UNIVERSIDAD NACIONAL AUTÓMONA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

DESEMPEÑO PRÁCTICO DE ALGUNOS MÉTODOS DE ESTIMACIÓN DE VARIANZA CON DATOS DE ENCUESTAS

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS

P R E S E N T A

EMILIO LÓPEZ ESCOBAR

DIRECTORA DE TESIS:
DRA. GUILLERMINA ESLAVA GÓMEZ

MÉXICO, D.F. **ABRIL, 2006**

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mis papás,

y a mi hermana.

Un especial agradecimiento a mi directora de tesis y sinodal la Dra. Guillermina Eslava Gómez.

Agradezco también a mis sinodales:

Dr. Ignacio Méndez Ramírez
Dr. Raúl Rueda Díaz del Campo
Dr. José Rubén Hernández Cid
Dr. Martín Romero Martínez

Índice

1. Resumen

La presente tesis se concentra en la evaluación empírica del desempeño práctico de algunos métodos de estimación de la varianza de algunos estimadores. Esta evaluación se lleva a cabo comparando resultados numéricos y gráficos obtenidos de un ejercicio de simulación. En él se utilizaron datos provenientes de la muestra con cuestionario ampliado del XII Censo General de Población y Vivienda del año 2000 por parte del Instituto Nacional de Estadística, Geografía e Informática (INEGI).

Se tomaron en consideración métodos conocidos de estimación de varianza y uno nuevo denominado Método del Vector Post-Diseño propuesto por Ollila (2004); método que adopta la forma de abordar el muestreo en términos de vectores diseño, metamuestras y el diseño muestral (y metamuestral) como una distribución multivariada. Figuran entonces: el de los estimadores π ó de Horwitz-Thompson, los Métodos de Linealización de Taylor, Woodruff para Cuantiles, Jackknife, el Bootstrap y el del Vector Post-Diseño. Todos ellos implementados en el contexto de estimación de varianza de estimadores lineales (medias y proporciones) y no lineales (mediana y razón); tomando como población los datos muestrales mencionados de INEGI asociados a las variables: Ingresos mensuales totales por hogar, Total de personas por hogar, Total de cuartos por hogar y disponibilidad de teléfono en la vivienda. Medidas en la entidad federativa de Aguascalientes, bajo un diseño muestral de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.) y con dos tamaños de muestra, $n = 350$ y $n = 1,000$.

Los criterios utilizados para la evaluación empírica del desempeño de los métodos se basaron en la medición de propiedades relativas a características de exactitud y precisión. Esta medición se realizó utilizando estimaciones de los estadísticos correspondientes a dichas propiedades por medio del Método de Monte Carlo. Complementariamente al cotejo de cifras obtenidas para cada propiedad, estimador, método y tamaño de muestra; se prestó atención a las representaciones gráficas, que en repetidas veces clarificaron de manera más crítica la evaluación.

Para el ejercicio de simulación se requirió de programación de las rutinas asociadas a cada método. Éstas fueron creadas en *MATLAB*® Versión 6.0 Release 12 y también se hizo uso del paquete estadístico *SPSS*® 13.0 para la creación de las gráficas y cálculo de estadísticos u operaciones con las variables. Adicionalmente se empleo el paquete especializado de muestreo *SUDAAN*® Versión 7.5 para la verificación de algunos cálculos.

Como conclusión y comentario general, resultado del actual ejercicio de evaluación, se encontraron las esperadas dificultades en la estimación de varianza de estimadores como la media, mediana y razón ante la presencia de una alta asimetría en la distribución de la variable de interés (en el numerador para el caso del estimador de razón) bajo muestreo aleatorio simple. También se encontró que, frente a las situaciones y criterios aquí contemplados, de manera colectiva el Método de Taylor mostró un mejor desempeño que el resto de los métodos usados. No obstante, se obtuvieron resultados particulares casuísticos de la evaluación de desempeños en la estimación de varianza de estimadores para cada circunstancia examinada.

2. Introducción

La variabilidad es un elemento importante presente en todo momento en el muestreo probabilístico. Originalmente, se encuentra en la variable asociada a la característica de interés de la población de estudio. Luego, esta característica poblacional es representada de manera resumida por determinado parámetro. Ulteriormente, la antes mencionada variabilidad se traduce en la aleatoriedad que hereda el estimador del parámetro al considerar únicamente información parcial de la población, resultado de un proceso de selección probabilística de individuos (muestra) y de sus correspondientes mediciones de la variable relativa al parámetro.

En la generación de estimaciones del parámetro, la varianza del estimador provee de elementos relativos a la precisión de las estimaciones; que en términos generales se traducen en eficacia del estimador y del esquema de muestreo utilizados. Pero subyace un problema, la varianza del estimador en cuestión es usualmente en términos prácticos desconocida pues para su cálculo se requiere de información específica de la población de estudio completa. No obstante, a partir de los datos disponibles de la muestra se puede obtener una estimación de dicha varianza; o bien, bastaría conocer los primeros cuatro momentos de la distribución de la variable de interés.

Existen varios métodos propuestos en bibliografía para la estimación de varianza de un estimador. Sin embargo no se tiene la supremacía de alguno de estos métodos, se tiene que algunos funcionan mejor que otros bajo determinadas circunstancias y para ciertos casos. Ya en la implementación práctica, se tiene que algunos de ellos llegan a necesitar bastantes recursos de cómputo; haciendo menos atractivo su aplicación que la de otros. Esto último se piensa como un brete que será desvanecido dramáticamente con la evolución de los equipos de cómputo.

La presente tesis realiza una comparación empírica del desempeño práctico de los métodos de estimación de varianza de estimadores, contemplando los métodos siguientes: el de los estimadores π ó de Horwitz-Thompson, el Método de Woodruff para Cuantiles, el Método de Linealización de Taylor, el Método de Replicaciones Repetidas de Jackknife[1], el Método de Replicaciones de Bootstrap y el Método del Vector Post-Diseño[2]. El desempeño en la estimación de varianza es evaluado para cada método calculando estadísticos asociados a ciertas propiedades utilizando simulaciones con el Método de Monte Carlo y también observando las simulaciones resultantes en representaciones gráficas.

Para implementar por medio de simulación un método de estimación de varianza hay numerosos factores que tienen que ser considerados: el marco muestral, la distribución de los datos, los parámetros de interés, los estimadores de los parámetros y su tipo (lineal, no lineal), los tamaños de muestra, el diseño muestral, el número de simulaciones; y para algunos métodos: el número de remuestras, el tamaño de remuestra, el diseño de remuestreo y algunas especificaciones propias de cada método. Por lo tanto, para hacer comparables los resultados obtenidos de la simulación para cada método, se procedió a ir fijando cada uno de esos factores. En lo que atañe a esto último, se piensa que el mantener la simplicidad en las comparaciones las hace más fuertes, pues las diferencias encontradas no serán circunstanciales o innumerablemente casuísticas.

Entonces, la simulación del uso de los métodos de estimación de varianza a comparar serán de

[1]En dos modalidades: quitando uno y dos elementos en la muestra.

[2]Propuesto recientemente por Ollila (2004), en el que se adopta la novedosa forma de abordar el muestreo en términos de vectores diseño, metamuestras y el diseño muestral (y metamuestral) como una distribución multivariada.

estimadores lineales (media, proporción) y no lineales (mediana, razón), bajo un diseño muestral de muestreo aleatorio simple sin reemplazo[3](SRS, SI o m.a.s.) y para dos tamaños de muestra 350 y $1,000$; utilizando como marco muestral los datos muestrales de la entidad federativa de Aguascalientes de la muestra con cuestionario ampliado del XII Censo General de Población y Vivienda por parte del Instituto Nacional de Geografía, Estadística e Informática (INEGI); el número de simulaciones será de $1,000$ y el número de remuestras quedará fijado en $1,000$ también[4].

La estructura de la presente tesis es como sigue. En el Capitulo 3 se describe la teoría y notación necesaria para la comprensión del Método del Vector Post-Diseño; luego, en el Capítulo 4 se presenta la teoría correspondiente a algunos métodos de estimación de varianza ya conocidos; posteriormente, en el Capitulo 5 se presentaran resultados numéricos obtenidos de la simulación y la comparación de estos para los métodos utilizados en cada caso; seguido, en el Capítulo 6 se dan conclusiones; y finalmente, en el Apéndice se incluyen las rutinas de programación utilizadas en el ejercicio de simulación.

Como una guía al lector se sugiere la lectura sistemática y completa del Capítulo 3 únicamente si se está interesado en conocer con cierto detalle el Método del Vector Post-Diseño. Similarmente, el Capítulo 4 si se desea una presentación breve de los métodos ya conocidos (Taylor, Woodruff, Jackknife, Bootstrap). De otro modo, se remite al lector al Capítulo 5 en donde se realiza la evaluación empírica del desempeño de los métodos. No obstante, se sugiere la lectura del Capítulo 6, donde se resumen los hallazgos del capítulo anterior, siempre que no se desee entrar en detalles relativos al ejercicio de comparación de los métodos. El Apéndice es sólo para aquellos lectores interesados en las rutinas de programación de los métodos considerados.

[3]Aunque su utilidad se encuentra muy restringida en la práctica, es el diseño de muestreo contra el cual se comparan todos los diseños muestrales. Al respecto véase la parte 5.3.2 en la página 42, inclusive se refieren fuentes y comentarios de P. K. Ollila sobre el uso del Método del Vector Post-Diseño, para diseños muestrales diferentes.

[4]Mayores detalles se describen en el Capítulo 5 que inicia en la página 38.

3. Método de Estimación de Varianza usando el Vector Post-Diseño

3.1. Introducción

Una manera reciente de abordar el concepto de muestreo probabilístico a partir de una población es el *Enfoque del Vector Diseño*, en el que tanto la población como la muestra son tratados como vectores de dimensión igual al número de individuos que integran la población. Este enfoque conlleva una definición alternativa de conceptos básicos esenciales de la teoría de muestreo como por ejemplo, el diseño muestral. Se permite entonces, un manejo diferente de los desarrollos matemáticos necesarios para efectos de estimación de características de una población a partir de una muestra. Finalmente, el enfoque en cuestión servirá para el ajuste de criterios en la reutilización de datos de la muestra por medio del remuestreo con el objeto de lograr estimaciones de varianzas.

En este capítulo se describen de manera breve algunas definiciones básicas, el Enfoque del Vector Diseño estudiado por Ollila (2004), apoyado en los trabajos de Traat (2000) y Traat *et al.* (2000); así como algunas de sus correspondientes propiedades teóricas, definiciones del muestreo a partir de un conjunto de muestras o remuestras, la varianza condicional como estimador de la varianza, posteriormente la definición del Vector Post-Diseño y la obtención de un método de estimación de varianza a partir de este enfoque propuesto por Ollila (2004)[Cap. 5].

3.2. Definiciones Básicas, Nomenclatura y Notación

Se tiene una *población* con N *individuos* (o *elementos*) denotada por $U = \{u_1, \ldots, u_N\}$, para cada individuo u_i se tiene la medición x_{ij} de la variable $j = 1, 2, \ldots, p$. Obtenemos entonces, una *matriz de datos de la población* X de tamaño $N \times p$. A la función de esta matriz de datos, $g(X) = \theta$, se le denominará *parámetro de la población* o simplemente *parámetro.*

Una *muestra ordenada* de n individuos se denota como $os = (u_{(1)}, u_{(2)}, \ldots, u_{(n)})$ donde $u_{(1)}$ es el primer elemento extraído de la población para conformar la muestra, $u_{(2)}$ el segundo y así sucesivamente (Cassel *et al.* (1977)). Una *muestra vector* (o *vector diseño*[5]) es denotada por $\underline{k} = (k_1, k_2, \ldots, k_N)$ donde $k_i \in \{0, 1, \ldots\}$ es el número de veces que el elemento u_i, con $(i = 1, 2, \ldots, N)$, aparece en la muestra. Así, $\underline{k}$ es un punto en el espacio N-dimensional de enteros no negativos. Entonces, se tienen n^N muestras ordenadas diferentes y para cada muestra vector se tienen $n!/\prod_{i=1}^{N} k_i!$ muestras ordenadas diferentes. Nótese que en la muestra vector se pierde el orden de extracción de los individuos.

Para el tamaño de la población N fijo, el tamaño de muestra n puede ser cualquier entero no negativo, entonces se tiene un conjunto infinito[6] de muestras ordenadas denominado *espacio muestral* $\mathcal{S} = \{os_1, os_2, \ldots\}$. Se define el *diseño muestral* o *diseño de muestreo* como la función $p(\cdot)$ en $\mathcal{S}$,

[5]Nomenclatura que se utilizará posteriormente junto con la definición de vector post-diseño.

[6]Es infinito pues no hay restricciones en el número de veces en que determinado elemento o individuo aparezca en la muestra.

$$p(\cdot) : \{os_1, os_2, \ldots\} \longrightarrow \{p(os_1), p(os_2), \ldots\},$$

tal que,

$$p(os) \geq 0, \, \forall os \in \mathcal{S} \quad \text{y} \quad \sum_{os \in \mathcal{S}} p(os) = 1.$$

Esta función proporciona la probabilidad $p(os)$ de seleccionar una muestra os de $\mathcal{S}$. Aquí la muestra os es un valor tomado por la variable aleatoria OS, cuya distribución está dada por $p(\cdot)$.

En la práctica se toma el *espacio muestral restringido* que toma únicamente las muestras con probabilidad positiva bajo el diseño muestral considerado o con un tamaño de muestra n restringido a un intervalo o a un valor fijo. De hecho se puede uno dar cuenta de que el diseño muestral define al espacio muestral restringido.

El estimador correspondiente al parámetro de la población θ, será denotado por $\widehat{\theta}$ y está definido sobre el espacio muestral $\mathcal{S}$, mismo que a su vez está definido por el diseño muestral $p(\underline{k})$. La estimación sobre una muestra específica $\underline{k}$ se denotará $\widehat{\theta}_{\underline{k}}$. $\widehat{\theta}$ normalmente puede arrojar estimaciones para todas las muestras del espacio muestral aunque puede suceder que no, debido al diseño muestral (e.g. muestras con reemplazo en donde todos los individuos seleccionados sean el mismo, de modo que no sería posible calcular el coeficiente de correlación).[7]

Se denomina *esquema muestral* a las reglas de cómo incluir individuos en la muestra para la obtención de una muestra a partir de una población, de modo que la definición de probabilidades descritas por el diseño muestral sean cumplidas. Por ejemplo puede hablarse de un esquema muestral *con* o *sin reemplazo*: en el primer caso las probabilidades de extracción de cada individuo no varía de extracción en extracción i.e. la extracción es independiente del resto de individuos; en el segundo caso las probabilidades de extracción varían de acuerdo a la extracción previa.

3.3. El Enfoque del Vector Diseño

En Ollila (2004), basándose en los trabajos de Traat (2000) y Traat *et al.* (2000), se describe el muestreo en términos de vectores y se considera la naturaleza distribucional del diseño muestral. El vector $\underline{I} = (I_1, I_2, \ldots, I_N)$ es un *vector diseño aleatorio* donde I_i representa el número de veces que el individuo $i \in \{1, 2, \ldots, N\}$ fue seleccionado. La realización correspondiente de dicho vector aleatorio es el *vector diseño* (o muestra vector) $\underline{k} = (k_1, k_2, \ldots, k_N)$ donde, como ya se dijo, $\underline{k}$ es un punto en el espacio N-dimensional de enteros no negativos. Bajo este enfoque, a la distribución multivariada del vector $\underline{I}$ se le denomina *diseño muestral*. La función de probabilidad asociada al vector diseño aleatorio $\underline{I}$ es,

$$p(\underline{k}) = Pr\{\underline{I} = \underline{k}\}, \quad \text{con} \sum_{\underline{k}} p(\underline{k}) = 1,$$

donde $\underline{I} = \underline{k}$ cuando $I_i = k_i \; \forall i \in \{1, 2, \ldots, N\}$, $k_i \in \{0, 1, \ldots\}$. Entonces, las sumas

$$\sum_{i=1}^{N} I_i \quad \text{y} \quad \sum_{i=1}^{N} k_i \; ,$$

[7] Ollila (2004).

proporcionan el tamaño de muestra aleatorio y de su realización respectivamente.

A continuación se presentan tres diseños muestrales comunes que se utilizarán posteriormente en la descripción del método de estimación de varianza que utiliza el vector post-diseño[8], propuesto por Ollila (2004)[Cap. 5].

Primero, el *diseño SI* [9], o en términos de distribuciones tradicionales, la *distribución Bernoulli multivariada simple*, cuya función de probabilidad es,

$$p(\underline{k}) = \binom{N}{n}^{-1}, \tag{1}$$

donde $k_i \in \{0, 1\}$ y $\sum_{i=1}^{N} k_i = n$; $p(\underline{k}) = 0$ en otro caso.

El denominado *diseño multinomial*, denotado $M(n; p_1, \ldots, p_N)$, que es un diseño de muestreo con probabilidades proporcionales al tamaño de alguna variable auxiliar (o en términos de distribuciones tradicionales, una *distribución multinomial*), es un diseño muestral con reemplazo de la forma,

$$p(\underline{k}) = n! \prod_{i=1}^{N} p_i^{k_i} / k_i!, \tag{2}$$

donde $k_i \in \{0, 1, \ldots, n\}$ y $\sum_{i=1}^{N} k_i = n$, $\sum_{i=1}^{N} p_i = 1$; $p(\underline{k}) = 0$ en otro caso.

Un caso especial del diseño muestral anterior con probabilidades iguales de extracción es el *diseño SIR* [10], denominado también *diseño multinomial simple*, y es de la forma,

$$p(\underline{k}) = n! / \left[N^n \prod_{i=1}^{N} k_i! \right], \tag{3}$$

donde $k_i \in \{0, 1, \ldots\}$ y $\sum_{i=1}^{N} k_i = n$; $p(\underline{k}) = 0$ en otro caso. Nótese que en general cuando se utiliza muestreo con reemplazo (WR) $k_i \in \{0, 1, \ldots\}$ mientras que cuando el muestreo es sin reemplazo (WOR) se tiene la restricción $k_i \in \{0, 1\}$.

Posteriormente será necesario el estudio de la partición específica del espacio muestral creada por los *conteos de ocurrencia* c_g, donde c_g representa el número de individuos o elementos de la población que fueron seleccionados g veces, $g \in \{0, 1, \ldots, n\}$. En otras palabras y en otros términos (vector diseño), el conteo de ocurrencia c_g denota el número de apariciones del valor g en cierta muestra vector (o vector diseño) $\underline{k}$. Acorde con lo anterior, se tiene que

$$\sum_{g=0}^{n} g c_g = n. \tag{4}$$

Entonces para un muestreo WOR se tiene $c_0 = N - n$ y $c_1 = n$.

[8] El concepto de Vector Post-Diseño así como el método asociado a éste serán definidos más adelante.
[9] SI = Muestreo Aleatorio Simple.
[10] SIR = Muestreo Aleatorio Simple con Reemplazo.

3.4. Muestreo a partir de un Conjunto de Muestras o Remuestras

Menciona Ollila (2004) que el muestrear a partir de una población puede considerarse como un caso particular de *muestrear a partir del conjunto de todas las muestras*, denominado (por él mismo) *metamuestreo*. También que es raro en la práctica seleccionar varias muestras del conjunto de muestras, aunque para efectos de estimación de varianza este esquema es ya conocido, por ejemplo el *Método de Grupos Aleatorios Independientes*, cuya idea tuvo su origen en trabajos de Mahalanobis (1939, 1944, 1946) durante su estancia en el Instituto de Estadística de la India; y también con el trabajo de Deming (1956)[11].

En el caso de *muestrear a partir del conjunto de todas las remuestras* (donde una remuestra es aquella submuestra extraída a partir de la muestra), uno llega a métodos familiares como el de *Replicaciones de Bootstrap*[12], algunos de *Grupos Aleatorios* y el *Jackknife*.

También, Ollila (2004) menciona que: En el contexto de remuestreo donde algunos de los métodos tienen restricciones imposibles de manejar en términos de la terminología ordinaria de diseños, el concepto de *metamuestrear* resulta muy útil.

Acorde con los principios de la teoría de muestreo, se define una *metamuestra* $os^* = S_1, S_2, \ldots, S_A$, donde $S_1, S_2, \ldots, S_A$ son muestras del conjunto de muestras. El *diseño metamuestral* (o *diseño de metamuestreo*) $p(os^*)$ define la distribución de probabilidad de la metamuestra os^*. El *espacio metamuestral* $\mathcal{S}^*$ es un conjunto de metamuestras tales que $p(os^*) > 0$. Entonces, análogamente para el conjunto de remuestras condicionadas por la muestra S_a se tiene el diseño metamuestral $p(os^*|S_a)$ y para el espacio metamuestral la condición $p(os^*|S_a) > 0$. El muestrear de manera repetida de la población (e.g. grupos aleatorios independientes) es un caso especial del diseño metamuestral en donde $p(os^*) = p(S_1)p(S_2)\ldots p(S_A)$. En el contexto del remuestreo, de manera análoga, se sostiene lo mismo.

3.5. El uso de la Varianza Condicional sobre el Espacio de Remuestras para la Estimación de la Varianza

Una posibilidad en el reuso de datos que proporciona una muestra S_a para la estimación de la varianza $V(\widehat{\theta}_a)$ es la utilización de la *varianza condicional del estimador sobre el espacio remuestral (o espacio de remuestras)* $\widehat{\theta}_b$ *dada la muestra vector* $\underline{k}_a$, donde a y b se refieren a la primera y segunda fase de muestreo respectivamente (muestreo y remuestreo), i.e.

$$V(\widehat{\theta}_b|\underline{k}_a) = \sum_{\underline{k}_b|\underline{k}_a} p(\underline{k}_b|\underline{k}_a)\left(\widehat{\theta}_{\underline{k}_b} - E[\widehat{\theta}_b|\underline{k}_a]\right)^2, \tag{5}$$

donde $\widehat{\theta}_{\underline{k}_b}$ es el estimador $\widehat{\theta}_b$ evaluado en la remuestra $\underline{k}_b$, $\underline{k}_b|\underline{k}_a$ es una posible remuestra que puede ser tomada a partir de $\underline{k}_a$, la suma se efectúa sobre todas estas remuestras, $p(\underline{k}_b|\underline{k}_a)$ es el diseño remuestral (o diseño de remuestreo), y

$$E[\widehat{\theta}_b|\underline{k}_a] = \sum_{\underline{k}_b|\underline{k}_a} p(\underline{k}_b|\underline{k}_a)\widehat{\theta}_{\underline{k}_b}, \tag{6}$$

[11] Más sobre este método en Särndal *et al.* (1992)[pag. 423-430] y Wolter (1985).

[12] Detalles del uso de este método se presentan posteriormente en la parte 4.5 y también en la implementación del Método del Vector Post-Diseño, en la parte 3.7.2.

es la *esperanza condicional del estimador* $\widehat{\theta}_b$ *dada la muestra vector* $\underline{k}_a$.

La *esperanza de la varianza condicional*

$$\begin{aligned} E_a[V(\widehat{\theta}_b|\underline{I}_a)] &= \sum_{\underline{k}_a} p(\underline{k}_a) V(\widehat{\theta}_b|\underline{k}_a) \\ &= \sum_{\underline{k}_a} p(\underline{k}_a) \sum_{\underline{k}_b|\underline{k}_a} p(\underline{k}_b|\underline{k}_a) \left(\widehat{\theta}_{\underline{k}_b} - E[\widehat{\theta}_b|\underline{k}_a]\right)^2, \end{aligned} \tag{7}$$

y el *error cuadrático medio condicional del estimador* $\widehat{\theta}_b$ *dada la muestra vector* $\underline{k}_a$

$$MSE(\widehat{\theta}_b|\underline{k}_a) = \sum_{\underline{k}_b|\underline{k}_a} p(\underline{k}_b|\underline{k}_a) \left(\widehat{\theta}_{\underline{k}_b} - \widehat{\theta}_a\right)^2. \tag{8}$$

3.5.1. Metamuestreo a partir del Espacio de Remuestras para la Estimación de la Varianza Condicional de un Estimador Utilizando el Método de Replicaciones de Bootstrap

Un caso particular de metamuestrear es la selección de remuestras independientes (Ollila (2004)). Este tipo de selección puede derivar en el uso del *Método de Replicaciones de Bootstrap* en donde la selección de la metamuestra os^* se lleva a cabo utilizando el diseño de remuestreo original $p(S_b|S_a)$ en cada extracción de la remuestra S_b. De modo que, un estimador (insesgado) de la varianza condicional es,

$$\widehat{V}(\widehat{\theta}_b|S_a) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}_{S_b} - \bar{\widehat{\theta}}\right)^2}{A-1}, \tag{9}$$

donde A (idealmente muy grande para una buena estimación) es el número de remuestras independientes de la metamuestra os^*, $\widehat{\theta}_{S_b}$ es el estimador evaluado en la remuestra S_b, $\widehat{\theta}_b$ es el estimador evaluado en todo el espacio remuestral y

$$\bar{\widehat{\theta}} = \sum_{S_b \in os^*} \frac{\widehat{\theta}_{S_b}}{A}, \tag{10}$$

es la media metamuestral de los estimadores evaluados en todas las remuestras de la metamuestra os^*.

La expresión (9) es un estimador insesgado de la varianza condicional pues como las remuestras S_b's son independientes entonces las estimaciones $\widehat{\theta}_{S_b}$'s son también independientes, así se tiene entonces que

$$E_a[\widehat{V}(\widehat{\theta}_b|S_a)] = V(\widehat{\theta}_b|S_a).$$

Es posible utilizar (10) como estimador puntual (Bootstrap) de θ, cuya varianza (según Särndal *et al.* (1992)[pag. 52] y Wolter (1985)[pag. 33]) es,

$$\widehat{V}(\bar{\widehat{\theta}}) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}_{S_b} - \bar{\widehat{\theta}}\right)^2}{A(A-1)}. \tag{11}$$

Notar que la ecuación anterior es igual a la ecuación (9) excepto por una A adicional que divide.

El estimador del error cuadrático medio condicional del estimador $\widehat{\theta}_b$ dada la muestra S_a es,

$$\widehat{MSE}(\widehat{\theta}_b|S_a) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}_{S_b} - \widehat{\theta}_a\right)^2}{A-1}, \tag{12}$$

donde $\widehat{\theta}_a$ es la estimación sobre la muestra S_a.

3.5.2. Corrección de la Varianza Condicional sobre el Espacio de Remuestras para Utilizarla como Estimador de la Varianza

Menciona Ollila (2004) que casi todos los métodos de remuestreo planteados en la literatura relacionada con muestreo requieren que la *condición del caso lineal* se cumpla. Esto es que, *para el caso de estimadores lineales*, la varianza de remuestreo del estimador $\widehat{V}_{res}(\widehat{\theta}_a)$ y la varianza analítica del estimador basada en la muestra, $\widehat{V}(\widehat{\theta}_a)$, sean iguales, i.e.

$$\widehat{V}_{res}(\widehat{\theta}_a) = \widehat{V}(\widehat{\theta}_a). \tag{13}$$

Existen diversas estrategias a seguir para alcanzar esta condición. En Ollila (2004)[Sec. 3.6] se lista una relación completa de éstas. Un coeficiente de escala comúnmente utilizado para la corrección de la varianza condicional, basada en el espacio remuestral, es el denominado *Coeficiente del Caso Lineal* y es el siguiente:

$$\widehat{Q}_{lin} = \frac{\widehat{V}(\widehat{\theta}_a)}{V(\widehat{\theta}_b|\underline{k}_a)}, \tag{14}$$

que resulta (como se verá en la Tabla 3.5.2 siguiente) una constante cuando el estimador $\widehat{\theta}$ es lineal y no depende de la variable de mediciones de interés en la que está basado el parámetro θ.

Por ejemplo, para el caso en el que $\widehat{\theta} = \bar{y}$, de acuerdo con (14), se tiene que

$$\widehat{Q}_{lin} = \frac{\widehat{V}(\bar{y}_a)}{V(\bar{y}_b|\underline{k}_a)}. \tag{15}$$

A continuación se tienen en la tabla[13] siguiente los coeficientes del caso lineal $\widehat{Q}_{lin}$ para algunas combinaciones de diseños de la primera y segunda fase, i.e. muestreo y remuestreo respectivamente

[13]En la tabla, SI = *Muestreo Aleatorio Simple* ; SIR = *Muestreo Aleatorio Simple con Reemplazo.*

Tabla 3.5.2 Coeficientes del Caso Lineal

Muestreo	*Remuestreo*	$\widehat{V}(\bar{y}_a)$	$V(\bar{y}_b\|\underline{k}_a)$	$\widehat{Q}_{lin} = \frac{\widehat{V}(\bar{y}_a)}{V(\bar{y}_b\|\underline{k}_a)}$
SI	SI	$\frac{N-n_a}{Nn_a}S^2_{y,a}$	$\frac{n_a-n_b}{n_an_b}S^2_{y,a}$	$\frac{(N-n_a)n_b}{N(n_a-n_b)}$
SI	SIR	$\frac{N-n_a}{Nn_a}S^2_{y,a}$	$\frac{n_a-1}{n_an_b}S^2_{y,a}$	$\frac{(N-n_a)n_b}{N(n_a-1)}$
Multinomial	SI	$\frac{1}{N^2n_a}\dot{S}^2_{y,a}$	$\frac{n_a-n_b}{N^2n_an_b}\dot{S}^2_{y,a}$	$\frac{n_b}{n_a-n_b}$
Multinomial	SIR	$\frac{1}{N^2n_a}\dot{S}^2_{y,a}$	$\frac{1}{N^2n_an_b}\dot{S}^2_{y,a}$	$\frac{n_b}{n_a-1}$

con $n_a - n_b > 0$, donde:

$$S^2_{y,a} = \frac{1}{n_a - 1}\sum_{i\in S_a}\left(y_i - \bar{y}_{S_a}\right)^2, \tag{16}$$

con $\bar{y}_{S_a} = \frac{1}{n_a}\sum_{i\in S_a} y_i$, y:

$$\dot{S}^2_{y,a} = \frac{1}{n_a - 1}\sum_{i\in S_a}\left((y_i/p_i) - \frac{1}{n_a}\sum_{i\in S_a}(y_i/p_i)\right)^2. \tag{17}$$

Se tiene que tener presente el hecho de que el diseño SIR es un caso particular del diseño Multinomial.

También hay que resaltar que cuando se utilizan combinaciones de los diseños SI y SIR, en muestreo o remuestreo, tenemos que

$$\widehat{Q}_{lin} = \frac{V(\bar{y}_a)}{E_a[V(\bar{y}_b|\underline{k}_a)]}, \tag{18}$$

ya que $E_a[\widehat{V}(\bar{y}_a)] = V(\bar{y}_a)$. Esto se justifica porque, acorde con la expresión (16), bajo los diseños SI o SIR se tiene que:

$$E_a\left[\frac{1}{n_a - 1}\sum_{i\in S_a}\left(y_i - \bar{y}_{S_a}\right)^2\right] = \frac{1}{N-1}\sum_{i\in U}\left(y_i - \bar{y}_U\right)^2. \tag{19}$$

En Ollila (2004)[Sec. 3.6] se hace mención de que la varianza condicional, basada en el espacio remuestral, de estimadores no lineales puede no comportarse de manera similar a la varianza condicional de estimadores lineales. Este fenómeno ocurre muy frecuentemente en poblaciones y muestras pequeñas, o en la práctica al tener estratos pequeños. También se hace notar que con poblaciones y tamaños de muestra grandes la importancia del escalamiento y la corrección del diseño disminuye bastante y entonces la principal atención se centra en el diseño de remuestreo y en su efecto para la estimación de varianza.

La corrección que se requiere en la varianza condicional, basada en el espacio de remuestras, para satisfacer la condición del caso lineal puede hacerse de manera externa a la varianza condicional (*escalamiento externo*). Otra posibilidad es hacer esta corrección internamente (*escalamiento interno*) i.e. ajustando las variables en las que se basa θ o ajustando los pesos de muestreo. El método del vector post-diseño utiliza un escalamiento interno original, evitando los problemas usuales (véase Ollila (2004)[Secs. 3.7 y 5]) a los que se enfrenta este tipo de escalamiento, que será descrito a continuación.

3.6. El Vector Post-Diseño

Acorde con las definiciones asociadas al enfoque del vector diseño de la sección 3.3, la muestra vector (o vector diseño) $\underline{k}$ representa la realización del vector diseño aleatorio $\underline{I}$ con tamaño de muestra n. Es posible realizar una modificación posterior de la muestra aumentando ciertas $k_j > 0$ de la muestra vector $\underline{k}$ i.e. aumentar la frecuencia de estas observaciones en la muestra; consecuentemente, estas alteraciones repercutirán en el proceso de estimación basado en la muestra.

Sea $\underline{k} = (k_1, k_2, \ldots, k_N)$ una muestra vector (o vector diseño) realizada de tamaño de muestra n, se expandirá (i.e. se aumentará la aparición de elementos en muestra) con el vector $\underline{d} = (d_1, d_2, \ldots, d_N)$, $d_i \geq 1$ con d_i no necesariamente entero. Se define al *vector post-diseño* como $\underline{k}^* = (d_1k_1, d_2k_2, \ldots, d_Nk_N)$ (Ollila (2004)[Sec. 5.2]) cuyo tamaño de muestra es

$$n^* = \sum_{i=1}^{N} d_i k_i. \tag{20}$$

En el contexto de remuestreo se traduce en que se tiene una remuestra $\underline{k}_b$ y esta es convertida en la remuestra $\underline{k}_b^*$ *para efecto de estimación de varianza*. La expansión resultante del uso del vector post-diseño puede llevarse a cabo de cualquier forma, en particular en Ollila (2004) se alteran únicamente las primeras $n-1$ observaciones en una *muestra ordenada* (de modo que azarosamente queda una observación sin alteración), así los criterios de expansión de frecuencias de los k-valores en muestra escogidos a ser alterados no quedan determinados subjetivamente. Por ejemplo, se tiene una muestra de tamaño n, se fija $d_j = 1$ para alguna $j \in \{1, 2, \ldots, N\}$ y para el resto se multiplica por q, entonces los nuevos conteos de ocurrencia (acorde con las definiciones dadas en la sección 3.3) del vector post-diseño son $c_0^* = N - n$, $c_r^* = q(c_r - 1) + 1$, y para el resto qc_r, con $r \in \{0, 1, \ldots, n\}$[14].

Una observación muy importante por parte de Ollila (2004) es que la aplicación del vector post-diseño no está diseñada para arrojar una estimación puntual del estimador $\widehat{\theta}_a$, pues la expansión llevada a cabo crea una pérdida de balance en los estimadores lineales y en la mayoría de los no lineales, aumentando su varianza y haciéndolos menos efectivos. *Su principal aplicación es en el contexto de remuestreo para la estimación de varianza $V(\widehat{\theta}_a)$, del estimador $\widehat{\theta}_a$.*

[14]El *conteo de ocurrencia*, c_g, representa el número de elementos en la población que fueron seleccionados g veces, con $g \in \{0, 1, \ldots, n\}$. Es decir, el número de apariciones del valor g en cierta muestra vector (o vector diseño) $\underline{k}$.

3.7. Estimación de Varianza con Estimadores del Vector Post-Diseño

Se pretende estimar la varianza $V(\widehat{\theta}_a)$ del estimador $\widehat{\theta}_a$ calculado sobre la muestra vector $\underline{k}_a$ de tamaño de muestra n_a. Si se quiere utilizar remuestreo es necesario corregir las diferencias entre el diseño de muestreo y el diseño de remuestreo. Aún en el caso en que ambos diseños fueran el mismo, por ejemplo SI, estos de entrada son diferentes por el hecho de que el primero tiene el tamaño de muestra n_a y el segundo n_b. Esto finalmente causa diferencias entre el espacio muestral y el espacio remuestral. Como ya se hizo notar anteriormente, una común corrección implementada el coeficiente del caso lineal $\widehat{Q}_{lin} = \frac{\widehat{V}(\bar{y}_a)}{V(\bar{y}_b|\underline{k}_a)}$. Adicionalmente se tiene el hecho de que $\widehat{Q}_{lin}$ no depende de la variable de mediciones de interés y.

Una estrategia es ajustar el diseño de remuestreo de modo que $\widehat{Q}_{lin} = 1$. Así la varianza condicional $V(\widehat{\theta}_b|\underline{k}_a)$ o su estimador Bootstrap serán el estimador de $V(\widehat{\theta}_a)$. Para lograr esto usualmente basta con encontrar el tamaño de remuestra n_b apropiado. Este tamaño de muestra de remuestreo casi siempre es un número no entero por lo que se requieren tomar los valores enteros vecinos de n_b, es decir $n_{b,l}$ (el entero menor más cercano a n_b) y $n_{b,u}$ (el entero mayor más cercano a n_b)y llevar a cabo un *proceso de aleatorización* i.e. utilizar los dos tamaños de remuestra siguiendo una estructura de probabilidad predeterminada.

El método de la presente sección, propuesto por Ollila (2004), *evita el proceso de aleatorización* utilizando un vector post-diseño específico para la fase de remuestreo (denotada con el subíndice b). El resultado del método es que se tiene un tamaño de remuestra fijo y que la varianza condicional $V(\widehat{\theta}^*_b|\underline{k}_a)$ o su estimador Bootstrap

$$\widehat{V}(\widehat{\theta}^*_b|\underline{k}_a) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}^*_{S_b} - \bar{\widehat{\theta}}^*\right)^2}{A-1}, \tag{21}$$

son estimadores insesgados (véase lo comentado en la página 11) de la varianza $\widehat{V}(\widehat{\theta}_a)$, donde $\widehat{\theta}^*_{S_b}$ es el estimador $\widehat{\theta}$ evaluado en el vector post-diseño $\underline{k}^*_b = (d_1 k_{b1}, d_2 k_{b2}, \ldots, d_N k_{bN})$; A es el número de remuestras independientes de la metamuestra os^*, y

$$\bar{\widehat{\theta}}^* = \sum_{S_b \in os^*} \frac{\widehat{\theta}^*_{S_b}}{A}. \tag{22}$$

Entonces Ollila (2004)[Sec. 5.3] sintetiza que el objetivo del método es crear una situación donde la varianza condicional del estimador post-diseño $\bar{y}^*_b$ que utiliza $\underline{k}^*_b = (d_1 k_{b1}, d_2 k_{b2}, \ldots, d_N k_{bN})$ sea igual a la varianza condicional del estimador remuestral $\bar{y}_b$ que use el tamaño de muestra n_b, en otras palabras, satisfacer la condición del caso lineal asegurando que

$$V(\bar{y}^*_b|\underline{k}_a) = V(\bar{y}_b|\underline{k}_a, n_b), \tag{23}$$

donde ulteriormente $V(\bar{y}^*_b|\underline{k}_a)$ será utilizado para estimar la varianza del estimador basado en el diseño muestral de la primera fase (muestreo) $\widehat{V}(\bar{y}_a)$. De modo que es posible encontrar los valores $d_i \geq 1$ que componen al vector $\underline{d} = (d_1, d_2, \ldots, d_N)$ necesarios para la creación de un vector post-diseño $\underline{k}^*_b = (d_1 k_{b1}, d_2 k_{b2}, \ldots, d_N k_{bN})$ que satisface (23).

Un ejemplo de los principios descritos anteriormente sería el siguiente: Tómese la forma de

expansión en la que $d_j = 1$ para alguna j y $d_i = q$, con $i \neq j$, bajo un esquema de muestreo WOR[15]. Se tiene entonces

$$\underline{k}_b^* = (k_{b1}^*, k_{b2}^*, \ldots, k_{bN}^*) = (qk_{b1}, \ldots, qk_{bj-1}, 1, qk_{bj+1}, \ldots, qk_{bN}), \tag{24}$$

en donde j es la unidad o elemento de la remuestra S_b con restricción en la expansión. Se toma el tamaño de remuestra $n_{b,u}$, que es el redondeo al entero mayor más cercano a n_b. Los conteos de ocurrencia son entonces: $c_0 = N - n_{b,u}$, $c_1 = 1$, $c_q = n_{b,u} - 1$. Por lo tanto se tiene entonces una solución para la construcción del estimador de la varianza,

$$\widehat{V}(\widehat{\theta}_a) = V(\widehat{\theta}_b^* | \underline{k}_a). \tag{25}$$

Nótese que para un esquema de remuestreo WOR está contemplada la situación en que $n_{b,u} = n_a$ pues, siguiendo las definiciones de la sección 3.2, se tiene que hay $n_a! / \prod_{g=0}^{n_a} c_g! = n_a!/1!(n_a - 1)! = n_a$ estimaciones.

Entonces se utiliza la varianza condicional $V(\widehat{\theta}_b^* | \underline{k}_a)$ para la estimación de $\widehat{V}(\widehat{\theta}_a)$ o la aproximación por Bootstrap de ésta, $\widehat{V}(\widehat{\theta}_b^* | \underline{k}_a) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}_{S_b}^* - \overline{\widehat{\theta}}^*\right)^2}{A-1}$, de acuerdo a la ecuación (9) en la página 11, donde A es el número de remuestras independientes que componen la metamuestra os^*, $\widehat{\theta}_{S_b}^*$ es el estimador que utiliza el vector post-diseño $\underline{k}_b^* = (d_1 k_{b1}, d_2 k_{b2}, \ldots, d_N k_{bN})$, y $\overline{\widehat{\theta}}^* = \sum_{S_b \in os^*} \frac{\widehat{\theta}_{S_b}^*}{A}$.

3.7.1. Obtención de un Factor q de Expansión para el Vector Post-Diseño

A continuación se mostrará la obtención de un vector post-diseño que satisfaga (25), para las combinaciones de diseño SI en la primera fase (muestreo) y diseños SI y SIR en la segunda (remuestreo).

Para la obtención del tamaño de remuestra necesario de modo que sea satisfecha la condición del caso lineal se requiere despejar el término n_b de la expresión correspondiente al diseño muestral y remuestral de acuerdo a la Tabla 3.5.2, en la parte 3.5.2 de la presente tesis, que contiene los coeficientes del caso lineal.

En el caso SI/SI (diseño muestral SI y diseño remuestral SI) se tiene que el coeficiente del caso lineal y la respectiva expresión para el término n_b, que es de nuestro interés, son:

$$\begin{aligned} & & \widehat{Q}_{lin} &= \frac{(N - n_a) n_b}{N(n_a - n_b)} \\ & & &= \frac{(1 - f_a) n_b}{(1 - f_b) n_a}, \\ \widehat{Q}_{lin} = 1 \quad \Longrightarrow \quad & & n_b &= \frac{N n_a}{2N - n_a} \\ & & &= \frac{(1 - f_b)}{(1 - f_a)} n_a. \end{aligned} \tag{26}$$

[15] WOR = Sin reemplazo.

Análogamente para el caso SI/SIR (diseño muestral SI y diseño remuestral SIR) son:

$$\begin{aligned} \widehat{Q}_{lin} &= \frac{(N-n_a)n_b}{N(n_a-1)}, \\ \widehat{Q}_{lin}=1 \quad \Longrightarrow \quad n_b &= \frac{N(n_a-1)}{N-n_a}. \end{aligned} \tag{27}$$

Ahora, tómese una remuestra $\underline{k}_b$ de tamaño $n_{b,u}$ (entero mayor más cercano a n_b). En este caso en la expansión a utilizar, uno de los elementos o individuos de la remuestra no será expandido por el factor q. Entonces para el vector post-diseño $\underline{k}_b^*$, resultado de la expansión de $\underline{k}_b$, el *tamaño de remuestra artificial* será

$$n_b^* = \sum_{i=1}^{N} k_{bi}^* = (n_{b,u}-1)q+1. \tag{28}$$

Adicionalmente, se tiene el término siguiente (que será utilizado posteriormente en la expresión de la varianza),

$$\sum_{i=1}^{N} k_{bi}^{*2} = (n_{b,u}-1)q^2 + 1\cdot 1^2 + (N-n_{b,u})\cdot 0^2 = (n_{b,u}-1)q^2+1. \tag{29}$$

La remuestra tomada $\underline{k}_b$ tiene asociados sus correspondientes valores de los conteos de ocurrencia $c_0 = N-n_{b,u}, c_1, c_2, \ldots, c_{n_{b,u}}$ para un diseño SIR de remuestreo (WR) o $c_0 = N-n_{b,u}, c_1 = n_{b,u}$ en el caso de un diseño SI de remuestreo (WOR).

Por otro lado, de acuerdo a los desarrollos mostrados por Ollila (2004)[Sec. 5.1, Eq.(5.1.6)], se tiene que

$$V(\bar{y}_b|c_0,\ldots,c_{n_{b,u}};\underline{k}_a) = \left(\frac{\sum_{i=1}^{N} k_{bi}^2}{n_{b,u}^2} - \frac{1}{n_a}\right)\widehat{S}_{y_a}^2. \tag{30}$$

Entonces, utilizando (30) para lograr satisfacer (23) y recordando que se está utilizando un diseño SI en la primera fase (denotada con el subíndice a) se obtiene el siguiente desarrollo,

$$\begin{aligned} V(\bar{y}_b^*|\underline{k}_a) &= \widehat{V}(\bar{y}_a) \\ \Longrightarrow \quad \left(\frac{\sum_{i=1}^{N} k_{bi}^{*2}}{n_b^{*2}} - \frac{1}{n_a}\right)\widehat{S}_{y_a}^2 &= \left(\frac{1}{n_b}-\frac{1}{n_a}\right)\widehat{S}_{y_a}^2 \\ \Longrightarrow \quad \frac{\sum_{i=1}^{N} k_{bi}^{*2}}{n_b^{*2}} &= \frac{1}{n_b}. \end{aligned} \tag{31}$$

Ahora, sustituyendo (28) y (29) en (31) obtenemos

$$\frac{(n_{b,u}-1)q^2+1}{\left[(n_{b,u}-1)q+1\right]^2} = \frac{1}{n_b}, \tag{32}$$

y finalmente resolviendo (32) para q se tiene que,

$$q = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}, \tag{33}$$

donde

$$A = (n_{b,u} - 1)^2 - n_b(n_{b,u} - 1), \tag{34}$$
$$B = 2(n_{b,u} - 1), \tag{35}$$
$$C = 1 - n_b, \tag{36}$$

y se toma la mayor de las raíces de q (véase Ollila (2004)[Sec. 5.1]).

Entonces, el vector post-diseño que satisface (25) para un diseño con reemplazo de remuestreo (SIR) queda de la forma $\underline{k}^* = (qk_1, \ldots, q(k_j - 1) + 1, \ldots, qk_N)$ y con los conteos de ocurrencias del vector post-diseño $c_0^* = N - n_{b,u}$, $c_r^* = q(c_r - 1) + 1$, y para el resto qc_r. Análogamente para un diseño sin reemplazo de remuestreo (SI), $\underline{k}^* = (qk_1, \ldots, qk_{j-1}, 1, qk_{j+1}, \ldots, qk_N)$, con los conteos de ocurrencias $c_0^* = N - n_{b,u}$, $c_1^* = 1$, $c_q^* = n_{b,u} - 1$.

En la práctica, para la expansión vista y para el factor q calculado, se expanden con el factor q los primeros $n_b - 1$ elementos de la remuestra ordenada y la última unidad se deja sin alterar.

3.7.2. Implementación del Método del Vector Post-Diseño en la Estimación de Varianza de Diferentes Estimadores

A continuación se detalla la estimación de la varianza de diferentes estimadores lineales y no lineales para un diseño SI en la primera fase (muestreo), y diseños SI y SIR en la segunda (remuestreo). Se utilizará un factor q único y la restricción de la expansión vista anteriormente. Los estimadores a implementar son: Total, Media, Mediana y Razón de los totales de dos variables; asociados a los correspondientes parámetros de la población para la variable de estudio y o el par de variables y y x según sea el caso.

En todos los casos se requiere del conocimiento del tamaño de la población N y de una muestra (no importa si es o no ordenada) S_a, de tamaño n_a, que contiene las mediciones de la variable y. También para todos los estimadores se calcula primero el tamaño de remuestra n_b que satisfaga la condición del caso lineal, esto se logra despejando el término de la expresión (véase la Tabla 3.5.2), $\widehat{Q}_{lin} = 1$, acorde con el diseño muestral y remuestral. Se obtiene $n_{b,u}$, el entero mayor más cercano a n_b. Posteriormente se calcula el factor q de acuerdo a la expresión (33) con ayuda de las ecuaciones (34), (35) y (36) de la página 18.

Posteriormente se procede a calcular la estimación de la varianza condicional $V(\widehat{\theta}_b^*|\underline{k}_a)$ por medio de su aproximación de Bootstrap (siguiendo la expresión (9) en la página 11),

$$\widehat{V}(\widehat{\theta}_b^*|\underline{k}_a) = \sum_{S_b \in os^*} \frac{\left(\widehat{\theta}_{S_b}^* - \bar{\widehat{\theta}}^*\right)^2}{A - 1}, \tag{37}$$

donde A es el número de remuestras independientes que componen la metamuestra os^*, $\widehat{\theta}_{S_b}^*$ es el estimador[16] que utiliza el vector post-diseño $\underline{k}_b^* = (k_{b1}^*, k_{b2}^*, \ldots, k_{bN}^*) = (d_1 k_{b1}, d_2 k_{b2}, \ldots, d_N k_{bN})$

[16] Adelante se detalla la expresión correspondiente a cada estimador.

asociado a la remuestra S_b, y

$$\bar{\widehat{\theta}}^* = \sum_{S_b \in os^*} \frac{\widehat{\theta}^*_{S_b}}{A}. \tag{38}$$

Para la Estimación de la Varianza del Estimador de la Media y del Total

Se sigue el procedimiento descrito en párrafos anteriores y se define el *estimador del vector post-diseño para la media*, $\bar{y}^*_b$, necesario en las fórmulas (37) y (38) anteriores, como:

$$\widehat{\theta}^*_{S_b} = \bar{y}^*_b = \frac{\underline{k}^*_b U_y^T}{n^*_b} = \frac{\sum_{i=1}^{N} k^*_{bi} y_i}{n^*_b}, \tag{39}$$

donde $\underline{k}^*_b = (k^*_{b1}, k^*_{b2}, \ldots, k^*_{bN})$ es el vector post-diseño, U_y^T es el vector transpuesto de $U_y = (y_1, \ldots, y_N)$ que es el vector de mediciones de la variable y en la población y donde n^*_b es el tamaño de remuestra artificial posterior a la expansión.

Alternativamente, una forma más sencilla de aplicar y programar la fórmula anterior (39) es la misma ecuación pero no en términos del vector de mediciones en la población y del vector post-diseño sino en términos de las mediciones de la variable y para la remuestra *ordenada*[17] S_b de tamaño $n_{b,u}$ i.e. $S_{b_y} = (y_{b1}, \ldots, y_{bn_{b,u}})$, de la expansión en curso (con la restricción sobre un individuo) y del factor de expansión único q del vector post-diseño,

$$\widehat{\theta}^*_{S_b} = \bar{y}^*_b = \underline{z} S^T_{b_y} = \frac{y_{bn_{b,u}} + q \sum_{i=1}^{n_{b,u}-1} y_{bi}}{n^*_b}. \tag{40}$$

donde $\underline{z} = \frac{1}{n^*_b}(q, \ldots, q, 1)$ es un vector de tamaño $n_{b,u}$, $S^T_{b_y}$ es el vector transpuesto de S_{b_y} y n^*_b es el tamaño de remuestra artificial posterior a la expansión.

En lo que respecta al *estimador del vector post-diseño para el total*, y^*_b, basta con multiplicar las expresiones (39) y (40) por el tamaño de la población N,

$$\widehat{\theta}^*_{S_b} = y^*_b = N\bar{y}^*_b. \tag{41}$$

[17] Acorde con el concepto de muestra ordenada definido en la parte 3.2 en la página 7 de la presente tesis, aunque no con la misma notación (os) para no confundir con la notación de metamuestra (os^*). Es ordenada con el objeto de que la restricción en la expansión, de tener un individuo cualquiera sin expandir, se aplique de manera aleatoria al seleccionar el individuo que corresponda al último elemento extraído en la remuestra ordenada.

Para la Estimación de la Varianza del Estimador de la Mediana

Al igual que para el estimador de la media se sigue el procedimiento descrito en los primeros párrafos de la parte 3.7.2. Se define a la *mediana* como el percentil 50, es decir el valor de la variable de interés en aquel elemento o individuo en el que se acumula el 50 % de los datos al ordenar los individuos de manera ascendente respecto a la variable en cuestión. En caso de que no sea posible reportar este valor se tomó aquel más cercano que no supere el 50 %, de acuerdo a la definición. Es importante notar que, aunque se sigue muy cercanamente la teoría de estimación de la mediana descrita en Särndal *et al.* (1992)[pag. 197-204], no se toma a la mediana como el punto medio del intervalo de valores que satisfacen la condición (44) descrita más adelante y en la referencia, sino que se toma al extremo inferior de ese intervalo, acorde con la definición de mediana utilizada en la presente tesis. También se realizaron ajustes en los procedimientos de Särndal *et al.* (1992) pues en el método del presente capítulo se tienen expansiones no enteras de frecuencias.

Entonces, se requiere de una estimación $\widehat{F}(y)$ de la función de distribución acumulada $F(y)$ de la variable de interés y. Ésta se obtiene con el procedimiento siguiente:

1. Se toman las mediciones de la variable y para la remuestra *ordenada*[17] S_b de tamaño $n_{b,u}$, i.e. $S_{b_y} = (y_{b1}, \ldots, y_{bn_{b,u}})$, perteneciente a la metamuestra os^* extraída y se crea un vector $\underline{z} = \frac{1}{n_b^*}(q, \ldots, q, 1)$ de tamaño $n_{b,u}$ en donde $z_i \in \underline{z}$ contiene la frecuencia relativa del valor y_{bi}, posterior a la expansión indicada por el método del vector post-diseño. Y donde n_b^* es el tamaño de remuestra artificial una vez ya realizada la expansión.

2. Se ordenan ambos vectores con respecto a los valores del vector S_{b_y} de manera ascendente. Una vez ordenado el vector S_{b_y} se denotará como $S_{b_y}^{'} = (y_{b1}^{'}, \ldots, y_{bn_{b,u}}^{'})$.

3. Luego, se genera otro vector $\underline{z}^{'}$ que contiene las frecuencias relativas acumuladas correspondientes a los valores y_{bi}'s, y es de la forma $\underline{z}^{'} = (z_1^{'}, z_2^{'}, \ldots, z_{n_{b,u}}^{'})$ donde,

$$z_i^{'} = \begin{cases} z_1 & i = 1 \\ z_{i-1}^{'} + z_i & 1 < i \leq n_{b,u} \end{cases} \tag{42}$$

   de modo que la primera entrada del vector $\underline{z}^{'}$ es el valor de la primera entrada del vector $\underline{z}$ y la última debe ser 1, i.e. $z_1^{'} = z_1$, y $z_{n_{b,u}}^{'} = 1$.

4. Entonces, la función de distribución acumulada estimada $\widehat{F}(y)$ de la variable y queda definida como,

$$\widehat{F}(y) = \begin{cases} 0 & y < y_{b1}^{'} \\ z_1^{'} & y_{b1}^{'} \leq y < y_{b2}^{'} \\ z_2^{'} & y_{b2}^{'} \leq y < y_{b3}^{'} \\ \vdots & \vdots \\ z_i^{'} & y_{b(i)}^{'} \leq y < y_{b(i+1)}^{'} \\ \vdots & \vdots \\ z_{n_{b,u}-1}^{'} & y_{b(n_{b,u}-1)}^{'} \leq y < y_{bn_{b,u}}^{'} \\ 1 & y_{bn_{b,u}}^{'} \leq y \end{cases} \tag{43}$$

Una vez obtenida $\widehat{F}(y)$ se define (acorde con la notación utilizada en Särndal *et al.* (1992)) el *estimador del vector post-diseño para la mediana*, $\widehat{M}^*_{y_b}$, necesario en (37) y (38) como,

$$\widehat{\theta}^*_{S_b} = \widehat{M}^*_{y_b} = \widehat{F}^{-1}(0.5), \tag{44}$$

donde $\widehat{F}^{-1}(\cdot)$ es la función inversa de $\widehat{F}(\cdot)$.

En lo que respecta a hallar $\widehat{M}^*_{y_b}$ utilizando $\widehat{F}^{-1}(\cdot)$, en la presente tesis se utilizó el procedimiento siguiente:

1. Se ubica el máximo de los elementos del vector (ordenado en forma ascendente) $S^{'}_{b_y} = (y^{'}_{b1}, \ldots, y^{'}_{bn_{b,u}})$, tal que $\widehat{F}(y^{'}_{bj}) \leq 0.5$, para $i \in \{1, \ldots, n_{b,u}\}$.

2. Luego, como el factor q expande todas las frecuencias de los elementos de $S^{'}_{b_y}$ exceptuando a un elemento (esto por la restricción de expansión usada en la presente tesis) y como $q \geq 1$, se revisa si el elemento $y^{'}_{b(j+1)}$ de $S^{'}_{b_y}$ es la mediana.

Es decir,

$$\widehat{M}^*_{y_b} = \begin{cases} y^{'}_{bj} & y^{'}_{bj} = max\{y^{'}_{bj} \in S^{'}_{b_y} | \widehat{F}(y^{'}_{bj}) \leq 0.5 < \widehat{F}(y^{'}_{bj}) + \frac{1}{n^*_b}\} \\ y^{'}_{b(j+1)} & y^{'}_{bj} = max\{y^{'}_{bj} \in S^{'}_{b_y} | \widehat{F}(y^{'}_{bj}) + \frac{1}{n^*_b} \leq 0.5\} \end{cases} \tag{45}$$

Todo este procedimiento puede ser utilizado para cualquier percentil, basta con sustituir en cada ocasión la cantidad 0.5 por la correspondiente al percentil deseado[18].

Para la Estimación de la Varianza del Estimador de la Razón de los Totales de Dos Variables

De nueva cuenta, al igual que para los estimadores anteriores, se sigue el procedimiento descrito en los primeros párrafos de la parte 3.7.2. Una vez hecho esto, se define el *estimador del vector post-diseño para la razón* (de los totales de dos variables), $\widehat{R}^*_{(y/x)b}$, como:

$$\widehat{\theta}^*_{S_b} = \widehat{R}^*_{(y/x)b} = \frac{\bar{y}^*_b}{\bar{x}^*_b} = \frac{\underline{k}^*_b U^T_y}{\underline{k}^*_b U^T_x} = \frac{\sum_{i=1}^{N} k^*_{bi} y_i}{\sum_{i=1}^{N} k^*_{bi} x_i}, \tag{46}$$

donde $\underline{k}^*_b = (k^*_{b1}, k^*_{b2}, \ldots, k^*_{bN})$ es el vector post-diseño, U^T_y y U^T_x son los vectores transpuestos de $U_y = (y_1, \ldots, y_N)$ y $U_x = (x_1, \ldots, x_N)$ que son los vectores de mediciones de las variables y y x en la población respectivamente.

Al igual que en el caso del estimador para la media, una forma alternativa de la expresión anterior (46) es en términos de las mediciones de las variables y y x para la remuestra *ordenada*[19]

[18] La definición de la mediana utilizada en la presente tesis es diferente a la utilizada en Särndal *et al.* (1992). Véase también la parte 4.3.1 en la página 26 de la presente tesis.

[19] Véase la nota al pie número 17.

S_b de tamaño $n_{b,u}$, i.e. $S_{b_y} = (y_{b1}, \ldots, y_{bn_{b,u}})$ y $S_{b_x} = (x_{b1}, \ldots, x_{bn_{b,u}})$, de la expansión en curso (con la restricción descrita anteriormente) y del factor de expansión único q del vector post-diseño,

$$\widehat{\theta}^*_{S_b} = \widehat{R}^*_{(y/x)b} = \frac{\underline{z} S^T_{b_y}}{\underline{z} S^T_{b_x}} = \frac{y_{bn_{b,u}} + q \sum_{i=1}^{n_{b,u}-1} y_{bi}}{x_{bn_{b,u}} + q \sum_{i=1}^{n_{b,u}-1} x_{bi}}, \tag{47}$$

donde $\underline{z} = \frac{1}{n^*_b}(q, \ldots, q, 1)$ es un vector de tamaño $n_{b,u}$, $S^T_{b_y}$ y $S^T_{b_x}$ son los vectores transpuestos de S_{b_y} y S_{b_x} respectivamente.

Entonces, (46) y (47) se utilizan en las expresiones (37) y (38) anteriores.

4. Estimación de Varianza Utilizando Algunos Métodos Conocidos

4.1. Introducción

La varianza de un estimador tiene gran importancia en la generalidad de procesos científicos que infieren acerca de una población a partir de la disponibilidad incompleta de datos de esta población. Su importancia fundamentalmente reposa en que la varianza de un estimador provee de elementos o nociones de la calidad del estimador mismo. Adicionalmente, el conocimiento de la varianza de un estimador es crucial en el diseño de un esquema de muestreo para la estimación de determinado parámetro de interés en una población.

El problema de la estimación de varianza de un estimador surge, como se menciona en la introducción de la presente tesis y en cualquier bibliografía que considere temas relacionados al muestreo probabilístico, de la imposibilidad de calcular la varianza del estimador pues se requiere de información de toda la población. Como respuesta a esta complicación se han propuesto diversos métodos de estimación de varianza. De hecho, el problema de estimación de varianza ocupa un gran espacio en la teoría del muestreo probabilístico y es común encontrarlo aludido como un área de interés o linea de investigación per se.

En el presente capítulo se presentan de manera resumida los métodos de *Linealización de Taylor*, *Woodruff* (1952), *Jackknife* y de replicaciones de *Bootstrap*. En cada caso se da una breve introducción, una presentación sintética del método y se apuntan de manera explícita las expresiones necesarias para su implementación.

4.2. Estimación de la Varianza con el Método de Linealización de Taylor

La técnica de aproximación por *Linealización de Taylor*[20], método conocido en matemáticas, puede de manera relativamente fácil ser utilizada para la estimación de varianza de estimadores en el muestreo probabilístico. Esta técnica consiste en *hacer una aproximación lineal de estimadores no lineales, permitiendo así el empleo de teoría utilizada para estimadores lineales.*

De acuerdo con lo expuesto en Wolter (1985)[Cap. 6] y en Särndal *et al.* (1992)[pag. 172-176], el hecho de utilizar un pseudo-estimador, aproximación lineal por Taylor del estimador no lineal original, y emplear en él la teoría de muestreo conocida para la estimación de varianza *puede conducir a estimadores sesgados, pero usualmente consistentes*[21], de la varianza del estimador no lineal. Un interesante tratado de consideraciones y ventajas del uso de este método se encuentra en Kish (1987)[pag. 131-137].

En Wolter (1985)[pag. 221] se hace especial énfasis en que *el método de linealización de Taylor*

[20] En algunas fuentes también denominado Método Delta, e.g. Kish (1987)[pag. 133].

[21] Brevemente, acorde con Särndal *et al.* (1992)[pag. 166-169], sea τ un parámetro estimado por el estimador $\widehat{\tau}_n$, una función de n variables aleatorias independientes idénticamente distribuidas $\xi_1, \xi_2, \ldots, \xi_n$. El estimador $\widehat{\tau}_n$ se dice *consistente* para τ si, para cualquier $\epsilon > 0$ fija, se tiene que $lim_{n\to\infty} Pr\left\{|\widehat{\tau}_n - \tau| > \epsilon\right\} = 0$.

no produce estimadores de varianza per se, no puede actuar solo en la estimación de varianza; sino que éste, como ya se mencionó, produce aproximaciones lineales de la estadística muestral de interés y que entonces posteriormente se necesita de otros métodos o teoría para estimar la varianza de la aproximación lineal obtenida.

A continuación se hace una breve presentación del método de linealización de Taylor para estimación de varianza de estimadores no lineales en poblaciones *finitas*, tomando como base lo expuesto en Särndal *et al.* (1992)[pag. 172-176]. Posteriormente se hacen explícitas las expresiones a utilizar en la estimación de varianza del estimador del cociente de los totales de dos variables para un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

4.2.1. Definición del Método de Linealización de Taylor

Se considera solamente a un parámetro θ que puede ser expresado como una *función* de q totales, $t_1, \ldots, t_q$, es decir $\theta = f(t_1, \ldots, t_q)$ donde,

$$t_j = \sum_{k \in U} y_{jk} \tag{48}$$

con $j = 1, \ldots, q$. El parámetro θ se estima por medio del estimador $\widehat{\theta} = f(\widehat{t}_1, \ldots, \widehat{t}_q)$, donde,

$$\widehat{t}_j = \sum_{k \in S} \frac{y_{jk}}{\pi_k} \tag{49}$$

para $j = 1, \ldots, q$, y donde $\pi_k = Pr\{k \in S\}$, es la *probabilidad de inclusión*[22] del individuo k en la muestra S.

Considérese el caso en el que f es no lineal, entonces el *Método de Linealización de Taylor* lo que hace es aproximar el estimador no lineal $\widehat{\theta}$ con un *pseudo-estimador*, denotado $\widehat{\theta}_0$, que es una *función lineal* de $\widehat{t}_1, \ldots, \widehat{t}_q$. ($\widehat{\theta}_0$ *no es un estimador en sí mismo, sino una aproximación de* $\widehat{\theta}$.) Si la aproximación es buena, entonces la variable aleatoria lineal $\widehat{\theta}_0$ *emulará lo mejor posible* al estimador $\widehat{\theta}$ y por lo tanto se podrá utilizar a $V(\widehat{\theta}_0)$ como *aproximación* de $V(\widehat{\theta})$. Entonces, de acuerdo a la expansión de primer orden de la serie de Taylor de la función f alrededor del punto $(t_1, \ldots, t_q)$ y desechando los términos restantes, tenemos,

$$\widehat{\theta} \doteq \widehat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j \left(\widehat{t}_j - t_j\right), \tag{50}$$

con

$$a_j = \left.\frac{\partial f}{\partial \widehat{t}_j}\right|_{(\widehat{t}_1, \ldots, \widehat{t}_q) = (t_1, \ldots, t_q)}. \tag{51}$$

Se tiene entonces que,

$$V(\widehat{\theta}) \doteq V(\widehat{\theta}_0) = V\left(\sum_{j=1}^{q} a_j \widehat{t}_j\right), \tag{52}$$

[22] Esta probabilidad depende del esquema o diseño de muestreo utilizado en la obtención de la muestra S. Véase Särndal *et al.* (1992)[pag. 30].

por lo tanto,

$$\widehat{V}(\widehat{\theta}) \doteq \widehat{V}(\widehat{\theta}_0). \tag{53}$$

Obsérvese que $\widehat{\theta}_0$ depende de valores poblacionales del parámetro y de valores estimados.

4.2.2. Implementación del Método de Linealización de Taylor para la Razón de los Totales de Dos Variables

En la presente sección se hacen explícitas las expresiones a utilizar para la estimación de varianza del estimador de la razón de los totales de dos variables, la pareja de variables y y x, bajo un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

Sea S una muestra de tamaño n subconjunto de una población $U = \{1, \ldots, N\}$ de tamaño N, y sean y_i y x_i las mediciones de la variables y y x, respectivamente, asociadas al individuo i-ésimo. El parámetro de interés es,

$$\theta = \frac{\sum\limits_{k \in U} y_k}{\sum\limits_{k \in U} x_k} = \frac{\bar{y}_{_U}}{\bar{x}_{_U}}, \tag{54}$$

y su correspondiente estimador,

$$\widehat{\theta} = \frac{\sum\limits_{k \in S} \frac{y_k}{\pi_k}}{\sum\limits_{k \in S} \frac{x_k}{\pi_k}} = \frac{\sum\limits_{k \in S} N \frac{y_k}{n}}{\sum\limits_{k \in S} N \frac{x_k}{n}} = \frac{\sum\limits_{k \in S} y_k}{\sum\limits_{k \in S} x_k} = \frac{\bar{y}_{_S}}{\bar{x}_{_S}}, \tag{55}$$

donde $\bar{y}_{_U}$ es la media de la variable y sobre el conjunto de elementos que conforman la población U y $\bar{y}_{_S}$ es la media de la variable y sobre la muestra S; análogamente sucede con $\bar{x}_{_U}$ y $\bar{x}_{_S}$ para la variable x.

Se tiene entonces las fórmulas siguientes (Ver Särndal *et al.* (1992)[pag. 179-180]), resultado del seguimiento de las definiciones expuestas anteriormente en la presentación del método de linealización de Taylor,

$$\widehat{\theta}_0 = \theta + \frac{1}{\bar{x}_{_U}} \frac{1}{n} \sum_{k \in S} (y_k - \theta x_k) = \theta + \frac{\bar{y}_{_S} - \theta \bar{x}_{_S}}{\bar{x}_{_U}}, \tag{56}$$

$$V(\widehat{\theta}_0) = \frac{1}{\bar{x}_{_U}^2} \frac{1 - \frac{n}{N}}{n} \frac{1}{N-1} \sum_{k \in U} (y_k - \theta x_k)^2, \tag{57}$$

$$\widehat{V}(\widehat{\theta}) = \frac{1}{\bar{x}_{_S}^2} \frac{1 - \frac{n}{N}}{n} \frac{1}{n-1} \sum_{k \in S} \left(y_k - \widehat{\theta} x_k \right)^2. \tag{58}$$

Obsérvese que son expresiones no difíciles de calcular. En la expresión (58) puede usarse $\bar{x}_{_S}$ o $\bar{x}_{_U}$, se prefiere esta última cuando se conoce con precisión.

4.3. Estimación de la Varianza del Estimador de la Mediana con el Método de Woodruff

El *Método de Woodruff para Cuantiles* (1952), se encuentra fundamentalmente basado en la ingeniosa idea de utilizar expresiones analíticas correspondientes a la creación de intervalos de confianza empleando una estimación de la función de distribución acumulada de la variable de interés.

Kish (1965)[pag. 495] menciona que el uso de la mediana es de especial interés al trabajar con variables altamente asimétricas, cuando la mediana difiere considerablemente de la media (e.g. el ingreso económico). Se plantea también que en tales distribuciones la varianza de la mediana puede ser menor a la varianza de la media pues esta última está fuertemente influenciada por valores muy elevados de las colas de la distribución. Esto último se verifica más adelante en la presente tesis en el capítulo de resultados de la simulación.

En la presente sección se hace una breve presentación del método de Woodruff para la mediana en particular, tomando como base lo expuesto en Woodruff (1952) y en Särndal *et al.* (1992)[pag. 197-204]. Adicionalmente se muestran los procedimientos necesarios para su implementación y se explicitan las fórmulas a utilizar bajo un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

4.3.1. Definición e Implementación del Método de Woodruff para la Mediana

Se define, en la presente tesis, a la *mediana* como el percentil 50, es decir el valor de la variable de interés asociado a aquel elemento en la población en el que se acumula el 50 % de los datos al ordenar los individuos de manera ascendente respecto a las mediciones reportadas por la variable en cuestión. En caso de que no sea posible reportar este valor (debido a que no se obtenga exactamente el 50 %) se toma aquel *más cercano que no supere* el 50 %. En otros términos, sea y la variable de interés, M_y la correspondiente mediana poblacional de la variable y y $\widehat{M}_y$ su correspondiente estimador (basado en datos muestrales):

$$M_y = F^{-1}(0.5), \tag{59}$$

$$\widehat{M}_y = \widehat{F}^{-1}(0.5), \tag{60}$$

donde $F(y)$ es la función de distribución acumulada de la variable y (basada en datos de la población U) y $\widehat{F}(y)$ es su correspondiente estimador (basado en datos de la muestra S).

Es importante resaltar que, aunque se sigue muy cercanamente la teoría de estimación de la mediana descrita en Särndal *et al.* (1992)[pag. 197-204], *no se toma a la mediana como el punto medio* del intervalo de valores que satisfacen la condición (59) o (60) según sea el caso, sino que se toma al *extremo inferior* de ese intervalo pues cualquier valor contenido en él es válido como mediana.

Si se tomase el punto medio del intervalo se perdería información sobrestimando, pues como se verá más adelante en el capítulo de resultados de las simulaciones, se utilizará una variable asociada al ingreso de hogares en México y los intervalos obtenidos de valores que satisfacen (59) o (60) (según corresponda) serán muy amplios. Por consiguiente, se reportarían valores de medición de la variable de interés *muy altos* respecto al correspondiente al extremo inferior de los intervalos

y además se reportarían valores *inexistentes* en la muestra o población según sea el caso. Por lo tanto utilizar el punto medio *no refleja o describe* (en estas circunstancias) las condiciones precisas de la variable en la muestra o en la población (véase el segundo gráfico de la figura 5.1 en Särndal *et al.* (1992)[pag. 198].).

Para la obtención de $\widehat{F}(y)$ se hace lo siguiente: Se ordenan los elementos de la muestra $S = \{1, \ldots, n\}$ con respecto a los valores de la variable y y se obtiene el vector $S_y = (y_1, \ldots, y_n)$ con $y_i \leq y_{i+1}$ y donde y_i representa la medición de la variable y asociada al individuo i-ésimo. Luego a cada valor y_i se le asocia el valor z_i del vector de frecuencias relativas acumuladas de la forma $\underline{z} = \frac{1}{n}(z_1, \ldots, z_n)$ donde,

$$z_i = \begin{cases} 1 & i = 1 \\ z_{i-1} + 1 & 1 < i \leq n \end{cases} \tag{61}$$

Entonces, la función de distribución acumulada estimada $\widehat{F}(y)$ de la variable y queda definida como,

$$\widehat{F}(y) = \begin{cases} 0 & y < y_1 \\ \frac{1}{n} z_1 & y_1 \leq y < y_2 \\ \vdots & \vdots \\ \frac{1}{n} z_i & y_i \leq y < y_{i+1} \\ \vdots & \vdots \\ \frac{1}{n} z_{n-1} & y_{n-1} \leq y < y_n \\ 1 & y_n \leq y \end{cases} \tag{62}$$

Luego, para obtener $\widehat{M_y}$ a partir de $\widehat{F}^{-1}(\cdot)$,

$$\widehat{M_y} = \widehat{F}^{-1}(0.5) = max \left\{ y_j \in S_y | \widehat{F}(y_j) \leq 0.5 < \widehat{F}(y_j) + \frac{1}{n} \right\}. \tag{63}$$

Este procedimiento puede ser utilizado para cualquier *cuantil* o *percentil*, basta con sustituir en (63) la cantidad 0.5 por la correspondiente.

Para la estimación de varianza del estimador de la mediana (o cualquier otro cuantil) se utiliza la idea siguiente,

$$Pr\{c_1 \leq \widehat{F}(M_y) \leq c_2\} \doteq Pr\{\widehat{F}^{-1}(c_1) \leq M_y \leq \widehat{F}^{-1}(c_2)\}, \tag{64}$$

para cualesquiera dos constantes c_1 y c_2. Luego, se iguala (64) con la cantidad 0.95 y entonces el intervalo $[\widehat{F}^{-1}(c_1), \widehat{F}^{-1}(c_2)]$ es aproximadamente un intervalo al 95 % de confianza para M_y. Asumiendo que $\widehat{F}(M_y)$ se distribuye aproximadamente de forma normal alrededor de su valor esperado $E[\widehat{F}(M_y)] = F(M_y) \doteq 0.5$; supuesto que se cumple para *tamaños de muestra grandes*. Entonces,

$$c_1 = 0.5 - Z_{(0.975)} \sqrt{V(\widehat{F}(M_y))} \quad , \quad c_2 = 0.5 + Z_{(0.975)} \sqrt{V(\widehat{F}(M_y))}, \tag{65}$$

y entonces,

$$
\begin{aligned}
V(M_y) &= \left(\frac{\widehat{F}^{-1}(c_2) - \widehat{F}^{-1}(c_1)}{2(Z_{(0.975)})} \right)^2 \\
&= \left(\frac{\widehat{F}^{-1}\left(0.5 + Z_{(0.975)}\sqrt{V(\widehat{F}(M_y))}\right) - \widehat{F}^{-1}\left(0.5 - Z_{(0.975)}\sqrt{V(\widehat{F}(M_y))}\right)}{2(Z_{(0.975)})} \right)^2 \qquad (66)
\end{aligned}
$$

por lo tanto, para estimar $V(\widehat{M}_y)$ se utiliza,

$$
\widehat{V}(\widehat{M}_y) = \left(\frac{\widehat{F}^{-1}\left(0.5 + Z_{(0.975)}\sqrt{\widehat{V}(\widehat{F}(\widehat{M}_y))}\right) - \widehat{F}^{-1}\left(0.5 - Z_{(0.975)}\sqrt{\widehat{V}(\widehat{F}(\widehat{M}_y))}\right)}{2(Z_{(0.975)})} \right)^2, \qquad (67)
$$

donde para un intervalo al 95 % de confianza $Z_{(0.975)}$ es el valor inverso de una función de distribución acumulada Normal con parámetros 0 y 1, evaluada en 0.975. Entonces, $Z_{(0.975)} = 1.96$.

Nótese que $\widehat{F}(\widehat{M}_y) \doteq \widehat{F}(M_y) \doteq 0.5$ y que $E[\widehat{F}(M_y)] = F(M_y) \doteq 0.5$. Entonces, para un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.) se tiene que,

$$
V(\widehat{F}(M_y)) = \frac{N-n}{N-1}\frac{1}{n}F(M)\{1-F(M)\} \doteq \frac{\left(1-\frac{n}{N}\right)}{n}0.25, \qquad (68)
$$

entonces,

$$
\widehat{V}(\widehat{F}(\widehat{M}_y)) = \frac{N-n}{N-1}\frac{1}{n}\widehat{F}(\widehat{M})\{1-\widehat{F}(\widehat{M})\} \doteq \frac{\left(1-\frac{n}{N}\right)}{n}0.25, \qquad (69)
$$

pues $n\widehat{F}(M_y)$ y $n\widehat{F}(\widehat{M}_y)$ se distribuyen como *variables aleatorias hipergeométricas* (véase Särndal *et al.* (1992)[pag. 203] y la observación en Mood *et al.* (1974)[pag. 92]).

Todo este mismo procedimiento para la estimación de la varianza de la mediana o cualquier cuantil viene también descrito de manera sucinta en Kish (1965)[Sec. 12.9 pag. 495-496]. Ahí mismo se presenta una extensión del método para la diferencia de dos medianas.

En Mood *et al.* (1974)[pag. 255-259] se da la definición usual de la mediana muestral y también se describen resultados sobre la distribución asintótica de ésta.

4.4. Estimación de Varianza con el Método Jackknife

Uno de los métodos más populares de estimación de varianza y que recientemente ha tenido gran ecumenicidad por su creciente presencia en los paquetes de cómputo comerciales de estadística es el *Método de Replicaciones Repetidas de Jackknife* o comúnmente denominado *Jackknife*. Éste, a diferencia de otros métodos de remuestreo de cómputo intensivo como el de *Replicaciones de Bootstrap* o el de *Medias Muestras Balanceadas* (o *Replicaciones Repetidas Balanceadas*)[23], ha

[23]En Inglés: *Balanced Half Samples* o *Balanced Repeated Replications*.

sido muy estudiado en gran cantidad de bibliografía.

La idea original del Método Jackknife fue desarrollada por Quenouille (1949), originalmente para la corrección o ajuste del sesgo de estimadores, pero es en Quenouille (1956) y en trabajos posteriores de otros investigadores cuando esta idea toma la forma y utilidad presentada en la presente tesis. El Jackknife ha gozado de la atención de varios investigadores que han generalizado el método y lo han extendido para diseños complejos que involucran estratificación y más de una etapa de muestreo. Una descripción detallada del Jackknife se encuentra en Wolter (1985). También, en Efron (1979)[pags. 6,12] se habla de su carencia de consistencia asintótica en la estimación de varianza de la mediana muestral y se demuestra que el Método Jackknife es una aproximación lineal del Método Bootstrap.

En esta sección se hace una breve presentación del Método Jackknife y se hacen explícitas las fórmulas utilizadas en la implementación de éste para algunos estimadores, lineal (media) y no lineal (razón de los totales de dos variables), bajo un diseño de muestreo aleatorio simple sin reemplazo.

4.4.1. Definición del Método Jackknife

Acorde con la notación utilizada en Särndal *et al.* (1992)[Sec. 11.5 pag. 437-442] se define a S como una muestra de tamaño n, subconjunto de una población con N elementos o individuos, $U = \{1, \ldots, k, \ldots, N\}$. Luego, sea θ el parámetro de interés de la población, cuyo estimador correspondiente basado en datos de la muestra S será denotado por $\widehat{\theta}$.

El método Jackknife particiona la muestra S, de tamaño fijo n, en A *grupos aleatorios* disjuntos, $S_1, S_2, \ldots, S_a, \ldots, S_A$, de *igual tamaño* $m = n/A$, es decir:

$$S = \bigcup_{a=1}^{A} S_a \quad , \quad S_a \cap S_b = \phi, a \neq b. \tag{70}$$

La partición de S se realiza conforme a un esquema aleatorio, de modo que cada grupo aleatorio S_a tiene el mismo diseño de muestreo que la muestra original S. Luego, por la ecuación (70) es fácil notar que estos grupos aleatorios no son independientes. Se obtienen entonces A *grupos aleatorios dependientes*.

También se tiene que, para cualquier muestra S dada, cada uno de los grupos S_a es una muestra(remuestra) obtenida con un diseño SI (*Muestreo Aleatorio Simple Sin Reemplazo*) de la muestra original S, *aún cuando S no hubiese sido una muestra SI originalmente*. Posteriormente se calcula $\widehat{\theta}_{(a)}$ para cada $a = 1, \ldots, A$, que es el estimador de θ de la misma forma funcional que $\widehat{\theta}$, pero basado en datos de la muestra S omitiendo el grupo aleatorio S_a.

Se definen entonces los *pseudo-valores*,

$$\widehat{\theta}_a = A\widehat{\theta} - (A-1)\widehat{\theta}_{(a)}, \tag{71}$$

luego se obtiene el estimador Jackknife $\widehat{\theta}_{JK}$ de θ,

$$\widehat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a, \tag{72}$$

y el estimador de varianza Jackknife se define como,

$$\widehat{V}_{JK1} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\widehat{\theta}_a - \widehat{\theta}_{JK} \right)^2 . \tag{73}$$

Särndal *et al.* (1992)[pag. 438] apunta que en la práctica se utiliza $\widehat{V}_{JK1}$ como estimador de $V(\widehat{\theta})$ y de $V(\widehat{\theta}_{JK})$. Una alternativa de $\widehat{V}_{JK1}$ es,

$$\widehat{V}_{JK2} = \frac{1}{A(A-1)} \sum_{a=1}^{A} \left(\widehat{\theta}_a - \widehat{\theta} \right)^2 . \tag{74}$$

También Särndal *et al.* (1992)[pag. 438] señala que $\widehat{V}_{JK2} \geq \widehat{V}_{JK1}$, también que los pseudo-valores $\widehat{\theta}_a$ están *correlacionados* y que por consiguiente el estimador de la varianza *no es insesgado*. En algunos casos, dependiendo del estimador $\widehat{\theta}$ en cuestión, es posible obtener expresiones analíticas exactas del sesgo y por lo tanto es posible eliminarlo (corregirlo). De cualquier forma, cuando esto último no es posible, se sabe (véase Särndal *et al.* (1992)[pag. 440] y Wolter (1985)[pag. 169]) que *el sesgo es despreciable*.

En la práctica, la modalidad del método Jackknife más utilizada es cuando $m = 1$. En este trabajo se explorará más adelante, en el capítulo siguiente, el desempeño del método para $m = 1, 2$, con algunos estimadores.

4.4.2. Implementación del Método Jackknife en la Estimación de Varianza de Algunos Estimadores

A continuación se hacen explícitas las fórmulas a utilizar para la estimación de varianza de algunos estimadores, lineal (Media) y no lineal (Razón de los totales de dos variables), para la variable de estudio y o el par de variables y y x según sea el caso. Se utilizará un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

Para la Estimación de la Varianza del Estimador de la Media

De acuerdo con las definiciones básicas del método Jackknife presentados en la parte 4.4.1, se tienen las fórmulas siguientes cuando el parámetro de interés θ es la media de la variable y en la

población,

$$\theta = \frac{1}{N}\sum_{k\in U} y_k, \tag{75}$$

$$\widehat{\theta} = \frac{1}{n}\sum_{k\in S} y_k = \bar{y}_S, \tag{76}$$

$$\widehat{\theta}_{(a)} = \frac{1}{n-m}\sum_{k\in S-S_a} y_k = \bar{y}_{S-S_a}, \tag{77}$$

$$\widehat{\theta}_a = \frac{1}{m}\sum_{k\in S_a} y_k = \bar{y}_{S_a}, \tag{78}$$

$$\widehat{\theta}_{JK} = \widehat{\theta} = \bar{y}_S, \tag{79}$$

$$\widehat{V}_{JK1} = \widehat{V}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\bar{y}_{S_a} - \bar{y}_S\right)^2, \tag{80}$$

y eliminando el sesgo (véase Särndal *et al.* (1992)[pag. 440] y Wolter (1985)[pag. 169]) se tiene que,

$$\widehat{V}(\widehat{\theta}) = \left(1 - \frac{n}{N}\right)\widehat{V}_{JK1}, \tag{81}$$

es un estimador insesgado de $V(\widehat{\theta})$.

Nótese que en las expresiones anteriores si $A = n$ y $m = 1$ se obtienen las mismas fórmulas, para la estimación de varianza del estimador de la media, que las que se obtendrían desarrollando las fórmulas de los estimadores π ó de Horwitz y Thompson (véase Särndal *et al.* (1992)[pag. 68]).

Para la Estimación de la Varianza del Estimador de la Razón de los Totales de Dos Variables

Acorde con lo presentado en la parte 4.4.1, se tienen las siguientes expresiones cuando el parámetro de interés θ es el cociente del total de la variable y entre el total de la variable x, en la población,

$$\theta = \frac{\sum\limits_{k\in U} y_k}{\sum\limits_{k\in U} x_k}, \tag{82}$$

$$\widehat{\theta} = \frac{\sum\limits_{k\in S} \frac{y_k}{\pi_k}}{\sum\limits_{k\in S} \frac{x_k}{\pi_k}} = \frac{\sum\limits_{k\in S} N\frac{y_k}{n}}{\sum\limits_{k\in S} N\frac{x_k}{n}} = \frac{\sum\limits_{k\in S} y_k}{\sum\limits_{k\in S} x_k} = \frac{\bar{y}_S}{\bar{x}_S}, \tag{83}$$

$$\widehat{\theta}_{(a)} = \frac{\sum\limits_{k\in S-S_a} y_k}{\sum\limits_{k\in S-S_a} x_k} = \frac{\bar{y}_{S-S_a}}{\bar{x}_{S-S_a}}, \tag{84}$$

$$\widehat{\theta}_a = A\widehat{\theta} - (A-1)\widehat{\theta}_{(a)} = A\frac{\bar{y}_S}{\bar{x}_S} - (A-1)\frac{\bar{y}_{S-S_a}}{\bar{x}_{S-S_a}}, \tag{85}$$

$$\widehat{\theta}_{JK} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a, \tag{86}$$

$$\widehat{V}_{JK1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\widehat{\theta}_a - \widehat{\theta}_{JK}\right)^2, \tag{87}$$

$$\widehat{V}_{JK2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\widehat{\theta}_a - \widehat{\theta}\right)^2, \tag{88}$$

y ajustando con el factor de corrección por finitud (Särndal *et al.* (1992)[pag. 440] y Wolter (1985)[pag. 169]) se tiene que,

$$\widehat{V}(\widehat{\theta}) = \left(1 - \frac{n}{N}\right)\widehat{V}_{JK1}, \tag{89}$$

o alternativamente,

$$\widehat{V}^*(\widehat{\theta}) = \left(1 - \frac{n}{N}\right)\widehat{V}_{JK2}, \tag{90}$$

son estimadores de $V(\widehat{\theta})$.

En particular, acorde con Särndal *et al.* (1992)[pag. 438], se prefiere la expresión (89) a la expresión (90), pues los cálculos resultan más económicos en términos de programación.

4.5. Estimación de Varianza con el Método Bootstrap

De los métodos de remuestreo utilizados en la estimación de varianza que, a consideración del autor de la presente tesis, puede considerarse como *el más fascinante*, por la naturaleza sumamente intuitiva de la idea básica que fundamenta el método (es decir, el uso de una población artificial o pseudo-población construida a partir de la muestra original expandida y la posterior extracción de remuestras que imitan al proceso de extracción de la primera)[24]; es el denominado *Método de Replicaciones de Bootstrap*, desarrollado por Efron en 1979, y también conocido simplemente con el nombre de *Bootstrap*.

Llevar a la práctica el Bootstrap requiere de grandes recursos de cómputo, de hecho son pocos los paquetes de cómputo estadístico que lo incluyen y por otro lado, aquellos que de alguna manera lo contemplan no son capaces de aplicarlo considerando mayores complejidades en los diseños muestrales o estimadores.

En esta sección se hace una breve presentación del Método Bootstrap y posteriormente se hacen explícitas las expresiones utilizadas en la implementación de éste para algunos estimadores: lineal (media) y no lineales (mediana y razón de los totales de dos variables), bajo un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

4.5.1. Definición del Método Bootstrap

Se sigue el breve, pero a su vez muy completo, sumario de la idea principal del Método Bootstrap descrito en Särndal *et al.* (1992)[pag. 442]. Se tiene una muestra S de tamaño n, subconjunto de una población con N elementos o individuos, $U = \{1, \ldots, k, \ldots, N\}$. Luego, sea $\widehat{\theta}$ el estimador (basado en datos de la muestra) del parámetro poblacional θ, se pretende estimar $V(\widehat{\theta})$.

Primero, se construye una *población artificial* o *pseudo-población* U^* a partir de los datos de la muestra S y de los correspondientes *factores de expansión*[25] de cada individuo que compone la muestra.

Para un diseño de muestreo aleatorio simple (SRS, SI ó m.a.s.) la probabilidad de inclusión $\pi_k = n/N$ para toda $k \in U$, el correspondiente factor de expansión de cada elemento en muestra es N/n. La construcción de U^* se realiza en la presente tesis de la forma siguiente: Se tiene que,

$$N = qn + r, \tag{91}$$

con $n \leq N$ y $r < n$, donde $q, r \in \{0, 1, 2, \ldots\}$. Se construye un vector en el que cada individuo $k \in S$ aparece q veces. Posteriormente se eligen al azar, sin reemplazo, r elementos de la muestra S y se incluyen en el vector, éste será entonces la pseudo-población U^*, que se piensa *replicará o imitará* las características de la población desconocida U.

Segundo, se extraen A muestras independientes o *remuestras de Bootstrap* de tamaño n a partir de la población artificial U^*, utilizando el diseño o esquema de muestreo utilizado inicialmente

[24] Hay que anotar que esta idea en Efron (1979)[pag. 4] es considerada solamente una forma o caso particular de *"Bootstrapear"* en la estimación de la función de distribución asociada a lo que es de interés.

[25] El factor de expansión $1/\pi_k$ del individuo k se define como el inverso de la *probabilidad de inclusión* $\pi_k = Pr\{k \in S\}$ del individuo k en la muestra S, $\forall k \in U$.

en la extracción de la muestra original S a partir de la población U. La independencia se logra utilizando reemplazo en la extracción de las remuestras (o *metamuestreo con reemplazo*, acorde con las definiciones y nomenclatura de Ollila[26]). Luego, se calcula la estimación $\widehat{\theta}_a^*$ del parámetro θ, con $a = 1, \ldots, A$ utilizando la misma forma funcional del estimador $\widehat{\theta}$.

Tercero, se considera que la distribución resultante de los A estimadores $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_A^*$ del parámetro θ es una estimación de la distribución muestral del estimador $\widehat{\theta}$, y entonces $V(\widehat{\theta})$ es estimado por,

$$\widehat{V}_{BS} = \sum_{a=1}^{A} \frac{\left(\widehat{\theta}_a^* - \widehat{\theta}^*\right)^2}{A-1}, \tag{92}$$

donde

$$\widehat{\theta}^* = \sum_{a=1}^{A} \frac{\widehat{\theta}_a^*}{A}. \tag{93}$$

Computacionalmente, si U^* es muy grande lo que se puede hacer es conservar un vector que contenga a la identificación de los individuos $k \in S$, junto con otro vector de igual dimensión que contenga el número entero de veces que aparece cada individuo en la pseudo-población, acorde con los valores de N, q, n y r de la expresión (91). Y la extracción de las remuestras se harían tomando en cuenta, para cada individuo, el número máximo de veces que puede aparecer en la remuestra según su frecuencia en la población artificial.

En lo que atañe al valor de A, éste no está dado, es una de las constantes a determinar en cualquier caso en el que vaya a usarse el Método Bootstrap; no hay una guia, depende del caso específico y a veces se encuentra empíricamente probando algunos valores a pesar de que se piensa que tiene que ser muy grande para obtener mejores resultados pero esto hace lento el cómputo de estimaciones. En Lehtonen & Pahkinen (2004)[pag. 162] se menciona que A debe estar preferentemente entre 500 y $1,000$. En este trabajo, en el siguiente capítulo, se utilizará $A = 1,000$.

El Método de Bootstrap puede llegar a ser un método de *verdadero computo intensivo* para tamaños de muestra y población grandes, y valores elevados de A, dependiendo de los recursos de cómputo disponibles. No obstante ante la creciente velocidad de las máquinas de cómputo esto en un futuro no muy lejano no será un obstáculo para llevar el método a la práctica cotidiana. Actualmente para su aplicación existen: macros de *SAS*® (e.g. extensión en Internet de Lehtonen & Pahkinen (2004) -sitio Web en la URL de John Wiley & Sons, Ltd.- o en el Apéndice 3 de la edición anterior de Lehtonen & Pahkinen del año 1994), Sintaxis de *SPSS*® (e.g. http://www.spsstools.net), algunas librerías en *MATLAB*® (e.g. StatLib en http://lib.stat.cmu.edu) y el paquete *WesVar*® de Westat (en http://www.westat.com) permite el uso de Bootstrap (de hecho considerando estratificación y conglomeración) si se le proporcionan los pesos de replicación correspondientes.

[26]Véase la parte 3.4 en la página 10.

4.5.2. Implementación del Método Bootstrap en la Estimación de Varianza de Algunos Estimadores

En la presente sección se hacen explícitas las expresiones a utilizar para la estimación de varianza de algunos estimadores, lineal (Media) y no lineales (Mediana y razón de los totales de dos variables), para la variable de estudio y o el par de variables y y x según sea el caso. Se utilizará un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.).

Para la Estimación de la Varianza del Estimador de la Media

Acorde con la definición del Método Bootstrap presentada en la parte 4.5.1, se tienen las fórmulas siguientes cuando el parámetro de interés θ es la media de la variable y en la población,

$$\theta = \frac{1}{N}\sum_{k\in U} y_k, \tag{94}$$

$$\widehat{\theta} = \frac{1}{n}\sum_{k\in S} y_k = \bar{y}_S, \tag{95}$$

y una vez construida la pseudo-población U^* a partir de la información de la muestra S y extraídas las A remuestras de Bootstrap (con reemplazo) de la población artificial, se procede a calcular,

$$\widehat{\theta}^*_a = \frac{1}{n}\sum_{k\in S_a} y_k = \bar{y}_{S_a}, \tag{96}$$

para $a = 1, \ldots, A$. Luego,

$$\widehat{\theta}^* = \sum_{a=1}^{A} \frac{\bar{y}_{S_a}}{A}, \tag{97}$$

y entonces, el estimador Bootstrap de $V(\widehat{\theta})$ es:

$$\widehat{V}_{BS} = \sum_{a=1}^{A} \frac{\left(\bar{y}_{S_a} - \widehat{\theta}^*\right)^2}{A-1}. \tag{98}$$

Para la Estimación de la Varianza del Estimador de la Mediana

Cuando el parámetro de interés θ es la mediana de la variable y en la población, siguiendo la definiciones del Método Bootstrap y la definición de la mediana[27] utilizada en todo momento en la presente tesis, se tienen entonces las siguientes expresiones,

$$\theta = M_y = F^{-1}(0.5), \tag{99}$$

$$\widehat{\theta} = \widehat{M}_y = \widehat{F}^{-1}(0.5), \tag{100}$$

[27] Véase la parte 4.3.1 en la página 26.

donde $F(y)$ es la función de distribución acumulada de la variable y (basada en datos de la población U) y $\widehat{F}(y)$ es su correspondiente estimador (basado en datos de la muestra S).

Para la obtención de $\widehat{F}(y)$ se sigue el procedimiento descrito en la parte 4.3.1 de la presente tesis. Luego, para obtener $\widehat{M}_y$ a partir de $\widehat{F}^{-1}(\cdot)$,

$$\widehat{M}_y = max\left\{y_j \in S_y | \widehat{F}(y_j) \leq 0.5 < \widehat{F}(y_j) + \frac{1}{n}\right\}. \tag{101}$$

Lo anterior es relacionado con la forma funcional del estimador $\widehat{\theta}$. Continuando con la implementación del Método Bootstrap en la estimación de varianza del estimador de la mediana, se construye la pseudo-población U^* a partir de la información de la muestra S y se extraen las A remuestras de Bootstrap (con reemplazo) de la población artificial, y se procede a calcular,

$$\widehat{\theta}_a^* = \widehat{M}_{y,a}^* = \widehat{F}_a^{*-1}(0.5), \tag{102}$$

para cada remuestra S_a con $a = 1, \ldots, A$. Posteriormente,

$$\widehat{\theta}^* = \sum_{a=1}^{A} \frac{\widehat{M}_{y,a}^*}{A}, \tag{103}$$

y entonces, el estimador Bootstrap de $V(\widehat{\theta})$ es:

$$\widehat{V}_{BS} = \sum_{a=1}^{A} \frac{\left(\widehat{M}_{y,a}^* - \widehat{\theta}^*\right)^2}{A-1}. \tag{104}$$

Para la Estimación de la Varianza del Estimador de la Razón de los Totales de Dos Variables

De acuerdo con la definición del Método Bootstrap presentada en la parte 4.5.1, se tienen las fórmulas siguientes cuando el parámetro de interés θ es la razón de los totales de dos variables y y x en la población,

$$\theta = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k}, \tag{105}$$

$$\widehat{\theta} = \frac{\sum_{k\in S} \frac{y_k}{\pi_k}}{\sum_{k\in S} \frac{x_k}{\pi_k}} = \frac{\sum_{k\in S} N\frac{y_k}{n}}{\sum_{k\in S} N\frac{x_k}{n}} = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \frac{\bar{y}_S}{\bar{x}_S}, \tag{106}$$

$$\tag{107}$$

y una vez construida la pseudo-población U^* a partir de la información de la muestra S y extraídas las A remuestras de Bootstrap (con reemplazo) de la población artificial, se procede a calcular,

$$\widehat{\theta}_a^* = \frac{\sum_{k\in S_a} y_k}{\sum_{k\in S_a} x_k} = \frac{\bar{y}_{S_a}}{\bar{x}_{S_a}}, \tag{108}$$

para $a = 1, \ldots, A$. Luego,

$$\widehat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\frac{\bar{y}_{S_a}}{\bar{x}_{S_a}}, \tag{109}$$

y entonces, el estimador Bootstrap de $V(\widehat{\theta})$ es:

$$\widehat{V}_{BS} = \frac{1}{A-1}\sum_{a=1}^{A}\left(\frac{\bar{y}_{S_a}}{\bar{x}_{S_a}} - \widehat{\theta}^*\right)^2. \tag{110}$$

5. Evaluación por Simulación del Uso de Algunos Métodos de Estimación de Varianza

5.1. Introducción

En el presente capítulo se muestran los resultados obtenidos de la simulación del uso de los métodos de estimación de varianza descritos en los capítulos anteriores. Esto último con el objeto de observar su desempeño ante el uso de diversos estimadores asociados a algunas variables o combinaciones de éstas.

Como una primera parte, se especifican detalles relativos a los marcos muestrales y a las variables a utilizar: su fuente, algunos estadísticos descriptivos, tamaño de la población representada por el marco muestral correspondiente y en su caso una representación gráfica de la distribución de las mediciones de interés en la población. Posteriormente se entra en la materia asociada al título del presente capítulo.

5.2. Descripción de las Variables y Especificación de los Marcos Muestrales

Las cuatro variables que se listan a continuación fueron tomadas del Sistema Contar 2000 de la muestra con cuestionario ampliado del XII Censo General de Población y Vivienda del 2000 por parte del Instituto Nacional de Estadística, Geografía e Informática (INEGI).

Tabla 5.2.a Variables a Utilizar

Nombre de la Variable	*Descripción de la Variable*
INGTOHOG	Ingresos [mensuales] totales por hogar.
TOTPERS	Total de personas por hogar.
TOTCUART	Total de cuartos por hogar.
TELEFONO	Disponibilidad de teléfono por vivienda.

Se consideró únicamente la información del estado de Aguascalientes y se tomaron los datos de la muestra realizada en el censo por el INEGI como la población de interés. Se consideraron únicamente los registros válidos conjuntamente, por hogar, para las variables INGTOHOG (con valores diferentes a *cero* y *no especificado*), TOTPERS (con valores diferentes a *no especificado*) y TOTCUART (con valores diferentes a *no especificado*). Por su parte, para la variable TELEFONO (por vivienda) se consideraron los registros con valores diferentes a *no especificado*. De modo que las bases de datos recortadas para esa entidad conformarán los dos marcos muestrales a utilizar en las simulaciones; una base de datos para las tres primeras variables y otra base para la variable TELEFONO.

Tabla 5.2.b Registros Válidos Conjuntamente para los Dos Marcos

Variable	*Unidad*	*Número de Unidades*	*Num. No Especif.*	*Num. cero en INGTOHOG*	*Num. Válidos Conjuntamente*
INGTOHOG	Hogar	19,132	427	1,007	17,492
TOTPERS	Hogar	19,132	0	no aplica	17,492
TOTCUART	Hogar	19,132	221	no aplica	17,492
TELEFONO	Vivienda	18,326	218	no aplica	18,108

A continuación se mostrarán los estadísticos descriptivos asociados a éstas variables.

Tabla 5.2.c Estadísticos Descriptivos de las Variables a Utilizar

Variable	*Número de Casos*	*Mínimo*	*Máximo*	*Media*	*Desviación Estándar*
INGTOHOG	17,492	15	999,998	5,369.311	15,831.653
TOTPERS	17,492	1	28	4.628	2.241
TOTCUART	17,492	1	17	4.143	1.726
TELEFONO	18,108	0	1	0.353	0.478

Como se comentará posteriormente, la distribución de algunas variables (ingresos [mensuales] totales por hogar) requerirán de mayor atención; en consecuencia, se presenta a continuación una tabla con más estadísticos descriptivos.

Tabla 5.2.d Más Estadísticos Descriptivos

	INGTOHOG	TOTPERS	TOTCUART
Estadístico			
Asimetría	52.077	1.143	0.935
Percentiles			
Percentil 5	857	2	2
Percentil 10	1,286	2	2
Percentil 25	2,079	3	3
Percentil 50	3,429	4	4
Percentil 75	6,200	6	5
Percentil 90	10,441	7	6
Percentil 95	14,500	9	7
Valores Extremos Mayores			
1	999,998	28	17
2	999,998	23	16
3	999,998	22	15
4	859,714	20	15
5	223,314	19	15

En la Tabla 5.2.d se puede notar la alta asimetría que tiene la distribución de la variable INGTOHOG en la población de estudio; esto se confirma observando el estadístico asimetría con un valor de 52.077, muy alejado del valor 1 que representaría simetría perfecta. Así también, observando con detenimiento los percentiles en conjunto con la Tabla 5.2.c, se puede observar que el percentil 95 es 14,500, valor que dista mucho del valor máximo 999,998 (valor que por cierto es

el límite superior establecido por parte de INEGI en la recopilación de la información). En otras palabras, 95 por ciento de los individuos en la población de estudio tienen un valor inferior a 14,500 en la variable INGTOHOG.

Adicionalmente, si se observan los valores extremos mayores de la variable INGTOHOG en la Tabla 5.2.d, se puede notar que sólo 3 individuos en la población alcanzan el valor máximo, y de ahí se percibe el abrupto decrecimiento de la variable INGTOHOG a valores del orden de 223,314 en el quinto valor extremo mayor.

A continuación, se muestra de manera gráfica la distribución correspondiente a las variables utilizadas. Primero se muestra en la Figura 1 la distribución del ingreso mensual total por hogar. Se puede observar que tiene una distribución altamente asimétrica como ya se señaló en párrafos anteriores. Resultará entonces muy interesante observar el comportamiento de los diferentes estimadores de la varianza de los estimadores utilizados en el presente texto con esta variable. Hay que señalar que, aunque no se incluyeron tales gráficos, se procedió a representar gráficamente la transformación logarítmica de la variable INGTOHOG (con el objeto de observar valores extremos de manera más fácil) pero no se logró gran cosa; lo mismo sucedió al eliminar los 3 individuos que alcanzaron el máximo 999,998. Por otro lado, posteriormente se muestran las Figura 2 y 3, en ellas se muestran gráficamente las distribuciones asociadas a las variables de total de personas por hogar y total de cuartos por hogar.

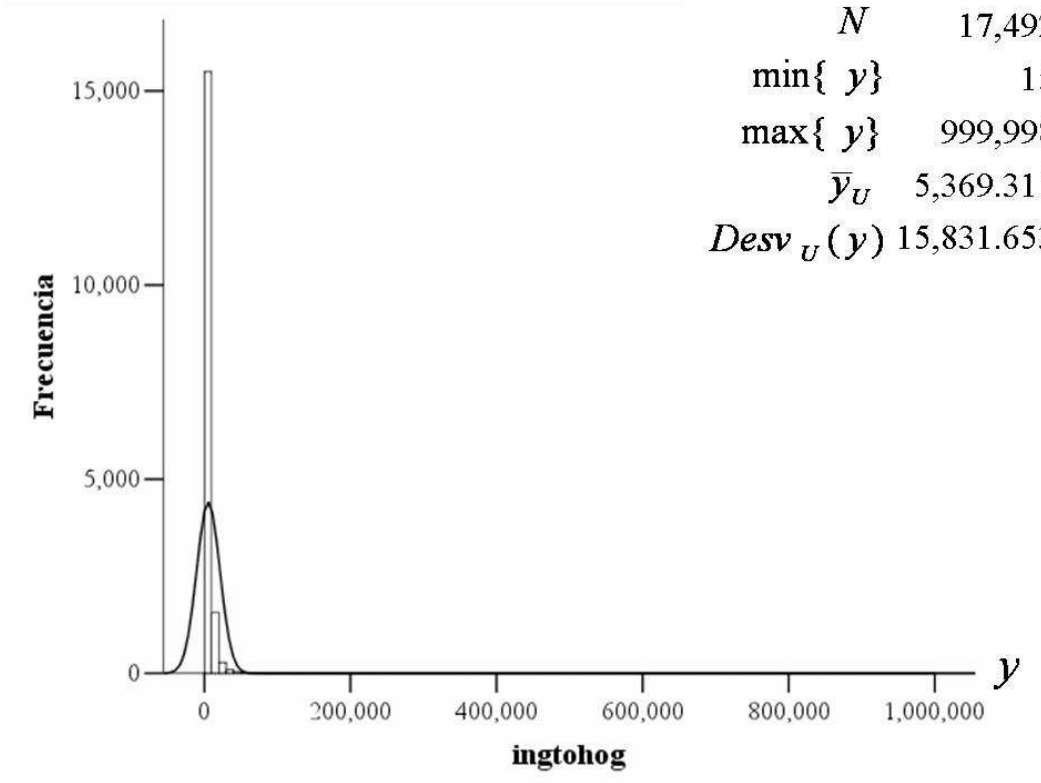

Figura 1: **Histograma del Ingreso Mensual Total por Hogar.**

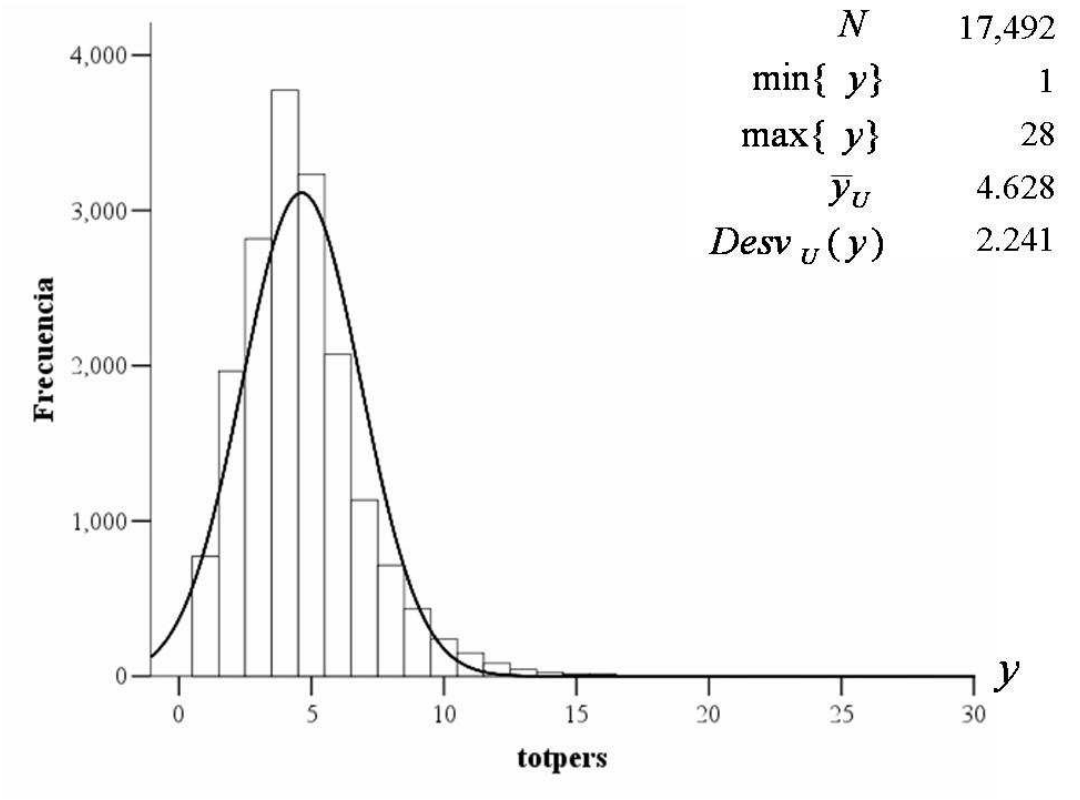

Figura 2: **Histograma del Total de Personas por Hogar.**

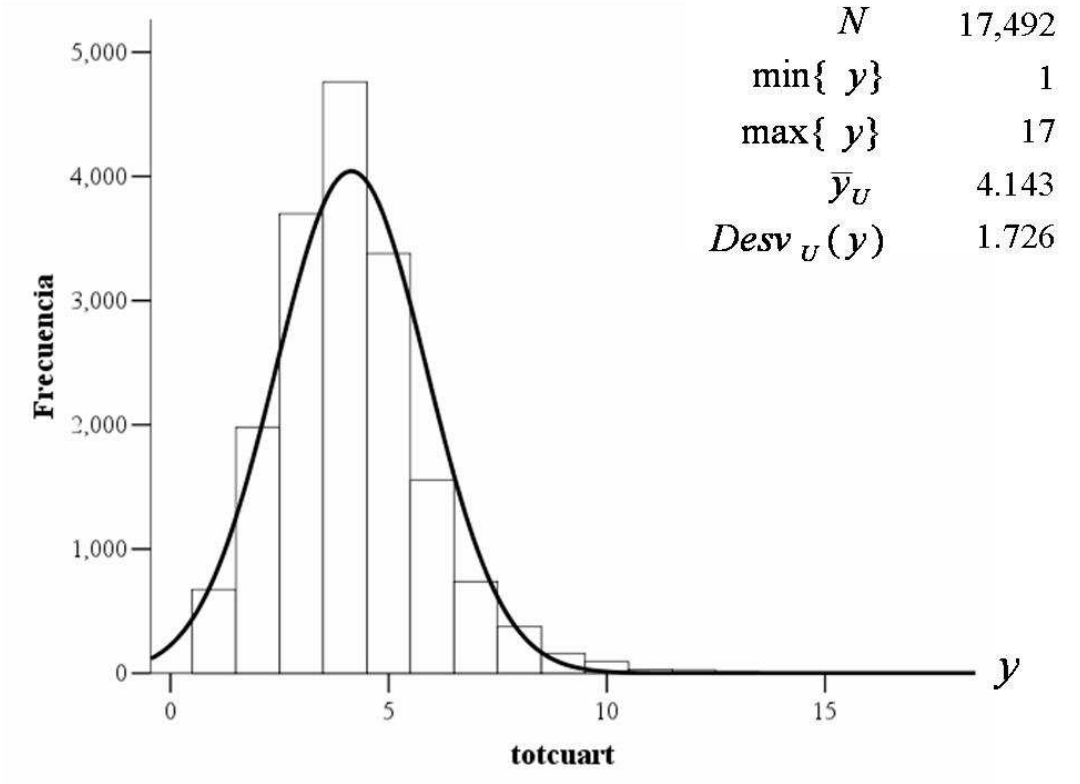

Figura 3: **Histograma del Total de Cuartos por Hogar.**

5.3. Especificaciones del Ejercicio de Simulación

En esta sección se hará mención de los detalles relativos al ejercicio de realizar simulaciones para observar el desempeño de algunos métodos de estimación de varianza.

5.3.1. Métodos de Estimación de Varianza a Simular

Se utilizarán los métodos siguientes de estimación de varianza:

- Método de Linealización de Taylor.
- Método de Woodruff para la Mediana.
- Método Jackknife (En dos modalidades: Quitando uno y dos elementos en la muestra).
- Método del Vector Post-Diseño.
- Método Bootstrap.

Por supuesto que se considerará también el uso de las expresiones de los estimadores π ó de Horwitz y Thompson para la estimación de varianza de estimadores de la media y proporciones cuando corresponda (véase Särndal *et al.* (1992)[pag. 68]).

5.3.2. Diseño de Muestreo y Tamaños de Muestra

En lo que atañe al diseño muestral a utilizar, se hará consideración únicamente de un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.) pues se piensa que si hay métodos de estimación de varianza que no tengan buen desempeño bajo este diseño, existe la idea intuitiva de que bajo diseños de muestreo más complejos su actuación será más pobre; pues es precisamente este diseño contra el cual se comparan todos los diseños muestrales. Claro que esto último no se puede sostener categóricamente. Se apela también, al principio de mantener la simplicidad (*parsimonia*) en las comparaciones para subrayar las diferencias de los métodos en su desempeño[28].

Se pueden encontrar en Kish & Frankel (1974) algunos resultados sobre desempeño de estimación de varianza con algunos de los métodos aquí utilizados para diseños de muestreo diferentes al SI (considerando estratificación y conglomeración). En esta referencia se concluye que ninguno de los métodos allí utilizados (Taylor, Balanced Repeated Replications y Jackknife) mostró ser el mejor o el peor, bajo los criterios allí también utilizados; se menciona también que finalmente la elección entre cuál método utilizar depende de costos, simplicidad, situaciones y los estadísticos utilizados. Otra fuente donde se consideran diseños de muestreo más complejos que el SI es Wolter (1985). Al respecto, el Dr. Pauli K. Ollila comenta que el Método del Vector Post-Diseño que propuso en Ollila (2004) podría ser empleado en estratificación pues, como se sabe, la estimación de varianza dentro de cada estrato puede tratarse por separado; aunque externó que éste está limitado en lo que se refiere a conglomeración pues el desarrollo analítico empleado en la obtención de ecuaciones propias del método no pueden ser llevadas al caso en el que se conglomera.

[28]De esto ya se habló en el Capítulo 2 página 5, Introducción de la presente tesis.

Los tamaños de muestra a emplear serán de $n = 350$ y $n = 1,000$. Esto debido a que son tamaños de muestra comúnmente utilizados en la práctica.[29]

5.3.3. Detalles Específicos de Algunos Métodos

En el ejercicio de simulación se emplearon 1,000 simulaciones para la estimación del estadístico correspondiente, i.e. Método de Monte Carlo con $N_{sim} = 1,000$.

Para el caso del Método Bootstrap el número de remuestras utilizado fue $A = 1,000$, acorde con la notación y definiciones de la parte 4.5.1 en las páginas 33 a 34 de la presente tesis y conforme a lo sugerido en Lehtonen & Pahkinen (2004).

En lo que atañe al Método del Vector Post-Diseño, siguiendo la teoría y la notación descrita en el apartado 3.7 en las páginas 15 a 18, se tiene que los valores correspondientes de n_b, $n_{b,u}$, q y n_b^* para los valores de N y $n = n_a$ utilizados en la presente tesis son acorde con la Tabla 5.3.3 siguiente:

Tabla 5.3.3 Parámetros de Implementación del Método del Vector Post-Diseño

	$N = 17,492$		$N = 18,108$	
Valor	$n_a = 350$	$n_a = 1,000$	$n_a = 350$	$n_a = 1,000$
n_b	176.7685	514.7128	176.7077	514.1981
$n_{b,u}$	177	515	177	515
q	1.9287	2.1553	2.1783	9.5696
n_b^*	340.4529	1,108.7993	384.3832	4,919.7919

Observar en la Tabla 5.3.3 que los valores que toma n_b^*, en términos porcentuales con respecto a los valores de n_a, en cada caso son: 97.3 %, 110.9 %, 109.8 % y 492.0 % correspondientemente.

5.3.4. Parámetros de Interés a Utilizar en las Simulaciones

A continuación, se listan los parámetros de interés de los cuales se calculará la estimación de la varianza de su estimador, y también se indica la referencia de las expresiones utilizadas.

[29] En Ollila (2004) se hacen simulaciones con datos finlandeses para una población de tamaño $N = 9$ y tamaño de muestra $n = 6$; con diseños muestrales de muestreo aleatorio simple con y sin reemplazo.

Tabla 5.3.1 Parámetros y Estimadores de los Parámetros

Parámetro	*Número de la Expresión del Parámetro*	*Número de la Expresión del Estimador del Parámetro*	*Página o Páginas*
Media de los ingresos (mensuales) totales por hogar	(75)	(76)	31
Mediana de los ingresos (mensuales) totales por hogar	(59)	(60)	26
Ingreso (mensual) per cápita	(54)	(55)	25, 25
Número de personas por cuarto en el hogar	(54)	(55)	25, 25
Proporción de viviendas con teléfono	(75)	(76)	31
Proporción de viviendas con teléfono	(54)	(55)	25, 25

Ahora se lista en términos prácticos u operativos, acorde con la presente tesis, el estimador a utilizar para cada parámetro de interés[30].

Tabla 5.3.2 Parámetros y Estimadores en Términos Prácticos

Parámetro	*Expresión Práctica u Operativa del Estimador del Parámetro*
Media de los ingresos (mensuales) totales por hogar	Media muestral de la variable INGTOHOG
Mediana de los ingresos (mensuales) totales por hogar	Mediana muestral de la variable INGTOHOG
Ingreso (mensual) per cápita	Razón de los totales muestrales de las variables INGTOHOG y TOTPERS
Número de personas por cuarto en el hogar	Razón de los totales muestrales de las variables TOTPERS y TOTCUART
Proporción de viviendas con teléfono	Media muestral de la variable TELEFONO
Proporción de viviendas con teléfono	Razón de los totales muestrales de las variables TELEFONO y UNOS

5.3.5. Sobre la Programación Empleada

Las rutinas de simulación utilizadas fueron programadas en *MATLAB*® Versión 6.0 Release 12 y también se hizo uso del paquete estadístico *SPSS*® 13.0 para la creación de las gráficas y cálculo de estadísticos u operaciones con las variables. Adicionalmente se empleó el paquete especializado de muestreo *SUDAAN*® Versión 7.5 para la verificación de algunos cálculos realizados por las rutinas programadas para la simulación. En la parte de Anexos de la presente tesis, se adjuntan los códigos fuente de las rutinas programadas.

[30]La variable UNOS es una variable artificial donde todos los individuos en el marco tienen el valor 1.

5.4. Propiedades a Estudiar de los Estimadores de Varianza

Los criterios a utilizar para determinar la calidad de la estimación de varianza de los métodos o para comparar su desempeño entre sí se basarán en la medición de las siguientes propiedades para cada estimador. En todos los casos en la Tabla 5.4.1, el subíndice *sim* se refiere a que se utilizaron procedimientos de simulación para la estimación del correspondiente estadístico, i.e. Método de Monte Carlo; se omitirá esta especificación en la nomenclatura de las propiedades. Adicionalmente, en las siguientes tablas y resultados $V(\hat{\theta})$, la varianza del estimador en cuestión, se denotará como V; mientras que $\widehat{V}(\hat{\theta})$, el estimador de la varianza del estimador, se denotará como $\widehat{V}$. Esto último para hacer más cortas las expresiones pues en todo momento el presente texto se enfoca principalmente en la estimación de varianza del estimador y no en el estimador.

Tabla 5.4.1 Propiedades a Revisar de los Estimadores de Varianza

Nombre	*Expresión*
Esperanza del Estimador de la Varianza	$E_{sim}[\widehat{V}] = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \widehat{V}_i$
Error Estándar del Estimador de la Varianza	$\sqrt{V_{sim}(\widehat{V})} = \sqrt{\frac{1}{N_{sim}-1} \sum_{i=1}^{N_{sim}} \left(\widehat{V}_i - E_{sim}[\widehat{V}]\right)^2}$
Coeficiente de Variación del Estimador de la Varianza	$CV_{sim}(\widehat{V}) = \frac{\sqrt{V_{sim}(\widehat{V})}}{V}$
Sesgo Relativo del Estimador de la Varianza	$SesgoR_{sim}(\widehat{V}) = \frac{E_{sim}[\widehat{V}]-V}{V}$
Raíz del E.C.M. Relativo del Estimador de la Varianza	$RECMR_{sim}(\widehat{V}) = \frac{\sqrt{V_{sim}(\widehat{V})+\{E_{sim}[\widehat{V}]-V\}^2}}{V}$

Donde, $\widehat{V}_i$ denota el estimador de la varianza del estimador para la simulación i-ésima con $i = 1, \ldots, N_{sim}$.[31]

Nótese que los nombres de las propiedades anteriores y las expresiones no coinciden con las usualmente presentadas en algunas referencias[32] pues habitualmente al tratarse de relativizar errores o sesgos, etc; esto se hace con respecto al valor esperado del estadístico en cuestión. Es decir, en la presente tesis, las expresiones anteriores (aquellas en las que corresponda) debiesen tener en el denominador a $E_{sim}[\widehat{V}]$ en lugar de V ya que se asume el desconocimiento de V, no obstante se prefirió utilizar las listadas[33] en la Tabla 5.4.1 pues así se evita el acarreo de errores numéricos (en

[31] A la *Raíz del Error Cuadrático Medio* se le conoce también como *Error Total* (véase Kish (1965)[pag. 510]). También, al *Coeficiente de Variación* se le conoce como *Error Estándar Relativo* (véase Särndal *et al.* (1992)[pag. 42]).

[32] Por ejemplo: Särndal *et al.* (1992) y Lehtonen & Pahkinen (2004).

[33] Mismas que coinciden con las definidas por Kish (1965)[pag. 47,535], además del hecho de que en nuestro caso no desconocemos a V.

la evaluación del desempeño) asociados a los sesgos que pudiesen tener los estimadores utilizados. Se justifica el abuso de nomenclatura aduciendo familiaridad en la comprensión.

5.5. Resultados del Ejercicio de Simulación

A continuación se muestran los resultados obtenidos del ejercicio de simulación efectuado para finalmente realizar una comparación de algunos métodos de estimación de varianza de ciertos estimadores, ambos (los métodos de varianza y los estimadores de interés) listados anteriormente. Para la comparación del desempeño asociado a cada método y estimador se considerará la información obtenida para cada propiedad a estudiar de cada estimador acorde, también, con lo antes mencionado. Adicionalmente se hará uso, en algunas ocasiones, de representaciones gráficas para mostrar de manera más sencilla y clara ciertos detalles del desempeño de determinado método para cierto estimador.

En aquellas ocasiones en que se muestren representaciones gráficas (histogramas) de las estimaciones de varianza resultantes de la simulación, se exhibirá la curva de la distribución Normal asociada a la media y varianza correspondiente a los datos representados. Se indicará también, por medio de una linea vertical continua la media de tales datos, i.e. lo que para efectos de la presente tesis representa al estadístico $E_{sim}[\widehat{V}]$; mientras que con una linea vertical punteada el verdadero valor de la varianza (i.e. V) que se quiere estimar con diferentes métodos y que en este caso no desconocemos.

5.5.1. Para la Media de los Ingresos Mensuales Totales por Hogar

Los resultados obtenidos para la estimación de la varianza del estimador de la Media de los Ingresos Mensuales Totales por Hogar[34] se resumen en la Tabla 5.5.1 [35].

Como una primera observación se puede notar que todos los métodos presentados en la Tabla 5.5.1 mostraron un desempeño muy parecido pues todas las cifras listadas son muy similares a excepción del Método Bootstrap en términos de las cantidades registradas para los estadísticos $E_{sim}[\widehat{V}]$, $\sqrt{V_{sim}(\widehat{V})}$, $CV_{sim}(\widehat{V})$ y el resto. Se puede decir que el peor desempeño registrado hasta este momento es el del Método Bootstrap.

Como era de esperarse, un aumento en n (de 350 a $1,000$) incrementa la precisión, o bien disminuye V y por tanto también $\widehat{V}$. Esto se puede verificar observando los valores del estadístico $\sqrt{V_{sim}(\widehat{V})}$ para cada tamaño de muestra empleado. Ahora, relativizando esto último, es decir observando lo obtenido para $CV_{sim}(\widehat{V})$ con cada tamaño de muestra y cada método, se tiene que la mayor ganancia en precisión se logra con el uso del Método Jackknife(2).

Aparte de la ganancia en términos de precisión con el aumento del tamaño de muestra, se puede

[34] Acorde con lo establecido en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.

[35] Únicamente se presenta la modalidad del Método Jackknife con $m = 2$ (acorde con la notación y expresiones utilizadas en la presente tesis) para el estimador en cuestión pues se tiene que bajo un diseño de muestreo aleatorio simple (SRS, SI ó m.a.s.) el estimador de varianza Jackknife con $m = 1$ es idéntico al de Horwitz-Thompson. Véase la observación 11.2.1. en Särndal *et al.* (1992)[pag. 422].

notar que hablando de exactitud (i.e. sesgo) se logran mayores ganancias con el Método Jackknife(2) y el Método Bootstrap; es decir, estos dos métodos mostraron las mayores reducciones[36] en el estadístico $SesgoR_{sim}(\widehat{V})$.

Finalmente, combinando precisión y exactitud, es decir observando el error total (i.e. el estadístico $RECMR_{sim}(\widehat{V})$) se tiene que para $n = 350$ todos los métodos tienen un desempeño similar a excepción del Método Bootstrap. Al aumentar el tamaño de muestra a $n = 1,000$, el método que mostró tener mayor precisión y exactitud simultáneamente fue el Método Jackknife(2) y los tres métodos restantes obtuvieron valores similares en el error total; aunque el Método Bootstrap mostró mejorías superiores en términos de sesgo, mientras los otros dos en términos de precisión.

Tabla 5.5.1 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador de la Media de los Ingresos (Mensuales) Totales por Hogar, $\bar{y}_S = n^{-1}\sum_{k\in S} y_k$

	Horwitz-Thompson	*Jackknife(2)*	*Vector Post-Diseño*	*Bootstrap*
$N_{sim} = 1,000$		$N = 17,492$	$n = 350$	
V	701,788.92	701,788.92	701,788.92	701,788.92
$Min_{sim}\{\widehat{V}\}$	37,844.59	35,779.07	38,439.69	38,896.95
$Max_{sim}\{\widehat{V}\}$	13,870,850.09	13,828,260.48	13,770,248.72	16,191,894.23
$E_{sim}[\widehat{V}]$	717,102.97	717,227.48	717,300.76	795,205.69
$\sqrt{V_{sim}(\widehat{V})}$	2,073,549.10	2,073,953.53	2,074,011.16	2,257,271.01
$CV_{sim}(\widehat{V})$	2.954662	2.955238	2.955320	3.216453
$SesgoR_{sim}(\widehat{V})$	0.021821	0.021999	0.022103	0.133112
$RECMR_{sim}(\widehat{V})$	2.954743	2.955320	2.955403	3.219206
$N_{sim} = 1,000$		$N = 17,492$	$n = 1,000$	
V	236,312.33	236,312.33	236,312.33	236,312.33
$Min_{sim}\{\widehat{V}\}$	18,002.53	16,903.47	18,090.94	16,868.37
$Max_{sim}\{\widehat{V}\}$	2,579,842.34	1,921,452.54	2,620,235.14	2,799,822.68
$E_{sim}[\widehat{V}]$	241,588.14	236,184.28	241,629.38	248,184.14
$\sqrt{V_{sim}(\widehat{V})}$	419,837.60	398,964.74	420,325.36	427,054.49
$CV_{sim}(\widehat{V})$	1.776622	1.684486	1.778686	1.807161
$SesgoR_{sim}(\widehat{V})$	0.022326	-0.000542	0.022500	0.050238
$RECMR_{sim}(\widehat{V})$	1.776762	1.684486	1.778828	1.807859

Hay que anotar que en general, si se comparan los valores del estadístico $RECMR_{sim}(\widehat{V})$ de la Tabla 5.5.1 con los obtenidos en las tablas subsecuentes, se tienen desempeños pobres de todos los métodos. Esto último se debe en gran parte por la distribución altamente asimétrica que posee la variable asociada a los ingresos [mensuales] totales por hogar (INGTOHOG)[37]. Adicionalmente, el uso de un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.) para variables con

[36] En valor absoluto.

[37] Véase la parte 5.2. Descripción de las Variables y Especificación de los Marcos Muestrales, que comienza en la página 38 de la presente tesis.

distribución asimétrica obligaría a utilizar tamaños de muestra más grandes. Alternativamente, para tratar el problema de asimetría de la variable de interés, es conveniente: utilizar un diseño de muestreo con probabilidades de inclusión proporcionales al tamaño (coloquialmente conocido como PPT), estratificar a la población en grupos de niveles o valores de la variable de interés (si el ejercicio práctico lo permite), post-estratificar acorde con información poblacional asociada a lo que se está midiendo en la muestra (si en la práctica no es posible llevar a cabo una estratificación de inicio), o finalmente retirar y tratar por separado aquellos elementos en muestra con valores extremos.

Con el objeto de complementar lo hasta ahora observado en las cifras de la Tabla 5.5.1 relativo al desempeño de los métodos de estimación de varianza, se presentan a continuación las Figuras 4, 5, 6 y 7; correspondientes al ejercicio de simulación con tamaño de muestra $n = 1,000$[38].

De manera generalizada se puede observar en las Figuras correspondientes que todos los métodos, que aunque registraron valores pequeños[39] del $SesgoR_{sim}(\widehat{V})$, presentaron una distribución bimodal[40] asimétrica. Esto describe el alto riesgo de obtener una desmedida subestimación o sobrestimación de la varianza del estimador en cuestión cuando subyace una distribución con alta asimetría en la variable de estudio[41].

Adicionalmente en relación a los valores pequeños obtenidos del estadístico $SesgoR_{sim}(\widehat{V})$, también hay que agregar que, aunque se observan en todos los casos valores esperados de las estimaciones ($E_{sim}[\widehat{V}]$, linea vertical continua) muy cercanos a los verdaderos valores que queremos estimar (V, linea vertical punteada), la frecuencia asociada a los valores vecinos de V en los histogramas resulta gráficamente nula; no obstante se observa una especie de balance conseguido entre las estimaciones y la frecuencia de éstas en las simulaciones que ulteriormente se traducen en cuantías reducidas relativas al sesgo. Situaciones de este tipo son aquellas a las que regularmente recurren algunos estadísticos (usualmente Bayesianos) para socavar la atención o importancia otorgada al insesgamiento de los estimadores en general. Y tienen razón, pero sólo en el caso en que esta propiedad se piense como la única que se debe buscar o revisar.

[38] Los gráficos con $n = 350$ no se presentan pues exhiben características muy similares con $n = 1,000$.

[39] En valor absoluto.

[40] De igual manera sucedió en las representaciones gráficas con una tamaño de muestra $n = 350$.

[41] Comentarios respecto a los problemas con distribuciones altamente asimétricas en la aproximación Normal de las distribuciones muestrales de ciertos estadísticos en Kish (1965)[pag. 17]. Y una sucinta pero rica discusión sobre posibles alternativas para enfrentar poblaciones con alta asimetría se encuentra en Kish (1965)[pag. 410-412].

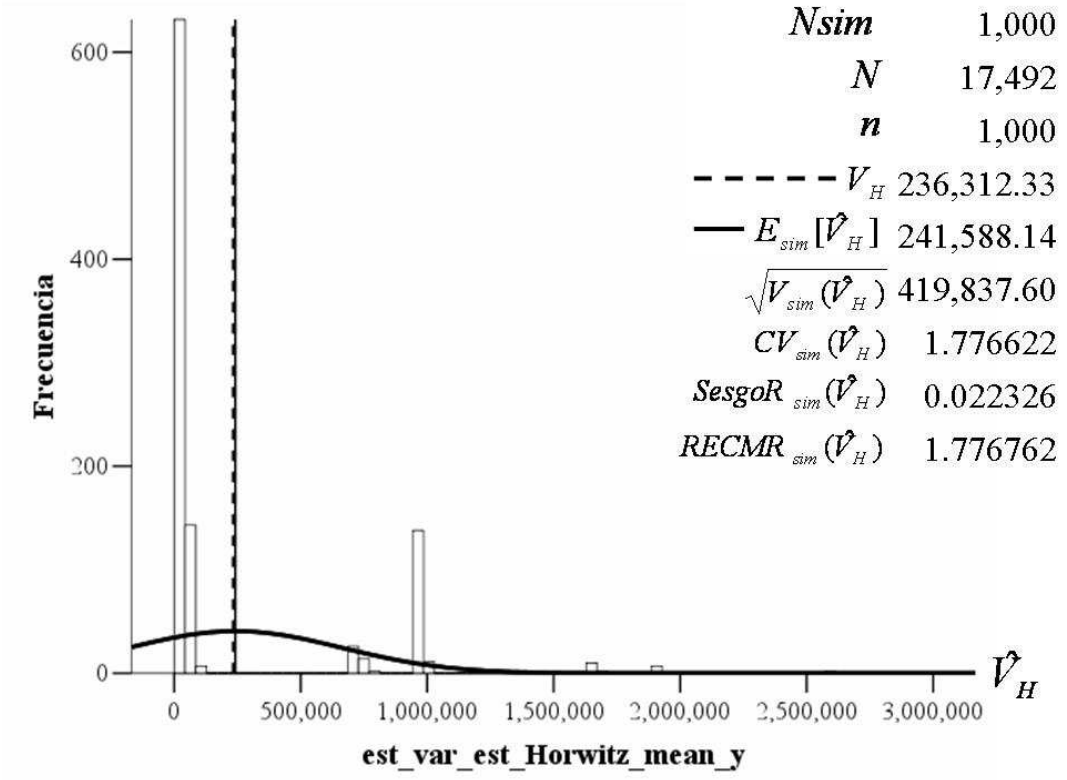

Figura 4: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Media de los Ingresos Mensuales Totales por Hogar utilizando los estimadores π ó de Horwitz-Thompson.**

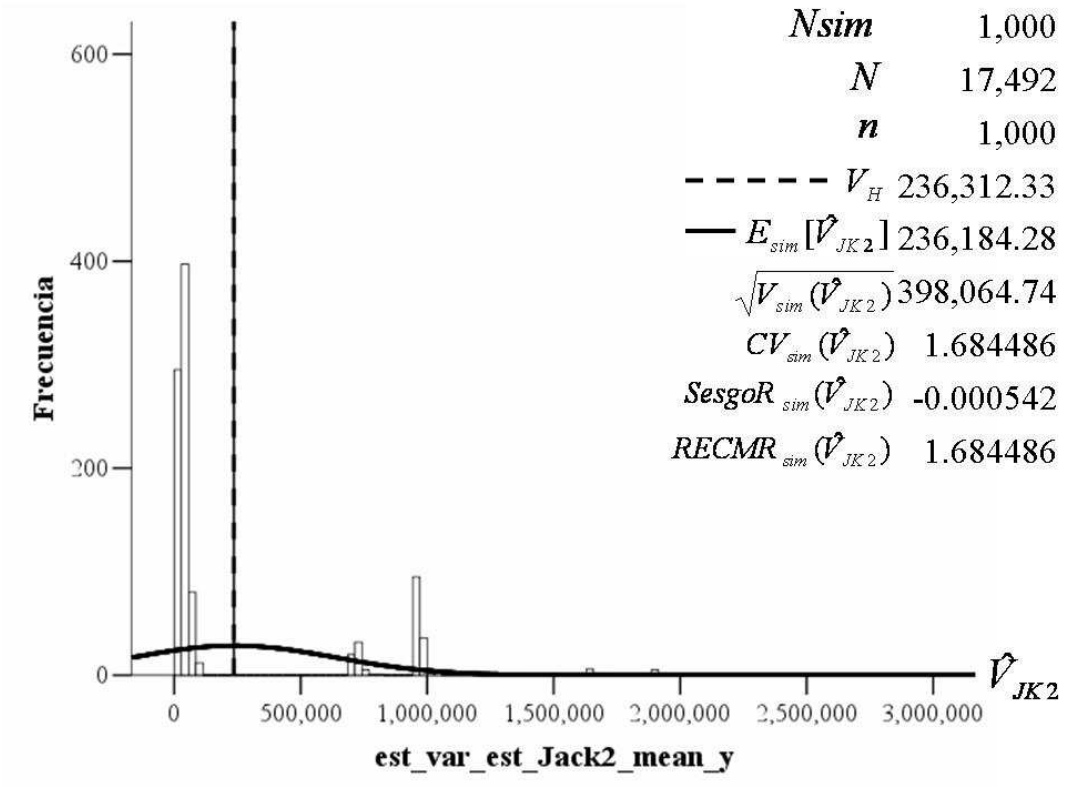

Figura 5: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Media de los Ingresos Mensuales Totales por Hogar utilizando el Método Jackknife con $m = 2$.**

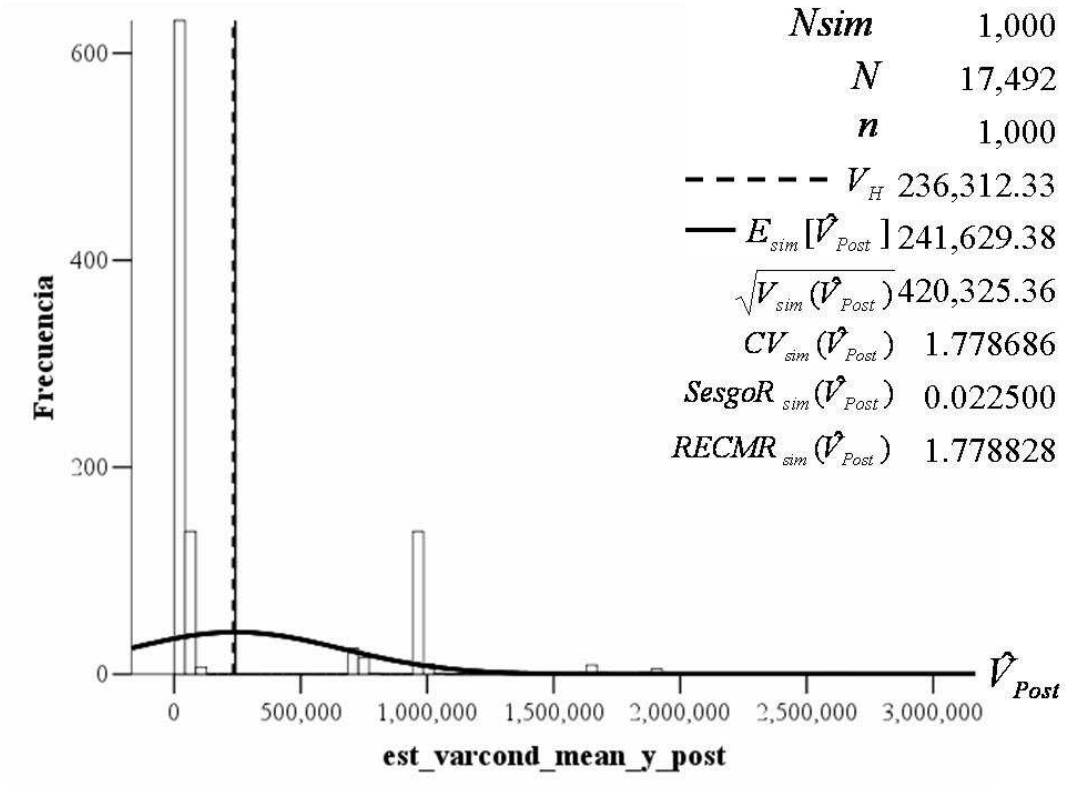

Figura 6: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Media de los Ingresos Mensuales Totales por Hogar utilizando el Método del Vector Post-Diseño.**

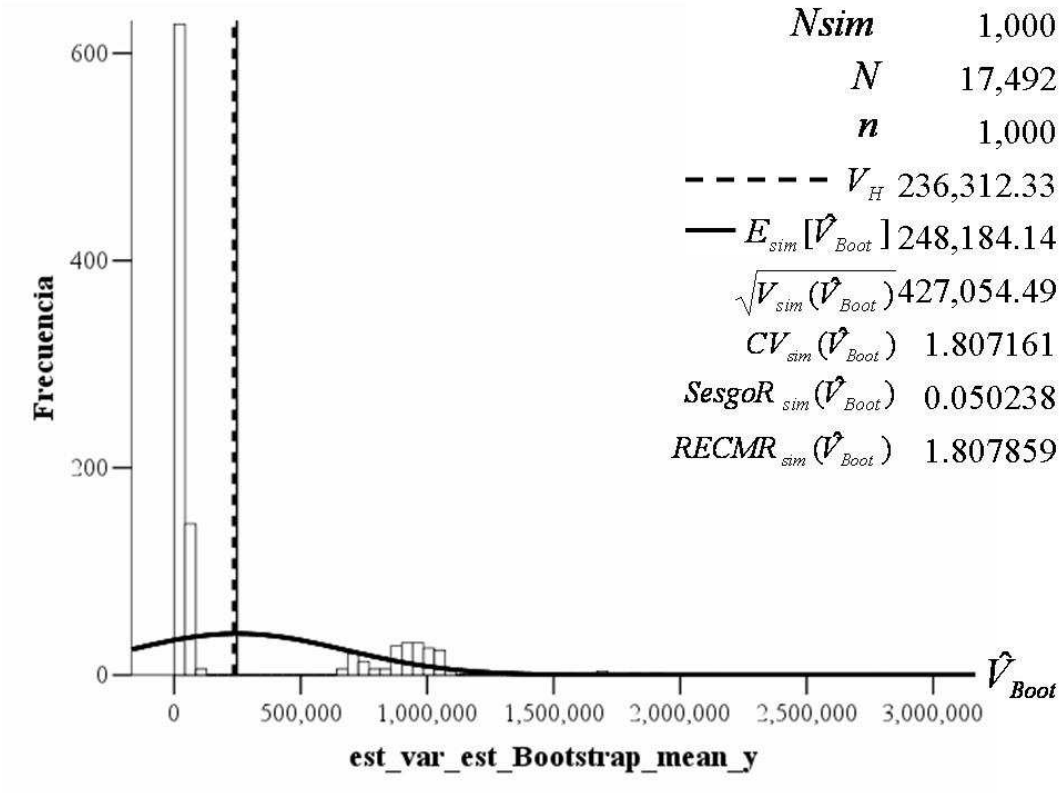

Figura 7: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Media de los Ingresos Mensuales Totales por Hogar utilizando el Método Bootstrap.**

5.5.2. Para la Mediana de los Ingresos Mensuales Totales por Hogar

En lo que toca a los resultados derivados de la simulación, para la estimación de la varianza del estimador de la Mediana de los Ingresos Mensuales Totales por Hogar[42], estos se encuentran resumidos en la Tabla 5.5.2 siguiente.

No se reporta lo relativo al Método Jackknife pues, conforme a lo mencionado anteriormente, se sabe de sus limitaciones en la estimación de varianza para el estimador de la mediana (véase Efron (1979)[pag. 6]); también en Särndal *et al.* (1992)[pag. 442] se menciona que se encontró en estudios empíricos que éste tiene un mal desempeño en la estimación de varianza de estimadores de cuantiles en general[43].

[42]Acorde con la nomenclatura y definiciones de la presente tesis mostradas en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.

[43]Esto también se señala en los comentarios o *Advertencias* del paquete especializado de muestreo *SUDAAN*® Versión 7.5.

Tabla 5.5.2 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador de la Mediana de los Ingresos (Mensuales) Totales por Hogar, $\widehat{M_y} = \widehat{F}^{-1}(0.5)$

	Woodruff	*Vector Post-Diseño*	*Bootstrap*
$N_{sim} = 1,000$	$N = 17,492$		$n = 350$
V	49,483.55	49,483.55	49,483.55
$Min_{sim}\{\widehat{V}\}$	11,976.85	12,481.76	11,325.64
$Max_{sim}\{\widehat{V}\}$	127,551.02	120,051.95	106,355.01
$E_{sim}[\widehat{V}]$	43,684.52	44,462.42	44,312.46
$\sqrt{V_{sim}(\widehat{V})}$	14,782.36	14,730.41	14,809.24
$CV_{sim}(\widehat{V})$	0.298733	0.297683	0.299276
$SesgoR_{sim}(\widehat{V})$	-0.117191	-0.101471	-0.104501
$RECMR_{sim}(\widehat{V})$	0.320897	0.314502	0.316996
$N_{sim} = 1,000$	$N = 17,492$		$n = 1,000$
V	15,307.75	15,307.75	15,307.75
$Min_{sim}\{\widehat{V}\}$	3,842.73	4,376.05	3,906.55
$Max_{sim}\{\widehat{V}\}$	30,625.00	31,992.12	34,154.48
$E_{sim}[\widehat{V}]$	14,606.38	14,809.73	14,680.19
$\sqrt{V_{sim}(\widehat{V})}$	4,577.12	4,705.19	4,643.39
$CV_{sim}(\widehat{V})$	0.299007	0.307373	0.303336
$SesgoR_{sim}(\widehat{V})$	-0.045818	-0.032534	-0.040996
$RECMR_{sim}(\widehat{V})$	0.302497	0.309090	0.306093

En términos numéricos conforme al estadístico $CV_{sim}(\widehat{V})$ los métodos de Woodruff, Vector Post-Diseño y Bootstrap obtuvieron cifras similares en la simulación, aún para los dos tamaños de muestra utilizados $n = 350$ y $n = 1,000$. El sesgo relativo, $SesgoR_{sim}(\widehat{V})$, se redujo[44] considerablemente de manera semejante con el aumento de tamaño de muestra de 350 a $1,000$ en todos los métodos presentados. No obstante, al igual que los resultados obtenidos por Ollila (2004) relativos al método de Woodruff, siempre se obtuvo un sesgo negativo; lo que habla de una posible subestimación de la varianza prevaleciente en todos los métodos. Luego, combinando estas dos características (precisión y exactitud), i.e. observando el error total relativo, se tienen desempeños similares para todos los métodos presentados. No obstante, al igual que en el caso de la media, se integrará lo mostrado en la Tabla 5.5.2 indagando en el comportamiento exhibido en las Figuras 8, 9, 10 y 11 siguientes.

Respecto al método de Woodruff se puede observar que para el caso con $n = 350$ (Figura 8) se obtuvo en la simulación una distribución más o menos normal, lo que refleja el posible riesgo de tener estimaciones disímiles al verdadero valor de la varianza del estimador (V, linea vertical punteada) de la mediana[45]; de hecho bastante diferentes al valor esperado del estimador de la varianza del estimador ($E_{sim}[\widehat{V}]$, linea vertical continua). Esto deja de ocurrir, como se puede verificar en la

[44] En valor absoluto.

[45] Recordar que se están utilizando datos altamente asimétricos (ingresos económicos), precisamente el caso en el que se destaca la utilidad de la mediana.

Figura 9, cuando se utiliza el método de Woodruff con tamaño de muestra $n = 1,000$. Se coteja y confirma entonces lo mencionado en Woodruff (1952)[pag. 642] relativo a que el método funciona correctamente[46] con tamaños de muestra grandes en general[47] y lo también astutamente advertido en Särndal *et al.* (1992)[pag. 203], concerniente al hecho de que el Método de Woodruff dado que descansa en diversas aproximaciones debe ser utilizado con precaución, al menos cuando se tiene un tamaño de muestra pequeño; y precisamente el tener un tamaño de muestra de 350 puede en nuestro caso (diseño de muestreo aleatorio simple sin reemplazo) pensarse pequeño si consideramos que se está abordando la estimación de la mediana de una variable que presenta una crecida asimetría. En Sitter & Wu (2001) se demuestra el buen desempeño del Método de Woodruff a pesar de sus limitaciones cuando se tiene un tamaño de muestra moderado.

El problema antes mencionado no se presentó con los otros dos métodos, el del Vector Post-Diseño (Figura 10) y Bootstrap (Figura 11). Se pueden apreciar las distribuciones suaves y con la usualmente deseable forma de campana de una función de densidad Normal.

Hasta este momento del desarrollo de resultados de la simulación de la presente tesis es importante resaltar lo valioso que resultan las representaciones gráficas en la evaluación del desempeño de los métodos vistos aquí.

[46] Inclusive para cualquier diseño de muestreo.

[47] Pues finalmente el método descansa en la estimación de la varianza de la proporción de elementos menores a el valor correspondiente para la mediana. En Kish (1995) se menciona que de hecho se ha conjeturado y encontrado que el efecto de diseño, $\widehat{Deft}(\widehat{\theta}) = deft(\widehat{\theta}) = [\widehat{V}(\widehat{\theta})/\widehat{V}_{m.a.s.}(\widehat{\theta})]^{1/2}$, para medianas tendría que ser similar al $deft$ para proporciones cercanas a 0.5.

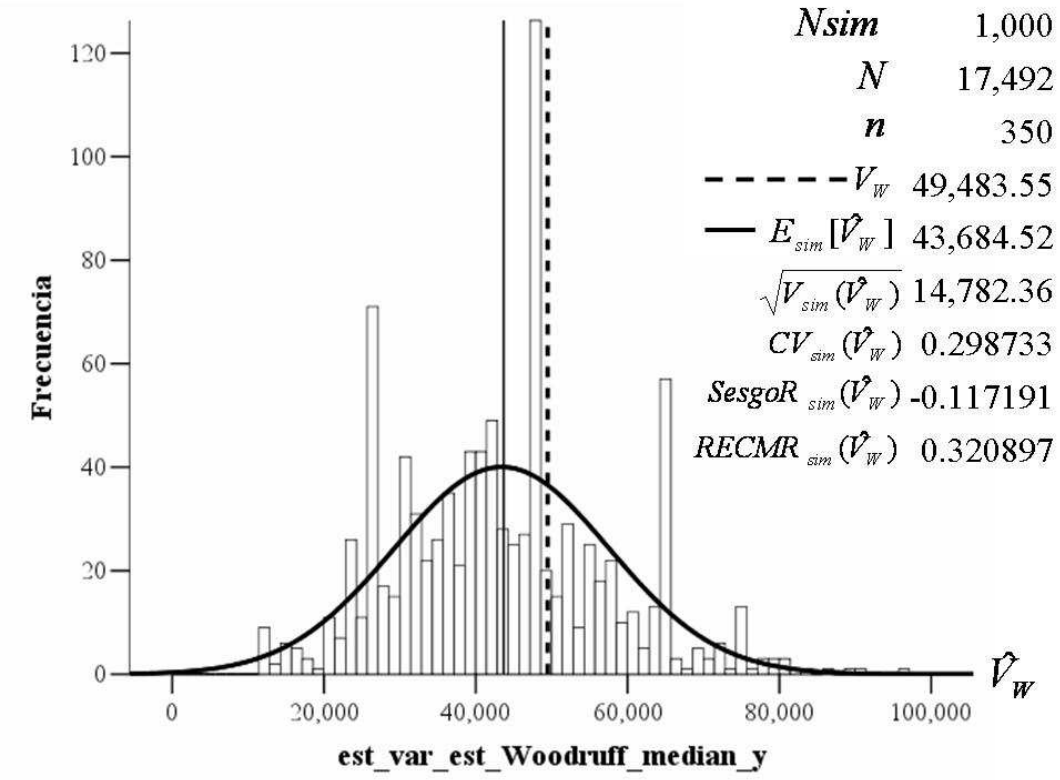

Figura 8: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Mediana de los Ingresos Mensuales Totales por Hogar utilizando el Método de Woodruff con** $n = 350$**.**

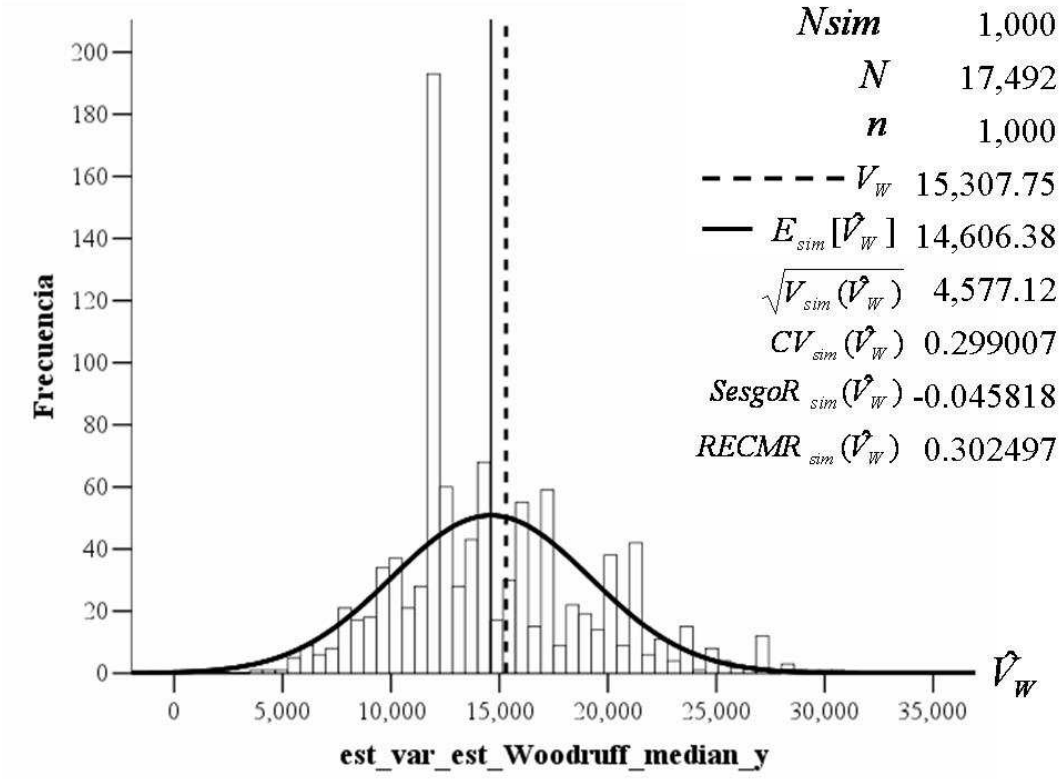

Figura 9: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para la Mediana de los Ingresos Mensuales Totales por Hogar utilizando el Método de Woodruff con $n = 1,000$.**

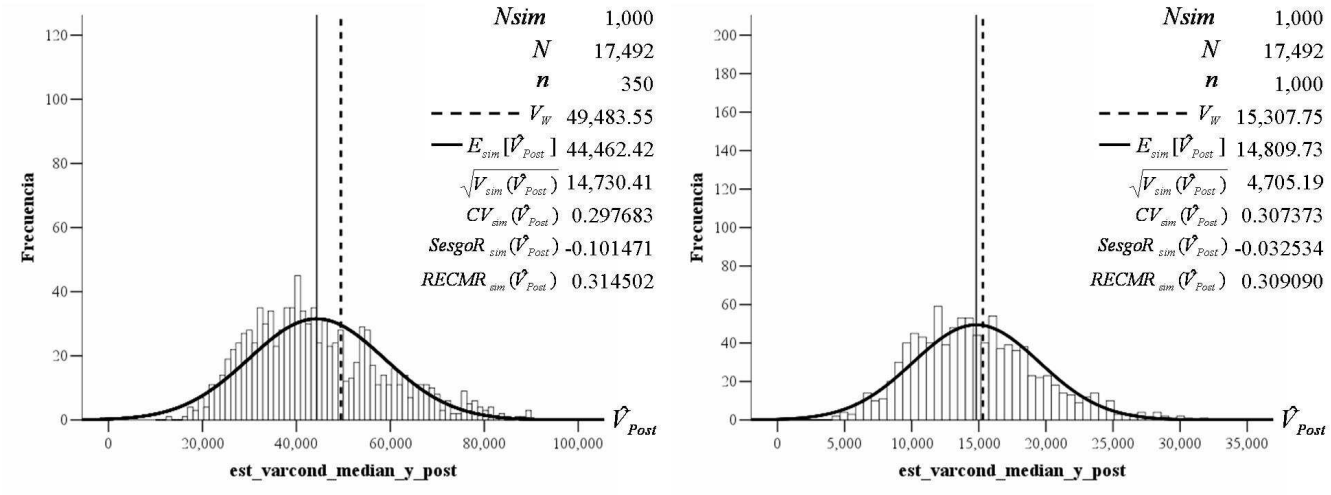

Figura 10: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Mediana de los Ingresos Mensuales Totales por Hogar utilizando el Método del Vector Post-Diseño con $n = 350$ y con $n = 1,000$.**

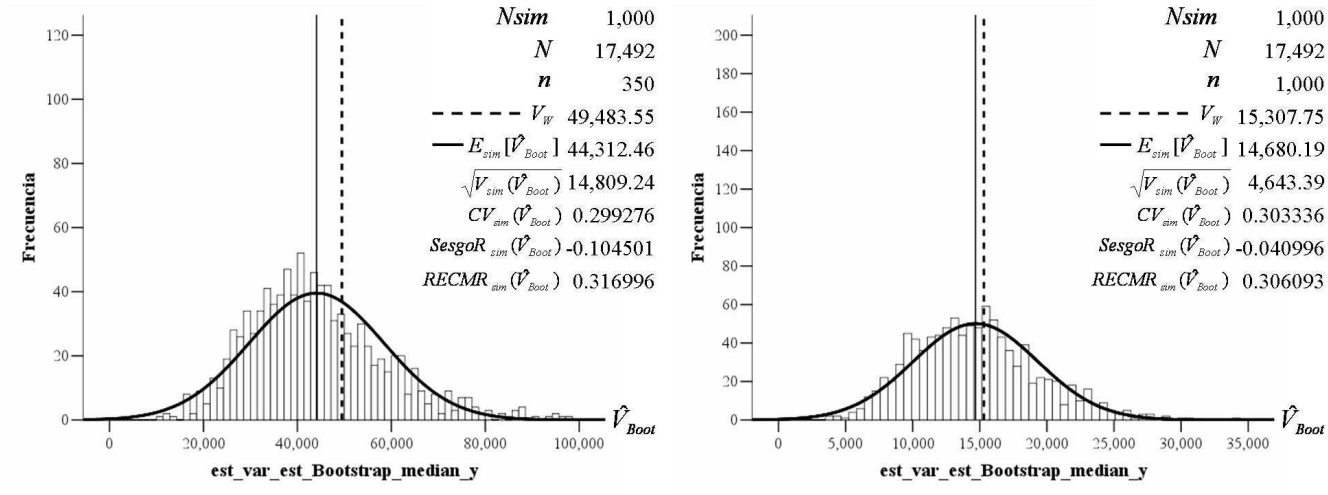

Figura 11: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Mediana de los Ingresos Mensuales Totales por Hogar utilizando el Método Bootstrap con** $n = 350$ **y con** $n = 1,000$**.**

5.5.3. Para el Ingreso Mensual Per Cápita

A continuación se presentan de manera sintética, en la Tabla 5.5.3 siguiente, los resultados conseguidos en la simulación para la estimación de la varianza del estimador del Ingreso Mensual Per Cápita, estimado con la razón de los totales muestrales de las variables INGTOHOG y TOTPERS[48].

Lo primero que salta a la vista cuando uno observa las cifras de la Tabla 5.5.3 son los elevados valores del estadístico $Max_{sim}\{\widehat{V}\}$ para el método Jackknife con $m = 1$ y para los dos tamaños de muestra utilizados, comparado con el resto de los métodos. Esto último se ve reflejado consecuentemente en las cuantías registradas para los estadísticos: $E_{sim}[\widehat{V}]$, $\sqrt{V_{sim}(\widehat{V})}$, $CV_{sim}(\widehat{V})$, $SesgoR_{sim}(\widehat{V})$ y $RECMR_{sim}(\widehat{V})$; siendo entonces el método Jackknife con $m = 1$ aquel con peor desempeño, seguido del método Jackknife con $m = 2$.

No obstante, aunque en este caso pobres en precisión i.e. valores elevados de $CV_{sim}(\widehat{V})$ con respecto a los demás métodos, nótese que los dos métodos Jackknife son los que obtienen mayor ganancia (relativa al valor de V) en términos de exactitud (valores pequeños de $SesgoR_{sim}(\widehat{V})$) al aumentar el tamaño de muestra de 350 a $1,000$.

Ahora observando los tres métodos restantes: Linealización de Taylor, Vector Post-Diseño y Bootstrap, se tiene que al ir de un tamaño de muestra de 350 a $1,000$ todos presentan una reducción del $CV_{sim}(\widehat{V})$. No obstante, los dos primeros sufren un aumento[49] en el $SesgoR_{sim}(\widehat{V})$ en lugar de reducirlo, además de que este par de métodos obtuvo valores negativos en el $SesgoR_{sim}(\widehat{V})$ i.e. que en términos relativos presentaron una subestimación de la varianza; mientras que el Bootstrap exhibe la intuitivamente esperada disminución. Se puede decir que de entre estos tres métodos, el de Linealización de Taylor y el Vector Post-Diseño mostraron tener desempeños similares y de mejores resultados para tamaños de muestra de 350; por su parte el Bootstrap presenta mejorías

[48]De acuerdo con lo establecido en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.
[49]En valor absoluto.

congruentes en términos de $CV_{sim}(\widehat{V})$, $SesgoR_{sim}(\widehat{V})$ y $RECMR_{sim}(\widehat{V})$ (error total relativo) con el aumento del tamaño de muestra a $1,000$.

De nueva cuenta, es importante recordar que la variable INGTOHOG utilizada en el numerador del cociente o razón utilizado como estimador posee una distribución altamente asimétrica.

Tabla 5.5.3 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador del Ingreso (Mensual) Per Cápita, $\bar{y}_S/\bar{x}_S = \left(\sum_{k\in S} y_k\right) / \left(\sum_{k\in S} x_k\right)$

	Aprox. Lin. Taylor	*Jackknife(1)*	*Jackknife(2)*	*Vector Post-Diseño*	*Bootstrap*
	$N_{sim} = 1,000$	$N = 17,492$		$n = 350$	
V	33,083.91	33,083.91	33,083.91	33,083.91	33,083.91
$Min_{sim}\{\widehat{V}\}$	1,994.60	1,506.47	1,415.32	2,005.86	1,948.77
$Max_{sim}\{\widehat{V}\}$	762,525.78	2,044,452.18	1,864,947.46	764,924.91	748,017.85
$E_{sim}[\widehat{V}]$	30,688.39	43,140.73	30,720.38	30,742.02	37,452.76
$\sqrt{V_{sim}(\widehat{V})}$	92,297.84	190,488.61	131,416.97	92,471.61	105,231.64
$CV_{sim}(\widehat{V})$	2.789811	5.757742	3.972232	2.795063	3.180750
$SesgoR_{sim}(\widehat{V})$	-0.072408	0.303979	-0.071440	-0.070786	0.132054
$RECMR_{sim}(\widehat{V})$	2.790750	5.765761	3.972875	2.795959	3.183490
	$N_{sim} = 1,000$	$N = 17,492$		$n = 1,000$	
V	11,140.30	11,140.30	11,140.30	11,140.30	11,140.30
$Min_{sim}\{\widehat{V}\}$	942.08	807.61	809.50	942.70	893.18
$Max_{sim}\{\widehat{V}\}$	125,203.78	275,559.12	183,120.60	125,401.15	132,973.44
$E_{sim}[\widehat{V}]$	9,558.20	12,140.30	11,201.44	9,563.83	11,676.46
$\sqrt{V_{sim}(\widehat{V})}$	17,480.09	29,449.76	26,826.95	17,482.25	19,837.83
$CV_{sim}(\widehat{V})$	1.569087	2.643535	2.408101	1.569281	1.780727
$SesgoR_{sim}(\widehat{V})$	-0.142015	0.089748	0.005489	-0.141510	0.048128
$RECMR_{sim}(\widehat{V})$	1.575500	2.645058	2.408107	1.575648	1.781377

En lo que respecta a las representaciones gráficas, esta vez se obtuvieron histogramas bimodales muy asimétricos de las simulaciones para todos los métodos. Esto representa, nuevamente, altos riesgos de subestimación de la varianza del estimador en cuestión. De hecho, es gráficamente nula la frecuencia de valores cercanos al valor verdadero. Al respecto, véase la parte 5.5.1 (que comienza en la página 46) con comentarios para la media de los ingresos mensuales totales por hogar.

A continuación sólo se presentan dos gráficos, Figuras 12 y 13, de las distribuciones para los métodos de Linealización de Taylor y Bootstrap, con $n = 1,000$. No se exhibe más, pues los gráficos restantes son muy parecidos.

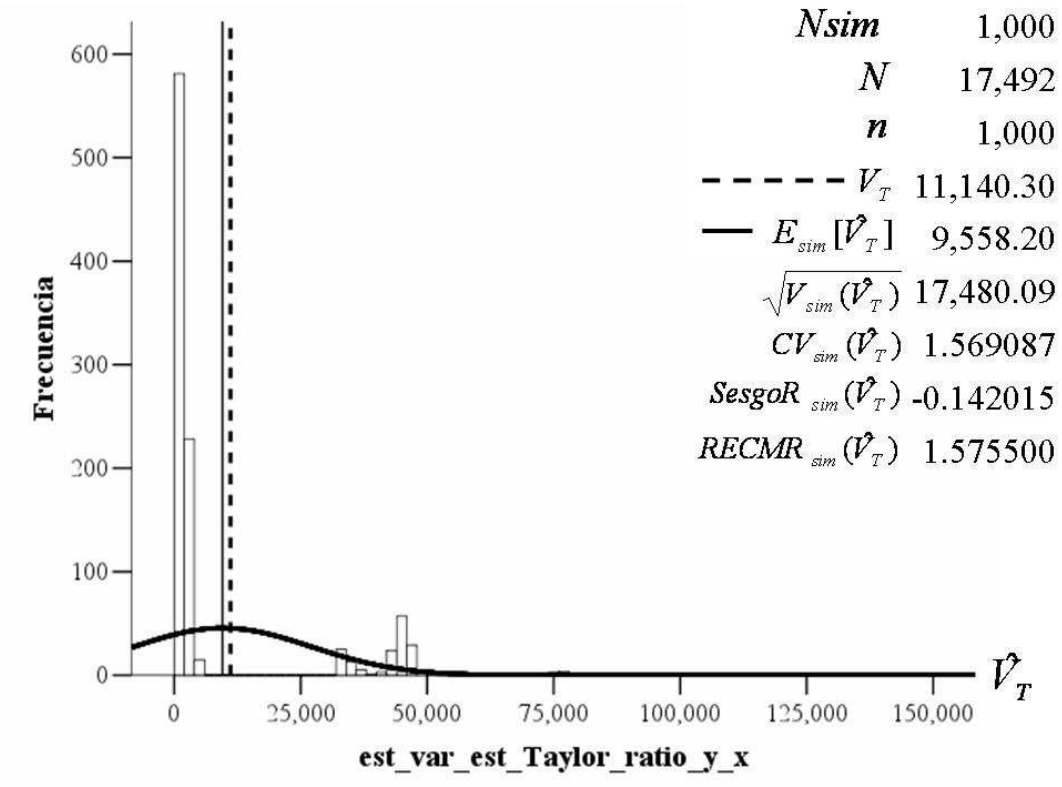

Figura 12: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para el Ingreso Mensual Per Cápita utilizando el Método de Linealización de Taylor con $n = 1,000$.**

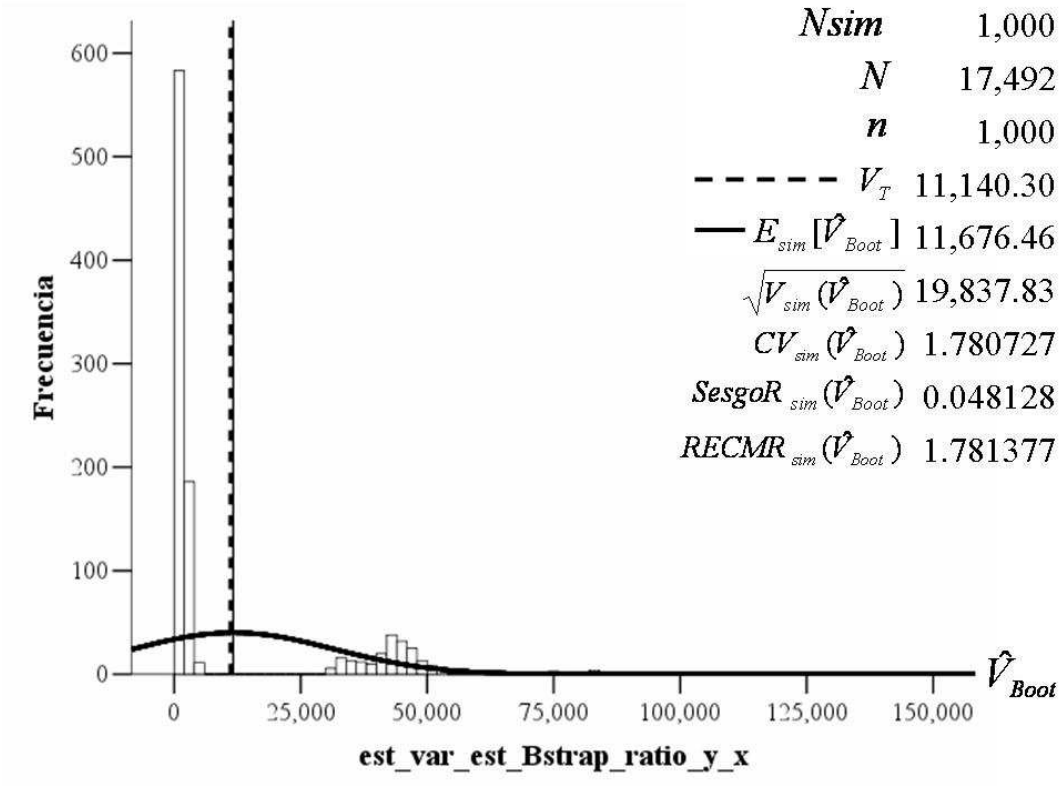

Figura 13: **Histograma de las estimaciones de varianza del estimador obtenidas en la simulación, para el Ingreso Mensual Per Cápita utilizando el Método Bootstrap con $n = 1,000$.**

5.5.4. Para el Número de Personas por Cuarto en el Hogar

Los resultados obtenidos de la simulación de la estimación de varianza del estimador del Número de Personas por Cuarto en el Hogar, se resumen en la Tabla 5.5.4 siguiente. Éste se construyó a partir del cociente de los totales muestrales de las variables TOTPERS y TOTCUART[50].

Tabla 5.5.4 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador del Número de Personas por Cuarto en el Hogar, $\bar{y}_S/\bar{x}_S = \left(\sum_{k\in S} y_k\right) / \left(\sum_{k\in S} x_k\right)$

	Aprox. Lin. Taylor	*Jackknife(1)*	*Jackknife(2)*	*Vector Post-Diseño*	*Bootstrap*
	$N_{sim} = 1,000$	$N = 17,492$		$n = 350$	
V	0.001304	0.001304	0.001304	0.001304	0.001304
$Min_{sim}\{\widehat{V}\}$	0.000915	0.000833	0.000791	0.000909	0.000918
$Max_{sim}\{\widehat{V}\}$	0.001970	0.002586	0.002364	0.001984	0.001922
$E_{sim}[\widehat{V}]$	0.001297	0.001294	0.001287	0.001298	0.001302
$\sqrt{V_{sim}(\widehat{V})}$	0.000147	0.000195	0.000211	0.000150	0.000155
$CV_{sim}(\widehat{V})$	0.112733	0.149765	0.161747	0.114896	0.119014
$SesgoR_{sim}(\widehat{V})$	-0.005781	-0.007776	-0.013459	-0.004425	-0.001600
$RECMR_{sim}(\widehat{V})$	0.112881	0.149966	0.162306	0.114981	0.119025
	$N_{sim} = 1,000$	$N = 17,492$		$n = 1,000$	
V	0.000439	0.000439	0.000439	0.000439	0.000439
$Min_{sim}\{\widehat{V}\}$	0.000367	0.000343	0.000331	0.000369	0.000331
$Max_{sim}\{\widehat{V}\}$	0.000529	0.000624	0.000607	0.000539	0.000542
$E_{sim}[\widehat{V}]$	0.000438	0.000440	0.000441	0.000439	0.000439
$\sqrt{V_{sim}(\widehat{V})}$	0.000028	0.000038	0.000043	0.000028	0.000034
$CV_{sim}(\widehat{V})$	0.062695	0.086298	0.098769	0.064329	0.077554
$SesgoR_{sim}(\widehat{V})$	-0.001604	0.003084	0.004293	-0.000194	-0.000388
$RECMR_{sim}(\widehat{V})$	0.062716	0.086353	0.098862	0.064329	0.077555

Se aprecia en la Tabla 5.5.4 que todos los métodos mostraron en general un desempeño similar. Se obtuvieron cifras que denotan buen desempeño en términos de precisión ($CV_{sim}(\widehat{V})$) y exactitud ($SesgoR_{sim}(\widehat{V})$), así como la combinación de estos atributos en el estadístico asociado al error total relativo ($RECMR_{sim}(\widehat{V})$). También, prevaleció un consistente comportamiento general esperado en todos los métodos para ambos tamaños de muestra n utilizados. Todo esto quizá, debido a la distribución que guardan los datos de las variables TOTPERS[51] y TOTCUART[52].

Ahora, agudizando la evaluación, el método con menor $CV_{sim}(\widehat{V})$ fue el Método de Taylor seguido del Método del Vector Post-Diseño, mientras que los que obtuvieron mayores valores fueron el par de Métodos Jackknife; esto para ambos tamaños de muestra. En cuanto al estadístico

[50] De acuerdo con la nomenclatura y definiciones de la presente tesis mostradas en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.

[51] Véase la Figura 2 en la página 41.

[52] Véase la Figura 3 en la página 41.

$SesgoR_{sim}(\widehat{V})$ todos los métodos presentaron valores negativos para $n = 350$ y en este caso el Método Bootstrap fue el que obtuvo mayor exactitud, seguido del Método del Vector Post-Diseño. Para el caso $n = 1,000$ el Método del Vector Post-Diseño fue quien mejor desempeño obtuvo en lo que se refiere al estadístico $SesgoR_{sim}(\widehat{V})$, seguido del Método Bootstrap. No obstante, en términos del error total relativo ($RECMR_{sim}(\widehat{V})$) el método que para ambos tamaños de muestra obtuvo el mejor desempeño fue el de Taylor, seguido del Método del Vector Post-Diseño y posteriormente el Método Bootstrap. Y, como se puede constatar en la Tabla 5.5.4, el par de Métodos Jackknife fue aquel con desempeño más pobre (para los dos tamaños de muestra $n = 350$ y $n = 1,000$) comparado con el resto, aunque por una diferencia pequeña.

En lo que respecta a las representaciones gráficas de las simulaciones se obtuvieron histogramas muy parecidos con todos los métodos y para los dos tamaños de muestra. Lo único interesante que puede concluirse de la observación de los gráficos es que en todos los casos se exhibió una distribución con forma de campana similar a la de una distribución de densidad Normal. Esto último, como ya se mencionó, debido quizás a la distribución de las variables utilizadas, visiblemente suave y con forma de campana también.

A continuación se exhibe únicamente la Figura 14 correspondiente a los métodos de Taylor y Jackknife con $m = 2$, para el caso con $n = 350$. No se presentan más pues los gráficos restantes son muy parecidos. (Nota: En la Figura 14 los valores de los ejes horizontales están multiplicados por 10^3.)

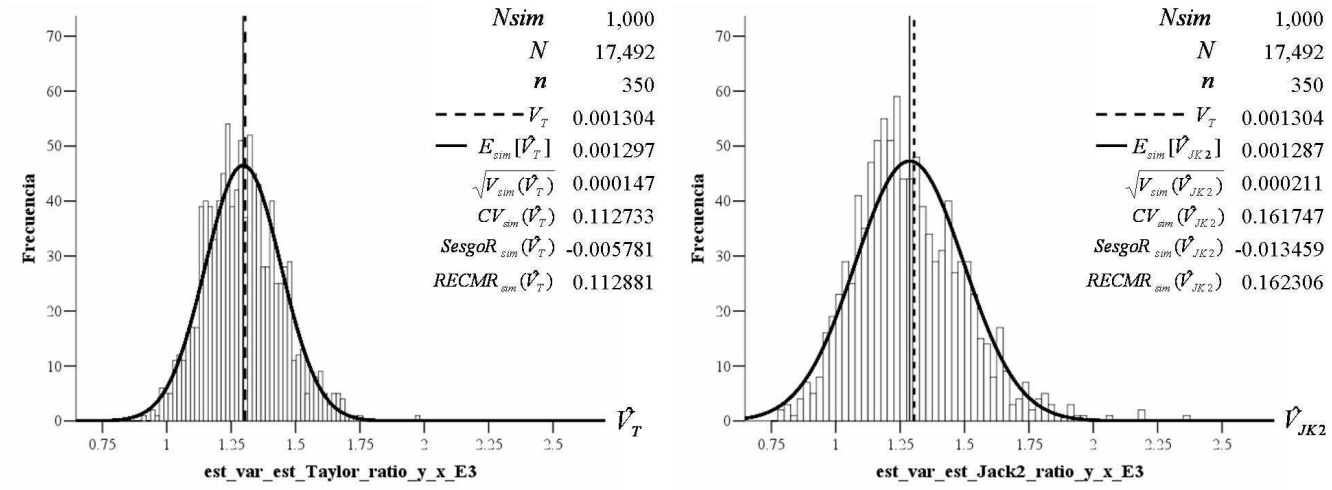

Figura 14: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para el Número de Personas por Cuarto en el Hogar utilizando el Método de Linealización de Taylor y el Método Jackknife con $m = 2$, para el caso $n = 350$.**

5.5.5. Para la Proporción de Viviendas con Teléfono (Usando la Media Muestral Como Estimador)

En esta sección se tienen los resultados obtenidos en la simulación de las estimaciones de la varianza del estimador de la varianza de la Proporción de Viviendas con Teléfono, haciendo uso de la media muestral de la variable indicadora TELEFONO[53] como estimador. Estos resultados se encuentran resumidos en la Tabla 5.5.5 siguiente[54].

Tabla 5.5.5 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador de la Proporción de Viviendas con Teléfono (Usando la Media Muestral como Estimador), $\bar{y}_S = n^{-1}\sum_{k\in S} y_k$

	Horwitz-Thompson	*Jackknife(2)*	*Vector Post-Diseño*	*Bootstrap*
$N_{sim} = 1,000$		$N = 18,108$	$n = 350$	
V	0.000641	0.000641	0.000641	0.000641
$Min_{sim}\{\widehat{V}\}$	0.000566	0.000506	0.000560	0.000529
$Max_{sim}\{\widehat{V}\}$	0.000693	0.000833	0.000706	0.000748
$E_{sim}[\widehat{V}]$	0.000640	0.000641	0.000640	0.000640
$\sqrt{V_{sim}(\widehat{V})}$	0.000022	0.000052	0.000023	0.000035
$CV_{sim}(\widehat{V})$	0.033772	0.080435	0.036594	0.055142
$SesgoR_{sim}(\widehat{V})$	-0.001152	0.001211	-0.000524	-0.000788
$RECMR_{sim}(\widehat{V})$	0.033792	0.080445	0.036598	0.055147
$N_{sim} = 1,000$		$N = 18,108$	$n = 1,000$	
V	0.000216	0.000216	0.000216	0.000216
$Min_{sim}\{\widehat{V}\}$	0.000202	0.000186	0.000199	0.000182
$Max_{sim}\{\widehat{V}\}$	0.000226	0.000253	0.000229	0.000252
$E_{sim}[\widehat{V}]$	0.000216	0.000216	0.000216	0.000215
$\sqrt{V_{sim}(\widehat{V})}$	0.000004	0.000011	0.000005	0.000011
$CV_{sim}(\widehat{V})$	0.017984	0.049203	0.022791	0.050335
$SesgoR_{sim}(\widehat{V})$	-0.000023	0.001913	-0.000470	-0.002936
$RECMR_{sim}(\widehat{V})$	0.017984	0.049240	0.022796	0.050421

[53] Acorde con lo establecido en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.

[54] Únicamente se presenta la modalidad del Método Jackknife con $m = 2$ (acorde con la notación y expresiones utilizadas en la presente tesis) para el estimador en cuestión pues se tiene que bajo un diseño de muestreo aleatorio simple (SRS, SI ó m.a.s.) el estimador de varianza Jackknife con $m = 1$ es idéntico al de Horwitz-Thompson. Véase la observación 11.2.1. en Särndal *et al.* (1992)[pag. 422].

Observando las cifras presentadas en la Tabla 5.5.5, lo primero que salta a la vista es que el método con el desempeño más pobre es el Método Jackknife(2). Esto se ve reflejado en los elevados valores para los estadísticos $CV_{sim}(\widehat{V})$ y $SesgoR_{sim}(\widehat{V})$[55] que ulteriormente se traducen en cantidades grandes en el estadístico $RECMR_{sim}(\widehat{V})$; esto siempre con respecto a los demás métodos y para un tamaño de muestra $n = 350$. Cuando se tiene un tamaño de muestra $n = 1,000$, el método que menos responde al aumento de tamaño de muestra n fue el Método Bootstrap (de hecho empeoró en términos de $SesgoR_{sim}(\widehat{V})$[56]) y por lo tanto es éste junto con el Método Jackknife(2) (que sí mejoró pero no lo suficiente considerando los demás) quienes tienen el peor desempeño comparado con el resto.

De entre los métodos de Horwitz-Thompson, y Vector Post-Diseño para el caso $n = 350$, se tienen cifras similares (siendo ligeramente mejor el Método de Horvitz-Thompson) en el estadístico $RECMR_{sim}(\widehat{V})$ que guarda de manera combinada los atributos asociados a la precisión y a la exactitud; no obstante el de Horwitz-Thompson es mejor en términos del estadístico $CV_{sim}(\widehat{V})$ (i.e. precisión) y el Método del Vector Post-Diseño de acuerdo al estadístico $SesgoR_{sim}(\widehat{V})$ (i.e. exactitud). Ahora, al elevar el tamaño de muestra a $n = 1,000$ se percibe que el Método de Horvitz-Thompson es superior en desempeño para todos los estadísticos en comparación con el resto de los métodos y a su vez considerablemente mejor que consigo mismo con tamaño de muestra de $n = 350$.

Se exhiben a continuación en las Figuras 15 y 16 los gráficos correspondientes a los cuatro métodos presentados en esta sección, para el caso con $n = 350$ únicamente pues los gráficos asociados al caso con $n = 1,000$ son bastante semejantes. (Nota: En las Figuras 15 y 16 los valores de los ejes horizontales están multiplicados por 10^3.)

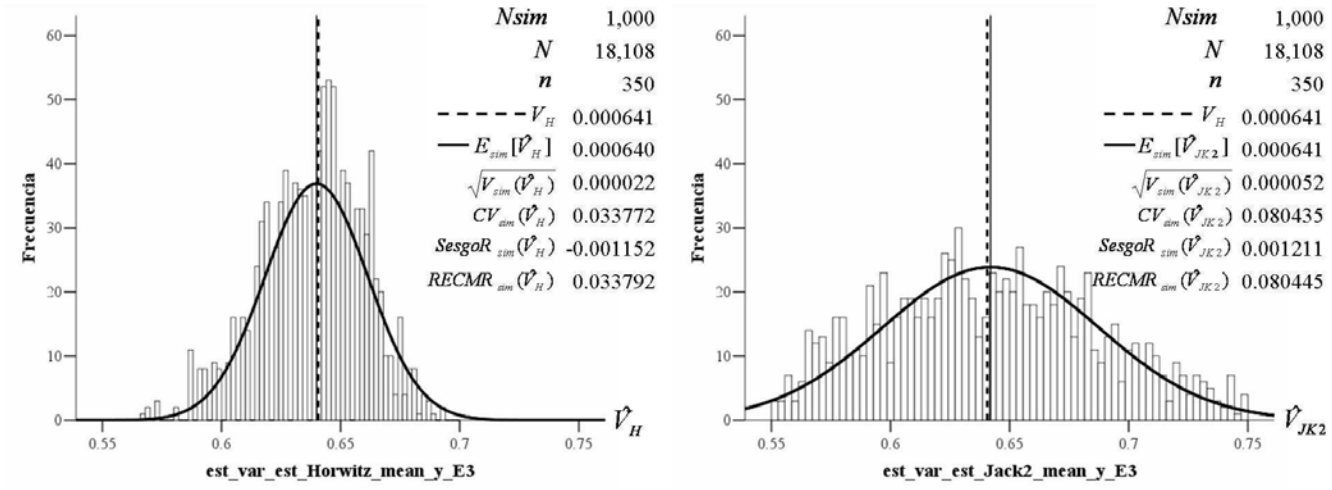

Figura 15: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Proporción de Viviendas con Teléfono (media muestral como estimador) utilizando los estimadores π ó de Horwitz-Thompson y el Método Jackknife con $m = 2$, para el caso $n = 350$.**

[55] En valor absoluto.
[56] *Ídem.*

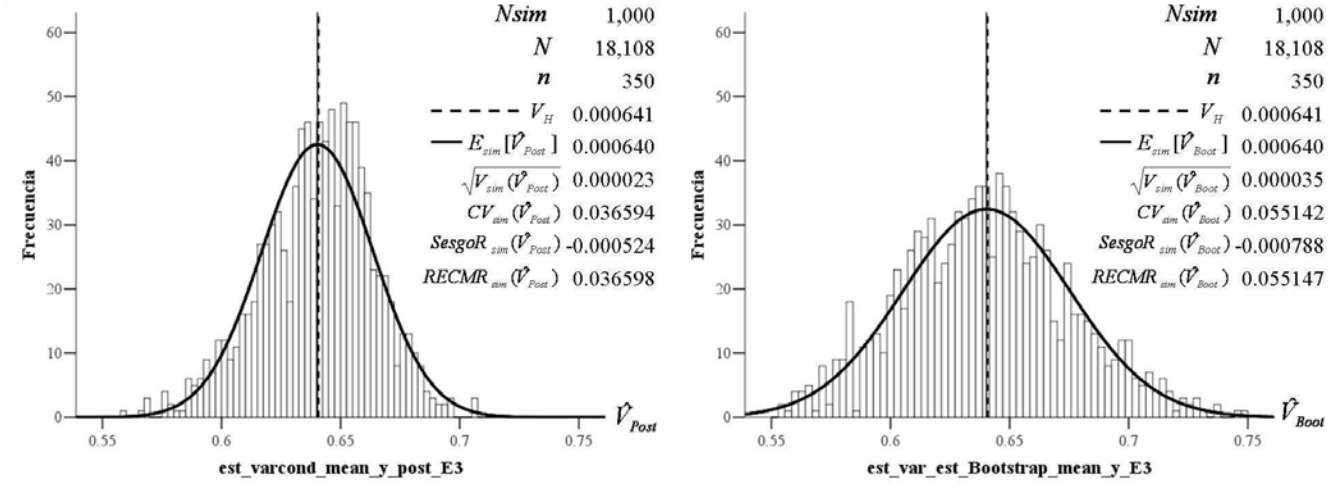

Figura 16: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Proporción de Viviendas con Teléfono (media muestral como estimador) utilizando el Método del Vector Post-Diseño y el Método Bootstrap, para el caso** $n = 350$**.**

5.5.6. Para la Proporción de Viviendas con Teléfono (Usando el Estimador de Razón)

Se presentan en la Tabla 5.5.6 siguiente los resultados de la simulación de estimaciones de la varianza del estimador de la Proporción de Viviendas con Teléfono, utilizando el estimador de razón construido a partir del cociente de los totales muestrales de las variables TELEFONO y UNOS[57].

De acuerdo a lo registrado en la Tabla 5.5.6 se tiene que para el caso $n = 350$ el método que mostró peor desempeño, con respecto al resto de métodos, fue el Método Jackknife(2) (i.e. Jackknife con $m = 2$), esto se puede notar observando sus valores obtenidos para los estadísticos $CV_{sim}(\widehat{V})$, $SesgoR_{sim}(\widehat{V})$ y $RECMR_{sim}(\widehat{V})$. Esto mismo sucede para el caso $n = 1,000$ pero acompañado también por el Método Bootstrap, pues este último en lugar de mejorar empeoró en términos del $SesgoR_{sim}(\widehat{V})$.

Los métodos con mejor desempeño resultaron ser el de Linealización de Taylor y el del Vector Post-Diseño, para el caso $n = 350$. Pero al aumentar el tamaño de muestra a $n = 1,000$ el método de Taylor supera claramente al del Vector Post-Diseño, pues este último no mejora tanto como el primero en lo atañe al estadístico $CV_{sim}(\widehat{V})$; cosa que se ve posteriormente reflejada en las cantidades que registraron para el error total relativo (i.e. $RECMR_{sim}(\widehat{V})$).

Algo interesante que también se puede observar en la Tabla 5.5.6 es que, para los dos tamaños de muestra utilizados, el par de métodos Jackknife obtuvieron cantidades negativas en el $SesgoR_{sim}(\widehat{V})$; igualmente sucedió con el Bootstrap.

Finalmente, se presentan en las Figuras 17 y 18 de manera gráfica por medio de histogramas los resultados de las simulaciones para los métodos: Linealización de Taylor, Jackknife(2), Vector Post-Diseño y Bootstrap, en el caso en el que el tamaño de muestra $n = 1,000$. No se presentan los gráficos para el caso $n = 350$ debido a que son gráficos análogamente muy parecidos. (Nota: En las Figuras 17 y 18 los valores de los ejes horizontales están multiplicados por 10^3.)

[57]Según lo establecido en las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.

Tabla 5.5.6 Resultados de las Simulaciones para el Estimador de la Varianza del Estimador de la Proporción de Viviendas con Teléfono (Usando el Estimador de Razón), $\bar{y}_S/\bar{x}_S = \left(\sum_{k\in S}\pi_k^{-1}y_k\right) / \left(\sum_{k\in S}\pi_k^{-1}\right)$

	Aprox. Lin. Taylor	*Jackknife(1)*	*Jackknife(2)*	*Vector Post-Diseño*	*Bootstrap*
	$N_{sim} = 1,000$	$N = 18,108$		$n = 350$	
V	0.000641	0.000641	0.000641	0.000641	0.000641
$Min_{sim}\{\widehat{V}\}$	0.000544	0.000491	0.000449	0.000553	0.000529
$Max_{sim}\{\widehat{V}\}$	0.000689	0.000699	0.000821	0.000720	0.000748
$E_{sim}[\widehat{V}]$	0.000641	0.000639	0.000636	0.000641	0.000640
$\sqrt{V_{sim}(\widehat{V})}$	0.000021	0.000030	0.000057	0.000023	0.000035
$CV_{sim}(\widehat{V})$	0.032699	0.046730	0.088889	0.035722	0.055142
$SesgoR_{sim}(\widehat{V})$	0.000902	-0.001793	-0.007424	0.000904	-0.000788
$RECMR_{sim}(\widehat{V})$	0.032712	0.046764	0.089198	0.035734	0.055147
	$N_{sim} = 1,000$	$N = 18,108$		$n = 1,000$	
V	0.000216	0.000216	0.000216	0.000216	0.000216
$Min_{sim}\{\widehat{V}\}$	0.000204	0.000195	0.000184	0.000202	0.000182
$Max_{sim}\{\widehat{V}\}$	0.000227	0.000231	0.000261	0.000229	0.000252
$E_{sim}[\widehat{V}]$	0.000216	0.000216	0.000216	0.000216	0.000215
$\sqrt{V_{sim}(\widehat{V})}$	0.000004	0.000006	0.000011	0.000005	0.000011
$CV_{sim}(\widehat{V})$	0.017642	0.027565	0.052984	0.022614	0.050335
$SesgoR_{sim}(\widehat{V})$	0.000325	-0.001333	-0.002145	0.000342	-0.002936
$RECMR_{sim}(\widehat{V})$	0.017645	0.027597	0.053028	0.022616	0.050421

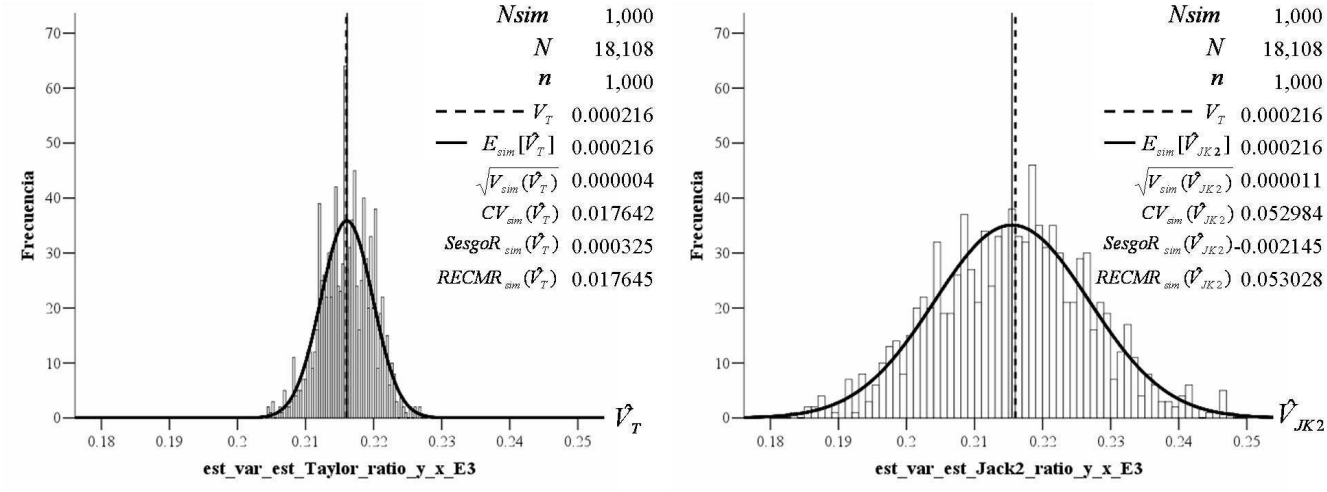

Figura 17: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Proporción de Viviendas con Teléfono (usando el estimador de razón) utilizando el Método de Linealización de Taylor y el Método Jackknife con $m = 2$, para el caso $n = 1,000$.**

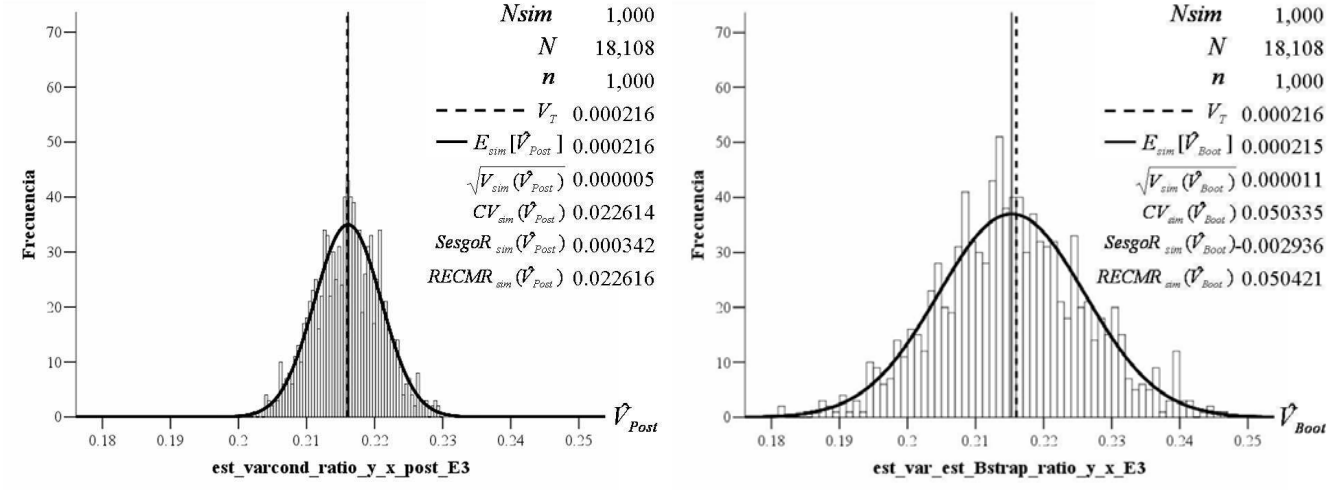

Figura 18: **Histogramas de las estimaciones de varianza del estimador obtenidas en la simulación, para la Proporción de Viviendas con Teléfono (usando el estimador de razón) utilizando el Método del Vector Post-Diseño y el Método Bootstrap, para el caso** $n = 1,000$**.**

Ahora, haciendo una comparación del uso de la media muestral de la variable indicadora TELEFONO como estimador de la proporción de viviendas con teléfono contra el uso de un estimador de razón de totales de la variable indicadora TELEFONO en el numerador y la variable artificial UNOS[58] (con valores de puros unos) en el denominador, se obtienen estimaciones más precisas y exactas utilizando el estimador de razón[59]. No obstante, en nuestro caso fue poco lo que se gana pues estamos calculando una proporción (por ello tenemos en el denominador una constante); y acorde con lo planteado por Kish (1987) esta ganancia depende en gran parte de las características de la variable utilizada en el denominador y por supuesto de la relación que tengan el par de variables involucradas.

Hay que anotar adicionalmente, al respecto de esta comparación, que los estimadores utilizados en ambos casos coinciden analíticamente bajo un diseño de muestreo aleatorio simple sin reemplazo (SRS, SI ó m.a.s.) y por lo tanto las ligeras diferencias numéricas obtenidas entre ambos casos tienen origen en los procedimientos de simulación utilizados para la evaluación del desempeño.

[58] Acorde con la nomenclatura y definiciones de las Tablas 5.2.a, 5.3.1 y 5.3.2, en las páginas 38 y 44.
[59] Como se anticipó en Kish (1987)[pag. 131-137].

6. Conclusiones

En este capítulo se presentan algunos comentarios y conclusiones de la investigación materia de la presente tesis. Conforme al título, el objetivo principal fue la evaluación del desempeño práctico de algunos métodos de estimación de varianza. Para llevar acabo esta evaluación se realizaron comparaciones de los métodos contemplados, sujetándolos a circunstancias iguales (véase la sección 5.3).

Se consideraron varios métodos conocidos de estimación de varianza: el Método de Linealización de Taylor (o Método Delta), el Método de Woodruff para cuantiles, el Método Jackknife en dos modalidades, el Método Bootstrap y también se utilizaron las expresiones correspondientes a los estimadores π ó de Horwitz-Thompson. Adicionalmente se incluyó un método nuevo propuesto en Ollila (2004) en el que se adopta un enfoque novedoso del muestreo probabilístico (como ya fue mencionado) en términos de *vectores diseño*, *metamuestras* y el diseño muestral (y *metamuestral*) como una distribución multivariada; cuya teoría fue incluida de manera resumida (Capítulo 3). En lo que respecta a los métodos conocidos también se incluyó una, también breve, presentación de estos (Capítulo 4) para su posterior implementación en el ejercicio de simulación utilizado para comparar y evaluar de sus desempeños.

Los criterios empleados para evaluar los métodos, cuya definición y nomenclatura fue establecida en la sección 5.4, pueden reducirse a tres propiedades básicas: $CV_{sim}(\widehat{V}(\widehat{\theta}))$ (precisión), $SesgoR_{sim}(\widehat{V}(\widehat{\theta}))$ (exactitud) y $RECMR_{sim}(\widehat{V}(\widehat{\theta}))$ (error total i.e. combinación de precisión y exactitud). Las tres son medidas relativizadas con el valor de $V(\widehat{\theta})$, usualmente desconocido pero que en el presente caso sí se conoce. Ahora, el cálculo de estas propiedades fue realizado mediante simulaciones utilizando el Método de Monte Carlo con $N_{sim} = 1,000$, número de simulaciones. Adicionalmente a los resultados numéricos obtenidos en la estimación de estas propiedades se emplearon representaciones gráficas (histogramas de frecuencias) de las estimaciones de varianza resultantes de la simulación; esto con el objeto de complementar la información exhibida por las propiedades estimadas.

Se identifican las situaciones siguientes, cuando se busca estimar la varianza del estimador de: 1) la media de una variable con alta asimetría, 2) la mediana de una variable con alta asimetría, 3) la razón de los totales de dos variables (con la variable numerador altamente asimétrica), 4) la razón de los totales de dos variables (con el par de variables no altamente asimétricas) y 5) la proporción (dos casos: usando la media muestral y el estimador de razón). Considerando dos tamaños de muestra $n = 350$ y $n = 1,000$; se contemplan entonces 5 pares de situaciones de estimación de varianza (bajo SRS, SI ó m.a.s.).

La forma en cómo serán presentadas las conclusiones será haciendo un listado de éstas situaciones, comentando sobre el desempeño de cada método y proporcionando algunas observaciones destacables cuando corresponda.

Para la **estimación de varianza del estimador de la media de una variable con alta asimetría** (i.e. media muestral de la variable INGTOHOG), se mostró en general que todos los métodos presentaron un mal desempeño bajo los criterios aquí utilizados y además si se consideran las distribuciones bimodales asimétricas de los histogramas de las estimaciones simuladas, la evaluación de su desempeño sufre gran menoscabo; pues se tienen frecuencias gráficamente nulas de estimaciones alrededor del valor verdadero de la varianza, lo que se traduce en un alto riesgo de

obtener subestimaciones o sobreestimaciones de la varianza por el alto $CV_{sim}(\widehat{V}(\widehat{\theta}))$ (al respecto, véase la parte 5.5.1, que comienza en la página 46, en donde se dan comentarios para la media de los ingresos mensuales totales por hogar y también se mencionan algunas posibles alternativas para enfrentar datos altamente asimétricos). También, es sabido que la *normalidad* de la *distribución asintótica del estimador del parámetro* de interés de una población, basados en muestras grandes depende del *Teorema Central del Límite*. Se sabe también que *en esta aproximación normal no se requiere de normalidad de las variables en su distribución poblacional*[60]. No obstante, la distribución del estimador de la varianza de la media muestral parece ser más afectada por la asimetría que la media muestral misma[61]. Adicionalmente hay que considerar que lo relativo a la distribución asintótica se refiere a la distribución límite, lo que en nuestro caso puede traducirse a que no se obtuvo un rápido acercamiento de la distribución de nuestro estimador a la normalidad con los valores de tamaño de muestra n finitos aquí utilizados. Volviendo a la evaluación, para el caso $n = 350$ se obtuvieron desempeños muy similares, aventajando ligeramente el método que utiliza los estimadores π ó de Horwitz-Thompson. Cuando se empleó un tamaño de muestra de $n = 1,000$, los desempeños dejan de ser tan parecidos de modo que el Método Jackknife con $m = 2$ obtuvo mejores cifras en términos del error total.

En la **estimación de varianza del estimador de la mediana de una variable con alta asimetría** (i.e. mediana muestral de la variable INGTOHOG), todos los métodos lograron un desempeño muy similar (aceptable si consideramos la distribución excesivamente asimétrica de los datos), inclusive para los dos tamaños de muestra empleados. Sin embargo, de nueva cuenta, observando los gráficos se tiene que el Método de Woodruff exhibió distribuciones (para el caso $n = 350$) alejadas de la deseable forma de campana de una función de densidad Normal[62]; no sucedió así con los Métodos del Vector Post-Diseño y el Bootstrap.

En la **estimación de varianza del estimador de la razón de los totales de dos variables (con la variable numerador altamente asimétrica)** (i.e. razón de los totales muestrales de las variables INGTOHOG y TOTPERS), se obtuvo que todos los métodos tuvieron un desempeño muy pobre, y observando además los histogramas correspondientes, todos reflejaron distribuciones bimodales muy asimétricas y (como ya se anotó en la sección 5.5.3) con frecuencias gráficamente nulas de estimaciones alrededor del valor verdadero, lo que representa muy altos riesgos de subestimación de la varianza. No obstante numéricamente resultó que, para ambos tamaños de muestra, el par de Métodos Jackknife fueron los que peor desempeño mostraron, mientras que el Método de Linealización de Taylor seguido del Método del Vector Post-Diseño fueron los que mejor desempeño mostraron.

Para la **estimación de varianza del estimador de la razón de los totales de dos variables (con el par de variables no altamente asimétricas)** (i.e. razón de los totales muestrales de las variables TOTPERS y TOTCUART), los métodos que resultaron con mejor desempeño, para los dos tamaños de muestra empleados, fueron el Método de Linealización de Taylor, seguido del Método del Vector Post-Diseño. Mientras que los que obtuvieron el peor desempeño (también para ambos tamaños de muestra) fueron el par de Métodos Jackknife contemplados. En lo que respecta a las representaciones gráficas se obtuvieron histogramas muy parecidos con todos los métodos y para los dos tamaños de muestra; en todos los casos se exhibieron distribuciones con forma de campana similar a una función de densidad Normal.

[60] Véase Kish (1965)[pag. 16].

[61] Véase Kish (1965)[pag. 411].

[62] En la sección 5.5.2 correspondiente a los resultados para la mediana se comenta sobre los posibles motivos por los que el Método de Woodruff obtuvo un desempeño pobre.

Relativo a la **estimación de varianza del estimador de la proporción**; primero, resumiendo lo ya comentado en la sección 5.5.5 sobre el uso de la media muestral, el método con mejor desempeño fue aquél que hace uso de las expresiones relativas a los estimadores π ó de Horwitz-Thompson seguido del Método del Vector Post-Diseño, para ambos tamaños de muestra. Mientras que el método con peor desempeño fue el Jackknife(2) para el tamaño de muestra $n = 350$; y el método con peor desempeño para $n = 1,000$ fue el de Bootstrap. Luego, acorde con lo encontrado en la sección 5.5.6 usando el estimador de razón, se obtuvo que fueron el Método de Linealización de Taylor, seguido del Método del Vector Post-Diseño los que mostraron mejor desempeño; y aquellos con peor desempeño fueron el Método Jackknife(2) y el Método Bootstrap, todo esto para los dos tamaños de muestra utilizados. Para todos los casos (i.e. con estimadores de razón y con la media muestral como estimador) se obtuvieron histogramas con la deseable forma de campana; se reflejó en estos que aquellos métodos con mejor desempeño mostraron curvas con forma de campana de menor varianza (i.e. más punta y extremos más angostos), mientras que los métodos con peor desempeño exhibieron curvas con mayor varianza, es decir extremos más gruesos. Se puede afirmar que se obtienen mejores estimaciones utilizando el estimador de razón en lugar de la media muestral (como se anticipa y se discute en Kish (1987)[pag. 131-137]), aunque no siempre se obtienen grandes beneficios como sucedió en el presente caso (pues estamos calculando una proporción y por ello tenemos en el denominador una constante -y aunado a esto el hecho de que bajo SRS, SI ó m.a.s. las expresiones de los dos casos coinciden-); las ventajas de utilizar el estimador de razón sobre la media muestral tienen que ver con la variables asociada al denominador de la razón y a la relación que guardan las dos variables del cociente.

Se puede notar, que en general el Método del Vector Post-Diseño obtuvo en general buenos resultados en comparación con el resto de métodos empleados, aunque no se puede afirmar su completa supremacía. No obstante, hay algunos comentarios al respecto, no hay que olvidar que en la presente tesis únicamente se utilizó una forma de expansión del *vector diseño* y que por tanto falta explorar la innumerable cantidad de expansiones posibles que el Método del Vector Post-Diseño pueda considerar. Aunque el Método Bootstrap no obtuvo una buena evaluación con respecto a los demás métodos, hay que recordar que el Método del Vector Post-Diseño en su interior contiene al Método de Replicaciones de Bootstrap para la estimación de la varianza condicional de las remuestras dada la muestra. También, falta explorar este método bajo otros diseños de muestreo diferentes. Se menciona en Ollila (2004)[pag. 109] que es posible utilizarlo cuando se tiene estratificación y se trata la estimación de varianza de manera independiente en cada estrato. No obstante esto puede tornarse complicado si se tienen muestras grandes con gran estratificación (i.e. muchos estratos) pues los métodos de corrección de la varianza condicional sobre el espacio de remuestras pueden llegar a ser difíciles de manejar. Adicionalmente, se menciona también que el método puede ser empleado en muestreo por conglomerados (unietápico) de manera similar tomando los conglomerados como individuos.

Finalmente, la evaluación aquí realizada del desempeño de los métodos de estimación de varianza puede llegar a ser muy útil en aquellas situaciones semejantes a las aquí contempladas. Se consideró: la estimación de parámetros poblacionales, variables, y tamaños de muestra usualmente empleados; aunque bajo un diseño de muestreo aleatorio simple que, como es sabido, tiene muy restringida su aplicación en la práctica, pero es sin embargo un diseño de referencia.

Apéndice. Programas de las Rutinas de Simulación

A continuación se exhiben las rutinas programadas para las tareas de simulación. Éstas fueron realizadas en *MATLAB*® Versión 6.0 Release 12.

Se presentan las rutinas empleadas para el estimador de la varianza del estimador de:

- la media (con los Métodos: Horwitz-Thompson, Jackknife y Vector Post-Diseño),
- la mediana (con los Métodos: Woodruff y Vector Post-Diseño) y
- la razón de los totales de dos variables (con los Métodos: Taylor, Jackknife(1), Jackknife(2) y Vector Post-Diseño)

Posteriormente se presenta por separado la rutina asociada al Método Bootstrap (para la media, mediana y razón). Y Finalmente, se muestra una pequeña subrutina utilizada para la obtención de muestras ordenadas sin reemplazo.

Para el Estimador de la Varianza del Estimador de la Media

```
%*************************************************************************
%nsim times simulation of the extraction of a sample of size na from a population
%of size N and use different methods for variance estimation of the estimator
%
%SI sampling / SI resampling
%
%Variance estimation methods:
%     Horwitz-Thompson
%     Jackknife2
%     Post-Design Vector (one q-factor)
%
%For the MEAN
%-------------------------------------------------------------------------
function SIM_SI_SI_mean(population_matrix,pop_mat_column,nsim,na,A);

tic %Elapsed time start point

%Takes as population the "pop_mat_column" column of the population matrix
population_vector_y = population_matrix(:,pop_mat_column);

population_matrix = []; %Cleaning up to free memory
pop_mat_column = []; %Cleaning up to free memory
N = length(population_vector_y); %Population size

%Direct calculation of the parameter and variance of the estimator
%--------------------------------------
```

```
%For the MEAN
population_mean_y = mean(population_vector_y);
var_est_Horwitz_mean_y = (1-(na/N))/na*var(population_vector_y);

%For the Post-Design Vector Method
%--------------------------------------
%Resample size that fulfils the linear case condition for SI/SI
nb = ( N * na ) / ( 2 * N - na );
%Upper integer resample size that fulfils the linear case condition
nbu = ceil(nb);
%For the calculation of the q-factor
Temp1 = ( nbu - 1 )^2 - nb * ( nbu - 1 );
Temp2 = 2 * ( nbu - 1 );
Temp3 = 1 - nb;
q1 = (-Temp2-realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
q2 = (-Temp2+realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
qfactor = max(q1,q2);
%Artificial resample size of the expanded resample
nb_artificial = qfactor*(nbu-1)+1;

nb = []; %Cleaning up to free memory
Temp1 = []; %Cleaning up to free memory
Temp2 = []; %Cleaning up to free memory
Temp3 = []; %Cleaning up to free memory
q1 = []; %Cleaning up to free memory
q2 = []; %Cleaning up to free memory

%Simulations
%--------------------------------------
sim_i = 1; %Initial value of the counter sim_i
conv_sim = 1; %Initial value of the convergence iteration counter
%Initialize a vector for the results of MEANS of the simulations by Horwitz
est_Horwitz_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Horwitz
est_var_est_Horwitz_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Jackknife2
est_Jackknife2_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Jackknife2
est_var_est_Jackknife2_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Bootstrap
est_boot_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Bootstrap
est_var_est_boot_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Post-Design
%Vector Method
est_varcond_mean_y_post = [];

%Simulation routines
```

```
while sim_i<=nsim
    %Convergence counter criteria
    if isempty(est_varcond_mean_y_post) |
    abs(mean(est_varcond_mean_y_post)-var_est_Horwitz_mean_y)
    >var_est_Horwitz_mean_y/10000
        conv_sim = conv_sim + 1;
    end
    %Take a sample of size na from the population
    sample_vector_y = population_vector_y(SI_os(na,N));
    %Horwitz calculation of the estimated variance of the estimator with
    %the sample
    %--------------------------------------
    %For the MEAN
    est_Horwitz_mean_y(sim_i) = sum(sample_vector_y) / na;
    est_var_est_Horwitz_mean_y(sim_i) = (1-(na/N))*var(sample_vector_y)/na;
        %Calculation of the Jackknife2 method estimates
    %--------------------------------------
    %For the MEAN
    m = 2;
    AA = na/m; %NOTE: In this case na should be an even number because m = 2
    Teta = est_Horwitz_mean_y(sim_i);
    pseudovalues2 = [];
    for a = 1:AA
        pseudovalues2(a) =
        AA*Teta-(AA-1)*((sum(sample_vector_y)-sample_vector_y(a)-
        sample_vector_y(a+AA))/(na-m));
    end
    est_Jackknife2_mean_y(sim_i) = mean(pseudovalues2);
    est_var_est_Jackknife2_mean_y(sim_i) = (1-(na/N))*var(pseudovalues2)/AA;
    pseudovalues2 = []; %Cleaning up to free memory

    %Calculation of the Bootstrap Estimator of the Conditional Variance of the
    %Resample Post-Design Vector Estimator given a sample
    %--------------------------------------
    %For the MEAN
    %Initialize a vector for the results of MEANS of the A bootstrap resamples
    bootmeans = [];
    %Calculating the post-design vector statistics of the A bootstrap resamples
    for boot_i = 1:A
        %Using SI resample design
        %Here "SI_os(nbu,na)" creates index vector of
        %an ordered SI bootstrap resample

        %Calculating the MEAN of the expanded resample
        bootmeans(boot_i) =
        (sample_vector_y(SI_os(nbu,na))'*
        [qfactor*ones(nbu-1,1);1])/nb_artificial;
    end
```

```
    sample_vector_y = []; %Cleaning up to free memory

    %Calculating the bootstrap estimation of the conditional variance of the
    %Resample Post-Design Vectors statistics given a sample
    est_boot_mean_y(sim_i) = mean(bootmeans);
    est_varcond_mean_y_post(sim_i) = var(bootmeans);
    est_var_est_boot_mean_y(sim_i) = est_varcond_mean_y_post(sim_i)/A;
    bootmeans = []; %Cleaning up to free memory

    sim_i = sim_i + 1; %Simulation iteration counter update
    disp(sim_i) %Displaying on the command window the simulation iteration
end

%For the Output on the Command Window
N
na
population_mean_y
var_est_Horwitz_mean_y
nbu
nb_artificial
qfactor
est_Horwitz_mean_y
est_var_est_Horwitz_mean_y
est_Jackknife2_mean_y
est_var_est_Jackknife2_mean_y
est_boot_mean_y
est_var_est_boot_mean_y
est_varcond_mean_y_post
nsim
conv_sim = conv_sim - 1
elapsed_time_mins = toc/60 %Reporting elapsed time in minutes

%For the Output on a file
fid = fopen('SIM_MEAN.txt','a');
fprintf(fid,'
N
na
population_mean_y
var_est_Horwitz_mean_y
nbu
nb_artificial
qfactor
est_Horwitz_mean_y
est_var_est_Horwitz_mean_y
est_Jack2_mean_y
est_var_est_Jack2_mean_y
est_boot_mean_y
```

```
est_var_est_boot_mean_y
est_varcond_mean_y_post
nsim
conv_sim
elapsed_time_mins\n');
for i = 1:nsim
fprintf(fid,'%30.0f',N);
fprintf(fid,'%30.0f',na);
fprintf(fid,'%30.8f',population_mean_y);
fprintf(fid,'%30.8f',var_est_Horwitz_mean_y);
fprintf(fid,'%30.0f',nbu);
fprintf(fid,'%30.4f',nb_artificial);
fprintf(fid,'%30.4f',qfactor);
fprintf(fid,'%30.8f',est_Horwitz_mean_y(i));
fprintf(fid,'%30.8f',est_var_est_Horwitz_mean_y(i));
fprintf(fid,'%30.8f',est_Jackknife2_mean_y(i));
fprintf(fid,'%30.8f',est_var_est_Jackknife2_mean_y(i));
fprintf(fid,'%30.8f',est_boot_mean_y(i));
fprintf(fid,'%30.8f',est_var_est_boot_mean_y(i));
fprintf(fid,'%30.8f',est_varcond_mean_y_post(i));
fprintf(fid,'%30.0f',nsim);
fprintf(fid,'%30.0f',conv_sim);
fprintf(fid,'%30.2f\n',elapsed_time_mins);
end
fclose(fid);
```

Para el Estimador de la Varianza del Estimador de la Mediana

```
%*************************************************************************
%nsim times simulation of the extraction of a sample of size na from a population
%of size N and use different methods for variance estimation of the estimator
%
%SI sampling / SI resampling
%
%Variance estimation methods:
%     Woodruff
%     Post-Design Vector (one q-factor)
%
%For the MEDIAN
%-------------------------------------------------------------------------
function SIM_SI_SI_median(population_matrix,pop_mat_column,nsim,na,A);

tic %Elapsed time start point

%Takes as population the "pop_mat_column" column of the population matrix
```

```
population_vector_y = population_matrix(:,pop_mat_column);

population_matrix = []; %Cleaning up to free memory
pop_mat_column = []; %Cleaning up to free memory
N = length(population_vector_y); %Population size

%Direct calculation of the parameter and variance of the estimator
%---------------------------------------
%For the MEDIAN
%Creating a sorted version of population_vector_y
x = sortrows(population_vector_y);
%Adding to the sorted population vector the cumulative distribution function
%F column of the population
x = [ x , cumsum(ones(N,1))/N ];
%Note: In finding the F's inverse at certain value "q" we took the maximum
%value "a" such that F(a)<=q
population_median_y = x(max(find(x(:,2) <= 0.50)),1);
c1=0.5-1.96*((N-na)/(N-1)*
 (1/na)*x(max(find(x(:,2)<=0.50)),2)*(1-x(max(find(x(:,2)<=0.50)),2)))^(1/2);
c2=0.5+1.96*((N-na)/(N-1)*
 (1/na)*x(max(find(x(:,2)<=0.50)),2)*(1-x(max(find(x(:,2)<=0.50)),2)))^(1/2);
F__c1 = x(max(find(x(:,2) <= c1)),1); %Finds the F's inverse at c1
F__c2 = x(max(find(x(:,2) <= c2)),1); %Finds the F's inverse at c2
var_est_Woodruff_median_y = ((F__c2 - F__c1)/(2*1.96))^2;
x = []; %Cleaning up to free memory
c1 = []; %Cleaning up to free memory
c2 = []; %Cleaning up to free memory
F__c1 = []; %Cleaning up to free memory
F__c2 = []; %Cleaning up to free memory

%For the Post-Design Vector Method
%---------------------------------------
%Resample size that fulfils the linear case condition for SI/SI
nb = ( N * na ) / ( 2 * N - na );
%Upper integer resample size that fulfils the linear case condition
nbu = ceil(nb);
%For the calculation of the q-factor
Temp1 = ( nbu - 1 )^2 - nb * ( nbu - 1 );
Temp2 = 2 * ( nbu - 1 );
Temp3 = 1 - nb;
q1 = (-Temp2-realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
q2 = (-Temp2+realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
qfactor = max(q1,q2);
%Artificial resample size of the expanded resample
nb_artificial = qfactor*(nbu-1)+1;

nb = []; %Cleaning up to free memory
Temp1 = []; %Cleaning up to free memory
```

```
Temp2 = []; %Cleaning up to free memory
Temp3 = []; %Cleaning up to free memory
q1 = []; %Cleaning up to free memory
q2 = []; %Cleaning up to free memory

%Simulations
%--------------------------------------
sim_i = 1; %Initial value of the counter sim_i
conv_sim = 1; %Initial value of the convergence iteration counter
%Initialize a vector for the results of MEDIANS of the simulations by Woodruff
est_Woodruff_median_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Woodruff
est_var_est_Woodruff_median_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Bootstrap
est_boot_median_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Bootstrap
est_var_est_boot_median_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Post-Design
%Vector Method
est_varcond_median_y_post = [];

%Simulation routines
while sim_i<=nsim
    %Convergence counter criteria
    if isempty(est_varcond_median_y_post) |
    abs(mean(est_varcond_median_y_post)-var_est_Woodruff_median_y)
    >var_est_Woodruff_median_y/10000
        conv_sim = conv_sim + 1;
    end
    %Take a sample of size na from the populations
    sample_vector_y = population_vector_y(SI_os(na,N));
    %Woodruff calculation of the estimated variance of the estimator with
    %the sample
    %--------------------------------------
    %For the MEDIAN
    %Using Sarndal et al. pages 197-204 and Ollila's dissertation pages 22-23
    %Creating a sorted version of sample_vector_y
    x = sortrows(sample_vector_y);
    %Adding to the sorted sample vector the cumulative distribution function
    %F column of the sample
    x = [ x , cumsum(ones(na,1))/na ];
    est_Woodruff_median_y(sim_i) = x(max(find(x(:,2) <= 0.50)),1);
    c1=0.5-1.96*((N-na)/(N-1)*(1/na)*
     x(max(find(x(:,2)<= 0.50)),2)*(1-x(max(find(x(:,2)<= 0.50)),2)))^(1/2);
    c2=0.5+1.96*((N-na)/(N-1)*(1/na)*
     x(max(find(x(:,2)<= 0.50)),2)*(1-x(max(find(x(:,2)<= 0.50)),2)))^(1/2);
    %Note: In finding the F's inverse at certain value "q" we took the maximum
    %value "a" such that F(a)<=q
```

```
F__c1 = x(max(find(x(:,2) <= c1)),1); %Finds the F's inverse at c1
F__c2 = x(max(find(x(:,2) <= c2)),1); %Finds the F's inverse at c2
est_var_est_Woodruff_median_y(sim_i) = ((F__c2 - F__c1)/(2*1.96))^2;

c1 = []; %Cleaning up to free memory
c2 = []; %Cleaning up to free memory
F__c1 = []; %Cleaning up to free memory
F__c2 = []; %Cleaning up to free memory
x = []; %Cleaning up to free memory

%Calculation of the Bootstrap Estimator of the Conditional Variance of the
%Resample Post-Design Vectors Estimator given a sample
%--------------------------------------
%For the MEDIAN
%Initialize a vector for the results of MEDIANS of the A bootstrap resamples
bootmedians = [];
%Calculating the post-design vector statistics of the A bootstrap resamples
for boot_i = 1:A
 %Initialize or clear previos matrix of a bootstrap resample
    boot_matrix_resample = [];
    %Using SI resample design
    %Here "SI_os(nbu,na)" creates index vector of an ordered
    %SI bootstrap resample

    %Calculating the MEDIAN of the expanded resample
    %Creates measures y column and factors column of the matrix of an
    %ordered bootstrap resample
    %and sorting the matrix resample with respect to the measures column y
    boot_matrix_resample =
     sortrows([sample_vector_y(SI_os(nbu,na)),[qfactor*ones(nbu-1,1);1]],1);
    %Overwriting the factors column with the cumulative distribution
    %function F* of the factor-expanded sorted resample
    boot_matrix_resample(:,2) =
     cumsum(boot_matrix_resample(:,2))/nb_artificial;
    %For calculate the median of the expanded resample we need F*'s inverse
    %at 0.5
    %(Following Sarndal et al., page 197 definitions but adjusting for
    %non-integer
     expansions and for the next note)
    %Note: In finding the F's inverse at certain value "q" we took the
    %maximum value "a" such that F(a)<=q
    %Finds the index line of the possible median value
    M = max(find(boot_matrix_resample(:,2) <= 0.5));
    %Checks if the M+1 indexed measure value of y is the median using the
    %fact that qfactor >= 1
    if boot_matrix_resample(M,2) + (1/nb_artificial) <= 0.5
        bootmedians(boot_i) = boot_matrix_resample(M+1,1);
    %Hence the M indexed measure value of y is the median
```

```
        else
            bootmedians(boot_i) = boot_matrix_resample(M,1);
        end
        M = []; %Cleaning up to free memory
    end

    boot_matrix_resample = []; %Cleaning up to free memory
    sample_vector_y = []; %Cleaning up to free memory

    %Calculating the bootstrap estimation of the conditional variance of the
    %Resample Post-Design Vectors statistics given a sample
    est_boot_median_y(sim_i) = mean(bootmedians);
    est_varcond_median_y_post(sim_i) = var(bootmedians);
    est_var_est_boot_median_y(sim_i) = est_varcond_median_y_post(sim_i)/A;
    bootmedians = []; %Cleaning up to free memory

    sim_i = sim_i + 1; %Simulation iteration counter update
    disp(sim_i)
end

%For the Output on the Command Window
N
na
population_median_y
var_est_Woodruff_median_y
nbu
nb_artificial
qfactor
est_Woodruff_median_y
est_var_est_Woodruff_median_y
%est_Jackknife2_median_y
%est_var_est_Jackknife2_median_y
est_boot_median_y
est_var_est_boot_median_y
est_varcond_median_y_post
nsim
conv_sim = conv_sim - 1
elapsed_time_mins = toc/60 %Reporting elapsed time in minutes

%For the Output on a file
fid = fopen('SIM_MEDIAN.txt','a');
fprintf(fid,'
N
na
population_median_y
var_est_Woodruff_median_y
nbu
nb_artificial
```

```
qfactor
est_Woodruff_median_y
est_var_est_Woodruff_median_y
est_boot_median_y
est_var_est_boot_median_y
est_varcond_median_y_post
nsim
conv_sim
elapsed_time_mins\n');
for i = 1:nsim
fprintf(fid,'%30.0f',N);
fprintf(fid,'%30.0f',na);
fprintf(fid,'%30.6f',population_median_y);
fprintf(fid,'%30.6f',var_est_Woodruff_median_y);
fprintf(fid,'%30.0f',nbu);
fprintf(fid,'%30.4f',nb_artificial);
fprintf(fid,'%30.4f',qfactor);
fprintf(fid,'%30.6f',est_Woodruff_median_y(i));
fprintf(fid,'%30.6f',est_var_est_Woodruff_median_y(i));
%fprintf(fid,'%30.6f',est_Jackknife2_median_y(i));
%fprintf(fid,'%30.6f',est_var_est_Jackknife2_median_y(i));
fprintf(fid,'%30.6f',est_boot_median_y(i));
fprintf(fid,'%30.6f',est_var_est_boot_median_y(i));
fprintf(fid,'%30.6f',est_varcond_median_y_post(i));
fprintf(fid,'%30.0f',nsim);
fprintf(fid,'%30.0f',conv_sim);
fprintf(fid,'%30.2f\n',elapsed_time_mins);
end
fclose(fid);
```

Para el Estimador de la Varianza del Estimador de la Razón de los Totales de Dos Variables

```
%*************************************************************************
%nsim times simulation of the extraction of a sample of size na from a population
%of size N and use different methods for variance estimation of the estimator
%
%SI sampling / SI resampling
%
%Variance estimation methods:
%     Taylor
%     Jackknife1
%     Jackknife2
%     Post-Design Vector (one q-factor)
%
```

```
%For the RATIO
%-------------------------------------------------------------------------------
function SIM_SI_SI_ratio(population_matrix,pop_mat_column_y,pop_mat_column_x,
 nsim,na,A);

tic %Elapsed time start point

%Takes as population the "pop_mat_column_y" column of variable y and
%"pop_mat_column_x" column of variable x of the population matrix
population_matrix_y_x = [ population_matrix(:,pop_mat_column_y) ,
 population_matrix(:,pop_mat_column_x) ];

population_matrix = []; %Cleaning up to free memory
pop_mat_column_y = []; %Cleaning up to free memory
pop_mat_column_x = []; %Cleaning up to free memory

N = length(population_matrix_y_x); %Population size

%Direct calculation of the parameter and variance of the estimator
%---------------------------------------
%For the RATIO
population_ratio_y_x =
 sum(population_matrix_y_x(:,1)) / sum(population_matrix_y_x(:,2));
var_est_Taylor_ratio_y_x =
 (1/(mean(population_matrix_y_x(:,2))^2))*((1-na/N)/na)*
 ((sum((population_matrix_y_x(:,1)-(population_ratio_y_x*
 population_matrix_y_x(:,2))).^2))/(N-1));

%For the Post-Design Vector Method
%---------------------------------------
%Resample size that fulfils the linear case condition for SI/SI
nb = ( N * na ) / ( 2 * N - na );
%Upper integer resample size that fulfils the linear case condition
nbu = ceil(nb);
%For the calculation of the q-factor
Temp1 = ( nbu - 1 )^2 - nb * ( nbu - 1 );
Temp2 = 2 * ( nbu - 1 );
Temp3 = 1 - nb;
q1 = (-Temp2-realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
q2 = (-Temp2+realsqrt(Temp2^2-4*Temp1*Temp3))/(2*Temp1);
qfactor = max(q1,q2);
%Artificial resample size of the expanded resample
nb_artificial = qfactor*(nbu-1)+1;

nb = []; %Cleaning up to free memory
Temp1 = []; %Cleaning up to free memory
Temp2 = []; %Cleaning up to free memory
Temp3 = []; %Cleaning up to free memory
```

```
q1 = []; %Cleaning up to free memory
q2 = []; %Cleaning up to free memory


%Simulations
%--------------------------------------
sim_i = 1; %Initial value of the counter sim_i
conv_sim = 1; %Initial value of the convergence iteration counter
%Initialize a vector for the results of RATIOS of the simulations by Horwitz
est_Horwitz_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Horwitz
est_var_est_Taylor_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Jackknife2
est_Jackknife1_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Jackknife2
est_var_est_Jackknife1_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Jackknife2
est_Jackknife2_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Jackknife2
est_var_est_Jackknife2_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Bootstrap
est_boot_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Bootstrap
est_var_est_boot_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by
%Post-Design Vector Method
est_varcond_ratio_y_x_post = [];

%Simulation routines
while sim_i<=nsim
    %Convergence counter criteria
    if isempty(est_varcond_ratio_y_x_post) |
    abs(mean(est_varcond_ratio_y_x_post)-var_est_Taylor_ratio_y_x)
    >var_est_Taylor_ratio_y_x/10000
        conv_sim = conv_sim + 1;
    end
    %Take a sample of size na from the populations
    sample_matrix_y_x = population_matrix_y_x(SI_os(na,N),:);
    %Taylor calculation of the estimated variance of the estimator
    %with the sample
    %--------------------------------------
    %For the RATIO
    est_Horwitz_ratio_y_x(sim_i) = sum(sample_matrix_y_x(:,1))
     / sum(sample_matrix_y_x(:,2));
    %Using Sarndal et al., Page 179
    est_var_est_Taylor_ratio_y_x(sim_i) =
     (1/(mean(sample_matrix_y_x(:,2))^2))*((1-(na/N))/na)*
     ((sum((sample_matrix_y_x(:,1)-(est_Horwitz_ratio_y_x(sim_i)*
```

```
 sample_matrix_y_x(:,2))).^2))/(na-1));
%Note: SUDAAN 8.0 calculate the Taylor Method variance with other
%formulae (See the User's Manual at page 376)

%Calculation of the Jackknife1 method estimates
%---------------------------------------
%For the RATIO
m = 1;
AA = na/m;
Teta = est_Horwitz_ratio_y_x(sim_i);
pseudovalues1 = [];
for a = 1:AA
    x_without_sa = sample_matrix_y_x; %See Wolter pp. 173
    x_without_sa(SI_os(m,na),:) = [];
    pseudovalues1(a) = AA*Teta-(AA-1)*(Teta+((1-(na/N))^(1/2))*
     ((sum(x_without_sa(:,1)) / sum(x_without_sa(:,2)))-Teta));
end
est_Jackknife1_ratio_y_x(sim_i) = mean(pseudovalues1);
est_var_est_Jackknife1_ratio_y_x(sim_i) = var(pseudovalues1) / AA;
pseudovalues1 = []; %Cleaning up to free memory

%Calculation of the Jackknife2 method estimates
%---------------------------------------
%For the RATIO
m = 2;
AA = na/m;
Teta = est_Horwitz_ratio_y_x(sim_i);
pseudovalues2 = [];
for a = 1:AA
    x_without_sa = sample_matrix_y_x; %See Wolter pp. 173
    x_without_sa(SI_os(m,na),:) = [];
    pseudovalues2(a) =
     AA*Teta-(AA-1)*(Teta+((1-(na/N))^(1/2))*
     ((sum(x_without_sa(:,1)) / sum(x_without_sa(:,2)))-Teta));
end
est_Jackknife2_ratio_y_x(sim_i) = mean(pseudovalues2);
est_var_est_Jackknife2_ratio_y_x(sim_i) = var(pseudovalues2) / AA;
pseudovalues2 = []; %Cleaning up to free memory
x_without_sa = []; %Cleaning up to free memory

%Calculation of the Bootstrap Estimator of the Conditional Variance of the
%Resample Post-Design Vectors Estimator given a sample
%---------------------------------------
%For the RATIO
%Initialize a vector for the results of RATIOS of the A bootstrap resamples
bootratios = [];
%Calculating the post-design vector statistics of the A bootstrap resamples
for boot_i = 1:A
```

```
        %Using SI resample design
        %Here "SI_os(nbu,na)" creates index vector of an ordered
        %SI bootstrap resample
        boot_matrix_resample =
         [ sample_matrix_y_x(SI_os(nbu,na),:),[qfactor*ones(nbu-1,1);1]];

        %Calculating the RATIO of the expanded resample
        bootratios(boot_i) =
         (boot_matrix_resample(:,1)'*boot_matrix_resample(:,3))/
         (boot_matrix_resample(:,2)'*boot_matrix_resample(:,3));
    end

    boot_matrix_resample = []; %Cleaning up to free memory
    sample_matrix_y_x = []; %Cleaning up to free memory

    %Calculating the bootstrap estimation of the conditional variance of the
    %Resample Post-Design Vectors statistics given a sample
    est_boot_ratio_y_x(sim_i) = mean(bootratios);
    est_varcond_ratio_y_x_post(sim_i) = var(bootratios);
    est_var_est_boot_ratio_y_x(sim_i) = est_varcond_ratio_y_x_post(sim_i)/A;
    bootratios = []; %Cleaning up to free memory

    sim_i = sim_i + 1; %Simulation iteration counter update
    disp(sim_i)
end

%For the Output on the Command Window
N
na
population_ratio_y_x
var_est_Taylor_ratio_y_x
nbu
nb_artificial
qfactor
est_Horwitz_ratio_y_x
est_var_est_Taylor_ratio_y_x
est_Jackknife1_ratio_y_x
est_var_est_Jackknife1_ratio_y_x
est_Jackknife2_ratio_y_x
est_var_est_Jackknife2_ratio_y_x
est_boot_ratio_y_x
est_var_est_boot_ratio_y_x
est_varcond_ratio_y_x_post
nsim
conv_sim = conv_sim - 1
elapsed_time_mins = toc/60 %Reporting elapsed time in minutes

%For the Output on a file
```

```
fid = fopen('SIM_RATIO.txt','a');
fprintf(fid,'
N
na
population_ratio_y_x
var_est_Taylor_ratio_y_x
nbu
nb_artificial
qfactor
est_Horwitz_ratio_y_x
est_var_est_Taylor_ratio_y_x
est_Jack1_ratio_y_x
est_var_est_Jack1_ratio_y_x
est_Jack2_ratio_y_x
est_var_est_Jack2_ratio_y_x
est_boot_ratio_y_x
est_var_est_boot_ratio_y_x
est_varcond_ratio_y_x_post
nsim
conv_sim
elapsed_time_mins\n');
for i = 1:nsim
fprintf(fid,'%30.0f',N);
fprintf(fid,'%30.0f',na);
fprintf(fid,'%30.8f',population_ratio_y_x);
fprintf(fid,'%30.8f',var_est_Taylor_ratio_y_x);
fprintf(fid,'%30.0f',nbu);
fprintf(fid,'%30.4f',nb_artificial);
fprintf(fid,'%30.4f',qfactor);
fprintf(fid,'%30.8f',est_Horwitz_ratio_y_x(i));
fprintf(fid,'%30.8f',est_var_est_Taylor_ratio_y_x(i));
fprintf(fid,'%30.8f',est_Jackknife1_ratio_y_x(i));
fprintf(fid,'%30.8f',est_var_est_Jackknife1_ratio_y_x(i));
fprintf(fid,'%30.8f',est_Jackknife2_ratio_y_x(i));
fprintf(fid,'%30.8f',est_var_est_Jackknife2_ratio_y_x(i));
fprintf(fid,'%30.8f',est_boot_ratio_y_x(i));
fprintf(fid,'%30.8f',est_var_est_boot_ratio_y_x(i));
fprintf(fid,'%30.8f',est_varcond_ratio_y_x_post(i));
fprintf(fid,'%30.0f',nsim);
fprintf(fid,'%30.0f',conv_sim);
fprintf(fid,'%30.2f\n',elapsed_time_mins);
end
fclose(fid);
```

Con el Método Bootstrap Para el Estimador de la Varianza del Estimador de la Media, Mediana y Razón de los Totales de Dos Variables

```
%*****************************************************************************
%nsim times simulation of the extraction of a sample of size na from a population
%of size N
%
%SI sampling
%
%Bootstrap variance estimation methods
%
%For the MEAN, MEDIAN and RATIO
%-----------------------------------------------------------------------------
function SIM_SI_BOOT(population_matrix,pop_mat_column_y,
 pop_mat_column_x,nsim,na,A);

tic %Elapsed time start point

%Takes as population the "pop_mat_column_y" column of variable y and
%"pop_mat_column_x" column of variable x of the population matrix
population_matrix_y_x = [ population_matrix(:,pop_mat_column_y) ,
 population_matrix(:,pop_mat_column_x) ];

population_matrix = []; %Cleaning up to free memory
pop_mat_column_y = []; %Cleaning up to free memory
pop_mat_column_x = []; %Cleaning up to free memory

N = length(population_matrix_y_x); %Population size

%Direct calculation of the parameter and variance of the estimator
%--------------------------------------

%For the MEAN
population_mean_y = mean(population_matrix_y_x(:,1));
var_est_Horwitz_mean_y = (1-(na/N))/na*var(population_matrix_y_x(:,1));

%For the MEDIAN
%Creating a sorted version of population_vector_y
x = sortrows(population_matrix_y_x(:,1));
%Adding to the sorted population vector the cumulative distribution function
%F column of the population
x = [ x , cumsum(ones(N,1))/N ];
%Note: In finding the F's inverse at certain value "q" we took the maximum
%value "a" such that F(a)<=q
population_median_y = x(max(find(x(:,2) <= 0.50)),1);
c1=0.5-1.96*((N-na)/(N-1)*(1/na)*
 x(max(find(x(:,2)<=0.50)),2)*(1-x(max(find(x(:,2) <= 0.50)),2)) )^(1/2);
c2=0.5+1.96*((N-na)/(N-1)*(1/na)*
```

```
 x(max(find(x(:,2)<=0.50)),2)*(1-x(max(find(x(:,2) <= 0.50)),2)) )^(1/2);
F__c1 = x(max(find(x(:,2) <= c1)),1); %Finds the F's inverse at c1
F__c2 = x(max(find(x(:,2) <= c2)),1); %Finds the F's inverse at c2
var_est_Woodruff_median_y = ((F__c2 - F__c1)/(2*1.96))^2;

x = []; %Cleaning up to free memory
c1 = []; %Cleaning up to free memory
c2 = []; %Cleaning up to free memory
F__c1 = []; %Cleaning up to free memory
F__c2 = []; %Cleaning up to free memory

%For the RATIO
population_ratio_y_x = sum(population_matrix_y_x(:,1)) /
 sum(population_matrix_y_x(:,2));
var_est_Taylor_ratio_y_x = (1/(mean(population_matrix_y_x(:,2))^2))*
 ((1-na/N)/na)*((sum((population_matrix_y_x(:,1)-
 (population_ratio_y_x*population_matrix_y_x(:,2))).^2))/(N-1));

%Simulations
%--------------------------------------
sim_i = 1; %Initial value of the counter sim_i
%Initialize a vector for the results of MEANS of the simulations by Bootstrap
est_Bootstrap_mean_y = [];
%Initialize a vector for the results of MEANS of the simulations by Bootstrap
est_var_est_Bootstrap_mean_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Bootstrap
est_Bootstrap_median_y = [];
%Initialize a vector for the results of MEDIANS of the simulations by Bootstrap
est_var_est_Bootstrap_median_y = [];
%Initialize a vector for the results of RATIOS of the simulations by Bootstrap
est_Bootstrap_ratio_y_x = [];
%Initialize a vector for the results of RATIOS of the simulations by Bootstrap
est_var_est_Bootstrap_ratio_y_x = [];

%Simulation routines

%For the Construction of the pseudopopulation
integer_expansion_factor = floor(N/na);
remain_expansion_factor = rem(N,na);

while sim_i<=nsim
%Take a sample of size na from the population
    sample_matrix_y_x = population_matrix_y_x(SI_os(na,N),:);

    %--------------------------------------
    %
    %Bootstrap calculation of the estimated variance of the estimators
    %--------------------------------------
```

```
%Following the summary given in Sarndal et al. (1992). pp. 442

%Construction of the pseudopopulation
%Integer expansion of the sample
pseudopopulation = [];
for i = 1 : integer_expansion_factor
    pseudopopulation =  [ pseudopopulation ; sample_matrix_y_x ];
end
%Adding the remaining elements to the pseudopopulation to fullfil the
%original population size
pseudopopulation =
 [pseudopopulation;sample_matrix_y_x(SI_os(remain_expansion_factor,na),:)];
%Give a random order to the elements of the pseudopopulation
pseudopopulation =
 pseudopopulation(randperm(N),:);

%Extraction of the A bootstrap resamples and calculation of the
%estimates
%Using SI resampling
%Initialize a vector for the MEANS of the bootstrap resamples
bootmeans_y = [];
%Initialize a vector for the MEDIANS of the bootstrap resamples
bootmedians_y = [];
%Initialize a vector for the RATIOS of the bootstrap resamples
bootratios_y_x = [];
for boot_i = 1 : A
    %Take a bootstrap resample of size na from the pseudopopulation
    resample_y_x = pseudopopulation(SI_os(na,N),:);

    %For the MEAN of the y variable
    bootmeans_y(boot_i) = mean(resample_y_x(:,1));

    %For the RATIO of the y numerator variable and x denominator variable
    bootratios_y_x(boot_i)=(na*bootmeans_y(boot_i))/sum(resample_y_x(:,2));

    %For the MEDIAN of the y variable
    %Overwriting and creating a sorted version of sample vector
    %of variable y
    x = sortrows(resample_y_x(:,1));
    %Adding to the sorted sample vector the cumulative distribution
    %function F column of the sample
    x = [ x , cumsum(ones(na,1))/na ];
    bootmedians_y(boot_i)=x(max(find(x(:,2)<=0.50)),1); %Finding the MEDIAN

end

est_Bootstrap_mean_y(sim_i) = mean(bootmeans_y);
est_var_est_Bootstrap_mean_y(sim_i) = var(bootmeans_y);
```

```
    est_Bootstrap_median_y(sim_i) = mean(bootmedians_y);
    est_var_est_Bootstrap_median_y(sim_i) = var(bootmedians_y);
    est_Bootstrap_ratio_y_x(sim_i) = mean(bootratios_y_x);
    est_var_est_Bootstrap_ratio_y_x(sim_i) = var(bootratios_y_x);

    disp(sim_i) %Displaying on the command window the simulation iteration
    sim_i = sim_i + 1; %Simulation iteration counter update
end

%For the Output on the Command Window
N
na
population_mean_y
var_est_Horwitz_mean_y
population_median_y
var_est_Woodruff_median_y
population_ratio_y_x
var_est_Taylor_ratio_y_x
mean_est_BOOT_mean_y = mean(est_Bootstrap_mean_y)
mean_est_var_est_BOOT_mean_y = mean(est_var_est_Bootstrap_mean_y)
mean_est_BOOT_median_y = mean(est_Bootstrap_median_y)
mean_est_var_est_BOOT_median_y = mean(est_var_est_Bootstrap_median_y)
mean_est_BOOT_ratio_y_x = mean(est_Bootstrap_ratio_y_x)
mean_est_var_est_BOOT_ratio_y_x = mean(est_var_est_Bootstrap_ratio_y_x)
nsim
elapsed_time_mins = toc/60 %Reporting elapsed time in minutes

%For the Output on a file
fid = fopen('SIM_BOOT.txt','a');
fprintf(fid,'
N
na
population_mean_y
var_est_Horwitz_mean_y
population_median_y
var_est_Woodruff_median_y
population_ratio_y_x
var_est_Taylor_ratio_y_x
est_Bootstrap_mean_y
est_var_est_Bootstrap_mean_y
est_Bootstrap_median_y
est_var_est_Bootstrap_median_y
est_Bootstrap_ratio_y_x
est_var_est_Bstrap_ratio_y_x
nsim
elapsed_time_mins\n');
for i = 1:nsim
fprintf(fid,'%30.0f',N);
```

```
fprintf(fid,'%30.0f',na);
fprintf(fid,'%30.8f',population_mean_y);
fprintf(fid,'%30.8f',var_est_Horwitz_mean_y);
fprintf(fid,'%30.8f',population_median_y);
fprintf(fid,'%30.8f',var_est_Woodruff_median_y);
fprintf(fid,'%30.8f',population_ratio_y_x);
fprintf(fid,'%30.8f',var_est_Taylor_ratio_y_x);
fprintf(fid,'%30.8f',est_Bootstrap_mean_y(i));
fprintf(fid,'%30.8f',est_var_est_Bootstrap_mean_y(i));
fprintf(fid,'%30.8f',est_Bootstrap_median_y(i));
fprintf(fid,'%30.8f',est_var_est_Bootstrap_median_y(i));
fprintf(fid,'%30.8f',est_Bootstrap_ratio_y_x(i));
fprintf(fid,'%30.8f',est_var_est_Bootstrap_ratio_y_x(i));
fprintf(fid,'%30.0f',nsim);
fprintf(fid,'%30.2f\n',elapsed_time_mins);
end
fclose(fid);
```

Pequeña Subrutina para Muestras Ordenadas

```
%*****************************************************************************
%Obtain a WOR ordered sample of size n from a population of size N (SI design)
%Reporting the indexes
%-----------------------------------------------------------------------------
function y = SI_os(n,N);
y = sortrows([ [ 1 : 1 : N ]' , rand(N,1) ],2);
y = y(1:n,1);
```

Índice de figuras

Referencias

[1] Cassel, C. M., Särndal, C. E. & Wretman, J. (1977). *Foundations of Inference in Survey Sampling.* John Wiley & Sons, Inc.

[2] Contar 2000, *Sistema para la consulta de tabulados y bases de datos de la muestra.* XII Censo General de Población y Vivienda 2000. INEGI, México.

[3] Deming, W. E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *Journal of the American Statistical Association* **51**, 24-53.

[4] Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Statistics* **7**, 1-26.

[5] Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman & Hall.

[6] Kalton, G. & Heeringa, S. G. (eds.) (2003). *Leslie Kish: Selected Papers.* John Wiley & Sons, Inc.

[7] Kish, L. (1965). *Survey Sampling.* John Wiley & Sons, Inc.

[8] Kish, L. (1987). *Statistical Design for Research.* John Wiley & Sons, Inc.

[9] Kish, L. (1995). Methods for design effects. *Journal of Official Statistics* **11**, 55-77.

[10] Kish, L. & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B* **36**, 1-37.

[11] Lehtonen, R. & Pahkinen, E. J. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second edition. John Wiley & Sons, Ltd.

[12] Mahalanobis, P. C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhya* **4**, 511-531.

[13] Mahalanobis, P. C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London B* **231**, 329-451.

[14] Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**, 325-370.

[15] *MATLAB® 6.0 Help Documents.* (2000). The MathWorks, Inc. [http://www.mathworks.com]

[16] Mood, A. M., Graybill, F. A. & Boes, D. C. (1974). *Introduction to the Theory of Statistics.* Third edition. McGraw-Hill, Inc.

[17] Méndez, I., Eslava, G. & Romero, P. (2004). *Conceptos Básicos de Muestreo.* Monografías Vol. 12, No. 27. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Universidad Nacional Autónoma de México.

[18] Oetiker, T., Partl, H., Hyna, I. & Schlegl, E. (2005). *The Not So Short Introduction to LaTeX 2ε (Or LaTeX 2ε in 133 minutes.)* Version 4.16. [http://www.tex.ac.uk]

[19] Ollila, P. K. (2004). *A Theoretical Overview for Variance Estimation in Sampling Theory with Some New Techniques for Complex Estimators.* Research Reports No. 240. Statistics Finland.

[20] Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics* **20**, 355-375.

[21] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353-360.

[22] Särndal, C. E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag New York, Inc.

[23] Sitter, R. R. & Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters* **52**, 353-358.

[24] *SPSS® Base 13.0 User's Guide.* (2004). SPSS, Inc. [http://www.spss.com]

[25] *SUDAAN® User's Manual, Release 8.0.* Research Triangle Institute (2001). Research Triangle Park, North Carolina: Research Triangle Institute. [http://www.rti.org/sudaan]

[26] Traat, I. (2000). Sampling design as a multivariate distribution. In: *New Trends in Probability and Statistics* Vol. 5, T. Kollo, E. M. Tiit & M. Srivastava (eds.), 195-207.

[27] Traat, I., Bondesson, L. & Meister, K. (2000). *Distribution Theory for Sampling Designs.* Research Report No. 2. Department of Mathematical Statistics. Umeå University.

[28] Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* **47**, 635-646.

[29] Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag New York, Inc.