen con ayuda del CVE que desee el usuario para su estudio.

3.2.6. Muestreos Bietápicos

Tal como su nombre lo indica es un muestreo que se realiza en dos etapas, tiene como propósito estimar el total de la caracteristica de interés t_i en cada uno de las agrupaciones tomadas a través de una submuestra tomada dentro de la agrupación seleccionada de la población. Bajo este esquema, este muestreo toma las siguientes consideraciones.

- UPM (Unidades Primarias de Muestreo), es la división que se hace sobre la población inicial, teniendo en cuenta carácterísticas de homogeneidad, agrupamientos naturales o estratificación con ayuda de información auxiliar disponible.
- USM (Unidades Secundarias de Muestreo), son cada una de las unidades muestrales dentro de cada una de las Unidades Primarias de muestreo.

El fundamento básico en el muestreo de dos etapas (y en general para un muestreo de n etapas), es hacer que las estimaciones vengan desde las unidades secundarias de muestreo hasta la primera etapa, para esto el muetreo en dos etapas debe cumplir las siguientes propiedades que Gutiérrez (2009) enuncia así:

1. **Invariancia:** La probabilidad de selección de una muestra de unidades en la segunda etapa no depende del diseño de muestreo de la primera etapa.
2. **Independencia:** Cuando se realiza el muestreo en la segunda etapa esta se llava a cabo de manera independiente con las otras unidades de muestreo, puede ser en la misma etapa o en la prtimera etapa.

En un muestreo en dos etapas se debe contar con un marco muestral compuesto es decir de antemano apra la promera etapa se tienen que la población U se dividió en N_I partes diferentes, es decir, $U_I = (U_1, ..., N_I)$, entonces la probabilidad de la realización de la muestra s_I de S_I es $Pr(S_I = s_I) = p_I(s_I)$. Para cada muestra escogida en la primera etapa se selecciona una muestra s_i dentro del conjunto de posibles muestras S_i, tal que la probabilidad de selección de dicha muestra es $Pr(S_i = s_i) = p_i(s_i)$.

Gracias a la propiedad de independencia tenemos que la probabilidad de escoger la muestra s_i es independiente de la selección de cualquier muestra en s_I de la primera etapa es decir:

$$Pr(S_i = s_i \mid S_I = s_I) = Pr(S_i = s_i). \tag{60}$$

Teniendo en cuenta esto el diseño de muestreo bietápico cumple con las dos condiciones para ser un diseño de muestreo, es decir:

1. $p(s) \geqslant 0$ para todo $s \in Q$
2. $\sum_{s \in Q} p(s) = 1$

Como en los demás casos el interés es obtener el total poblacional, para el muestreo bietápico tenemos:

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \sum_{i=1}^{N_I} t_{yi} \tag{61}$$

donde $t_{yi} = \sum_{k \in U_i} y_k$ es el total de la i-ésima unidad primaria de muestreo $i = 1, \ldots, N_I$.

Estimador de Horvitz y Thompson para muestreos bietápicos

Para determinar el Estimador HT en muestreos bietápicos Gutiérrez (2009) enmarca el hecho de hacer diferencia entre las probabilidades de selección de primer orden y segundo orden en cada etapa del muestreo. Para la primera etapa de muestreo se tiene:

$$\Delta_{Iij} = \begin{cases} \pi_{Iij} - \pi_{Ii}\pi_{Ij}, & \text{si } i, j \in U_I, \\ \pi_{Ii}(1 - \pi_{Ii}), & \text{si } i = j \in U_I. \end{cases} \tag{62}$$

Y para la segunda etapa tenemos:

$$\Delta_{kl|i} = \begin{cases} \pi_{kl|i} - \pi_{k|i}\pi_{l|i}, & \text{si } k \neq l, \\ \pi_{k|i}(1 - \pi_{k|i}), & \text{si } k = l. \end{cases} \tag{63}$$

Donde la probabilidad de inclusión de primer orden para el $k - ésimo$ elemto de la población esta dado por:

$$\begin{aligned} \pi_k = Pr(k \in S) &= Pr(k \in S_i \text{ y } i \in S_I) \\ &= Pr(k \in S_i | i \in S_I) Pr(i \in S_I) = \pi_{k|i}\pi_{Ii} \end{aligned} \tag{64}$$

La probabilidad de inclusión de segundo orden para los elementos k y l está dada por

$$\pi_{kl} = \begin{cases} \pi_{Ii}\pi_{k|i}, & \text{si } k = l \in U_i, \\ \pi_{Ii}\pi_{k|i}, & \text{si } k \neq l \in U_i, \\ \pi_{Iij}\pi_{k|i}\pi_{l|j}, & \text{si } k \in U, l \in U_j (i \neq l). \end{cases} \tag{65}$$

Bajo muestreo en dos etapas el estimador de Horvitz-Thompson toma la forma

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \sum_{k \in S_i} \frac{y_k}{\pi_{Ii}\pi_{k|i}} = \sum_{i \in S_I} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \tag{66}$$

La varianza de dicho estimador se expresa como

$$Var_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum_{U_I} \sum \Delta_{Iij} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}}}_{Var(UPM)} + \underbrace{\sum_{i \in U_I} \frac{Var_{p_i}(\hat{t_i})}{\pi_{Ii}}}_{Var(USM)} \tag{67}$$

La estimación insesgada para la varianza es

$$\widehat{Var}_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum_{S_I} \sum \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \frac{\hat{t}_{yj,\pi}}{\pi_{Ij}}}_{\widehat{Var}(UPM)} + \underbrace{\sum_{i \in S_I} \frac{\widehat{Var}(\hat{t}_{yi,\pi})}{\pi_{Ii}}}_{\widehat{Var}(USM)} \tag{68}$$

donde la varianza para las unidades secundarias de muestreo (USM) esta dada po

$$Var(\hat{t}_i) = \sum \sum_{U_i} \Delta_{kl|i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \tag{69}$$

$$\hat{t}_{yi,\pi} = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}$$

representando la estimación del total de la característica de interés en la i-ésima unidad primaria de muestreo y

$$\widehat{Var}(\hat{t}_i) = \sum \sum_{S_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \tag{70}$$

La variación del estimador se descompone en las para cada una de las dos etapas del muestreo

3.2.7. Muestreo bietápico πPT-MAS

Tal como su nombre lo indica este diseño de muestreo se caracteriza por que en su primera etapa utliza información de tipo auxiliar para las unidades primarias de muestreo (UPM) de tal manera que estas unidades tienen probabilidades de inclusión proporcional al tamaño de la característica de interés, luego de tener seleccionadas las unidades primarias que serán muestreadas, sobre cada una de dichas unidades se realizará un muestreo aleatorio simple (MAS) en las unidades secundarias de muestreo. Para obtener las estimaciones del total, la varianza de dicho estimador y el estimador de esta varianza Buitrago Chinchilla (2015) toma por aparte la estimación para la primera etapa y luego para la segunda etapa y las combina para dar el total estimado es el muestreo bietápico:

Sean $\hat{t}_i = \frac{N_i}{n_i} \sum_{s_i} y_k$ la estimación para el total de la primera en la segunda etapa y $\hat{t}_{y,\pi} = \sum_{s_I} \frac{\hat{t}_i}{\pi_{Ii}}$ el total estimado para este muetreo será:

$$\hat{t}_{\pi PT-MAS} = \frac{N_i}{n_i} \sum_{s_i} y_k \sum_{s_I} \frac{\hat{t}_i}{\pi_{Ii}} \tag{71}$$

De forma similar se presenta la forma para calcular la varianza de dicho estimador, empezando con la varianza de la segunda etapa:

$()=\mathrm{N}_i^2 \frac{}{n_i^2 (1-\frac{n_i}{N_i}) S^2_{s_{i,y}}}$ (72)

Donde $s_{i,y} = \frac{1}{n_i - 1} \sum_{s_i} (y_k - \hat{t})^2$, siendo $\hat{t} = \frac{1}{n_i} sum_{s_i} y_k$

El cálculo para la varianza en de la primera etapa junto con el de la segunda etapa formará la ecuación para el cálculo de la varianza del estimador:

$$Var_{\hat{t}_{\pi PT-MAS}} = \sum \sum_{U} \Delta_{Iij} \frac{t_j}{\pi_{Ij}} \frac{t_i}{\pi_{Ii}} + \sum_{U_I} \frac{V_i}{\pi_{Ii}} \tag{73}$$

Y el estimador para la varianza del estimador estará dada por:

$$\widehat{Var}_{\hat{t}_{\pi PT-MAS}} = \sum \sum_{U} \frac{\Delta_{Iij}}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}} \frac{t_i}{\pi_{Ii}} + \sum_{U_I} \frac{V_i}{\pi_{Ii}} \tag{74}$$

Como se mencionó anteriormente la expresión Δ_{Iij} requeriría una suma iterativa muy complicada incluso para las ordenadores, reemplazando esta expresión por alguna de los estimadores visto en el muestreo πPT tenemos:

$$\sum_U \frac{\Delta_{Iij}}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}} \frac{t_i}{\pi_{Ii}} = \frac{c_i(\hat{t}_{y,i} - \hat{t}_i^*)}{\pi_i^2}$$

Con $\hat{t}_i^* = \pi_i \frac{\frac{\sum_{j \in s_I} c_j \hat{t}_i}{\pi_{Ii}}}{\sum_{j \in s_I} c_j}$ y con $c_i = (1 - \pi_i)\frac{n_i}{n_i - 1}$

3.2.8. Muestreo bietápico ESTMAS-MAS

Este muestreo bietápico toma la información auxiliar para dividir las unidades primarias de la población en grupos o estratos, tomando con un diseño aleatorio simple muestras en cada estrato, una vez escogidas las unidades primarias de nuestreo con el diseño $ESTMAS$, tomará de cada unidad primaria seleccionada muestras por medio de un muestreo aleatorio simple MAS.

Sin más preambulos el estimador HT para el total en un diseño ESTMAS-MAS está dado por la expresión:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \left[\frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i}{n_i} \sum_{k \in S_i} y_k \right] \tag{75}$$

La varianza de este estimador esta denotada por:

$$\begin{aligned} Var_{EMM}(\hat{t}_{y,\pi}) &= \sum_{h=1}^{H} Var(\hat{t}_{yh,\pi}) \\ &= \sum_{h=1}^{H} \left[\frac{N_{Ih}^2}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{t_{yh} U_I} + \frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{y_{U_i}} \right] \end{aligned} \tag{76}$$

El estimador insesgado para esta varianza será:

$$\begin{aligned} \widehat{Var}_{EMM}(\hat{t}_{y,\pi}) &= \sum_{h=1}^{H} \widehat{Var}(\hat{t}_{yh,\pi}) \\ &= \sum_{h=1}^{H} \left[\frac{N_{Ih}^2}{n_{Ih}} \left(1 - \frac{n_{Ih}}{N_{Ih}}\right) S^2_{\hat{t}_{yh} S_I} + \frac{N_{Ih}}{n_{Ih}} \sum_{i \in S_{Ih}} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{y_{S_i}} \right] \end{aligned} \tag{77}$$

Särndal et al. (2003) asegura que existe la posibilidad de agrupar la población de acuerdo a una medida de tamaño, en la que dicha agrupación es dada por un comportamiento similar y asociadas en un mismo estrato, esta idea permite construir estratos que puedan llegar a ser mas homogéneos hacia el interior por la configuración de la variable o las observaciones que lo componen.

4. CONSTRUCCIÓN DE LOS ESCENARIOS

La base de datos de viviendas, hogares y personas (VIHOPE) es una base de datos creada por el DANE en 2005 después del último censo nacional, detalla a nivel de manzana censal (área geográfica), en esta base se detalla el total de viviendas que existen en dicha área, hogares que hay dentro del área geográfica y el total de personas ocupan el área ya mencionada. El registro por filas contienen el código del área geográfica, el código del departamento, el código del municipio, el código de clase de área geográfica (cabecera municipal, centro poblado o rural disperso), el código de la localidad (comuna) en el caso que aplique, la identificación de la manzana, el total de viviendas, el total de hogares y el total de personas por área geográfica entre otras.

El diccionario de los nombre codificados para las variables se encontraron en el informe de resultados-DANE (2014), la base de datos donde se encuentran los registros recibe el nombre *TOTEX7 VIHOPE 05MAR09* y la base donde se encuentran los códigos para departamentos, ciudades y demás se llama *División Político Administrativa de Colombia* mas conocida como *DIVIPOLA*, dichas bases pueden ser consultdas en la página del DANE[4].

4.0.1. Depuración de la base de datos y descripción de las variables

En principio la base de datos contaba con un total de 349.769 registros cuya llave I0E AG (Identficación del área geográfica) se encontraba repetida para varios registros, se decidió dejar un único registro por área geográfica que totalizara las viviendas, hogares y personas por hogares; en este paso se fusionaron 6.581 registros (1,88 %), quedando un total de 343.188 registros.

Tabla 1: *Total Variables de Interés en la Base de Datos.*

	Total
Áreas Geográficas	343.188
Total Viviendas	10.531.042
Total Hogares	10.673.684
Total Personas	41.834.837

En el reporte generado en marzo 17 de 2014 por la Dirección de Censos y Demografía - DCD en donde se muestra la información general del censo de 2005, se encuentra un diccionario para las variables que componen la base de datos.

Se presenta a continuaciónen la tabla 2, la distribución por clase de área geográfica de las variables total viviendas, total hogares y total personas.

Tabla 2: *Distribución de los datos por Clase de Área Geográfica*

	Total Áreas	Total Viviendas	Total Hogares	Total Personas
Cabecera municipal	277.597	7.927.866	8.764.555	31.762.408
Centro Poblado	47.255	587.850	878.555	2.403.848
Rural disperso	18.336	2.015.326	1.030.574	7.668.581

En el anexo A de este trabajo se podrá encontrar en la tabla 11 la distribución de las variables por departamento.

[4] http://www.dane.gov.co

Características de la Variable de total de personas por área geográfica

Se decidió escoger la variable total personas como variable auxiliar para las dos estrategias de muestreo que enmarcan este trabajo, dichas estrategias de muestreo corresponden a estrategias de muestreo bietápicas, en donde se consideraron como unidades primarias de muestreo (UPM) a los municipios y como unidades secundarias de muestreo (USM) a cada una de las áreas geográficas. Por un lado la variable de interés se utilizó para estratificar la población por municipios para la estrategia de muestreo EST-MAS y por otro lado fue la variable de información auxiliar que dará a cada municipio la probabilidad de selección proporcional a esta variable, dentro del muestreo πPT. Las estadísticas descriptivas de esta variable se muestran en la tabla 3:

Tabla 3: *Estadística descriptivas de la variable Total Personas*

Medidas de posicionamiento						
Mínimo	1er Cuartil	Media	Mediana	Moda	3er Cuartil	Máximo
1	35	121,9	73	5	135	61260
Medidas de Dispersion						
		Varianza	Curtosis	Asimetría		
		82129,12	8182,77	56		

En la tabla 3 se puede ver la altísima asimetría de esa variable, se apuntala a valores fuertemente en valores pequeños, junto con la asimetría la variable total de personas presenta una curtosis indicación de valores extremos grande. La figura 1 dará una idea general de dicho comportamiento

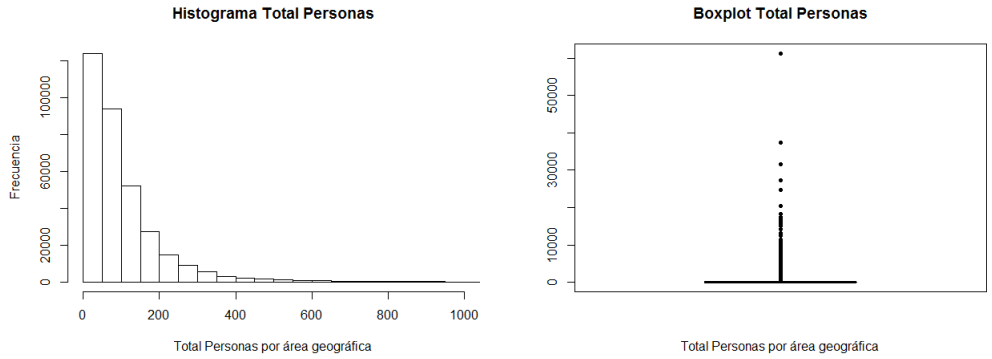

Figura 1: *Histograma y Boxplot de la variable total personas*

En la tabla 11 se puede encontrar la distribución de la variable total de personas por cada departamento del país.

4.0.2. Creación de los Escenarios

El objetivo de este trabajo como ya se ha mencionado es comparar dos estrategias de muestreo en diferentes escenarios creados. En primer lugar se construyeron 49 escenarios cada uno variando la correlación de la variable de interés (variable simulada) y la variable que utilizaremos como variable de información auxiliar (total personas por área), además de variar el intercepto del modelo lineal que las relaciona; cada escenario está representado por una combinación única de estos aspectos, como lo representa la tabla 4.

Tabla 4: Creación de Escenarios

	CORRELACION (ρ)								
INTERCEPTO (β_0)	99 %	97 %	95 %	93 %	90 %	70 %	50 %	30 %	10 %
10 %	E1_1	E1_2	E1_3	E1_4	E1_5	E1_6	E1_7	E1_8	E1_9
30 %	E2_1	E2_2	E2_3	E2_4	E2_5	E2_6	E2_7	E2_8	E2_9
50 %	E3_1	E3_2	E3_3	**E3_4**	E3_5	E3_6	E3_7	E3_8	E3_9
70 %	E4_1	E4_2	E4_3	E4_4	E4_5	E4_6	E4_7	E4_8	E4_9
90 %	E5_1	E5_2	E5_3	E5_4	E5_5	E5_6	E5_7	E5_8	E5_9

Antes de hablar de como se contruyeron los escenarios, es preciso aclarar que dentro de un muestreo bietápico existen, como se mencionó en el marco teórico unidades primarias de muestreo (UPM) y unidades secundarias de muestreo (USM), para el caso de este trabajo las unidades primarias serán los municipios reconocidos por el DANE como tales, un total de 1.114 municipios distribuidos en 33 departamentos. Las unidades secundarias de muestreo son en este trabajo cada una de las 343.188 áreas geográficas. En la tabla anterior el punto **E3_4** representa una correlación del 93 % y un intercepto al 50 %.

La figura 2 muestra como se realizan cada una de las estrategias de muestreo:

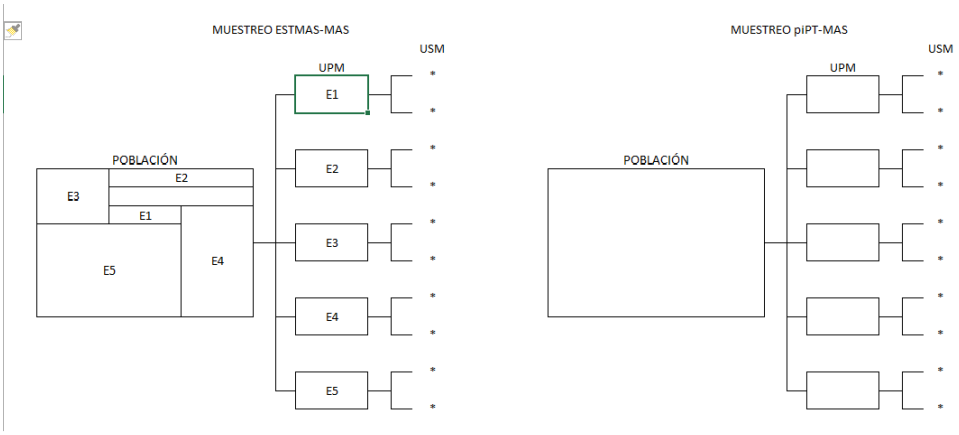

Figura 2: *Esquemas para cada una de las estrategias de muestreo utilizadas*

Los requerimientos más importantes para la estrategia de muestreo πPT es que la correlación entre variable auxiliar y variable auxiliar sea alta, ademas que el intercepto del modelo que las asocia sea pequeño, e esta manera los 49 escenarios representan cada una de las 49 posibles combinaciones entre correlación e intercepto, por ejemplo E3_4 representa el escenario en el cual las variables se correlaciónan en un 50 % y la razón entre el intercepto y la media de la variable de interés es del 70 %.

Para crear cada uno de los 49 escenarios se recurrió a un algoritmo por el cual se trazaba un modelo lineal acompañado de una variable aleatoria:

$$Y = \beta_0^{'} + \beta_1^{'} X + \epsilon \qquad con \quad \epsilon \sim Norm\left(\mu = 0, \sigma^2 = \alpha\right)$$

para evitar que la variable de interés tome valores negativos, se condiciona al algoritmo para que retome en la construcción un valor positivo

Se calibraron los valores del intercepto $\beta_0^{'}$ y la varianza α hasta conseguir los valores de la correlación e intercepto requeridos. En el Anexo C de este trabajo encontrará el código para el software estadístico R con el cual se crearon los escenarios.

Se muestran a continuación en la figura 3 el gráfico de las variables construidas para la correlación al 90 % y siguiendo este procedimiento podrá encontrarse en el Anexo D de este trabajo, los gráficos para las demás distribuciones.

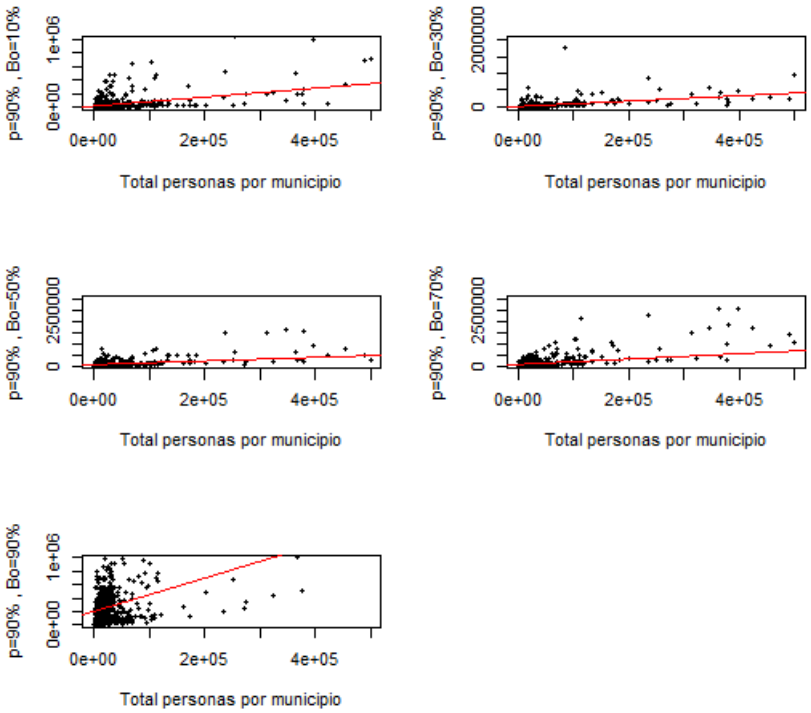

Figura 3: *Distribución de las variables de interés correlacionadas al 90 %*

5. MUESTREO BIETÁPICO ESTMAS-MAS

Se decidió empezar por esta estrategia de muestreo ya que será quien impondrá las pautas (a partir de esta se calculó el tamaño de muestra óptimo para la primer y segunda etapa). Para iniciar este muestreo se decidió dividir con ayuda de la información auxiliar las unidades primarias de muestreo en 5 estratos con ayuda del paquete ***stratification*** del software estadístic R, este paquete creado por Baillargeon & Rivest (2011), es una serie de funciones que estratifican variables unidimensionales por el método generalizado de Lavallee Hidiroglou, aunque si se desea tambien puede emplear los métodos de la raíz de las función de frecuencias acumulativas de Danelius y Hodge o el método de la regla geométrica de Gunning and Horgan.

En la figura 4, podrá verse la salida de la strata.lh para la estratificación de los municipios por total de personas.

La función LH requiere como entrada la variable de datos por al cual se va a estratificar, el tamaño de la muestra objetivo, el coeficiente de variación estimado (sea el tamaño de la muestra o el coeficiente de variación estimado deben ser introducidos). El número de estratos (por defecto la función toma como tamaño mínimo 3).

La salida del programa genera: bh vector de límites (límite superior) óptimos encontrados por la función.

Nh vector que genera el tamaño poblacional dentro de cada estrato.

nh vector que genera el tamaño de la muestra dentro de cada estrato.

$E(Y)$ y $Var(Y)$ vectores de las medias y varianzas dentro de cada estrato.

fh vector con la proporción de la muestra dentro de cada estraro $\frac{nh}{Nh}$.

```
Given arguments:
x = BaseI$Sum_Tot_PerHog
CV = 0.03, Ls = 5, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = none
algo = Kozak: minsol = 1000, idopti = nh, minNh = 2, maxiter = 10000,
              maxstep = 100, maxstill = 500, rep = 5, trymany = TRUE

Strata information:
          |      type rh |        bh      E(Y)       Var(Y)   Nh nh   fh
stratum 1 | take-some  1 |   10290.5   5792.45 6.708054e+06  511  5 0.01
stratum 2 | take-some  1 |   24170.0  15816.58 1.474817e+07  340  5 0.01
stratum 3 | take-some  1 |   60416.0  35510.09 8.317157e+07  172  6 0.03
stratum 4 | take-some  1 |  192918.5  98627.38 1.062807e+09   64  8 0.12
stratum 5 |  take-all  1 | 6740860.0 780643.41 1.606505e+12   27 27 1.00
Total                                                       1114 51 0.05

Total sample size: 51
Anticipated population mean: 37553.71
Anticipated CV: 0.02980551
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

Figura 4: *Salida de la función strata.lh para la estratificación de los municipios por total de personas*

En el marco teórico de este documento se puede encontrar la descripción de esta salida, por ahora basta con decir que en la división de la variable auxiliar agrupada por municipios en cinco estratos de los 1114 municipios se escogen 51 divididos de la siguiente manera:

- Estrato 1 ((stratum 5) municipios con más de 192.918 habitantes) hay 27 municipios, este estrato tiene todos sus elementos de elección obligatoria dentro del muestreo.
- Estrato 2 ((stratum 4) municipios que tienen entre 60.417 y 192.918 habitantes) se tienen 64 municipios de los cuales entran 8 dentro del muestreo.

- Estrato 3 ((stratum 3) municipios que tienen entre 24171 y 60.416 habitantes) se tienen 172 municipios de los cuales entran 6 dentro del muestreo.
- Estrato 4 ((stratum 2) municipios que tienen entre 10291 y 24170 habitantes) se tienen 340 municipios de los cuales entran 5 dentro del muestreo.
- Estrato 5 ((stratum 1) municipios que tienen menos de 10291 habitantes) se tienen 511 municipios de los cuales entran 5 dentro del muestreo.

Dado que en la primera etapa se retiene la mayor variabilidad del diseño bietápico se decide tomar el número de municipios sugerido por el algoritmo de Lavalleé Hidiroglou en las dos estrategias trabajadas. La variable de estratificación para este trabajo es el total de personas por área geográfica totalizada en cada municipio. Para la primera etapa de muestreo , el cardinal del tamaño de muestra esta dado por:

$$\binom{27}{27}\binom{64}{8}\binom{172}{6}\binom{340}{5}\binom{511}{5} = 1,53E+42$$

de las cuales por razones de tiempo solo se escogeran 5.000 muestras.

Una vez se calculó el tamaño de muestra para la primera etapa del muestreo, con ayuda del paquete *samplesize4surveys* Gutiérrez (2011), se caculó el tamaño de muestra para la segunda etapa de muestreo; dicho paquete requiere el tamaño de la población, el valor de la media estimada de la variable de interés, la desviación estandar estimada de la variable de interés, el error estandar permitido, el $deff$ que se tomó como 8 el cual garantizó la presición deseada dentro de la estrategia ESTMAS-MAS, el coeficiente de variación permitido por el usuario (se tomó para este trabajo como el 3 %), el nivel de confianza y el máximo error marginal permitido (se tomó del 95 %). La figura 5 muestra los valores ingresados y la salida de la función.

```
With the parameters of this function: N = 343188 mu = 121.9006 sigma = 121.9006 DEFF = 8   conf = 0.95 .

            The estimated sample size to obatin a maximun coefficient of variation of 2 % is n= 43642 .
            The estimated sample size to obatin a maximun margin of error of 3 % is n= 68362 .
```

Figura 5: *Salida de la función ss4m para cálculo del tamaño de muestra en la segunda etapa*

Una vez fueron calculados los tamaños de muestra para la primera y segunda etapa, se procedió a seleccionar aleatoriamente primero las muestras con tamaño de muestra fijo sin reemplazo para la primera etapa y posteriormente seleccioadas dichas muestras de municipios, crear los marcos muestrales de los municipios seleccionados y poceder a seleccionar las muestras aleatorioas dentro de cada municipio de tal forma que se dividió el tamaño de muestra para la segunda etapa (15382 en promedio por muestreo realizado) de forma equitativa[5].

Una vez se obtuvo la muestra para la segunda etapa, quedaba por estimar el total y la varianza estimada para este total, Lumley (2015) en la función *svytotal* del paquete *survey* del software estadístico R, ingresando previamente las carácterísticas del diseño de muestreo con ayuda de la función *svydesign* del mismo paquete, se obtuvieron las estimaciones requeridas.

Cabe destacar que se realizó una simulación de MonteCarlo determinado número de veces puesto que para poder determinar la convergencia hacia el valor estimado por la simulación. Se muestra en la tabla 5 una parte de los resultados obtenidos.

[5]si el tamaño del municipio era menor al tamaño de la razon de la muestra entre los 51 municipios, se procedía a realizar un censo en el municipio seleccionado en caso contrario se seleccionaba la razón.

Tabla 5: *Muestras de las 10 primeras salidas para ρ= 90 % y β_0=10 %*

ρ= 90 % y β_0=10 %1 ESTMAS-MAS			
CONTEO	$\widehat{t}_{y,\pi}$	$\widehat{Var}_{\widehat{t}_{y,\pi}}$	CVE
1	13075709	8,726E+15	37,78
2	7256697,81	2,687E+15	30,24
3	12982896,5	5,78E+14	16,32
4	20340628,9	1,877E+15	26,13
5	16418726,7	2,634E+15	26,98
6	18140889,9	6,402E+15	38,40
7	20410423,3	1,079E+15	17,20
8	51354232	4,041E+15	24,49
9	31082594,3	7,254E+15	31,05
10	11473004	4,094E+14	11,17

Para la realización de este muestreo se implantó la semilla en el software R; la tabla anterior muestra muestra la salida para las 10 primeras estimaciones de el escenario donde la correlación de las variables ρ= 90 % y el intercepto del modelo lineal β_0=10 %, (se realizaron en total 5.000 por cada uno de los escenarios propuestos). En el anexo B de este trabajo se podrá encontrar un resumen de las salidas para cada uno de los 49 escenario.

6. MUESTREO BIETÁPICO πPT-MAS

Una vez escogidos desde la estrategia de muestreo anterior los tamaños de muestra para la primera y segunda etapa, solo queda por asegurar que los tamaños de muestra sean iguales para etapa 1 (selecionar las unidades primarias de muestreo UPM, 51 municipios) y similar para etapa 2. La variable auxiliar total de personas por departamento se utilizó esta vez no para estratificar la población sino para dar las probabilidades de inclusión; con ayuda de la función ***S.piPS*** del paquete ***TeachingSamplin*** Gutiérrez (2015), esta función solo requiere el tamaño de muestra y la variable de infromación auxiliar arroja la muestra seleccionada y las probabilidades de inclusión de cada uno de los elementos de la muestra. Solo 6 de los 1.114 municipios son de inclusión forzosa a diferencia del muestreo estratificado donde 27 municipios del estrato 1 eran seleccionados en censo (inclusión forzosa).

Luego de constrir el marco muestral para la segunda etapa y escoger la muestra con la función *strata* del paquete *sampling*, del cual se habló anteriormente, quedaba por calcular las estimaciones del total y la varianza estimada, cabe destacar que un muestreo bietápico que incluya al diseño πPT debe calcular las probabilidades de inclusión de segundo orden sobre algún algoritmo que sea computacionalmente sencillo; la función *svydesign* del paquete *survey* trae cuando se cuenta con un diseño de muestreo proporcional a una variable de información auxiliar el argumento *pps*, dicho argumento se le puede indicar que el cálculo de la varianza se raliza por el método de "*brewer*" para utilizar la función de aproximación creada por Brewer. El algoritmo computacional con el cual se realizó este muestreo puede ser consultado en el anexo C de este documento.

Al igual que con el muestreo ESTMAS-MAS para cada uno de los 49 escenarios se realizaron 5.000 simulaciones con un ciclo for. Se muestra en la tabla 6 una parte de los resultados obtenidos.

Tabla 6: Muestra de las 10 primeras salidas para ρ=90 % y β_0=10 %

ρ=90 % y β_0=10 %			
CONTEO	$\widehat{t}_{y,\pi}$	$\widehat{Var}_{\widehat{t}_{y,\pi}}$	CVE
1	26210253,6	8,7554E+12	11,2892619
2	31908013,2	5,4874E+13	23,2157893
3	34225358,7	9,6776E+13	28,7432674
4	27494162,8	1,7365E+13	15,1562581
5	45536561,9	9,742E+13	21,6752393
6	29771294,2	1,3429E+13	12,3089702
7	22580118,3	1,1319E+13	14,8996369
8	27875672,5	6,4066E+13	28,7135877
9	25333941,1	1,9679E+13	17,5106055
10	28334828,7	1,8743E+13	15,2791936

Para la realización de este muestreo se implantó una semilla en el software **R** para hacer reproducibles los resultados; la tabla anterior muestra muestra la salida para las 10 primeras estimaciones de el escenario donde la correlación de las variables ρ= 90 % y el intercepto del modelo lineal β_0=10 %, (se realizaron en total 5.000 por cada uno de los escenarios propuestos). En el anexo B de este trabajo se podrá encontrar un resumen de las salidas para cada uno de los 49 escenarios.

7. COMPARACIÓN DE LOS RESULTADOS

A continuación, se presentan por niveles de correlación entre la variable de interés y la variable auxiliar, los resultados obtenidos para el total estimado $\widehat{t}_{y,\pi}$ y el coeficiente de variación estimado (CVE):

7.1. Correlación (ρ) > 90 %

Tabla 7: Correlación (ρ)= 99 %

	πPT-MAS		ESTMAS-MAS	
Intercepto	$\widehat{t}_{y,\pi}$	CVE	$\widehat{t}_{y,\pi}$	CVE
β_0=0,01	33930628	**2,83**	14584776	6,07
β_0=0,10	37579218	**3,03**	17931216	4,97
β_0=0,30	48411920	6,17	27865100	**3,37**
β_0=0,50	68366686	10,14	46164970	**2,44**
β_0=0,70	115687546	14,26	89560474	**2,05**
β_0=0,90	347162047	18,34	301834297	**2,00**

Una de las primeras impresiones que se tiene al ver este resultado es que aunque las variables estén altamente correlacionadas, si el intercepto de no es pequeño (en este caso menor al 10 %) el muestreo πPT-MAS es menos eficiente que el ESTMAS-MAS, quien por su parte parece tener una relación inversamente proporcional al intercepto del modelo. La figura 6 muestra estas relaciones.

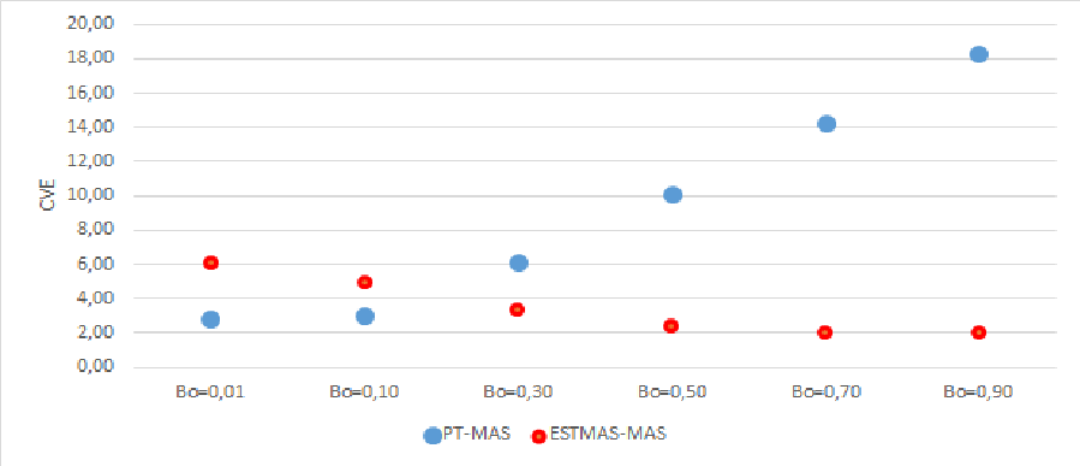

Figura 6: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=99 %*

Si se llegara a preguntar cual de las dos características, bien sea la alta correlación o el nivel bajo del intercepto es más importante, este primer punto arrojaría una mayor importancia al intercepto que a la correlación. Se resume a continuación en la tabla 8 para las correlaciones ρ=97 %, ρ= 95 % y ρ= 93 % los resultados obtenidos, el lector podrá darse cuenta que en estas correlaciones el comportamiento es similar a los resultados obtenidos para ρ=99 %, los gráficos 7, 8 y 9 describen la relación para las correlaciones descritas.

Tabla 8: Correlación (ρ)= 97 %, (ρ)= 95 % y (ρ)= 93 %

Variables		πPT-MAS		ESTMAS-MAS	
Correlación	Intercepto	$\widehat{t}_{y,\pi}$	CVE	$\widehat{t}_{y,\pi}$	CVE
ρ=97 %	β_0=0,01	30412497	**2,93**	15027454	5,78
	β_0=0,10	32217391	**2,09**	15316465	5,75
	β_0=0,30	42499064	7,01	23914981	**6,17**
	β_0=0,50	58150535	9,78	39155547	**7,26**
	β_0=0,70	105729733	17,32	87529535	**9,63**
	β_0=0,90	331200180	18,08	287016716	**4,15**
ρ=95 %	β_0=0,01	37268658	**2,31**	19695916	4,79
	β_0=0,10	38996406	**2,33**	21187547	4,69
	β_0=0,30	33279651	**2,79**	14584776	6,07
	β_0=0,50	77866475	17,86	17931216	**4,97**
	β_0=0,70	113605963	19,37	94076869	**14,82**
	β_0=0,90	406043725	18,53	355538111	**5,19**
ρ=93 %	β_0=0,01	148515688	**2,78**	211914571	6,19
	β_0=0,10	164023140	**3,01**	63877781	5,98
	β_0=0,30	22033003	**5,42**	13193436	5,56
	β_0=0,50	20985425	5,77	12745702	**5,58**
	β_0=0,70	318000881	10,29	277542781	**7,50**
	β_0=0,90	393203487	18,02	335316309	**8,52**

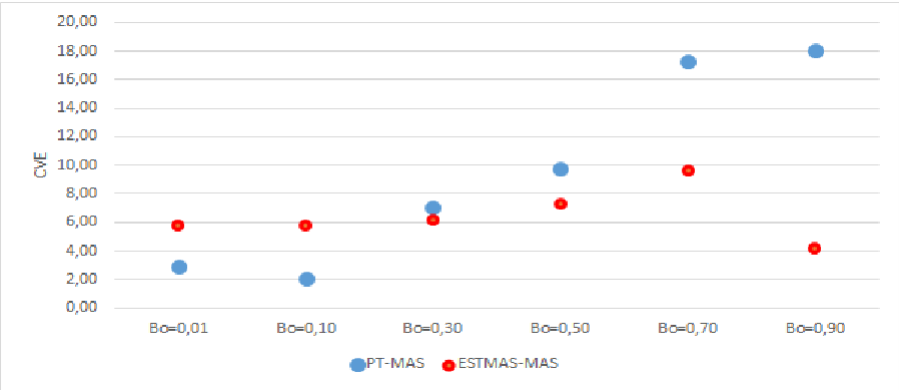

Figura 7: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=97 %*

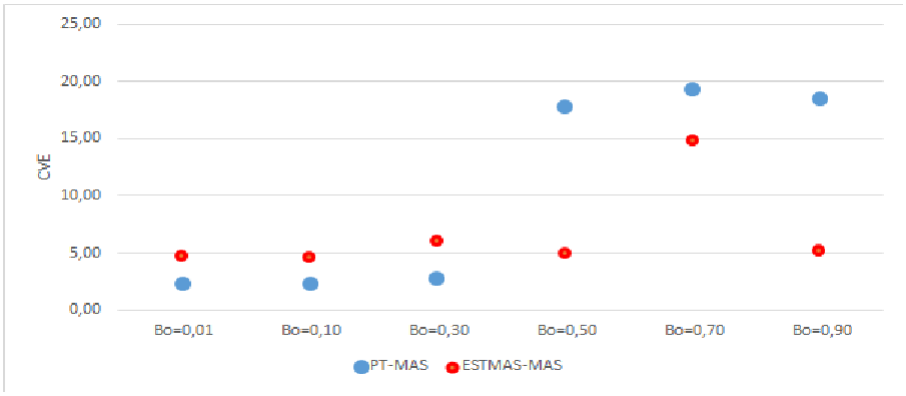

Figura 8: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=95 %*

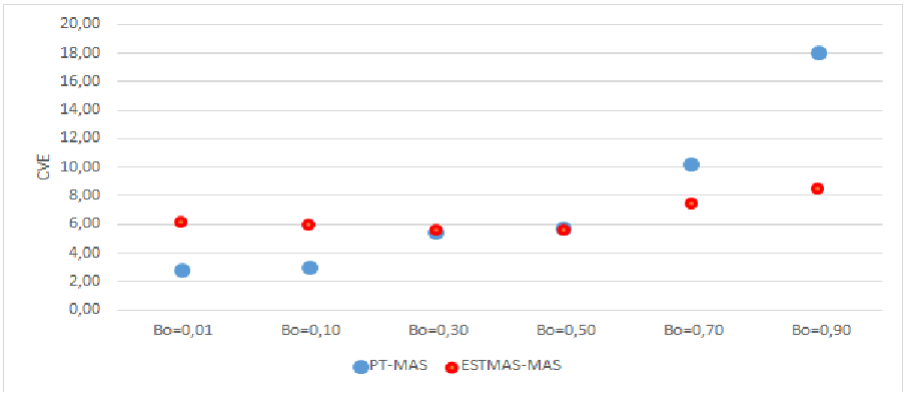

Figura 9: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=93 %*

7.2. Correlación (ρ) < 90 %

Ya se vio a través de los anteriores resultados que el intercepto tiene una altísima importancia a la hora de construir una estrategia de muestreo πPT; se mostrará a continuación que sucede al variar la correlación entre la variable de interés. Se muestra que sucede en una correlación del 90 % en la tabla 9.

Tabla 9: Correlación (ρ)= 90 %

	πPT-MAS		ESTMAS-MAS	
Intercepto	$\hat{t}_{y,\pi}$	CVE	$\hat{t}_{y,\pi}$	CVE
β_0=0,10	335883810	**13,28**	219582853	24,21
β_0=0,30	113116852	**13,65**	69402195	22,89
β_0=0,50	66083431	**16,37**	36827831	22,97
β_0=0,70	43546126	**14,42**	24868622	19,31
β_0=0,90	216567589	**17,26**	21267082	37,20

Para esta correlación en todos los escenarios del intercepto el muestreo πPT-MAS tiene los menores CVE sin embargo superan con el tamaño de muestra dado el 10 %, el muestreo ESTMAS-MAS parece salir

del escenario duplicando en cada escenario el CVE obtenido por el πPT-MAS. La figura 10 ilustra mejor la situación.

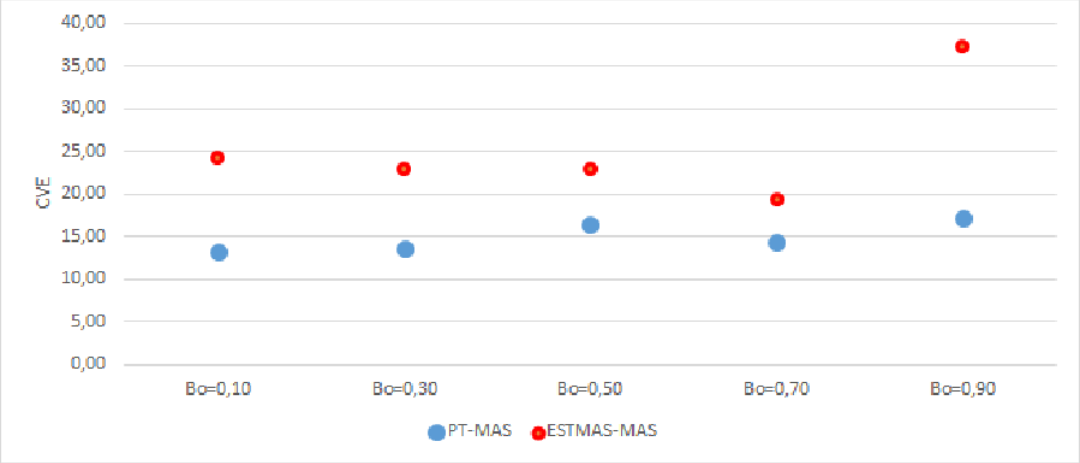

Figura 10: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=90 %*

Similar situación tienen las demás correlaciones, en la tabla 10, acompañada de los gráficos 11, 12, 13 y 14 se muestran los resultados obtenidos para ambas estrategias de muestreo en las correlaciones ρ= 70 %, ρ= 50 %, ρ= 30 % y ρ= 10 %.

Tabla 10: Correlación (ρ)= 70 %, (ρ)= 50 %, (ρ)= 30 % y (ρ)= 10 %

Variables		πPT-MAS		ESTMAS-MAS	
Correlación	Intercepto	$\widehat{t}_{y,\pi}$	CVE	$\widehat{t}_{y,\pi}$	CVE
ρ=70 %	β_0=0,10	334242081	14,17	232414038	26,30
	β_0=0,30	100017987	17,72	61468940	29,33
	β_0=0,50	39043595	17,98	22110591	31,14
	β_0=0,70	23442433	21,90	12270744	39,23
	β_0=0,90	11326969	18,31	5933392	25,44
ρ=50 %	β_0=0,10	249057639	15,66	5933392	25,44
	β_0=0,30	41737814	13,01	165675742	29,93
	β_0=0,50	65860020	18,07	25523756	20,15
	β_0=0,70	6108603	18,24	38214317	24,79
	β_0=0,90	227195	14,37	3721529	20,30
ρ=30 %	β_0=0,10	232682475	18,45	152390470	33,91
	β_0=0,30	34869465	26,23	19333722	43,45
	β_0=0,50	7946439	24,74	4787363	21,91
	β_0=0,70	4459165	16,22	3297767	22,64
	β_0=0,90	454038	15,54	410376655	40,33
ρ=10 %	β_0=0,10	676799062	23,73	410376655	40,33
	β_0=0,30	14691349	34,17	8856312	22,06
	β_0=0,50	2462057	20,37	2157788	25,24
	β_0=0,70	1993502	19,31	1659857	24,62
	β_0=0,90	1736014	19,26	1442400	24,60

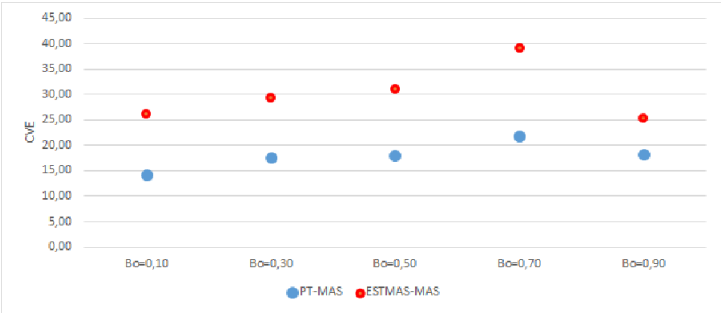

Figura 11: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=70 %*

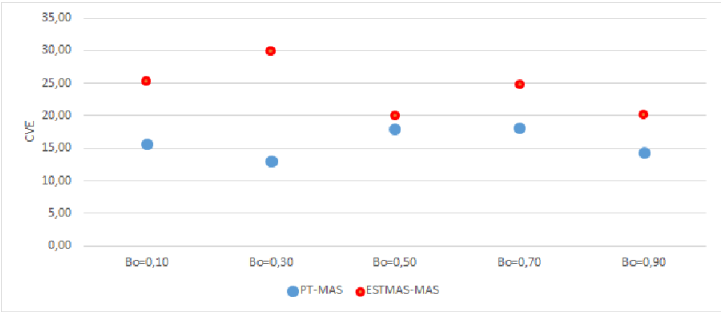

Figura 12: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=50 %*

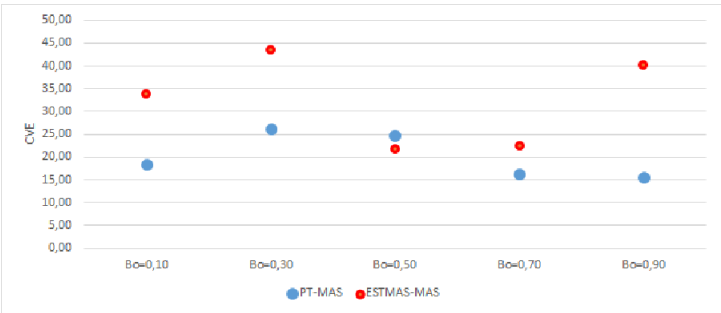

Figura 13: *Distribución de los CVE con respecto al Intercepto β_0. para las estrategias πPT-MAS y ESTMAS-MAS para ρ=30 %*

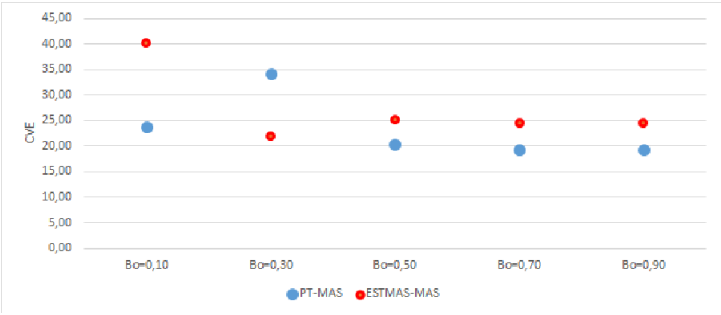

Figura 14: *Distribución de los CVE con respecto al Intercepto* β_0. *para las estrategias* π*PT-MAS y ESTMAS-MAS para* ρ*=10 %*

8. CONCLUSIONES

El objetivo principal de este trabajo era determinar cuales era los escenarios óptimos para cada estrategia de muestreo planteada, sin lugar a dudas el muestreo πPT-MAS es un muestreo poderoso, pero a la vez exigente, tanto para el criterio de correlación como para el criterio del intercepto.

El muestreo ESTMAS-MAS también es exigente con respecto a la correlación de la variable auxiliar y la variable de interés, pero flexible con respecto al intercepto del modelo lineal que relaciona las variables.

Infortunadamente las condiciones que se dieron sobre la variable de interés creada, no permitieron garantizar todos los supuestos sobre el modelo lineal para todos los escenarios, en especial con lo referente a la normalidad de los errores del modelo.

Si se desease saber cual es el mejor diseño bietápico de muestreo (πPT-MAS o ESTMAS-MAS) acompañado del estimador de Horvitz y Thompson, se tendría que revisar tanto la correlación de la variable auxiliar (tomada para estratififcar en el diseño ESTMAS o para dar las probabilidades proporcionales en el πPT) con respecto a la variable de interés, pueden surgir tres escenarios:

- Si tanto la correlación de las variables como la razón entre intercepto del modelo y la media de la variable de interés son altas, aplique un diseño de muestreo ESTMAS-MAS.
- Si la correlación entre la variable de interés es alta y la razón entre el intercepto del modelo y la media de la variable de interés es muy pequeña, aplique un diseño de muestreo πPT-MAS.
- Si la correlación es baja (menor a un 93 %), la estrategia de muestreo πPT-MAS puede presentar mejores resultados que la estrategia ESTMAS-MAS, siempre y cuando se aumente el tamaño de la muestra tanto en la primera como en la segunda etapa del muestreo.

9. ANEXOS

9.1. Anexo A.- Descripción de la Base de Datos

La distribución de las variables de interés por departamentos; dicha distribución se muestra en orden descendente por la variable total personas, variable que como se ha mencionado anteriormente será la variable auxiliar para uno de las estrategias de muestreo y servirá como soporte para generar las diferentes simulaciones de la variable de interés.

Tabla 11: *Distribución de los datos por Departamentos*

	Total Áreas	Total Viviendas	Total Hogares	Total Personas
BOGOTA, D.C.	37.627	1.762.641	1.931.372	6.740.859
ANTIOQUIA	32.335	1.504.961	1.458.193	5.562.812
VALLE DEL CAUCA	30.346	1.030.430	1.070.928	4.029.027
CUNDINAMARCA	16.073	595.745	594.167	2.200.790
ATLANTICO	19.608	436.830	473.037	2.107.781
SANTANDER	15.773	495.016	505.700	1.890.976
BOLIVAR	20.193	390.917	406.135	1.831.425
MAGDALENA	25.993	388.791	405.891	1.794.683
NARIÑO	8.365	347.024	341.835	1.491.026
CORDOBA	14.119	315.879	308.120	1.459.794
TOLIMA	12.168	36.3014	338.257	1.300.842
NORTE DE SANTANDER	12.125	295.513	293.500	1.197.590
BOYACA	9.834	336.224	322.850	1.192.983
CAUCA	6.676	294.036	290.936	1.176.026
HUILA	9.499	253.294	267.888	993.842
CALDAS	7.624	256.488	244.685	889.402
CESAR	12.572	204.225	199.110	874.797
RISARALDA	6.719	230.721	221.408	853.697
SUCRE	10.196	16.7713	171.383	759.026
META	8.803	179.423	18.7218	696.662
LA GUAJIRA	6.493	12.2974	120.543	654.454
QUINDIO	4.969	145.603	142.240	514.747
CHOCO	2.238	97.388	83.768	387.633
CAQUETA	2.965	80.036	80.320	332.326
CASANARE	5.172	73.598	71.908	275.311
PUTUMAYO	1.304	69.559	28.494	234.620
ARAUCA	1.950	3.5615	50.218	147.859
SAN ANDRES	430	16.291	14.669	59.010
GUAVIARE	340	13.631	21.004	56.203
AMAZONAS	191	9.999	8.867	46.182
VICHADA	253	8.959	9.181	43.982
VAUPES	89	3.382	3.321	19.709
GUAINIA	146	5.122	6.538	18.761
TOTAL	343.188	1.053.104	10.673.684	41.834.837

Cabe destacar que los primeros cinco departamentos tienen casi el 50 % del total de la población total del territorio nacional para el año 2005.

9.2. Anexo B.- Tablas de Resultados Para los Muestreos Realizados

Resultados del muestreo ESTMAS-MAS

Tabla 12: Resultados del muestreo ESTMAS-MAS

Correlación	Intercepto	$\hat{t}_{y,\pi}$	$\widehat{Var}_{\hat{t}_{y,\pi}}$	CVE	ni
	β_0=0,01	14584776	8,2235E+11	6,07	15382
	β_0=0,10	17931216	8,2727E+11	4,97	15382
ρ=99 %	β_0=0,30	27865100	9,1001E+11	3,37	15382
	β_0=0,50	46164970	1,3291E+12	2,44	15382
	β_0=0,70	89560474	3,7046E+12	2,05	15382
	β_0=0,90	301834297	4,3332E+13	2,00	15382
	β_0=0,01	15027454	7,85E+11	5,78	15382
	β_0=0,10	15316465	8,1232E+11	5,75	15382
ρ=97 %	β_0=0,30	23914981	2,2822E+12	6,17	15382
	β_0=0,50	39155547	8,5375E+12	7,26	15382
	β_0=0,70	87529535	7,4724E+13	9,63	15382
	β_0=0,90	287016716	1,5304E+14	4,15	15382
	β_0=0,01	19695916	1,0912E+13	4,79	15382
	β_0=0,10	21187547	1,1862E+14	4,69	15382
ρ=95 %	β_0=0,30	14584776	8,2235E+11	6,07	15382
	β_0=0,50	17931216	8,2727E+11	4,97	15382
	β_0=0,70	94076869	1,912E+14	14,82	15382
	β_0=0,90	355538111	3,6599E+14	5,19	15382
	β_0=0,01	211914571	2,0243E+14	6,19	14599
	β_0=0,10	63877781	1,532E+13	5,98	15382
ρ=93 %	β_0=0,30	13193436	5,6098E+11	5,56	15382
	β_0=0,50	12745702	5,2914E+11	5,58	15382
	β_0=0,70	277542781	8,3598E+14	7,50	15382
	β_0=0,90	335316309	8,6567E+14	8,52	15382
	β_0=0,10	219582853	3,5477E+15	24,21	15382
	β_0=0,30	69402195	3,1492E+14	22,89	15382
ρ=90 %	β_0=0,50	36827831	1,2741E+14	22,97	15382
	β_0=0,70	24868622	4,4676E+13	19,31	15382
	β_0=0,90	21267082	1,0055E+14	37,20	15382
	β_0=0,10	232414038	4,9632E+15	26,30	15382
	β_0=0,30	61468940	4,2185E+14	29,33	15382
ρ=70 %	β_0=0,50	22110591	1,1161E+14	31,14	15382
	β_0=0,70	12270744	6,3927E+13	39,23	15382
	β_0=0,90	5933392	1,1794E+13	25,44	15382
	β_0=0,10	5933392	1,1794E+13	25,44	15382
	β_0=0,30	165675742	3,1165E+15	29,93	15382
ρ=50 %	β_0=0,50	25523756	7,2041E+13	20,15	15382
	β_0=0,70	38214317	4,6288E+14	24,79	15382
	β_0=0,90	3721529	3,7918E+12	20,30	15382
	β_0=0,10	152390470	3,6897E+15	33,91	15382
	β_0=0,30	19333722	2,0904E+14	43,45	15382
ρ=30 %	β_0=0,50	4787363	1,3047E+13	21,91	15382
	β_0=0,70	3297767	8,4258E+12	22,64	15382
	β_0=0,90	410376655	5,0179E+16	40,33	15382
	β_0=0,10	410376655	5,0179E+16	40,33	15382
	β_0=0,30	8856312	9,8378E+13	22,06	15382
ρ=10 %	β_0=0,50	2157788	8,9257E+12	25,24	15382
	β_0=0,70	1659857	4,2613E+12	24,62	15382
	β_0=0,90	1442400	3,1822E+12	24,60	15382

Resultados del muestreo πPT-MAS

Tabla 13: Resultados del muestreo πPT-MAS

PiPT-MAS

Correlación	Intercepto	$\hat{t}_{y,\pi}$	$\widehat{Var}_{\hat{t}_{y,\pi}}$	CVE	ni
ρ=99 %	β_0=0,01	33930628	9,2399E+11	2,83	14599
	β_0=0,10	37579218	1,3479E+12	3,03	14599
	β_0=0,30	48411920	1,0465E+13	6,17	14599
	β_0=0,50	68366686	5,8031E+13	10,14	14599
	β_0=0,70	115687546	3,3029E+14	14,26	14599
	β_0=0,90	347162047	4,8946E+15	18,34	14599
ρ=97 %	β_0=0,01	30412497	7,9719E+11	2,93	14599
	β_0=0,10	32217391	4,5768E+11	2,09	14599
	β_0=0,30	42499064	8,8809E+12	7,01	15036
	β_0=0,50	58150535	3,8892E+13	9,78	14599
	β_0=0,70	105729733	4,0355E+14	17,32	14599
	β_0=0,90	331200180	4,2335E+15	18,08	14599
ρ=95 %	β_0=0,01	37268658	1,2387E+13	2,31	14599
	β_0=0,10	38996406	1,3465E+14	2,33	14599
	β_0=0,30	33279651	9,6089E+11	2,79	14599
	β_0=0,50	77866475	1,4082E+15	17,86	14599
	β_0=0,70	113605963	7,1325E+14	19,37	14599
	β_0=0,90	406043725	6,6564E+15	18,53	14599
ρ=93 %	β_0=0,01	148515688	1,7137E+13	2,78	14599
	β_0=0,10	164023140	2,5258E+13	3,01	14599
	β_0=0,30	22033003	1,6824E+12	5,42	14599
	β_0=0,50	20985425	1,7284E+12	5,77	14599
	β_0=0,70	318000881	3,08E+15	10,29	14599
	β_0=0,90	393203487	5,9591E+15	18,02	14599
ρ=90 %	β_0=0,10	335883810	2,5087E+15	13,28	14599
	β_0=0,30	113116852	2,6327E+14	13,65	14599
	β_0=0,50	66083431	1,9823E+14	16,37	14599
	β_0=0,70	43546126	8,5949E+13	14,42	14599
	β_0=0,90	216567589	1,6521E+15	17,26	14599
ρ=70 %	β_0=0,10	334242081	2,5448E+15	14,17	14599
	β_0=0,30	100017987	4,4894E+14	17,72	14599
	β_0=0,50	39043595	9,8503E+13	17,98	14599
	β_0=0,70	23442433	5,1776E+13	21,90	14599
	β_0=0,90	11326969	7,4866E+12	18,31	14599
ρ=50 %	β_0=0,10	249057639	1,8012E+15	15,66	14599
	β_0=0,30	41737814	5,0654E+13	13,01	14599
	β_0=0,50	65860020	2,7881E+14	18,07	14599
	β_0=0,70	6108603	3,1647E+12	18,24	14599
	β_0=0,90	227195	5880503930	14,37	14599
ρ=30 %	β_0=0,10	232682475	2,4028E+15	18,45	14599
	β_0=0,30	34869465	1,7485E+14	26,23	14599
	β_0=0,50	7946439	1,1367E+13	24,74	14599
	β_0=0,70	4459165	6,0627E+12	16,22	14599
	β_0=0,90	454038	4,9115E+10	15,54	14599
ρ=10 %	β_0=0,10	676799062	4,5445E+16	23,73	14599
	β_0=0,30	14691349	8,7826E+13	34,17	14599
	β_0=0,50	2462057	6,432E+12	20,37	14599
	β_0=0,70	1993502	3,0699E+12	19,31	14599
	β_0=0,90	1736014	2,2924E+12	19,26	14599

9.3. Anexo C.- Códigos en el Sofware Estadístico R requeridos en este trabajo

En este apéndice se encuentran los códigos que fueron utilizados para la creación de los escenarios, para el cálculo del tamaño de muestra y el total estimado y la varianza estimada del total estimado.

9.3.1. Código para la creación de los escenarios

```
# SIMULACIÓN DE LA VARIABLES #
# SELECCION BASE DE DATOS #
rm(list=ls(all=TRUE)) # borrar información almacenada dentro del programa
vihope = read.csv2("$:/XXXXX/vihope.csv", sep=";") #llamar la información de un archivo
colnames(vihope) # Nombre de las variables
library(sqldf) #Llamar la libreria sqldf
Basei=sqldf("select Mpios_n, AREA_GEOGRAFICA as ID,Tot_PerHog from vihope") #Selección de las variables requeridas 343.188
Ni=nrow(Basei) #Número de registros en la base
BaseI=sqldf('select Mpios_ n,sum(Tot_PerHog) as Sum_Tot_PerHog from Basei group by Mpios_ n') #se agrupan los datos por municipios
NI=nrow(BaseI) #Número de registros en la base por municipios 1.114


# SIMULACIÖN DE Ty CORRELACION 0.9 INTERCEPTO 0.90 (E1_1) #
set.seed(1234511) # plantar semilla de proceso aleatorio
tS1_1=matrix(round(β1'*BaseI$Sum_Tot_PerHog+β0'+rnorm(NI,mean=0,sd=α),digits=0)) # Modelo de inicio
minS1_1=min(S1_1) for (i in 1:NI) {
if(tS1_1[[i,1]]< 0)
tS1_1[[i,1]]=tS1_1[i]-minS1_1+1 else
tS1_1[[i,1]]=tS1_1[[i,1]]
} #Es ciclo for coloca los valores negativos dentro de otro modelo
cor(tS1_1,BaseI$Sum_Tot_PerHog) #Cálculo de la correlación
m1=mean(tS1_1) #Media de la varible respuesta
a1=as.matrix(lm(tS1_1 BaseI$Sum_Tot_PerHog)$coef,ncol=2) #intercepto del modelo
a1[1,1]/m1 razón entre el intercepto y la media de la varable


# Gráficos #
plot(BaseI$Sum_Tot_PerHog, tS1_1, pch = 20, xlim = c(0, 500000))
abline(lm(tS1_1  BaseI$Sum_Tot_PerHog ), col = red")


# distribucion de TY en BaseI a Basei
BaseI$tS1_1 = tS1_1
# Repartación de ty en las USM.
ls() # lista de objetos en memoria
Basei = merge(Basei,BaseI) # asociar a la base por municipios con la base por áreas geográficas #Contar cuantos USM (AG) hay en cada uno de los municipios
Tot_MG = sqldf("select Mpios_n, count(*) as 'N_AG' from Basei group by Mpios_n")
Basei = merge(Basei, Tot_MG)
Basei$repar_tyAG = Basei$tS1_1 / Basei$N_AG
set.seed(1234511)


# Repartir aleatoriamente el total de la variable de interés de cada municipio a cada una de las áreas geográficas
```

```
Basei$repar_tyAG_final = rnorm(n = nrow(Basei), mean = Basei$repar_tyAG, sd = Basei$repar_tyAG/3)
Basei$repar_tyAG_final = ifelse(Basei$repar_tyAG_final < 0, Basei$repar_tyAG, Basei$repar_tyAG_final)
table(Basei$repar_tyAG_final < 0)
Basei$S1_1 = round(Basei$repar_tyAG_final, digits = 0)
# Comprobación de que la correlación y la razón del modelo entre tx y ty se mantiene
cons2 = sqldf("select Mpios_n, sum(repar_tyAG_final) as 'S1_1', sum(Tot_PerHog) as 'tx' from Basei
group by Mpios_n")
cor(cons2$S1_1, cons2$tx)
m=mean(cons2$S1_1) # media de la variable respuesta
a=as.matrix(lm(cons2$S1_1 cons2$tx)$coef,ncol=2)
a[1,1]/m #razón entre el intercepto y la media de la varable
# borrar variables no útiles
Basei$repar_tyAG = NULL
Basei$Sum_Tot_PerHog = NULL
Basei$tS1_1 = NULL
Basei$repar_tyAG_final = NULL
Basei$N_AG = NULL
head(Basei)
vihope$S1_1=Basei$S1_1
colnames(vihope)
write.csv2(vihope, file = "$:/XXXXX/vihope.csv/vihope5.csv')
```

9.3.2. Código para el muestreo ESTMAS-MAS

```
# MUESTREO BIETAPICO ESTRATIFICADO-MAS (ESTMAS-MAS) gc(reset=T) #Eliminar la basura
rm(list=ls(all=TRUE))

vihope = read.csv2("D:/Rodrigo/vihope5.csv", sep=";")
#View(vihope)
colnames(vihope)
library(sqldf)

#Seleccionando las variables de interés
Basei=sqldf("select Mpios_n,AREA_GEOGRAFICA as ID,Tot_PerHog, S1_1 from vihope")
#View(Basei)
colnames(Basei)
Ni=nrow(Basei)# Número de USM

#Sumando a nivel de UPM la variable de interés y la auxiliar BaseI=sqldf('select Mpios_n,sum(Tot_PerHog) as Sum_Tot_PerHog, sum(S1_1) as SS1_1 from Basei group by Mpios_n')
#View(BaseI)
colnames(BaseI)
NI=nrow(BaseI)

# ESTRATIFICACION library(stratification)
set.seed(12345)
LH=strata.LH(BaseI$Sum_Tot_PerHog,CV=0.03,Ls=5)

cortes = c(min(BaseI$Sum_Tot_PerHog),LH$bh,max(BaseI$Sum_Tot_PerHog))
BaseI$Estrato=cut(x = BaseI$Sum_Tot_PerHog, breaks = cortes, labels = NULL, include.lowest = T, right = F)

BaseI$Estrato=cut(x=BaseI$Sum_Tot_PerHog, breaks = cortes, labels = c(".Estrato 5",".Estrato 4",".Estrato 3",".Estrato 2",".Estrato 1"),include.lowest = T, right = F)

table(BaseI$Estrato)

# Calcular el tamaño de muestra áreas geográficas

library(samplesize4surveys)
Muestra_ag = ss4m(N = Ni, mu = mean(Basei$Tot_PerHog),sigma = mean(Basei$Tot_PerHog), DEFF = 20, conf = 0.95, cve = 0.02, rme = 0.03,plot = FALSE)
tamai = Muestra_ag$n.cve

#Seleccionar la muestra de municipios
tam_pobXest = table(BaseI$Estrato)
tam_mueXest = as.numeric(as.character(LH$nh))

# for repetir esto 5000 veces
```

```
m=5200 # por fallos se cualculan más veces los muestreo
resultado=matrix(c(numeric(m),numeric(m),numeric(m)),ncol=3)
for(i in 1:m)
set.seed(1000+i)
library(sampling)
MuestraI = strata(data = BaseI, stratanames=.Esªtrato", size = tam_mueXest, description=F, method
= "srswor")
muestraI = getdata(BaseI, MuestraI)


marcomuestrai = merge(Basei, muestraI, all.y = T, by = "Mpios_n")

# SELECCIONAR LAS MUESTRAS POR ÁREA GEOGRÁFICAS

# APMS es la cantidad de Àreas Geográficas por municipios seleccionados
APMS = sqldf("select Mpios_n, count(*) as 'Num_areas_geogr' from marcomuestrai group by Mpios_n")


APMS$ni = tamai/sum(LH$nh)
APMS$ni = round(ifelse(APMS$ni ¿APMS$Num_areas_geogr, APMS$Num_areas_geogr, APMS$ni ))


set.seed(1000+i)
Muestrai = try(strata(data = marcomuestrai, stratanames="Mpios_n", size = APMS$ni, description=F,
method = "srswor"),silent=T)
muesfinal = try(getdata(marcomuestrai, Muestrai),silent=T)
dim(muesfinal)

#Verificar para garantizar que está bien: Calcular el número de municipios por estrato (Para tener los
tamaños en survey)
cons_tammueEstXMunic = data.frame(Estrato = paste0(.Esªtrato ", 1:5), NmunXEstrato = LH$Nh[c(5,4,3,2,1)])
cons_tammue_mzXmpio = APMS[c("Mpios_n", "ni")]
muefinal1 = try(merge(muesfinal, cons_tammueEstXMunic),silent=T) # Pegarle el tamaño de muestra
por estrato
muefinal = try(merge(muefinal1, cons_tammue_mzXmpio),silent=T) # Pegarle el tamaño de muestra
por municipio


# creación del diseño y estimación de las variable requeridas
library(survey)
diseno=try(svydesign(id= Mpios_n + ID, strata =  Estrato, fpc= NmunXEstrato +ni,data = muefi-
nal),silent=T)
try(svymean( S1_1, diseno, deff=TRUE),silent=T)
estimacion = try(svytotal( S1_1, diseno, deff=TRUE),silent=T)
options(sicipen = 1111)
result =try(matrix(c(ty = as.numeric(estimacion), var = (as.numeric(SE(estimacion)))^2, cve = 100 *
as.numeric(SE(estimacion))/as.numeric(estimacion)), ncol = 3), silent = T)
#guardarenlamatrizlosresultados
resultado[i, 1] = try(result[1, 1], silent = T)
resultado[i, 2] = try(result[1, 2], silent = T)
resultado[i, 3] = try(result[1, 3], silent = T)
print(paste("Muestreo", i, sep = ""))

write.csv2(resultado, file =' # : /???????/_1.csv2')
```

9.3.3. Código para el Muestreo Bietápico πPT-MAS

```
# MUESTREO BIETAPICO πPT-MAS

rm(list=ls(all=TRUE))

# La base contiene el VIHOPE junto con 15 variables que se simularon altamente
# correlacionadas con el total de personas por Unidad geográfica
# 90, 70, 50, 30, 10 y variando el intercepto (porcentaje respecto a la media)
# Documentar
vihope = read.csv2("D:/vihope5.csv", sep=";")
#View(vihope)
colnames(vihope)
library(sqldf)
Basei=sqldf("select Mpios_n,AREA_GEOGRAFICA as ID,Tot_PerHog,S2_4 from vihope")
#View(Basei)
colnames(Basei)
Ni=nrow(Basei)

# Agrupación por UPM (Municipio)
BaseI=sqldf('select Mpios_n,sum(Tot_PerHog)as Sum_Tot_PerHog, sum(S2_4) as SS2_4 from Basei group
by Mpios_n')
#View(BaseI)
colnames(BaseI)
NI=nrow(BaseI)

library(stratification)

# Tamaño de muestras primera etapa por estratificación
set.seed(12345)
# Se construyen 5 estratos y se calcula el tamaño de muestra por estrato de municipio
LH=strata.LH(BaseI$Sum_Tot_PerHog,CV=0.03,Ls=5)

# Calcular el tamaño de muestra para un MAS áreas geográficas (USM)
library(samplesize4surveys)
# Se coloca un deff de tal forma que el deff del bietápico pipt - mas satisfaga el 3# Para el 5Muestra_ag
= ss4m(N = Ni, mu = mean(Basei$Tot_PerHog),sigma = mean(Basei$Tot_PerHog),DEFF = 20, conf =
0.95, cve = 0.02, rme = 0.03, plot = FALSE)
#Cerca de 43000 encuestas para que satisfaga al final un cve del 3tamai = Muestra_ag$n.cve
nI=sum(LH$nh)

# ACA EMPIEZA EL FOR MARA 5000 MUESTRAS'
# Número de simulaciones
m=5200 # Se hace un exceso de simulaciones por aquellas manzanas que tienen un indviduo.
# Inicializo una matriz para guardar el total, la varianza
resultado=matrix(c(numeric(m),numeric(m),numeric(m)),ncol=3)

for(i in 1:m)
#### ESCOGER LOS MUNICIPIOS (MUESTRA PRIMERA ETAPA)
```

```
library(TeachingSampling)
#Se van a seleccionar municipios tomando como variable auxiliar los totales por munic (UPM)
set.seed(1000 + i)
MuestraIa = S.piPS(n = nI, x = BaseI$Sum_Tot_PerHog)


MuestraI = BaseI[MuestraIa[,1], ]
MuestraI$prob_inclu = MuestraIa[,2]


# SELECCIONAR LAS MUESTRAS POR ÁREA GEOGRÁFICAS (USM)
# Basei: base de todos los registros del VIHOPE
marcomuestrai = merge(Basei, MuestraI[,c("Mpios_n","prob_inclu")],
by = "Mpios_n", all.y = T)


# Quedan sólo la tabla VIHOPE de los municipios seleccionadas


# APMS es la cantidad de Àreas Geográficas (USM) por municipios seleccionados APMS = sqldf("select
Mpios_n, count(*) as 'num_mzXmpio'
from marcomuestrai group by Mpios_n")


APMS$ni = tamai/sum(LH$nh) # El número de manzanas (USM) que se deben tomar por mpio
# Si en algún municipio resulta un mayor tamaño de muestra que tamaño del mpio,
# Hacer censo
APMS$ni = round(ifelse(APMS$num_mzXmpio ¿APMS$ni,
APMS$ni, APMS$num_mzXmpio))
library(sampling)
set.seed(1000 + i)
Muestrai = try(strata(data = marcomuestrai, stratanames="Mpios_n", size = APMS$ni,
description=F, method = "srswor"),silent=TRUE)
# Extrae de la base del marco de USM de los municipios seleccionados lo resultante
# con la función strata
muefinal_final = try(getdata(marcomuestrai, Muestrai),silent=TRUE)


# cálculol de las estimaciones library(survey)
diseno =try(svydesign(id= Mpios_n + ID, fpc= prob_inclu +Prob, pps = "brewer", data = muefi-
nal_final),silent = TRUE)


try(svymean( S2_4, diseno, deff=TRUE),silent = TRUE)
estimacion = try(svytotal( S2_4, diseno, deff=TRUE),silent=TRUE)
# Notaciónd decimal
options(sicipen = 1111)
result =try(matrix(c(ty = as.numeric(estimacion), var = (as.numeric(SE(estimacion)))^2,
cve = 100 * as.numeric(SE(estimacion))/as.numeric(estimacion)), ncol = 3), silent = TRUE)
resultado[i, 1] = try(result[1, 1], silent = T)
resultado[i, 2] = try(result[1, 2], silent = T)
resultado[i, 3] = try(result[1, 3], silent = T)

print(paste("Muestreo", i, sep = ""))
```

9.4. Anexo D.- Gráficos de distribución de las variables creadas

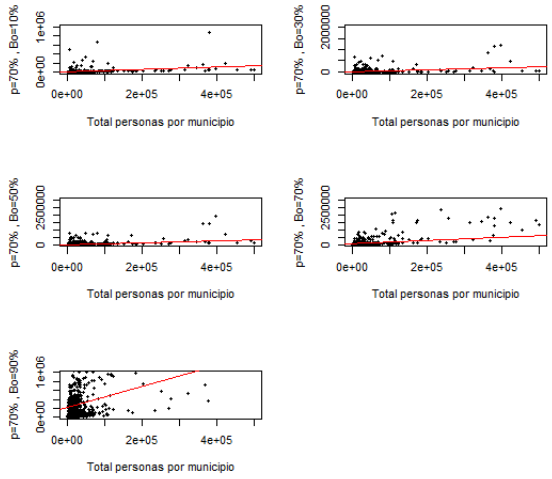

Figura 15: *Distribución de las variables de interés correlacionadas al 70 %*

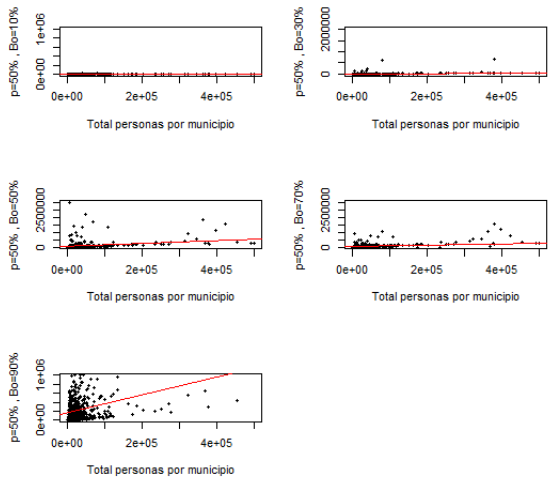

Figura 16: *Distribución de las variables de interés correlacionadas al 50 %*

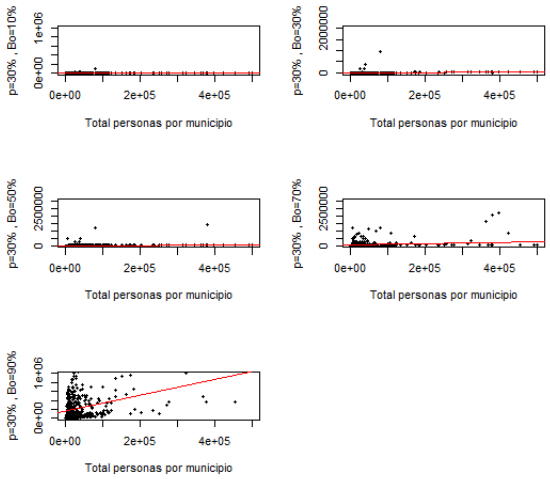

Figura 17: *Distribución de las variables de interés correlacionadas al 30 %*

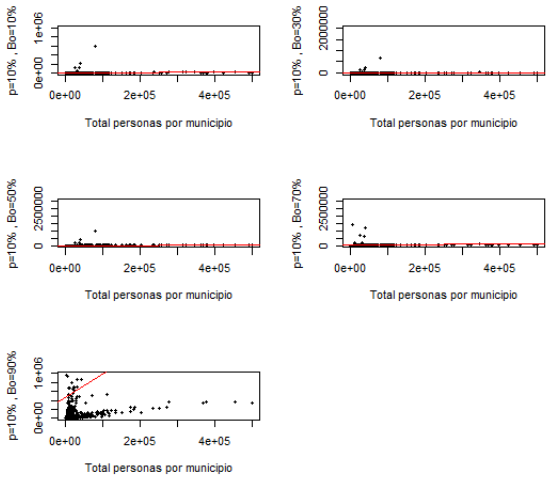

Figura 18: *Distribución de las variables de interés correlacionadas al 10 %*

10. REFERENCIAS BIBLIOGRÁFICAS

Referencias

Baillargeon, S. & Rivest, L.-P. (2011), 'The construction of stratified designs in r with the package stratification', *Survey Methodology* **37**(1), 53–65.

Buitrago Chinchilla, E. L. (2015), 'Comparación de diseños muestrales bietápicos con información auxiliar en la primera etapa'.

DANE (2014), Reporte generado en: March 17, 2014 colombia - censo general 2005.

Gunning, P. & Horgan, J. (2004), 'A new algorithm for the construction of stratum boundaries in skewed populations', *Survey Methodology* .

Gutiérrez, A. (2009), *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*, Colección de textos en estadística aplicada, Editorial de la Universidad Santo Tomás.

Gutiérrez, A. (2011), 'The required sample size for estimating a single mean, package samplesize4surveys', *R packages* **37**(1), 53–65.

Gutiérrez, Hugo Andres, M. H. A. G. (2015), 'Package teachingsampling'.

Lavalleé, P. & Hidiroglou, M. (1988), 'On the stratification of skewed populations', *Survey methodology* **14**(1), 33–43.

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.

Lumley, T. (2012), 'Describing, pps desing to r. version 3.30.3.'.

Lumley, T. (2015), 'Package âsurveyâ'.

Raj, D. (1980), *Teoría del muestreo*, Fondo de Cultura Económica.

Särndal, C.-E., Swensson, B. & Wretman, J. (2003), *Model assisted survey sampling*, Springer Science & Business Media.

Wackerly, D. D. M., Scheaffer, W., Muñoz, R. L. R. et al. (2010), *Estadística matemática con aplicaciones*, number 519.5 W3.