

Ajuste de los factores de expansión ante la presencia de observaciones atípicas

José Elías Rodríguez-Muñoz

Universidad de Guanajuato

Profesor Titular

+52 4737320006 ext. 1208

elias.rodriguez@ugto.mx

Luis Fernando Contreras-Cruz

Universidad Autónoma Chapingo

Profesor

lufecon@gmail.com

José Elías Rodríguez Muñoz es doctor en Ciencias con orientación en Probabilidad y Estadística por el Centro de Investigación en Matemáticas. Actualmente, es profesor del Departamento de Matemáticas de la Universidad de Guanajuato. Adicionalmente, en esta universidad ha desempeñado diversos cargos directivos. Ha sido consultor estadístico en proyectos para el sector público y privado. Su principal área de investigación es el Muestreo Estadístico de Poblaciones Finitas.

Luis Fernando Contreras Cruz es doctor en Ciencias en Estadística por el Colegio de Posgraduados. Actualmente, es profesor del Departamento de Fitotecnia de la Universidad Autónoma Chapingo. Su principal área de investigación es el Muestreo Estadístico de Poblaciones Finitas.

Resumen

En este trabajo se desarrolla una metodología de estimación robusta de parámetros poblacionales ante la presencia de observaciones atípicas y utilizando datos de encuestas. Esta metodología mejora la precisión, en términos del error cuadrático medio relativo, de las estimaciones con respecto a los métodos de estimación no robusta. Además, se presenta un experimento de simulación que muestra evidencia empírica del desempeño del estimador

propuesto. Finalmente se aplica la metodología propuesta a datos de una encuesta real para mostrar la factibilidad de su aplicabilidad.

Palabras clave

Datos atípicos, factores de expansión, muestreo de poblaciones finitas, estimación robusta

Abstract

In this work, an outlier-robust methodology to estimate population parameters using survey data is developed. This methodology improves accuracy regarding the relative mean square error of the estimator on non-robust estimation methods. Furthermore, a simulation experiment showing empirical evidence of performance of the proposed estimator is presented. Finally, the proposed methodology to a true survey data is applied to show its feasibility.

Keywords

Outliers, expansion factors, finite population sampling, robust estimation

Reconocimientos

La actividad de investigación del presente trabajo fue financiada por el Fondo Sectorial CONACYT-INEGI, México, con el convenio número 187564.

1. Introducción

Las observaciones atípicas se encuentran frecuentemente en el análisis de encuestas por muestreo. Las oficinas nacionales de estadística se enfrentan a este tipo de observaciones principalmente cuando las encuestas recaban información sobre variables económicas como ingresos, gastos y producción. Este tipo de variables tienen un comportamiento poblacional marcadamente sesgado por dicho tipo de observaciones.

Por observaciones atípicas nos referimos a los datos de la variable de interés cuyos valores están alejados significativamente del valor esperado, en un enfoque basado en modelos, o están alejados significativamente de la media poblacional, en un enfoque basado en diseños.

El impacto de las observaciones atípicas en la estimación de parámetros poblacionales, utilizando información de una encuesta, puede ser significativo. Esto es, estas observaciones atípicas en una muestra pueden dominar por completo el valor de la estimación resultante. Por ejemplo, las observaciones atípicas en una muestra pueden contribuir con un porcentaje alto en el valor de la estimación del total de la variable de interés. La presencia de valores atípicos grandes en la muestra produce valores altos en la estimación de promedios o totales aun cuando se utilice un estimador insesgado. Más aun, cuando se utiliza un estimador insesgado, la influencia de las observaciones atípicas es sobre la varianza de dicho estimador. Por lo tanto, métodos robustos de estimación deben utilizarse ante la presencia de observaciones atípicas con la finalidad de obtener estimadores más precisos de los parámetros poblacionales de interés. En particular, la ganancia en precisión de dichos métodos puede reflejarse en la disminución del error cuadrático medio, si se utilizaran estimadores sesgados.

Existe una extensiva literatura concerniente al problema de observaciones atípicas en muestreo de poblaciones finitas. Los primeros en considerar este problema en esta área fueron Kish (1965) y Searl (1966). Después importantes contribuciones fueron hechas por Hitirouglou y Srinath (1981) y Chambers (1982, 1986). Posteriormente este tema se puede encontrar en Chambers y Kokic (1993), Kokic y Bell (1994), Lee (1991, 1995), Hulliger (1995), Welsh y Rochetti (1998) y Duchesne (1999). La mayoría de los enfoques para tratar con observaciones atípicas en muestreo de poblaciones finitas se pueden agrupar en dos clases: i) el enfoque de factores de expansión modificados, donde se calibran los factores de expansión asociados con las observaciones atípicas y ii) el enfoque de valores modificados, donde los valores de la característica de interés asociados a las observaciones atípicas se modifican y se dejan intactos los factores de expansión.

Los dos enfoques tienen el mismo objetivo de reducir el impacto de las observaciones atípicas sobre las estimaciones. Sin embargo, estos pueden producir resultados diferentes en las propiedades de los respectivos estimadores. Por lo tanto, la elección de un enfoque dependerá del énfasis que se le dé a la reducción del sesgo o la reducción de la varianza. De acuerdo a

nuestra experiencia y conocimiento sobre el tema, no existe suficiente evidencia teórica para guiar a los usuarios en la elección de un estimador apropiado entre las posibles alternativas. Otro problema es que la mayoría de los estimadores robustos ante observaciones atípicas están diseñados para el muestreo aleatorio simple sin reemplazo y la generalización a diseños de muestreo más complejos no es inmediata.

En el presente trabajo se propone una metodología de estimación, basada principalmente en el artículo de Ren y Chambers (2002), que modifica los factores de expansión. Para esto, en la Sección 2 se especifica el concepto de valor atípico utilizado en este trabajo. Adicionalmente, se describe cómo valores de este tipo pueden influir en la varianza de un estimador. Después de esto, en la Sección 3 se expone el método para ajustar los factores de expansión ante la presencia de valores atípicos y el estimador resultante. Posteriormente, en la Sección 4 se presenta el estudio de simulación que pretende mostrar la bondad de la metodología propuesta en este trabajo. En la Sección 5 se muestra la aplicación de la metodología a la ENVIPE 2011. Por último, se exponen los comentarios finales en la Sección 6.

2. Valores atípicos en muestreo de poblaciones finitas

Un valor atípico a nivel poblacional de la variable de interés es un valor grande, o pequeño, con respecto a los valores correspondientes de la mayoría de los individuos de la población. Para el presente trabajo, los datos de una muestra etiquetados como atípicos son datos encontrados como tales en la población y no se consideran errores de recolección de la información.

En particular el efecto de los valores atípicos se refleja en la varianza poblacional:

$$\sigma^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu)^2, \tag{1}$$

donde U representa a la población de estudio de tamaño N , con media μ y los valores de la variable de interés son representados por y . Los datos atípicos, como se consideran aquí,

serán los valores más alejados de la media poblacional. Por lo tanto, éstos influyen más en el valor de la varianza como es de esperar.

Ahora para discutir cómo influyen los datos atípicos a nivel de una muestra, primero supóngase que se desea estimar el total poblacional:

$$t = \sum_{k \in U} y_k \tag{2}$$

con el estimador de Horvitz-Thompson, abreviado como HT,:

$$\hat{t}_{HT} = \sum_{k \in U} y_k \frac{S_k}{\pi_k}, \tag{3}$$

donde

$$S_k = \begin{cases} 1 & \text{si el } k\text{-ésimo individuo está en la muestra;} \\ 0 & \text{en otro caso} \end{cases}$$

y se le puede denominar la k -ésima coordenada del vector muestra $S = (S_1, \dots, S_N)$. Adicionalmente, π_k es la probabilidad de inclusión de primer orden. Así, para muestras de tamaño fijo, la varianza del anterior estimador es:

$$\sum_{j \in U} \sum_{k > j \in U} (\pi_j \pi_k - \pi_{jk}) \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right)^2,$$

donde π_{jk} representa a la probabilidad de inclusión de segundo orden. De aquí podemos observar que, si las π son aproximadamente proporcionales a los valores de la característica de interés, entonces la varianza del anterior estimador puede ser relativamente pequeña.

Esto nos indica en particular que, si el dato discordante tiene un valor relativamente alto, entonces debería tener una probabilidad de inclusión cercana a uno. En contraposición, si dicho valor fuera relativamente pequeño, su respectiva probabilidad de inclusión debería ser cercana a cero. Sin embargo, en una encuesta real es difícil controlar esta relación entre los valores de la variable de interés y las probabilidades de inclusión de primer orden. El no tener

control sobre la proporcionalidad mencionada puede inflar considerablemente la varianza del estimador del total.

En la siguiente sección se presenta una propuesta metodológica para ajustar los factores de expansión ante la presencia de valores etiquetados como atípicos.

3. Método para ajustar los factores de expansión

Es conveniente aclarar que para efectos de este trabajo los valores de la característica de interés se consideran escalares numéricos. En este mismo orden de ideas, la subpoblación con valores atípicos o discordantes de la variable de interés se denota por U_2 y la subpoblación con valores concordantes se denota por $U_1 = U \setminus U_2$.

Con esta distinción entre la subpoblación con valores atípicos y la de valores concordantes, el total t en (2) se puede reexpresar como:

$$t = \sum_{j \in U_1} y_j + \sum_{k \in U_2} y_k. \tag{4}$$

Obsérvese también que el vector muestra se puede ahora expresar como $S = S_1 + S_2$, donde las correspondientes coordenadas de S_1 y S_2 están dadas por:

$$S_{jk} = \begin{cases} S_k & \text{si } k \in U_j \\ 0 & \text{en otro caso,} \end{cases} \text{ para } j = 1, 2.$$

Por lo tanto, el estimador del total en (3) se puede reexpresar como:

$$\hat{t} = \sum_{j \in U_1} y_j \frac{S_{1j}}{\pi_j} + \sum_{k \in U_2} y_k \frac{S_{2k}}{\pi_k}.$$

Ahora, el punto de partida de la metodología de la presente propuesta es corregir la varianza en (1) de tal forma que disminuya la influencia de los valores atípicos en ésta, Ren y Chambers (2002). Una forma de hacer lo anterior es definir la varianza corregida de la población como el mínimo valor de

$$\sigma^{*2} = \frac{1}{N-1} \left(\sum_{j \in U_1} (g(\lambda)y_j - \mu)^2 + \sum_{k \in U_2} (\lambda y_k - \mu)^2 \right), \quad (5)$$

con respecto al parámetro λ y tal que

$$t = \sum_{j \in U_1} g(\lambda)y_j + \sum_{k \in U_2} \lambda y_k.$$

Cabe hacer notar que esta restricción garantiza que el total poblacional no se modifica con la inclusión de este último parámetro. Ahora, de la restricción anterior se obtiene que:

$$t - \sum_{j \in U_1} g(\lambda)y_j + \sum_{k \in U_2} \lambda y_k = 0$$

y de aquí se puede obtener la expresión para $g(\lambda)$ como:

$$\begin{aligned} g(\lambda) &= \frac{\sum_{k \in U_2} y_k}{\sum_{j \in U_1} y_j} (1 - \lambda) + 1 \\ &= \delta(1 - \lambda) + 1, \end{aligned}$$

con

$$\delta = \frac{\sum_{k \in U_2} y_k}{\sum_{j \in U_1} y_j}. \quad (6)$$

Así, se puede deducir el valor de λ para el cual la varianza corregida en (5) es mínima:

$$\lambda_{opt} = \frac{\delta(\delta + 1) \sum_{j \in U_1} y_j^2}{\delta^2 \sum_{j \in U_1} y_j^2 + \sum_{k \in U_2} y_k^2}.$$

Obsérvese que si

$$\frac{\sum_{k \in U_2} y_k}{\sum_{j \in U_1} y_j} < \frac{\sum_{k \in U_2} y_k^2}{\sum_{j \in U_1} y_j^2},$$

entonces $\lambda_{opt} < 1$ y se tendrá un factor de reducción para los valores atípicos.

Obsérvese también que si se tiene un estimador $\hat{\delta}$ de δ , entonces se puede construir un estimador de λ_{opt} como:

$$\hat{\lambda}_{opt} = \frac{\hat{\delta}(\hat{\delta} + 1) \sum_{j \in U_1} y_j^2 \frac{S_{1j}}{\pi_j}}{\hat{\delta}^2 \sum_{j \in U_1} y_j^2 \frac{S_{1j}}{\pi_j} + \sum_{k \in U_2} y_k^2 \frac{S_{2k}}{\pi_k}}.$$

Así, con los anteriores estimadores de δ y λ_{opt} , se obtiene un estimador del total t en (2), robusto ante valores atípicos y expresado como:

$$\hat{t}_{RA} = (\hat{\delta}(1 - \hat{\lambda}_{opt}) + 1) \sum_{j \in U_1} y_j \frac{S_{1j}}{\pi_j} + \hat{\lambda}_{opt} \sum_{k \in U_2} y_k \frac{S_{2j}}{\pi_k}, \quad (7)$$

donde $\hat{\delta}(1 - \hat{\lambda}_{opt}) + 1$ es el factor de corrección para los factores de expansión $1/\pi_j$, para $j \in U_1$, y $\hat{\lambda}_{opt}$ es el factor de corrección para $1/\pi_k$, $k \in U_2$. Sin embargo, este estimador no necesariamente será insesgado. Adicionalmente, obsérvese que el tamaño de la muestra debe ser más grande que el tamaño de la subpoblación con valores atípicos para que el estimador \hat{t}_{RA} esté bien definido.

Además, si se utiliza para δ el estimador

$$\tilde{\delta} = \frac{\sum_{k \in U_2} y_k \frac{S_{2k}}{\pi_k}}{\sum_{j \in U_1} y_j \frac{S_{1j}}{\pi_j}},$$

sugerido por la expresión (6), el estimador \hat{t}_{RA} se reduce al estimador de Horvitz-Thompson en (3). Por tal motivo Ren y Chambers (2002) sugirieron utilizar el estimador:

$$\hat{\delta} = \left(\frac{\hat{M}_2 \left[\sum_{k \in U_2} \frac{1}{\pi_k} \right] \left[\sum_{k \in U_2} y_k \frac{S_{2k}}{\pi_k} \right]}{\hat{M}_1 \left[\sum_{j \in U_1} \frac{1}{\pi_j} \right] \left[\sum_{j \in U_1} y_j \frac{S_{1j}}{\pi_j} \right]} \right)^{1/2},$$

del cual se ha mostrado empíricamente que tiene mayor estabilidad. En esta expresión, \hat{M}_1 es la mediana de los valores observados en la muestra de individuos con valores concordantes y \hat{M}_2 es la mediana de los valores atípicos en la muestra.

En este mismo orden de ideas, obsérvese que si la estimación de $\hat{\delta}$ es cero, lo cual sucede si no hay valores atípicos en la muestra observada, entonces la estimación del total t con \hat{t}_{RA} se reduce al valor que produciría el estimador de Horvitz-Thompson, como es de esperar.

Los experimentos de simulación realizados, los cuales se presentan en la siguiente sección, mostraron evidencia empírica sobre la reducción en el error cuadrático medio relativo y consecuentemente un aumento en la precisión del estimador propuesto \hat{t}_{RA} .

4. Estudio de Simulación

Con el objetivo de obtener evidencia empírica sobre el desempeño del estimador propuesto en (7), se realizaron los siguientes experimentos de simulación.

1. Se construyó una población artificial de tamaño $N = 10,000$ individuos, de tal forma que contuviera 0.5% de valores atípicos. Esto corresponde a $N_0 = 50$ valores atípicos. Más adelante se considerarán muestras de tamaño 200, si estas muestras se seleccionaran por muestreo aleatorio simple sin reemplazo, entonces la probabilidad de que una muestra contenga uno o más datos atípicos es aproximadamente 0.6367 con el tamaño de población y número de valores atípicos aquí establecidos. Obsérvese que el número de valores atípicos en una muestra, por el diseño de muestreo mencionado, se comporta como una distribución Hipergeométrica.
 - 1.1. Los valores concordantes fueron los percentiles de una distribución uniforme en $[0,100]$.
 - 1.2. Para los valores atípicos, se consideraron dos escenarios. Uno de éstos estuvo constituido por los percentiles de una distribución uniforme en $[195,205]$, para tener valores atípicos 100% aproximadamente por arriba del valor mayor de los valores concordantes. El otro escenario estuvo constituido por los percentiles de una uniforme en $[495,505]$ para tener valores atípicos 500% por arriba del máximo valor de los concordantes. A la población con los valores atípicos en el primer escenario se le denominará A1 y a la otra A5. Los valores generados de esta manera se pueden visualizar en la Gráfica 1.

[Gráfica 1, aquí]

2. Para el diseño de muestreo que se mencionará en el punto 3, se necesita generar los tamaños de las unidades de muestreo. Estos tamaños se generaron de tal forma que tuvieran una correlación de 0.75 con los valores concordantes generados en el punto 1. Para las unidades con valores atípicos se tomaron valores alrededor de la mediana de los tamaños de las unidades anteriores como sus respectivos tamaños. En una situación real, los individuos con valores atípicos tienen tamaños de unidad por debajo del valor que les correspondería, como ya se había mencionado al final de la Sección 2. Por esto, se asignaron los tamaños de unidad a los individuos con valores atípicos como se describe anteriormente y no más bajos para garantizar que las muestras resultantes tengan una alta probabilidad de contener individuos de este tipo.
3. El diseño de muestreo que se utilizó fue el muestreo con probabilidades proporcionales al tamaño de la unidad de muestreo.
4. Los tamaños de muestra considerados fueron 200 y 500, 2% y 5% del tamaño de la población respectivamente.
5. El parámetro poblacional de interés fue el total de los valores generados en el punto 1. Para cada muestra seleccionada se estimó este parámetro con el estimador de Horvitz-Thompson dado en (3) y el estimador propuesto en (7).
6. El número de simulaciones considerado fue de $M = 10,000$.
7. Para cada experimento de simulación se reporta:
 - 7.1. Porcentaje de muestras que contienen individuos con valores atípicos. Este porcentaje calculado de esta manera proporciona una aproximación a la probabilidad de que una muestra, seleccionada con el diseño de muestro mencionado en el punto 3, contenga al menos un dato atípico;
 - 7.2. El sesgo relativo de las estimaciones, utilizando la expresión:

$$SesgoR(\hat{\theta}) = \frac{\frac{1}{M} \sum_{j=1}^M \hat{\theta}_j - \theta}{\theta}.$$

La anterior cantidad proporciona una aproximación al valor del sesgo relativo del estimador. Esta forma de calcular el sesgo, permite interpretar dicha cantidad como una proporción del valor real del parámetro.

7.3. La raíz cuadrada del error cuadrático medio relativo de las estimaciones calculado con la expresión:

$$RECMR(\hat{\theta}) = \frac{\sqrt{\frac{1}{M} \sum_{j=1}^M (\hat{\theta}_j - \theta)^2}}{\theta};$$

Para el estimador de Horvitz-Thompson del total, expresión (3), la cantidad calculada de esta manera es una aproximación al valor del coeficiente de variación de dicho estimador, ya que éste es insesgado. Esta medida de variación se puede interpretar en términos proporcionales del valor real del parámetro.

7.4. La precisión del estimador expresada por:

$$P\left(\left|\frac{\hat{\theta} - \theta}{\theta}\right| \leq \epsilon\right) \geq 1 - \alpha.$$

Para un valor $\alpha = 0.05$, se reporta el percentil del 95% de los valores absolutos relativos de la diferencia entre el valor estimado del parámetro y el valor real del mismo. Dicho percentil se puede interpretar como una aproximación al error relativo de estimación ϵ .

De esta forma se realizaron cuatro experimentos de simulación, combinando las poblaciones A1 y A5 con los tamaños de muestra 200 y 500.

Antes de analizar los resultados de las simulaciones, cabe destacar el efecto que tienen los valores atípicos en el estimador de Horvitz-Thompson del total. En la Gráfica 2 se puede observar la contribución de los datos atípicos en las estimaciones, en este caso para la población A5 y 200 como tamaño de muestra. Esta contribución se deriva de la expresión:

$$\frac{\sum_{j \in U_2} y_j \frac{S_{2j}}{\pi_{2j}}}{\hat{t}_{HT}}.$$

Regresando a la gráfica citada, obsérvese por ejemplo que 1% de los datos, 2 valores atípicos, contribuye con 10% aproximadamente al valor de la estimación y 2.5%, 5 datos atípicos, contribuye con 20% aproximadamente. Esta contribución de los valores atípicos a las estimaciones puede producir valores altos de la varianza del respectivo estimador, como se mencionó anteriormente. Este efecto se observará en los resultados de las simulaciones.

[Grafica 2, aquí]

Ahora, los resultados de los cuatro experimentos de simulación se resumen en la Tabla 1. Cabe mencionar que 63% de las muestras simuladas de tamaño 200 tuvieron uno o más datos atípicos y 93% de las muestras de tamaño 500 tuvieron también datos de este tipo. Estas cantidades representan, como se mencionó anteriormente, una aproximación a la probabilidad de que una muestra contenga uno o más datos atípicos. Esta probabilidad depende de la cantidad de datos atípicos en la población, del diseño de muestreo y del tamaño de la muestra. Sin embargo, dicha probabilidad no depende de la distancia de los datos atípicos a los datos concordantes. Adicionalmente, esta distancia sí impacta en la precisión del estimador como es observado en los resultados de las simulaciones.

[Tabla 1, aquí]

En la Tabla 1 se puede observar primero que el sesgo relativo calculado para el estimador de Horvitz-Thompson del total es aproximadamente cero para todos los casos, como es de esperar, ya que este estimador es insesgado. Para el estimador propuesto dicho sesgo es negativo, pero por debajo del 1% para todos los casos, lo que muestra una ligera subestimación del total. El sesgo cercano a cero de un estimador puede ser importante para algunos usuarios de la metodología aquí propuesta. Adicionalmente, si un estimador tiene sesgo cercano a cero permite proponer un estimador de su varianza como estimador del error cuadrático medio.

En segundo lugar, los valores calculados de la raíz cuadrada del error cuadrático medio relativo disminuyen con el aumento del tamaño de muestra y con la menor distancia de los valores atípicos con respecto a los datos concordantes. En los cuatro experimentos de simulación, se puede observar que ésta medida de variabilidad disminuye un punto porcentual aproximadamente en el estimador propuesto con respecto a la correspondiente del estimador de Horvitz-Thompson.

Por último, se observa en todos los casos una ganancia de dos puntos porcentuales aproximadamente en la precisión del estimador propuesto con respecto al de Horvitz-Thompson. Como se mencionó anteriormente, esta precisión es una aproximación al error relativo de estimación con probabilidad del 95%.

Cabe mencionar que disminuir más la variabilidad de estimadores robustos ante la presencia de datos atípicos, sin tener un impacto significativo en el sesgo, y aumentar también más su respectiva precisión deja un área de oportunidad para futuros proyectos de investigación.

5. Aplicación de la metodología propuesta

Con el objetivo de mostrar la factibilidad del método en problemas reales, la metodología se aplicó a la Encuesta Nacional de Victimización y Percepción sobre seguridad pública, ENVIPE 2011 para el estado de Yucatán. Los datos de esta encuesta se pueden obtener de la página Web del INEGI.

Específicamente la metodología se aplicó para estimar la Incidencia Delictiva (número de eventos individuales de victimización delictiva reportados durante un periodo específico en el período de referencia). Para ese año en particular y para dicho estado, hubo una observación atípica cuya contribución a la estimación del total fue de 26% aproximadamente.

Ahora, derivado de la información de esta encuesta, se sabe que la estimación original de la Incidencia Delictiva fue de 40,448. Adicionalmente, al aplicarse la metodología propuesta en este trabajo, dicha estimación de la incidencia fue de 32,011. Esta estimación fue menor, como se podría anticipar. Sin embargo, lo más importante es presuponer que es más precisa, y por tanto más cercana al valor real. La anterior afirmación se sustenta en la evidencia empírica obtenida por los experimentos de simulación descritos anteriormente.

6. Comentarios finales

En la sección 3 se obtuvo un estimador robusto ante observaciones atípicas del total poblacional. Así mismo, el experimento de simulación mostrado en la Sección 4 y la

aplicación descrita en la Sección 5, muestran la bondad y factibilidad de aplicación de la metodología aquí propuesta.

En este mismo orden de ideas, para el muestreo en dos o más etapas se puede utilizar el método de conglomerados últimos para estimar el error cuadrático medio del estimador del total aquí propuesto, utilizando los factores de expansión ya ajustados.

Por otro lado, es importante destacar que el estimador propuesto aquí parte del supuesto que los datos atípicos ya han sido identificados en la muestra. Si en la práctica no se tienen identificados dichos datos, se tendría que realizar primero esta identificación. La detección de valores atípicos se puede hacer utilizando medidas de influencia. Para esto último, se puede consultar el trabajo de Beaumont et al. (2013).

Cabe mencionar también que el presente trabajo considera únicamente el caso univariado. Para este caso, la base de datos con la información de la muestra debería contener una columna adicional con los factores de expansión ajustados. Adicionalmente, la documentación metodológica que acompañe a dicha base de datos tendría que especificar cuándo utilizar dichos factores ajustados. En este mismo sentido, el desarrollo de una metodología para el caso multivariado es un área de oportunidad para futuros proyectos de investigación.

Por último, si se utiliza el diseño de muestreo estratificado, entonces se sugiere aplicar el estimador propuesto por separado en cada estrato. Por otro lado, si el diseño de muestreo es en dos o más etapas, dicho estimador puede ser aplicado a las unidades de la última etapa de muestreo y para las etapas anteriores se utilizan estimadores que no necesariamente consideran valores atípicos.

Fuentes

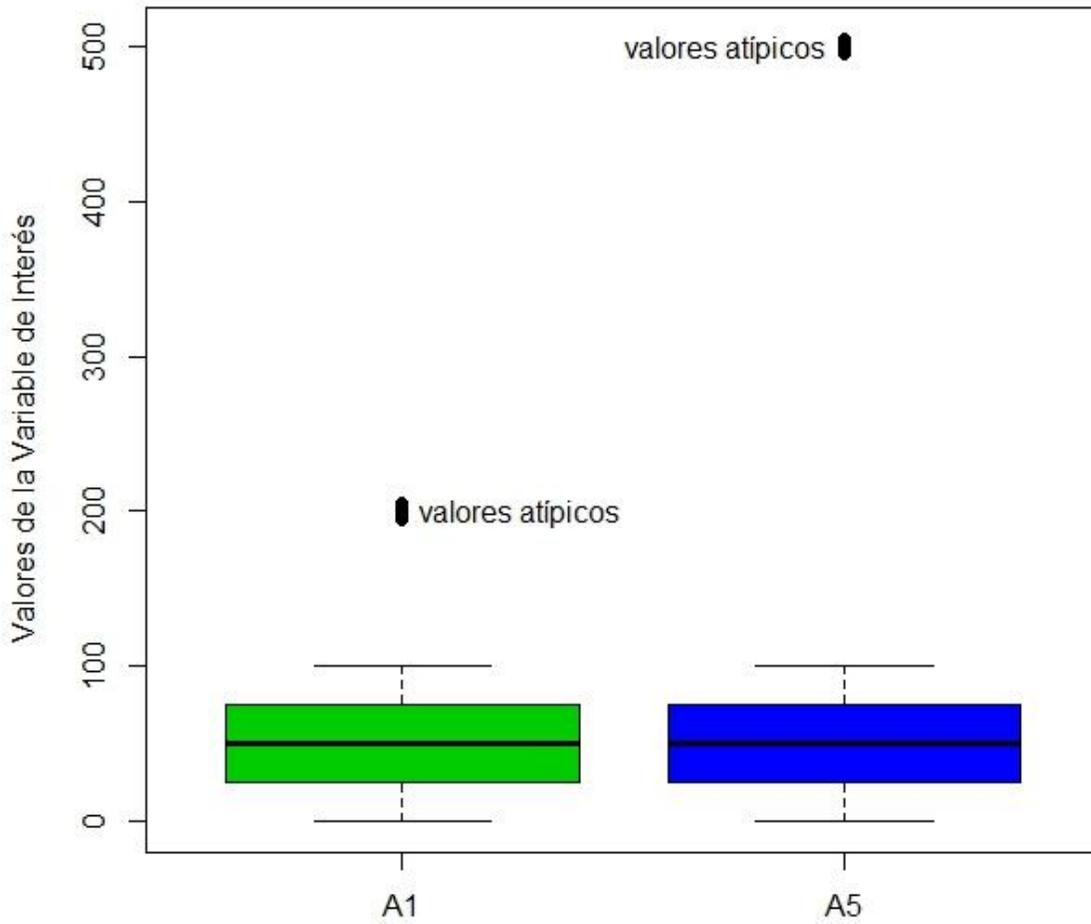
Beaumont, J. F., Haziza, D., & Ruiz-Gazen, A. “A unified approach to robust estimation in finite population sampling”, *Biometrika*. 100, 2013, págs. 555-569.

- Chambers, R. L. “Robust Finite Population Estimation”, *unpublished Ph.D. Thesis*, The Johns Hopkins University, Baltimore. 1982.
- Chambers, R. L. (1986). “Outlier robust finite population estimation”, *Journal of the American Statistical Association*. 81, 1986, págs. 1063-1069.
- Chambers, R. L. y Kokic, P. N. “Outlier robust sample survey inference”, *Proceedings of the ISI 49th Session*. 1993, págs. 55-72.
- Duchesne, P. “Robust calibration estimators”, *Survey Methodology*. 25, 1999, págs. 43-56.
- Hitiroglou, M. H. y Srinath, K. P. “Some estimators of the population total from simple random samples containing large units”, *Journal of American Statistical Association*. 76, 1981, págs. 690-695.
- Hulliger, B. “Outlier robust Horvitz-Thompson estimators” *Survey Methodology*. 21, 1995, págs. 79-87.
- Kish, L. *Survey Sampling*. John Wiley & Sons, New York. 1965.
- Kokic, P. N. y P. A. Bell (1994). “Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator”, *Journal of Official Statistics*. 10, 1994, págs. 419-435.
- Lee, H. “Model-based estimators that are robust to outliers”, *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of Census. 1991, págs. 178-202.
- Lee, H. “Outliers in business surveys”, *In Business Survey Methods*, (Eds. B.G. Box, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge y P. S. Kott), John Wiley & Sons, New York. 1995.
- Ren, R. y Chambers, R. “Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation by Reverse Calibration”. *Methods and Experimental Results from the EUREDIT Project*, J.R.H. Charlton (ed.). 2002.
- Searl, D. T. “An estimator which reduces large true observations”, *Journal of American Statistical Association*. 61, 1966, págs. 1200-1204.
- Welsh, A. H. y Ronchetti, E. “Bias-calibrated estimation from sample surveys containing outliers”, *Journal of the Royal Statistical Society. Serie B*, 60, 1998, págs. 413-428.

Tabla 1. Resultados de los experimentos de simulación. En la primera columna aparecen las poblaciones generadas, en la segunda los tamaños de muestra utilizados, en la tercera y cuarta aparece el sesgo relativo del estimador de Horvitz-Thompson y del estimador propuesto respectivamente, en la quinta y sexta está la raíz cuadrada del error cuadrático medio relativo de los estimadores del total y finalmente aparecen los errores relativos de estimación en las dos últimas columnas.

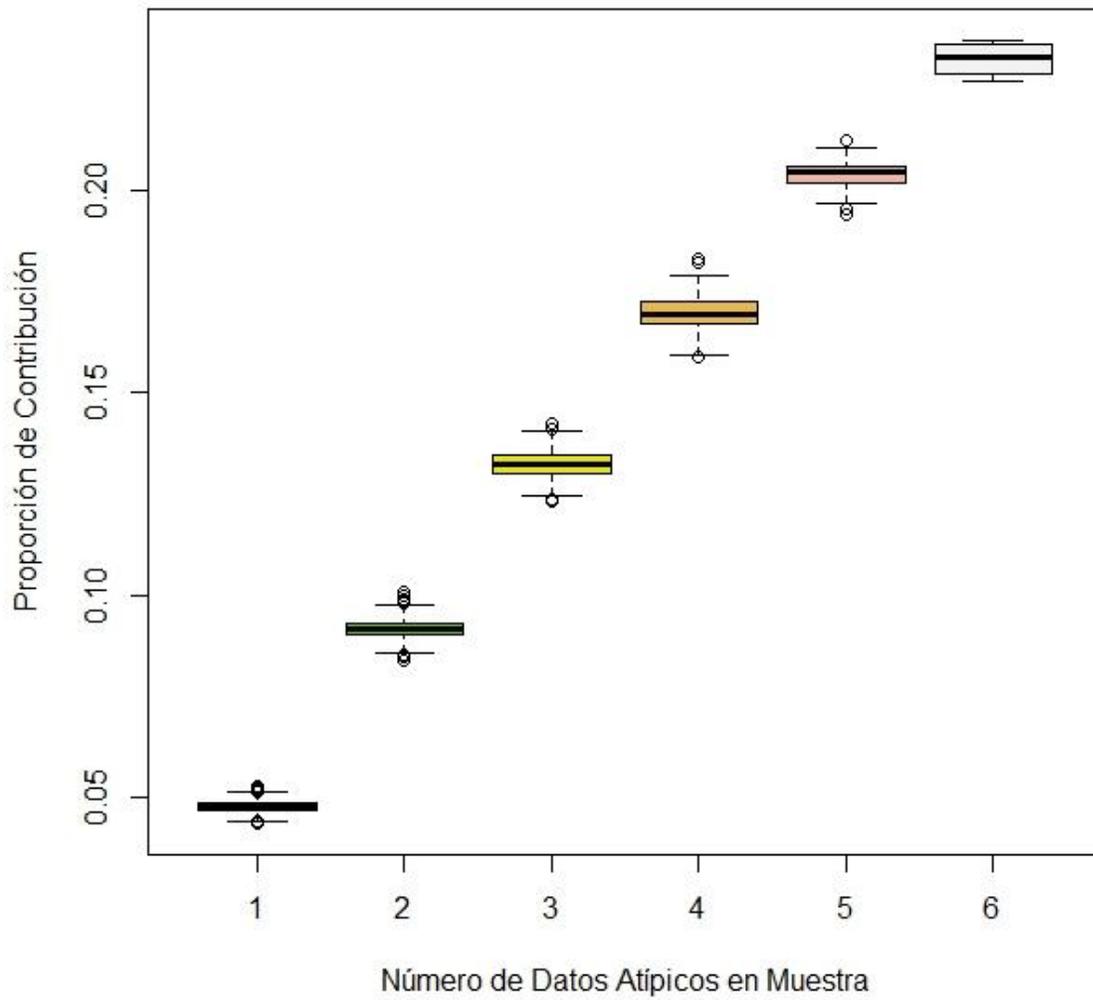
Población	Tamaño de Muestra	Sesgo Relativo		Raíz del Error Cuadrático Medio Relativo		Precisión	
		\hat{t}_{HT}	\hat{t}_{RA}	\hat{t}_{HT}	\hat{t}_{RA}	\hat{t}_{HT}	\hat{t}_{RA}
A1	200	1.3426×10^{-5}	-0.0014	0.0315	0.0209	0.0616	0.0405
	500	2.4790×10^{-4}	-0.0012	0.0193	0.0109	0.0378	0.0212
A5	200	5.2557×10^{-5}	-0.0043	0.0510	0.0377	0.0974	0.0707
	500	6.4813×10^{-5}	-0.0043	0.0308	0.0190	0.0598	0.0463

Poblaciones Generadas para los Experimentos de Simulación



Gráfica 1. Diagramas de caja de los valores generados para las poblaciones A1 y A5 utilizadas en los experimentos de simulación.

Contribución de los Datos Atípicos a la Estimación del Total



Grafica 2. Diagramas de caja de la contribución de los datos atípicos a la estimación del total para la población A5 y tamaño de muestra 200.