Bortolotto, Eugenia
Marí, Gonzalo
Instituto de investigaciones Teóricas y Aplicadas, Escuela de Estadística

ESTIMADORES ROBUSTOS DE TOTALES EN MUESTREO EN POBLACIONES FINITAS[1]

Resumen

Uno de los objetivos de las encuestas por muestreo es la estimación de parámetros de variables de interés. Una de las soluciones viene del lado de los estimadores clásicos que gozan de buenas propiedades distribucionales como por ejemplo el insesgamiento. El problema surge cuando en la encuesta se presentan en algunas variables, valores alejados del común de los datos. Ante esta situación, los estimadores clásicos presentan dificultades que se ven traducidas en un desempeño pobre respecto a medidas relacionadas a la precisión. Se presenta una serie de estimadores clásicos y sus versiones robustas para la estimación de totales poblacionales así como el estudio de sus propiedades a partir de simulaciones.

Palabras claves: muestreo en poblaciones finitas, estimador de Horvitz-Thompson, estimador de Hájek, estimadores robustos

Abstract

One of the objectives of sample surveys is the estimation of parameters of variables of interest. One solution is the use of the classical estimators that have good distributional properties such as, for example, unbiasedness. The problem arise when in a survey, values from some variables are far from the common data. Given this situation, the classical estimators present difficulties that are translated in a poor performance with respect to precision measures. We present classical estimators and their robust version for the estimation of population totals and the study of its property via simulation.

Keywords: sampling from finite populations, Horvitz-Thompson estimator, Hájek estimator, robust estimators

[1] Este trabajo se elaboró en el marco del Proyecto 1ECO199 Titulado “Métodos Estadísticos en el Ámbito Oficial”, dirigido por Gonzalo Marí

1. Introducción

Cuando se está interesado en conocer características de una determinada población, como totales, medias o proporciones, existen distintas formas de recolectar la información, pudiéndose mencionar entre las más importantes, a los censos y a las encuestas. Los primeros constituyen el método de recolección de datos más antiguo. Los mismos contemplan, en la mayoría de los casos, la enumeración completa de la población de interés, siendo los más conocidos los Censos de Población, Hogares y Viviendas, los Censos Agropecuarios, y los Censos Económicos, desarrollados en nuestro país por el Instituto Nacional de Estadística y Censos (INDEC) como entidad rectora del Sistema Estadístico Nacional (SEN). Debido al nivel de cobertura, los censos poseen la ventaja de obtener datos que permiten brindar información a niveles muy desagregados de la población en cuestión. Como desventaja, se puede mencionar que los mismos son operativos muy grandes, los cuales son costosos y muy difíciles de controlar, y como consecuencia de esta falta de control, brindan resultados que en ciertas situaciones pueden tener un nivel de precisión pobre.

La segunda fuente mencionada considera la recolección de datos a partir de encuestas por muestreo. A diferencia de los censos, las unidades sobre las cuales se recolectan los datos son un subconjunto de la población. El objetivo es, a partir de esta muestra, poder inferir a la misma. Existen dos tipos de muestras, las probabilísticas y las no probabilísticas. Las primeras son las que aseguran su representatividad y permiten realizar inferencias válidas para la población considerada. Esto se justifica a partir del hecho de asignar una probabilidad, a cada unidad de la población, no nula de ser seleccionada. Estas probabilidades son las que luego se utilizan en la etapa inferencial la cual debe contemplar el método de selección utilizado y diversas características como la no respuesta. En cambio, en el muestreo no probabilístico se desconoce la probabilidad de selección de las unidades, no se puede evaluar la precisión en términos probabilísticos y no garantiza la representatividad de las muestras sobre la población. En este trabajo, sólo se tienen en cuenta la recolección de datos a través de muestreos probabilísticos.

Entre los estimadores clásicos más utilizados para estimar medias y totales de las variables de interés se pueden mencionar el estimador de Horvitz-Thompson (Horvitz y Thompson, 1952) y el estimador de Hájek para el caso de la media. Con respecto a la estimación de la media, el estimador de Horvitz-Thomson se utiliza cuando el tamaño de la población sobre la cual se va a inferir es conocido, mientras que el estimador de Hájek se emplea generalmente cuando se desconoce esta cantidad.

Una de las dificultades que surgen en los datos recolectados en encuestas para una amplia gama de aplicaciones, es que los mismos contienen frecuentemente una o más observaciones atípicas llamadas outliers, que son observaciones que están separadas de la mayoría de los datos. En estos casos los estimadores clásicos de la media y el total, pueden estar muy influenciados por los outliers y no arrojar estimaciones precisas.

Una de las soluciones a este problema es la utilización de diseños muestrales que consideren información auxiliar que permita identificar estas unidades. Una opción es utilizar un muestreo estratificado, agrupando a la población de acuerdo al tamaño de los valores de la variable auxiliar correlacionado con la variable bajo estudio y relevando a todas las unidades del estrato que contiene a las observaciones de mayor tamaño. Sin embargo, algunas unidades con valores outliers en algunas variables aún pueden ser seleccionadas de forma inesperada en la muestra debido a información auxiliar poco precisa en el momento de extraer la muestra.

Una segunda solución al problema planteado proviene de la estadística robusta. La misma con-

templa un conjunto de técnicas y herramientas que resultan menos sensibles a la aparición de estas observaciones atípicas. Existe una serie de estimadores considerados robustos que son más apropiados que los estimadores clásicos ante la aparición de observaciones extremas. Entre sus características se puede mencionar que los mismos son generalmente no lineales, y precisan de la definición de términos constantes para su aplicación.

En el presente trabajo se considera la estimación del total poblacional de una variable. Para el mismo se presentan los estimadores clásicos de Horvitz-Thompson (Horvitz y Thompson, 1952) y de Hájek (Hájek, 1971). Luego, se introducirá una versión robusta del estimador de Horvitz-Thompson (HT) a través de M-estimadores, los cuales forman una clase de estimadores robustos simples y flexibles. Considerando al estimador HT como un funcional mínimo cuadrado, la robustificación del estimador HT se realiza en forma análoga a la realizada para los estimadores mínimo cuadrado en modelos lineales para poblaciones infinitas a través de los M-estimadores (Hampfel et al., 1986). Para este estimador, que se denomina Horvitz-Thompson Robusto (HTR), se consideran dos versiones del mismo, a un paso y M-estimador, a partir de la cantidad de iteraciones que se realizan.

En la sección 2 de este trabajo se detalla la teoría de los estimadores clásicos de Horvitz-Thompson y de Hájek. En la sección 3 se presentan los estimadores robustos mencionados anteriormente para la estimación de parámetros con sus correspondientes estimadores de variancia. En la sección 4 se muestran los resultados obtenidos por simulación. Por último en la sección 5 se brindan las conclusiones obtenidas del estudio por simulación y los estudios futuros que se plantean.

2. Estimadores Clásicos

Se presentan tres estimadores clásicos, que son los más difundidos en la actualidad para lo cual se presenta el escenario propuesto y sus supuestos ideales.

Sea una muestra s seleccionada a través de un diseño $p_d(.)$ de una población $U=\{u_1, u_2, \dots, u_N\} = \{1,2, \dots, N\}$. Se está interesado en estimar parámetros sobre la característica y, como el total poblacional $t = \sum_{k\epsilon U} y_k$, o la media poblacional $\bar{y}_U = \frac{t}{N} = \sum_U y_k/N$.

Una característica interesante que tienen las poblaciones finitas de N elementos, es que a los mismos se le pueden asignar distintas probabilidades de inclusión en la muestra. Sea $p_d(.)$ el diseño, la inclusión de un elemento k en la muestra es un evento aleatorio representado por una variable aleatoria I_k, definida como

$$I_k = \begin{cases} 1 & \text{si } k\epsilon S \\ 0 & \text{en otro caso} \end{cases}$$

En cualquier diseño muestral, π_k es la probabilidad de inclusión de primer orden, es decir, la probabilidad de que la unidad k-ésima sea incluida en una muestra y se obtiene como

$$\pi_k = P(k\epsilon S) = E(I_k) = P(I_k = 1) = \sum_{s \ni k} p(s)$$

donde $\sum s \ni k$ representa la suma de las probabilidades de todas las muestras que contienen la k-ésima unidad, y si el diseño es fijo tiene la propiedad que $\sum_{k=1}^{N} \pi_k = n$.

La probabilidad que las unidades k y l estén incluidas en la muestra se denota como π_{kl} y se obtiene del diseño $p_d(.)$ dado, de la siguiente manera

$$\pi_{kl} = P(k \wedge l \epsilon S) = E(I_k I_l) = P(I_k I_l = 1) = \sum_{s \ni k \& l} p(s)$$

Se tiene que $\pi_{kl} = \pi_{lk}$ para todo k y l. Además si $k = l$, entonces $\pi_{kk} = P({I^2}_k = 1) = P(I_k = 1) = \pi_k$.

Se asume que las $\pi_k > 0$, para todo $k \in U$, siendo esta condición necesaria para que el diseño muestral sea uno probabilístico. Otra propiedad importante de un diseño ocurre cuando $\pi_{kl} > 0$, para todo $k \neq l \in U$. Si la condición anterior y esta son satisfechas, el diseño muestral se dice medible, lo cual permite el cálculo de estimaciones de variancias e intervalos de confianza válidos basados en los datos observados.

Para cada elemento de $k \in s$ se tiene una ponderación $w_k = 1/\pi_k$ que refleja la probabilidad de inclusión en el diseño muestral.

Sea x_k una medida positiva no nula definida para todo $k \epsilon U$, la misma es utilizada para calcular las probabilidades de inclusión del diseño a emplear.

El estimador de Horvitz-Thompson del total poblacional es un estimador lineal que emplea las probabilidades de inclusión de primer orden (π_k). Para el caso de la media, este requiere conocer el tamaño de la población (N). Una alternativa para estimar parámetros poblacionales como la media cuando se desconoce N, es el estimador de Hájek. Este estimador es frecuentemente mejor que el estimador de Horvitz-Thompson, en el caso de que N sea conocido. El estimador de Hájek es no lineal, y por lo tanto aproximadamente insesgado, además se puede obtener una expresión aproximada para la variancia a través de linearización por series de Taylor.

2.1. Estimador de Horvitz-Thompson.

El estimador de Horvitz-Thompson surge al extender el concepto de Hansen y Hurwitz, quienes desarrollaron la teoría de muestreo con probabilidades proporcionales al tamaño con reemplazo, para muestras sin reemplazo de probabilidades desiguales. Horvitz y Thompson hicieron una contribución a la inferencia basada en diseños al formular tres clases de estimadores lineales y luego planteando la posibilidad que el mejor estimador (de mínima variancia) a través de todos los estimadores posibles lineales insesgados puede no existir para una muestra aleatoria simple.

El estimador de Horvitz-Thompson del total poblacional es

$$\hat{t}_{HT} = \sum_{S} \frac{y_k}{\pi_k}$$

mientras que la media poblacional es estimada a través del estimador

$$\bar{y}_\pi = \frac{\hat{t}_{HT}}{N} = \frac{1}{N} \sum_{k \epsilon S} \frac{y_k}{\pi_k}$$

Estos estimadores son insesgados para el total y la media poblacional respectivamente.

Si bien no forma parte en este trabajo el estudio de una medida del error muestral, se presen-

tan las formas correspondientes a la variancia y una estimación de la misma.

La variancia para el estimador del total es

$$V(\hat{t}_{HT}) = \sum\sum_{U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

La variancia para el estimador de la media es

$$V(\bar{y}_\pi) = \frac{V(\hat{t}_{HT})}{N^2} = \frac{1}{N^2}\sum\sum_{U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Siempre que $\pi_{kl} > 0$, para todo $k,l \in U$, un estimador insesgado de la variancia para el total es

$$\hat{V}(\hat{t}_{HT}) = \sum\sum_{S}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)y_k y_l$$

Y un estimador insesgado de la variancia para la media es

$$\hat{V}(\bar{y}_\pi) = \frac{1}{N^2}\sum\sum_{S}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)y_k y_l$$

2.2. Estimador de Hájek.

Una alternativa al estimador de Horvitz-Thompson para estimar la media poblacional $\bar{y}_U$ fue presentada por Hájek en 1971, quien desarrolló un estimador de la media en poblaciones finitas.

El estimador de Hájek para la media poblacional es

$$\tilde{y}_H = \frac{\hat{t}_{HT}}{\hat{N}} = \frac{\sum_S y_k/\pi_k}{\sum_S 1/\pi_k}$$

donde $\hat{N} = \sum_S\left(\frac{1}{\pi_k}\right)$ es el estimador HT de N. Se puede visualizar que este estimador es una razón entre dos estimadores para el total de y y z, donde $z_k = 1\ \forall k \in U$. Una aproximación de la variancia se obtiene aplicando el método de linearización por series de Taylor y está dada por

$$AV(\tilde{y}_H) = \frac{1}{N^2}\sum\sum_{U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k - \bar{y}_U}{\pi_k}\frac{y_l - \bar{y}_U}{\pi_l}$$

y un estimador de variancia es

$$\hat{V}(\tilde{y}_H) = \frac{1}{\hat{N}^2}\sum\sum_{S}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k - \tilde{y}_H}{\pi_k}\frac{y_l - \tilde{y}_H}{\pi_l}$$

Para algunos diseños los estimadores de Horvitz-Thompson y Hájek son idénticos, es decir, producen el mismo valor para todas las muestras como en los diseños simple al azar y muestreo estratificado simple al azar. Por ejemplo, para muestreo simple al azar, las probabilidades de inclusión son iguales a $\pi_k = n/N\ \forall k \epsilon U$ y por lo tanto

$$\tilde{y}_H = \frac{\hat{t}_{y\pi}}{\hat{N}} = \frac{\sum_S y_k/\pi_k}{\sum_S 1/\pi_k} = \frac{\sum_S y_k/\frac{n}{N}}{\sum_S 1/\frac{n}{N}} = \frac{\sum_S y_k/\frac{n}{N}}{\sum_S N/n} = \frac{\sum_S y_k/\frac{n}{N}}{N} = \frac{\hat{t}_{HT}}{N} = \bar{y}_\pi$$

Cuando el tamaño de la población es desconocido no es posible utilizar el estimador de Horvitz-Thompson para la media con lo cual se debe optar por el estimador de Hájek. Sin embargo si N es conocido, y los estimadores difieren, se debe elegir un estimador. Como ya se ha indicado es preferible usar el estimador de Hájek por tres motivos. En primer lugar, se ha visto que por la forma de la variancia de Hájek, éste es preferible cuando los valores $y_k - \tilde{y}_H$ son todos cercanos a cero.

En segundo lugar, tiene un mejor rendimiento en casos donde el diseño no es fijo, o sea, el tamaño de muestra es variable. Si el tamaño muestral es mayor que el promedio, la suma del numerador y la suma del denominador tendrán relativamente más términos. Análogamente, si el tamaño de la muestra es pequeño, ambas sumas tendrán pocos términos. La razón conserva de este modo una cierta estabilidad. Por el contrario el estimador de Horvitz-Thompson tiene denominador fijo y por lo tanto carece de esta propiedad.

Y finalmente en tercer lugar, en casos donde π_k no está correlacionado o lo está en forma negativa con los valores y_k. Si se supone que la muestra contiene un elemento con un gran valor de y_k pero un pequeño valor de π_k, la suma del numerador será muy grande. Sin embargo, será compensada hasta cierto punto por grandes valores de $1/\pi_k$ en el denominador. En este sentido, el estimador de Hájek es mejor que el de HT, donde el denominador de N permanece fijo.

En este trabajo se trabaja con los estimadores del total poblacional, por lo tanto se va a considerar un estimador que surge del estimador Hájek de la media y viene dado por $\hat{t}_H = N\tilde{y}_H$, con una variancia aproximada $AV(\hat{t}_H) = N^2 AV(\tilde{y}_H)$.

3. Estimadores Robustos

En muchas encuestas por muestreo, es común la aparición de valores que se alejan de la generalidad de los datos, denominados outliers. Los mismos pueden ser observaciones que se corresponden con valores observados y que resultan ser válidos. Estos valores afectan los estimadores tradicionales debido a que los mismos son sensibles ante la aparición de uno o más outliers. Una opción es la de ignorar esos valores. Desechar un valor outlier válido hará que los estimadores clásicos se vuelvan sesgados, pero mantenerlo con una ponderación completa hará al estimador altamente variable porque el valor o los valores atípicos se presentarán sólo en algunas de las muestras posibles.

Por otro lado el outlier puede ser una observación incorrecta, debido a mediciones o codificación errónea o derivada de una unidad fuera de la población objetivo. En este caso mantener el outlier con una ponderación completa puede implicar un gran sesgo sumado a una gran variabilidad para los estimadores clásicos. De este modo, descartar los valores atípicos incorrectos reduce tanto el sesgo como la variancia.

Ya que es frecuentemente difícil detectar outliers y decidir si estos son válidos o no, son necesarios los estimadores que se desempeñan bien en términos de sesgo y varianza con indepen-

dencia de la naturaleza y la detección de posibles valores atípicos. Se verán algunos de los estimadores robustos más difundidos desarrollados hasta la actualidad.

3.1. Estimador Horvitz-Thompson robusto.

Se supone que una medida positiva del tamaño x_k es conocida $\forall\, k \in U$ antes de sacar la muestra y que tiene correlación positiva con las variables de interés de la encuesta. Se denota con t_x el total poblacional de x_k. Sea la probabilidad de inclusión $\pi_k = nx_k/t_x\ \forall k \epsilon U$ y sea $w_k = 1/\pi_k$.

Como ya observamos en la sección 2, el estimador Horvitz-Thompson es $\hat{t}_{HT} = \sum_S \frac{y_k}{\pi_k} = \sum_S w_k y_k$. El modelo que inspira el estimador Horvitz-Thompson es $y_k = \beta x_k + \varepsilon_k$ con $E(\varepsilon_k) = 0$ y $Var(\varepsilon_k) = x_k\sigma^2$.

Al utilizar el estimador HT clásico, el razonamiento dado en la literatura de muestreo, es que tiene un error de muestreo o variancia igual a cero si las probabilidades de inclusión π_k son exactamente proporcionales a y_k. El estimador tendrá sesgo robusto pero no variancia robusta con respecto a desviaciones de la proporcionalidad entre π_k e y_k (Rao, 1966).

Por lo tanto, si se está interesado en formular para el estimador HT un estimador de variancia, se lo debe expresar como un funcional mínimo cuadrado de un estimador de la función de distribución poblacional de forma tal que el diseño sea incorporado en el estimador de la función de distribución poblacional mientras que la proporcionalidad entre y y x es tenida en cuenta por el funcional mínimo cuadrado.

La función de distribución poblacional conjunta de dos variables (y_k, x_k) es definida como

$$F_U(r,t) = \sum_{k\epsilon U} \mathbf{1}\{x_k \le r\}\mathbf{1}\{y_k \le t\}/N$$

donde

$$\mathbf{1}\{y_k \le t\} = \begin{cases} 1 & \text{si } y_k \le t \\ 0 & \text{en otro caso} \end{cases}$$

Existen varias posibilidades para la estimación de F_U, pero la más fácil y generalmente más utilizada es la función de distribución muestral.

$$F_S(r,t) = \sum_{k\epsilon S} \frac{1}{\pi_k}\mathbf{1}\{x_k \le r\}\mathbf{1}\{y_k \le t\} \Big/ \sum_{k\epsilon S} \frac{1}{\pi_k}$$

Para derivar un funcional mínimo cuadrado se considera el siguiente modelo de superpoblación para la proporcionalidad entre y_k y x_k, asumiendo que $\mathbf{y_U} = (y_1, \dots, y_N)$ es un vector de realizaciones de variables aleatorias independientes y_k con esperanza βx_k y variancia $\sigma^2 x_k$.

Se define el estimador mínimo cuadrado $\beta_{LS}(F_S)$ de β en el modelo anterior con respecto a la función de distribución muestral F_S de (x_k, y_k) $(k\epsilon S)$, el cual minimiza $\int \frac{(y-\beta x)^2}{x} dF_S(x,y)$ o, equivalentemente, es solución de

$$\sum_{k\epsilon S} \frac{1}{\pi_k}\left(\frac{y_k - \beta x_k}{\sqrt{x_k}}\right)\frac{x_k}{\sqrt{x_k}} = 0$$

Si s es una muestra obtenida de acuerdo a un diseño con probabilidad proporcional al tamaño (PPT), con probabilidades de inclusión $\pi_k = nx_k/\sum_{k\epsilon U} x_k$, luego el estimador HT es $\hat{t}_{HT} = t_x\beta_{LS}(F_S)$, donde $\beta_{LS}(F_S) = \frac{\sum_{k\epsilon U}\ y_k/\pi_k}{\sum_{k\epsilon U}\ x_k/\pi_k}$.

Se observa que la expresión $\hat{t}_{HT} = t_x\frac{\sum_{k\epsilon U}\ y_k/\pi_k}{\sum_{k\epsilon U}\ x_k/\pi_k}$ no depende del modelo de superpoblación. Sin embargo, en ese modelo la pendiente $\beta_{LS}(F_S)$ involucrada en el estimador HT es un estimador ponderado mínimo cuadrado que incorpora la información del diseño a través de F_S, así como la información en la variable auxiliar a través de la regresión.

Huber (1973) extiende los resultados del estimador robusto de parámetros de posición al caso de regresión lineal. Sea el modelo $y_k = \beta x_k + \varepsilon_k$ con $E(\varepsilon_k) = 0$ y $Var(\varepsilon_k) = \sigma^2$, el estimador mínimo cuadrado de β surge de minimizar la función $\sum_s((y_k - \beta x_k)/\sigma)^2$.

El mismo autor propone un estimador a partir de un método iterativo que considera estimadores mínimos cuadrados ponderados con pesos iguales a

$$\varphi_k = \min\{1, c/|r_k|\}$$

donde r_k es el k-ésimo residuo y c es una constante positiva. Estas ponderaciones no son fijas pero dependen del estimador. Más generalmente, Huber propuso definir al M-estimador $\hat{\beta}_M$ como

$$\Gamma\left(\hat{\beta}_M\right) = \min\{\Gamma(\beta)/\beta\epsilon B\}$$

donde

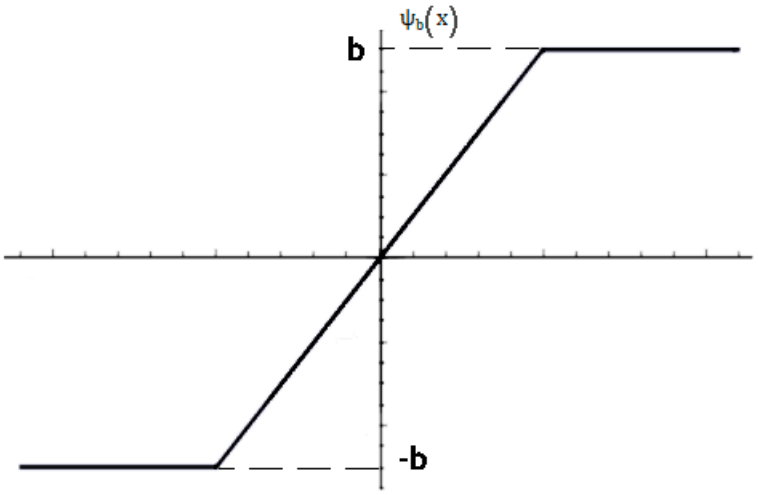
$$\Gamma(\beta) = \sum_{k=1}^{n}\rho\left(\frac{y_k - {x_k}^T\beta}{\sigma}\right)$$

para algunas funciones ρ reales y para un σ fijo. Si ρ tiene derivada $\frac{\partial}{\partial t}\rho(t) = \psi(t)$, $\hat{\beta}_M$ satisface el sistema de ecuaciones

$$\sum_{k=1}^{n}\psi\left(\frac{y_k - {x_k}^T\hat{\beta}_M}{\sigma}\right)x_k = 0$$

Si se elige ψ como la función de Huber (1964) que viene dada por

$$\psi_b(x) = \min\{b, \max\{x, -b\}\} = x\ \min\left(1, \frac{b}{|x|}\right)$$



para $0 < b < \infty$, se obtiene el estimador de Huber considerando los pesos φ_k mencionados anteriormente.

Luego de separar el diseño y la información auxiliar y expresarlo como un funcional mínimo cuadrado, la robustificación del estimador HT es análoga a la del estimador mínimo cuadrado en modelos lineales para una población finita a través del M-estimador.

La ecuación de estimación involucra algunas funciones η que dependen de los residuos estandarizados $(y_k - \beta x_k)/{x_k}^{1/2}$ y de x_k. Se define $x'_k = x_k/{x_k}^{1/2}$ y $r'_k(\beta) = (y_k - \beta x_k)/{x_k}^{1/2}$.

Sea $\beta(F_S, \eta)$ una solución de la ecuación

$$\sum_{k \epsilon S} \frac{1}{\pi_k} \eta\left(x'_k, r'_k(\beta)\right) x'_k = 0$$

El estimador HT robusto (HTR) viene dado por

$$\hat{t}_{HTR} = t_x \beta(F_S, \eta)$$

donde $\beta(F_S, \eta)$ es llamada la pendiente del estimador HT robusto.

En general η es de la forma $\eta(x, r) = v(x)\psi(r.u(x))$, donde $v(x)$ y $u(x)$ son dos funciones de ponderación y ψ es la función definida para el M-estimador de posición. Cuando $u(x) \equiv 1$, los estimadores resultantes le asignan un peso menor a los valores outliers de x y de los residuos en forma independiente.

Un caso particular es la función de Huber que se da cuando además los $v(x) \equiv 1$, ya que $\eta(x, r) = \psi_b(r) = \min(b, \max(r, -b))$ para alguna constante b. Si definimos la función $\eta(x, r) \equiv r$ $\forall\, x$, el estimador se transforma en el Horvitz-Thompson clásico. Al ajustar el término constante b en la función Huber los estimadores se vuelven más robustos.

El estimador HT robusto es un estimador no paramétrico. El modelo $E(y) = \beta x$ es simplemente utilizado para derivar la expresión del estimador HT como un funcional mínimo cuadrado. Ni el estimador HT ni el estimador HT robusto necesitan del modelo para ser aplicado, por lo tanto no es necesario comprobar los supuestos del mismo.

El estimador de Horvitz-Thompson robusto, asume que las ponderaciones no poseen valores outliers y por lo tanto sólo los residuos $(y_k - \beta x_k)\sqrt{x_k}$ deben robustificarse. Esto conduce a una nueva ecuación de estimación

$$\sum_{k=1}^{n} \frac{1}{\pi_k} \psi_b \left(\frac{y_k - \beta x_k}{\sqrt{x_k}} \right) \sqrt{x_k} = 0$$

donde ψ_b es la función Huber. La solución de esta ecuación es el estimador Horvitz-Thompson robusto. Este estimador puede ser expresado como uno ponderado, por lo tanto la solución puede ser obtenida con un algoritmo Reponderado Iterativo de Mínimos Cuadrados (IRLS, por sus siglas en inglés). Para el algoritmo IRLS se utiliza un valor inicial $\beta^{(0)} = \mathrm{med}(y_k, w_k)/\mathrm{med}(x_k, w_k)$, donde $\mathrm{med}(y_k, w_k)$ es la estimación de la mediana ponderada de y_k. Como paso siguiente se estima una desviación estándar robusta de los residuos estandarizados $r_k = \frac{(y_k - \beta^{(0)} x_k)}{\sqrt{x_k}}$. Una posibilidad viene dada por la desviación mediana absoluta (mad)

$$\hat{\sigma}_\varepsilon = \mathrm{mad}(r_k, w_k)$$

donde $\mathrm{mad}(r_k, w_k) = \mathrm{med}(|r_k - \mathrm{med}(r_k, w_k)|, w_k)$.

Luego, se obtiene una ponderación robusta u_k para cada observación a partir de

$$u_k = \frac{\psi_b \left(\frac{r_k}{\hat{\sigma}_\varepsilon} \right)}{\left| \frac{r_k}{\hat{\sigma}_\varepsilon} \right|}$$

La ponderación robusta toma el valor 1 para observaciones no outliers, mientras que para los valores atípicos es menor que 1 y puede ser 0 o cercano a 0 para outliers extremos. Una estimación de la pendiente del estimador Horvitz-Thompson robusto en la $(t+1)$-ésima iteración puede ser expresada como

$$\beta^{(t+1)} = \frac{\sum_S w_k u_k y_k}{\sum_S w_k u_k x_k}$$

donde u_k depende de $\beta^{(t)}$. El proceso iterativo es repetido hasta cumplirse los criterios de convergencia. Usando como covariable $x_k = 1 \ \forall k \epsilon U$, se obtiene un M-estimador que no tiene en cuenta la correlación entre las ponderaciones y la variable respuesta.

Si a este proceso de iteración solo lo realizamos una vez obtenemos el estimador Horvitz-Thompson robusto un paso que es igual a

$$\hat{t}_{HTR} = \frac{\sum_S w_k u_k y_k}{\sum_S u_k / n}$$

3.2. Estimador Hájek robusto.

El estimador HT robusto un paso para la media poblacional es igual a $\hat{\bar{y}}_{HTR} = \frac{1}{N} \frac{\sum_S w_k u_k y_k}{\sum_S u_k / n}$. Si se reemplaza N por $\sum_S w_k$, la formula corresponde al estimador Hájek robusto un paso de la media

$$\hat{\bar{y}}_{HR} = \frac{\sum_S w_k u_k y_k}{\sum_S w_k \sum_S u_k / n}$$

Y a partir del mismo se obtiene el estimador de Hájek robusto un paso del total $\hat{t}_{HR} = N\hat{\bar{y}}_{HR}$. Un razonamiento similar se realiza para obtener el M-estimador de Hájek robusto.

4. Estudio de propiedades de los estimadores a través de simulaciones

Para analizar cómo se comportan los distintos estimadores clásicos y robustos se lleva a cabo un estudio por simulación. El objetivo es investigar el desempeño de los estimadores robustos propuestos en términos de características propias de la distribución de los mismos como son la variancia, el sesgo y el ECM, para distintos escenarios. Para ello, se generaron seis poblaciones (Beaumont, Haziza y Ruiz-Gazen, 2013), cada una con una variable de interés, y, y una variable auxiliar, x, que tiene distribución Gamma con media 50 y variancia 500.

En las primeras tres poblaciones, de tamaños poblacionales 500, 1000 y 5000 respectivamente, se generan los valores de y considerando un modelo de razón

$$y_{1,k} = 2x_k + 3{,}7{x_k}^{1/2}\varepsilon_k$$

donde el término de error ε_k se generó de una distribución normal estándar. Estas poblaciones no contienen ningún valor outlier. En las otras tres poblaciones, también de tamaños 500, 1000 y 5000 respectivamente, los valores de y se generan de acuerdo a un modelo mixto

$$y_{2,k} = \tau_k\left(2x_k + 3{,}7{x_k}^{1/2}\varepsilon_k\right) + (1-\tau_k)z_k$$

donde los valores de z son generaciones independientes de una distribución normal con media 1200 y desvío estándar 200 y los $\tau_k's$ son generados en forma independiente de una distribución Bernoulli con probabilidad $p = 0.98$, o sea que estas poblaciones contienen aproximadamente 2% de valores outliers.

A continuación se presentan, a modo de ejemplo, los diagramas de dispersión entre la variable de interés y la variable auxiliar, para la población de tamaño 500, mientras que en el Anexo I se incluyen los gráficos correspondientes a las poblaciones de mayores tamaños.

Gráfico 1: Diagrama de dispersión entre la variable x y y_1 para tamaño poblacional 500.

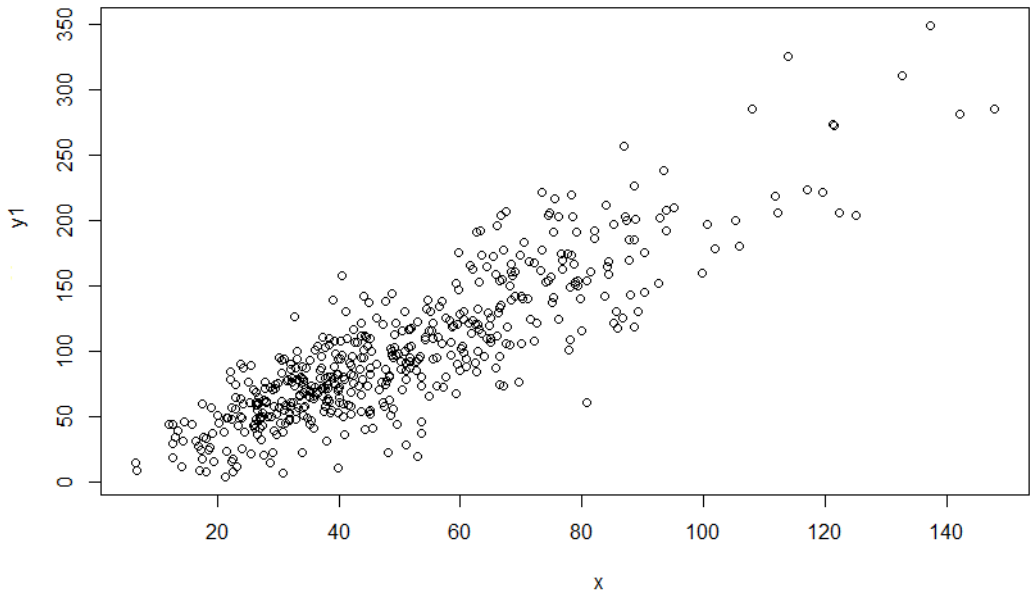
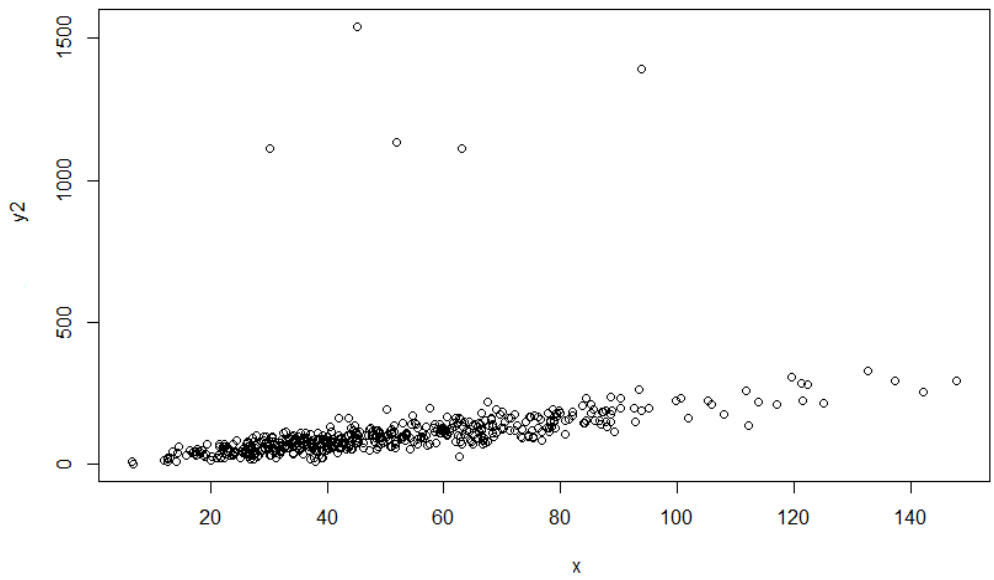


Gráfico 2: Diagrama de dispersión entre la variable x y y_2 para tamaño poblacional 500.



Para cada población se seleccionan $R = 1000$ muestras de acuerdo a un muestreo Poisson con probabilidad de inclusión π_k, proporcional a x_k , o sea $\pi_k = \tilde{n}x_k / \sum_{i \epsilon U} x_k$, donde $\tilde{n}$ representa la esperanza del tamaño muestral. Además se consideran dos fracciones de muestreo $f = \tilde{n}/N$,

0,02 y 0,10.

La elección del muestreo Poisson en el estudio de simulación se justifica por el hecho de facilitar la coordinación de muestras en encuestas de panel, situación que se presenta en la mayoría de encuestas económicas.

En cada muestra simulada se calcula los estimadores de Horvitz-Thompson, tanto clásicos como robustos, y los estimadores de Hájek, también en sus versiones clásicas y robustas. Para el cálculo de los estimadores clásicos se utilizó el paquete *survey* del programa estadístico R, mientras que para las versiones robustas se consideró el paquete *rhte* (Hulliger et al, 2011) que fue obtenido a partir de un pedido privado a uno de los autores dado que el mismo no está disponible en el *Comprehensive R Archive Network* (CRAN). Para las funciones utilizadas para el cálculo de los estimadores robustos, se utiliza la constante b de la función de Huber igual a 2.

Para finalizar se comparan los estimadores obtenidos en cada población. Esto se realiza a través de las características de la distribución de los estimadores::

- Variancia: $\frac{1}{R}\sum_{i=1}^{R}\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2$
- Sesgo relativo: $\frac{\frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i - \theta}{\theta} \times 100$
- Eficiencia relativa: $\frac{ECM}{ECM_{HT}}$

donde $\hat{\bar{\theta}} = \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i, ECM = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2$ y ECM_{HT} es el ECM correspondiente al estimador HT clásico.

En la Tabla 1 se presenta el resumen de los distintos escenarios considerados para el estudio por simulación.

Tabla1: Escenarios para el estudio por simulación.

Escenario	N	f	y
1	500	0,02	y_1
2	500	0,02	y_2
3	500	0,10	y_1
4	500	0,10	y_2
5	1000	0,02	y_1
6	1000	0,02	y_2
7	1000	0,10	y_1
8	1000	0,10	y_2
9	5000	0,02	y_1
10	5000	0,02	y_2
11	5000	0,10	y_1
12	5000	0,10	y_2

A continuación se presenta en la Tabla 2 los resultados correspondientes a los estimadores considerados en el escenario 1. En el Anexo II se incluyen las tablas II.1 a II.11 con los resultados obtenidos para las simulaciones de los escenarios 2 a 12 respectivamente.

Tabla 2: Resultados de la simulación para el escenario 1

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	252671792	1,1130	252729749	1,0000
HT robusto	282769360	-5,2727	289457942	1,1453
HT robusto 1 paso	297951662	-6,9445	309746920	1,2256
Hájek	104953023	2,6300	106582518	0,4217
Hájek robusto	116431948	-1,9314	117250925	0,4639
Hájek robusto 1 paso	124617561	-8,4617	142447419	0,5636

5. Conclusiones y Estudios Futuros

Se presentaron dos estimadores clásicos, el estimador HT y el estimador de Hájek, y sus versiones robustas a M y un paso para la estimación de parámetros de poblaciones finitas a partir de muestras seleccionadas en forma probabilística. Los primeros poseen el inconveniente de ser sensibles ante la aparición de valores atípicos, mientras que los estimadores robustos son sesgados.

Se realizó la evaluación de los estimadores clásicos y robustos a partir de simulaciones considerando diversos escenarios, donde varían los tamaños poblacionales, fracción de muestreo y cantidad de outliers en la variable de interés.

Luego de analizar los resultados obtenidos, se observa que para las poblaciones de tamaño 500, en los escenarios donde se trabajó con la variable y_1 que no contiene outliers la eficiencia relativa es menor en los estimadores clásicos comparado con sus versiones robustas. A su vez se observa que el estimador de Hájek arroja mejores resultados que el estimador HT, esto se debe a las propiedades que vimos anteriormente. Por otro lado, al analizar las poblaciones correspondientes a la variable y_2, los estadísticos robustos presentan eficiencias relativas menores a 1 lo cual implica un menor ECM para las versiones robustas. Para estos casos también el estimador de Hájek resulta mejor que el estimador HT.

A medida que aumenta el tamaño poblacional los estimadores robustos se vuelven más inestables. Cabe destacar que el comportamiento de los estimadores HTR y Hájek robusto son diferentes a medida que el tamaño de la población aumenta. En el primer caso, la ganancia respecto a los estimadores clásicos ante la presencia de outliers va disminuyendo, situación que referencia Beaumont et al (2013) en simulaciones con poblaciones similares. Para el estimador de Hájek, la situación es más extrema, llegándose a encontrar un comportamiento ineficiente de los estimadores robustos causado por un aumento del sesgo. La situación no presenta mejoras ante el cambio de la constante de robustez b, resultados que no son presentados en este

trabajo. Por otro lado, no se observan diferencias significativas entre los estimadores robustos M y un paso en ninguna de las poblaciones, si bien en todas las situaciones el estimador con más de una iteración se comporta levemente mejor que el que considera sólo un paso.

En estudios futuros se analizará el comportamiento de diversos estimadores tanto clásicos como robustos, como el estimador de razón y el estimador de Clark Winsorizado con distintos diseños muestrales y se estudiará el comportamiento del estimador de Hájek robusto en poblaciones grandes.

6. Referencias Bibliográficas

Beaumont, J.-F., Haziza, D., Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika,* 100 (3), 555-569.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., y Stahel, W.A. (1986). *Robust Statistics*. Ney York: Wiley.

Horvitz-D.G., y Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73-101.

Huber, P. J. (1973a). Robust regression: Asymptotics, conjectures, and Monte Carlo*. Annals of Statistics*, 1, 799-821.

Huber, P. J. (1973b). The use of Choquet capacities in statistics. *Proceedings of the 39th Session of the ISI*, Vol. 45, pp. 181-188.

Hulliger, B. (1995). Outlier Robust Horvitz-Thompson Estimators*. Survey Methodology*, 21, 79-87.

Hulliger, B. (1999). *Simple and Robust Estimators for Sampling*. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1999, 54-63

Hulliger, B. Alfons, A., Filzmoser, P., Meraner, A., Schoch, T., Templ, M. (2011) *R Programmes for Robust Procedures Including Manual.* AMELI Deliverable D4.1. AMELI Project.

Hulliger, B. Alfons, A., Filzmoser, P., Meraner, A., Schoch, T., Templ, M. (2011) *Robust Methodology for Laeken Indicators.* AMELI Deliverable D4.2. AMELI Project.

Lohr, S. (1999). *Sampling: Design and Analysis, 2nd Edition.* Cengage Lerning.

Lumley, T. (2010). *Complex Surveys: A guide to Analysis Using R*. New Jersey: Wiley & Sons.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). *Robust Statistics.* New York: John Wiley *&* Sons.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya A*, 28, 47-60.

Särdnal, C.E., Swensson, B., Wretman, J. (2003). *Model Assisted Survey Sampling.* Spring-er & Verlag.

ANEXO I

Gráfico I.1: Diagrama de dispersión entre la variable x y y_1 para tamaño poblacional 1000.

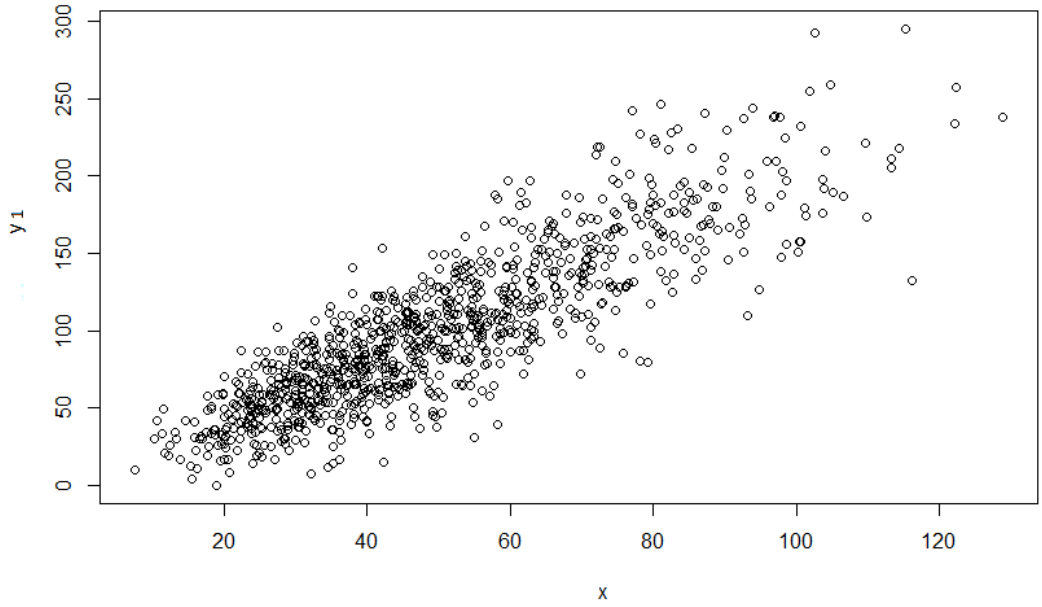


Gráfico I.2: Diagrama de dispersión entre la variable x y y_1 para tamaño poblacional 5000.

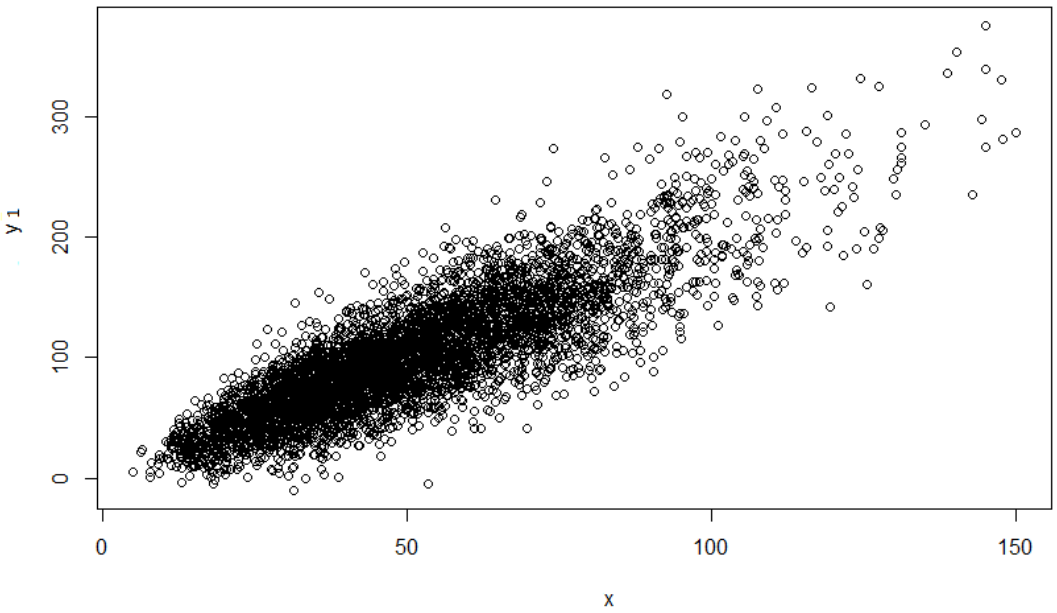


Gráfico I.3: Diagrama de dispersión entre la variable x y y_2 para tamaño poblacional 1000.

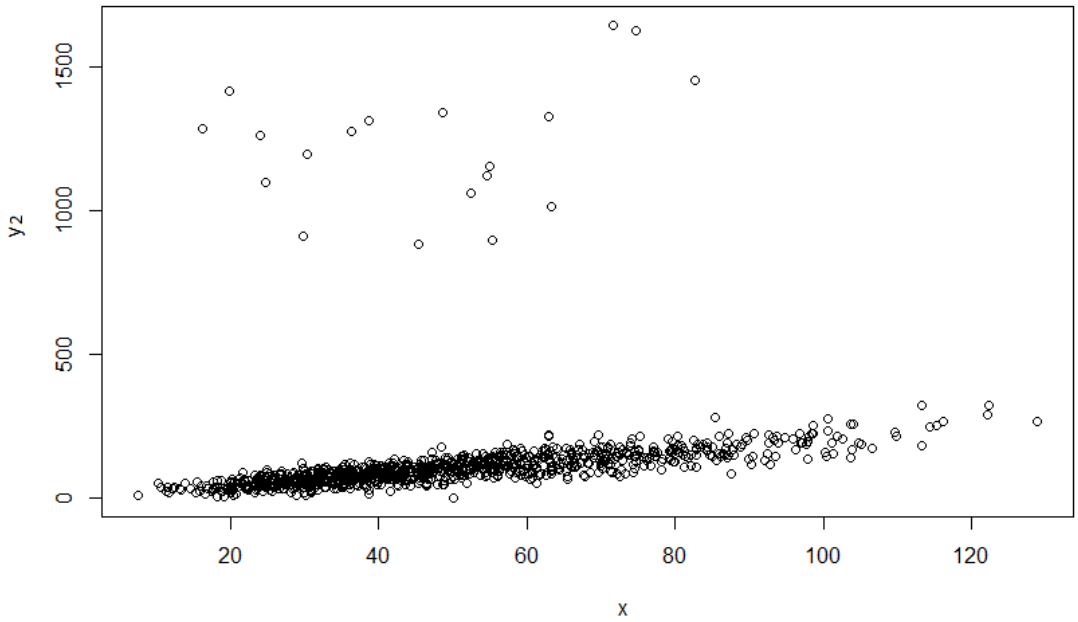
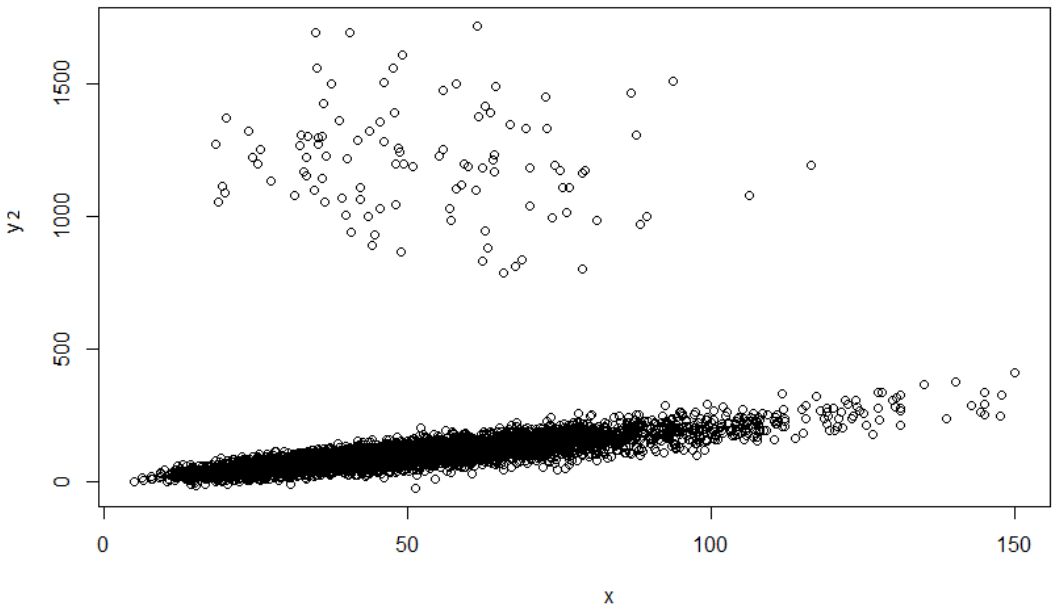


Gráfico I.4: Diagrama de dispersión entre la variable x y y_2 para tamaño poblacional 5000.



Facultad de Ciencias Económicas y Estadística

ANEXO II

Tabla II.1: Resultados de la simulación para el escenario 2

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	643111061	1,9446	643637375	1,0000
HT robusto	527811496	-6,3580	539785165	0,8386
HT robusto 1 paso	535903969	-7,9174	554754118	0,8619
Hájek	509262098	4,0350	513787839	0,7983
Hájek robus-to	160858548	-9,6266	189356720	0,2942
Hájek robus-to 1 paso	122623604	-15,0557	192602034	0,2992

Tabla II.2: Resultados de la simulación para el escenario 3

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	47525860,4	0,3989	47518234,7	1,0000
HT robusto	47937949,5	-1,2823	48302345,2	1,0165
HT robusto 1 paso	47890839,5	-1,3762	48317876,7	1,0168
Hájek	15424340,4	0,5065	15473238,7	0,3256
Hájek robusto	16691956,3	-1,9535	17632156	0,3711
Hájek robusto 1 paso	17498720,9	-3,6765	20870630,2	0,4392

Tabla II.3: Resultados de la simulación para el escenario 4

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	116258633	-0,4531	116205857	1,0000
HT robusto	89731059	-3,8519	94229927,4	0,8109
HT robusto 1 paso	89636106,6	-3,9033	94258176,1	0,8111
Hájek	81964466,9	0,6125	81998523,6	0,7056
Hájek robusto	17447168,8	-11,0826	55413870,1	0,4769
Hájek robusto 1 paso	17944834,7	-12,2673	64465740,9	0,5548

Tabla II.4: Resultados de la simulación para el escenario 5

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	497280584	0,7147	497290769	1,0000
HT robusto	506491066	-1,8027	509212772	1,0240
HT robusto 1 paso	509239591	-2,0578	512936679	1,0315
Hájek	145284451	1,4881	147338773	0,2963
Hájek robusto	158725448	-0,4234	158744835	0,3192
Hájek robusto 1 paso	186758865	-3,1292	196299048	0,3947

Tabla II.5: Resultados de la simulación para el escenario 6

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	2595337101	1,9917	2598589906	1,0000
HT robusto	2246031994	-2,6132	2253852581	0,8673
HT robusto 1 paso	2284704314	-2,5523	2292022926	0,8820
Hájek	1828598566	1,7349	1831207185	0,7047
Hájek robusto	223331003	-16,5353	626171354	0,2410
Hájek robusto 1 paso	210170074	-18,6872	724760309	0,2789

Tabla II.6: Resultados de la simulación para el escenario 7

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	84064249,7	0,2231	84029647,3	1,0000
HT robusto	82933536,7	-0,9225	83695878,1	0,9960
HT robusto 1 paso	82785810,6	-0,9009	83509201,3	0,9938
Hájek	28110017	0,4113	28249963,1	0,3362
Hájek robusto	30067473	-0,8434	30743993,1	0,3659
Hájek robusto 1 paso	32150974,4	-1,6502	34823952,5	0,4144

Tabla II.7: Resultados de la simulación para el escenario 8

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	423600235	-0,9329	424459711	1,0000
HT robusto	355224179	-3,7266	375341697	0,8843
HT robusto 1 paso	356421227	-3,7216	376482469	0,8870
Hájek	309095966	-0,3664	308984789	0,7279
Hájek robusto	30321266,9	-16,9042	451537808	1,0638
Hájek robusto 1 paso	30826462,9	-17,5158	483078846	1,1381

Tabla II.8: Resultados de la simulación para el escenario 9

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	2734700640	0,6298	2741871755	1,0000
HT robusto	2690521976	-0,4198	2692232684	0,9819
HT robusto 1 paso	2704722741	-0,4312	2706661442	0,9872
Hájek	764095082	0,0833	763504207	0,2785
Hájek robusto	770225590	-1,7660	847339877	0,3090
Hájek robusto 1 paso	778587801	-2,3195	912155168	0,3327

Tabla II.9: Resultados de la simulación para el escenario 10

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	11615248702	0,9465	11637526452	1,0000
HT robusto	10230116954	-1,8058	10343251454	0,8888
HT robusto 1 paso	10237273154	-1,8343	10354324759	0,8897
Hájek	8352281489	0,4952	8353206973	0,7178
Hájek robusto	981580271,1	-17,8692	13060635038	1,1223
Hájek robusto 1 paso	956808617,6	-18,5311	13947334949	1,1985

Tabla II.10: Resultados de la simulación para el escenario 11

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	503723644	-0,1009	503474148	1,0000
HT robusto	492976922	-0,9497	515005216	1,0229
HT robusto 1 paso	493014540	-0,9366	514425669	1,0218
Hájek	154833904	0,0600	154768980	0,3074
Hájek robusto	158329075	-1,6832	228923747	0,4547
Hájek robusto 1 paso	160201015	-2,0428	264252457	0,5249

Tabla II.11: Resultados de la simulación para el escenario 12

Estimadores	Variancia	Sesgo relativo	ECM	Eficiencia relativa
HT	2080901885	0,0041	2078821610	1,0000
HT robusto	1839015454	-2,5275	2078849480	1,0000
HT robusto 1 paso	1840132744	-2,5258	2079637864	1,0004
Hájek	1537229063	0,0314	1535729068	0,7387
Hájek robusto	170800291	-17,9492	1,2359E+10	5,9452
Hájek robusto 1 paso	172167832	-18,4528	1,3054E+10	6,2795