



Organización de las Naciones Unidas

Comisión Económica para América Latina y el Caribe (CEPAL)

Misión de Asistencia Técnica

Informe ejecutivo de misión a la Dirección General de Estadística y Censos (DIGESTYC)

Andrés Gutiérrez

Asesor Regional en Estadísticas Sociales

Ciudad Delgado – febrero de 2020



Tabla de contenidos

1. Resumen	3
2. Principales hallazgos y recomendaciones	4
3. Revisión de la estructura de ponderación en las encuestas	7
3.1. Creación de los pesos básicos	9
3.2. Descarte de las unidades no elegibles	10
3.3. Ajuste por elegibilidad desconocida	11
3.4. Ajuste por ausencia de respuesta	13
3.5. Calibración	15
3.6. Trimming de los pesos	18
Referencias	20



1. Resumen

La Dirección General de Estadística y Censos (DIGESTYC) de El Salvador solicitó asistencia técnica a la División de Estadísticas de la CEPAL como parte de las actividades que se vienen desarrollando en la implementación de mejoras en la planeación y análisis de las encuestas de hogares. La misión tuvo lugar en las instalaciones de la DIGESTYC en Ciudad Delgado entre el 3 y 7 de febrero del presente año. Por parte de la CEPAL estuvo presente el Asesor Regional en Estadísticas Sociales, Andrés Gutiérrez; mientras que, por parte de la DIGESTYC, estuvo presente Mabel Hernández, Gerente de Gestión y Análisis Estadístico, junto con tres funcionarios de su equipo.

El objetivo de la misión estuvo centrado en proveer lineamientos técnicos en la implementación de la Encuesta de Hogares de Propósito Múltiples (EHPM), así como en la imputación de las categorías que componen el ingreso disponible para la generación de indicadores de pobreza. De esta forma, se continuó con el programa de asistencias que CEPAL ha provisto a El Salvador propendiendo por darle continuidad al trabajo conjunto en la definición y actualización de las mejores metodologías en torno a las encuestas de hogares y a la medición de indicadores sociales.

Después de una breve introducción, se presenta un resumen de los principales hallazgos encontrados en el transcurso de la misión, así como un conjunto de recomendaciones técnicas y metodológicas enfocadas a afrontar estos retos. De la misma forma, este documento plantea una forma estandarizada de ajuste de los factores de expansión ante la ausencia de respuesta generada por los problemas de desactualización de marco de muestreo y/o los rechazos de los respondientes. Como resultado de esta asistencia técnica, se crearon dos rutinas computacionales en el software estadístico R, conteniendo más de 300 líneas de códigos computacionales.

3. Revisión de la estructura de ponderación en las encuestas

En el desarrollo de la misión se abordó el tema de la creación y definición de los factores de expansión de manera teórica y práctica, utilizando las bases de datos de la Encuesta de Hogares de Propósitos Múltiples (2018) y de la Encuesta Nacional de Violencia Contra la Mujer (2017). Se partió de la definición de marco de muestreo como un dispositivo utilizado para identificar y ubicar las unidades de la población finita sobre el cual se selecciona la muestra.

En condiciones ideales el marco de muestreo debería coincidir plenamente con la población finita; pero en general no siempre es posible contar con una lista de todos los elementos de la población. En el contexto de las encuestas a hogares, no existe una lista que enumere todos los hogares de un país de manera actualizada y la práctica estándar es construir el marco de muestreo en varias etapas,

seleccionando una muestra de áreas geográficas, luego realizando una actualización sobre las viviendas ocupadas por hogares particulares (empadronamiento exhaustivo de todos los hogares en las áreas seleccionadas), para finalmente seleccionar hogares. Este esquema de selección hace que el marco de muestreo de las encuestas a hogares presente imperfecciones. De esta manera, se introdujo el esquema de ponderación AAPOR (Valliant y Dever 2017) que consiste en tratar la ausencia de respuesta de manera diferenciada. Se recomienda clasificar a cada unidad en la muestra en algunas de las siguientes categorías:

- ER (unidades elegibles que fueron respondientes efectivos): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
- ENR (unidades elegibles no respondientes): casos elegibles para los cuales no se recolectó ningún dato o la información fue parcialmente recolectada.
- IN (unidades no elegibles): casos de miembros no elegibles que no hacen parte de la población de interés.
- UNK (unidades con elegibilidad desconocida): casos en donde no se puede conocer si la unidad es elegible o no.

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Descarte de las unidades no elegibles.
3. Ajuste por elegibilidad desconocida.
4. Ajuste por ausencia de respuesta.
5. Calibración por proyecciones poblacionales y variables auxiliares.

6. Identificación de valores extremos y recorte de los pesos (*trimming*).

Nótese que en el caso de la EHPM, el marco de la encuesta estableció la definición de 119 estratos geográficos, conformados en su totalidad por 12435 unidades primarias de muestreo (UPM). De este marco de muestreo fueron seleccionadas 1664 UPM. En total hubo 26829 estructuras que fueron seleccionados, pero que por distintos motivos no proveyeren una respuesta. La siguiente tabla permite discriminar cada uno de los códigos de disposición AAPOR para la ENIGH.

Estatus de elegibilidad	n
Elegible Respondiente	23867
Elegible No Respondiente	549
Inelegible	2242
Elegibilidad desconocida	171

3.1. Creación de los pesos básicos

Asociado a cada esquema particular de muestreo existe una única función que asocia a cada elemento con una probabilidad de inclusión en la muestra. De esta forma:

$$\pi_k = Pr(k \in s)$$

Por lo tanto, el primer paso del anterior esquema induce la creación de los pesos básicos d_k que se definen como el inverso multiplicativo de la probabilidad de inclusión

$$d_{1k} = \frac{1}{\pi_k}$$

Estos pesos son creados incluso para aquellas unidades que serán excluidas de la muestra porque son no elegibles o porque no proveyeron ninguna información y luego serán modificados convenientemente. La figura 1 muestra la distribución de estos pesos a lo largo de las estructuras seleccionadas.

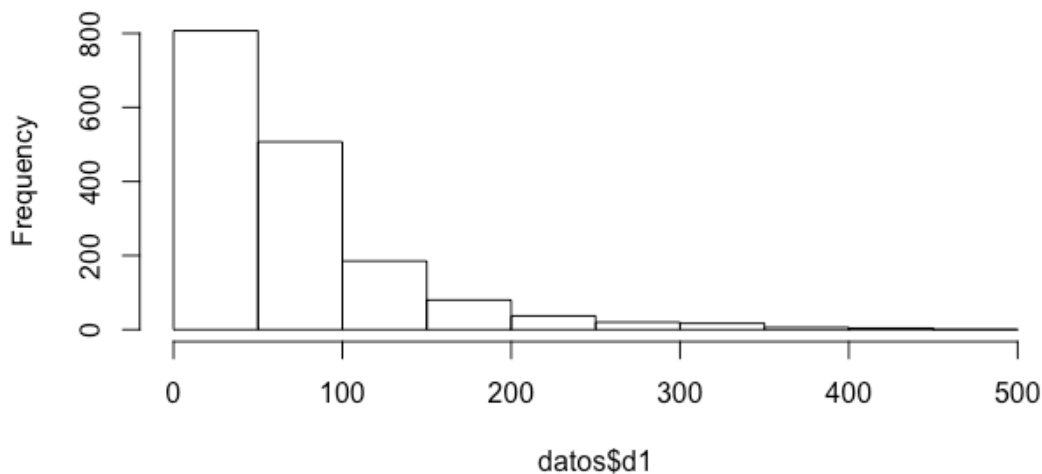


Figura 1. Distribución de los pesos de muestreo básicos muestra de dos etapas.

3.2. Descarte de las unidades no elegibles.

Si hay estructuras seleccionadas desde el marco de muestreo que han cambiado su estado de ocupación y ahora no contienen ningún hogar, entonces el segundo paso consiste en ajustar su peso básico de la siguiente manera:

$$d_{2k} = \begin{cases} 0, & \text{si la unidad } k \text{ no pertenece a la población objetivo} \\ d_{1k}, & \text{en otro caso} \end{cases}$$

La figura 2 muestra la distribución de estos pesos a lo largo de las viviendas seleccionadas.

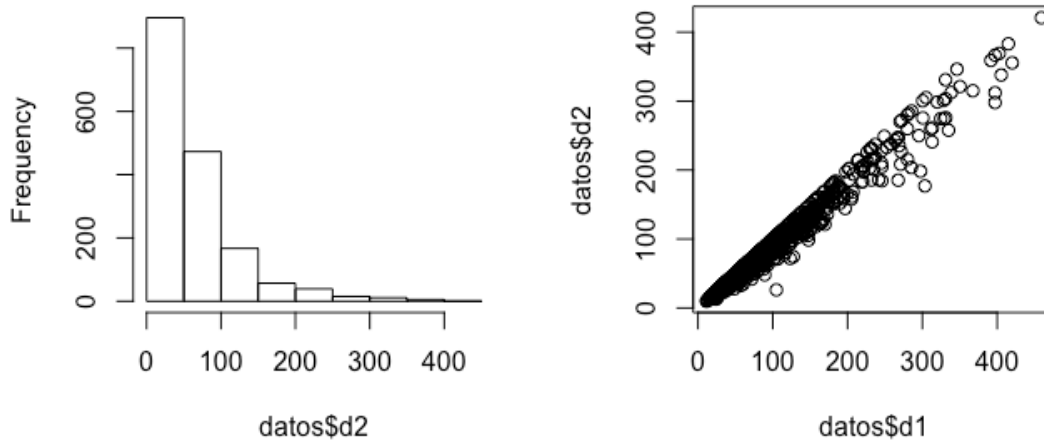


Figura 2. Distribución de los pesos ajustados por no elegibilidad (izquierda) y relación de ajuste con respecto a los pesos básicos (derecha).

3.3. Ajuste por elegibilidad desconocida.

El tercer paso consiste en redistribuir el peso de las unidades cuyo estado de elegibilidad es desconocido. Por ejemplo, si en una vivienda nunca se atendió el llamado del encuestador, o si la encuesta está enfocada en la población mayor de 15 años y hay personas que no proveen ninguna información acerca de su edad, entonces es necesario distribuir estos pesos. Esta situación se presenta a nivel de hogar cuando no puede ser contactado porque nadie nunca atendió el llamado del encuestador (nadie en casa). Se acostumbra a redistribuir los pesos de los UNK entre las unidades que sí disponen de su estatus de elegibilidad (ER, ENR, IN).

Luego, si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá el conjunto de las unidades elegibles en la muestra s_e , el conjunto de las unidades no elegibles en la

muestra s_n y el conjunto de las unidades con elegibilidad desconocida s_u . En este último caso, la elegibilidad de estos casos es desconocida, a no ser que de manera arbitraria sean clasificadas como ENR (elegibles no respondientes), o se tenga información auxiliar en el marco de muestreo que permita imputar su estado de elegibilidad.

Se recomienda formar B ($b = 1, \dots, B$) categorías¹ basadas en la información del marco de muestreo. Estas categorías pueden ser estratos o cruces de subpoblaciones. Siendo s_b la muestra de unidades en la categoría b (que incluye a ER, ENR y UNK), se define el factor de ajuste por elegibilidad como:

$$a_b = \frac{\sum_{s_b} d_{2k}}{\sum_{s_b \cap s_e} d_{2k}}$$

Para la categoría b , los pesos ajustados por elegibilidad desconocida para aquellas unidades cuya elegibilidad si pudo ser establecida (independientemente de su estado de respuesta) estarán dados por la siguiente expresión:

$$d_{3k} = a_b * d_{2k}$$

La figura 3 muestra la distribución de estos pesos a lo largo de las viviendas seleccionadas.

¹ Se acostumbra a formar categorías con al menos 50 casos.

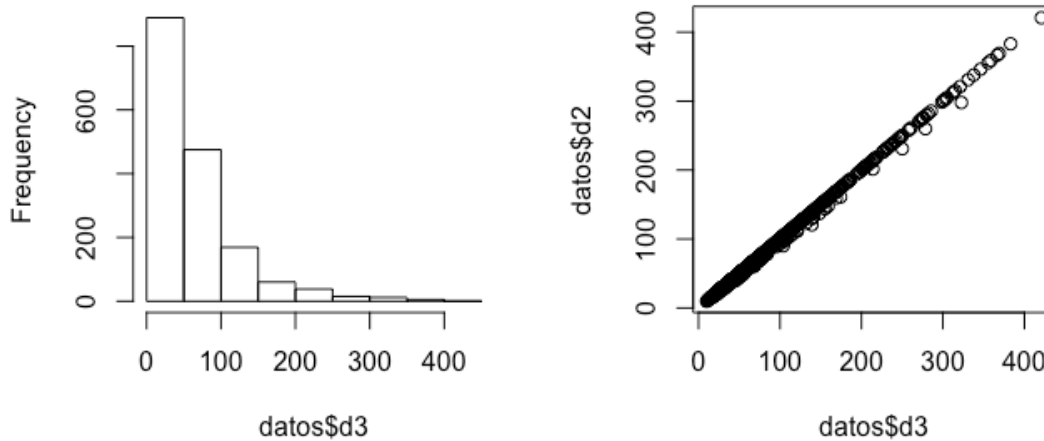


Figura 3. Distribución de los pesos ajustados por elegibilidad desconocida (izquierda) y relación de ajuste con respecto a los pesos ajustados por no elegibilidad (derecha).

3.4. Ajuste por ausencia de respuesta

En este paso los pesos básicos de los ER se ajustan para tener en cuenta a los ENR. Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de que algunas unidades elegibles no proveyeron información. Para el manejo efectivo de la ausencia de respuesta se consideran las siguientes variables aleatorias:

$$I_k = \begin{cases} 1, & \text{si } k \text{ pertenece a la muestra } s \\ 0, & \text{en otro caso.} \end{cases}$$

$$D_k = \begin{cases} 1, & \text{si } k \text{ pertenece al conjunto de respondientes } s_r \\ 0, & \text{en otro caso.} \end{cases}$$

Al suponer que la distribución de las respuestas puede ser estimada, entonces la probabilidad de respuesta, conocida como propensity score (Särndal y Lundström 2006), está dada por

$$Pr(k \in s_r | k \in s) = Pr(D_k = 1 | I_k = 1) = \phi_k$$

Si fuera posible tener acceso a las covariables \mathbf{x} , entonces es posible estimar el patrón de ausencia de respuesta mediante la siguiente expresión

$$\hat{\phi}_k = f(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$$

En el caso de la EHPM, se propuso un modelo para ausencia de respuesta por rechazo en función del estrato y la unidad primaria de muestreo. La Figura 4 muestra la distribución de las probabilidades de respuesta para los hogares incluidos en la muestra (respondientes y no respondientes), junto con el diagrama de dispersión de los nuevos pesos de muestreo ajustados en esta etapa. Bajo este escenario, es posible definir el siguiente estimador insesgado

$$\hat{t}_y = \sum_{k \in s_r} d_{4k} y_k$$

En donde $d_{4k} = \frac{d_{3k}}{\hat{\phi}_k}$.

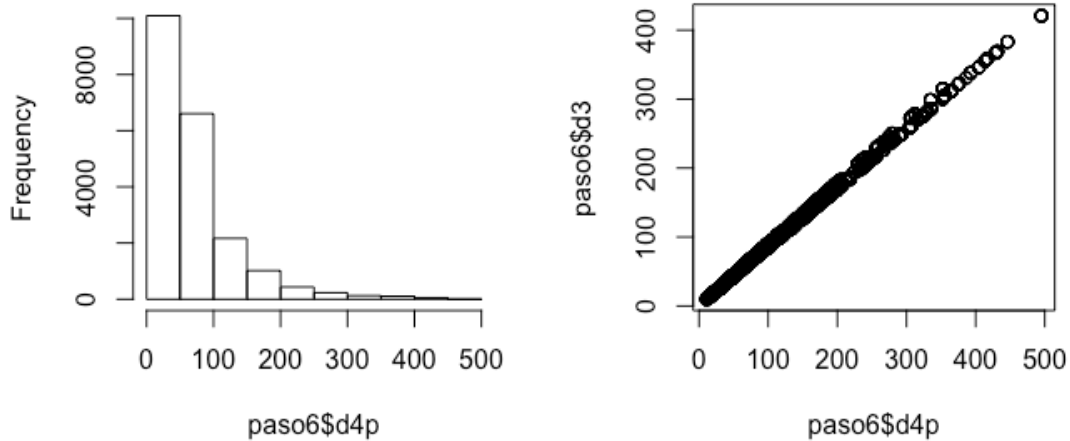


Figura 4. Distribución de los pesos ajustados por ausencia de respuesta (izquierda) y relación de ajuste con respecto a los pesos ajustados por elegibilidad desconocida (derecha).

3.5. Calibración

La calibración es un ajuste que se realiza a los pesos de muestreo con el propósito de que las estimaciones de algunas variables de control reproduzcan de forma perfecta los totales poblacionales de estas variables (Silva 2004). Esta propiedad de consistencia es deseable en un sistema de ponderadores. Cuando los estudios por muestreo están afectados por la ausencia de respuesta, es deseable tener las siguientes propiedades en la estructura inferencial que sustenta el muestreo: 1) sesgo pequeño o nulo, 2) errores estándares pequeños, 3) un sistema de ponderación que reproduzca la información auxiliar disponible, 4) un sistema de ponderación que sea eficiente al momento de estimar cualquier característica de interés en un estudio multipropósito.

La calibración es usualmente el último paso en el ajuste de los ponderadores y hace uso de información auxiliar que reduce la varianza y corrige los problemas de cobertura que no pudieron ser corregidos en los pasos previos. El objetivo de la calibración es obtener un nuevo sistema de ponderadores w_k que se encuentren cerca de los ponderadores básicos d_k , de tal forma que cuando los ponderadores sean usados para estimar los totales de las variables auxiliares, dichos totales sean reproducidos con exactitud. La idea de mantener cerca los nuevos pesos es para pedir prestada cualquier propiedad buena de estimación de los pesos básicos. Por ejemplo, si los ponderadores básicos producen estimadores insesgados, los nuevos ponderadores calibrados producirán también estimadores aproximadamente insesgados. Los nuevos pesos calibrados w_k se construyen a partir de minimizar una pseudo-distancia $G(w_k, d_k)$ sujeta a la siguiente restricción

$$\sum_s w_k \mathbf{x}_k = \mathbf{t}_x$$

En donde \mathbf{x}_k es el conjunto de variables auxiliares para la calibración, que debe ser preguntada en la encuesta. La forma final de los pesos calibrados no es cerrada y está dada por la siguiente expresión

$$w_k = d_k F(\boldsymbol{\lambda}' \mathbf{x}_k)$$

En donde F es una función real y $\boldsymbol{\lambda}$ es un vector de números. En algunos casos se requieren procedimientos numéricos iterativos para encontrar una solución final a los ponderadores de calibración.

En el caso particular de la EHPM, las variables de calibración correspondieron a los siguientes subgrupos de interés:

- Número de hombres en El Salvador.
- Número de mujeres en El Salvador

- Número de habitantes en cada uno de los 14 departamentos
- Número de habitantes en la región urbana.
- Número de habitantes en la región rural.

Luego de realizar el proceso de calibración, se verificó que efectivamente las ecuaciones de calibración se cumplieran a cabalidad. La Figura 5 muestra la distribución de los pesos calibrados finales y su relación con los pesos originales. Se puede notar que existe un sobreajuste de los pesos finales que permite llegar a tener la representatividad que se perseguía inicialmente y solucionar el problema de la ausencia de respuesta en la encuesta.

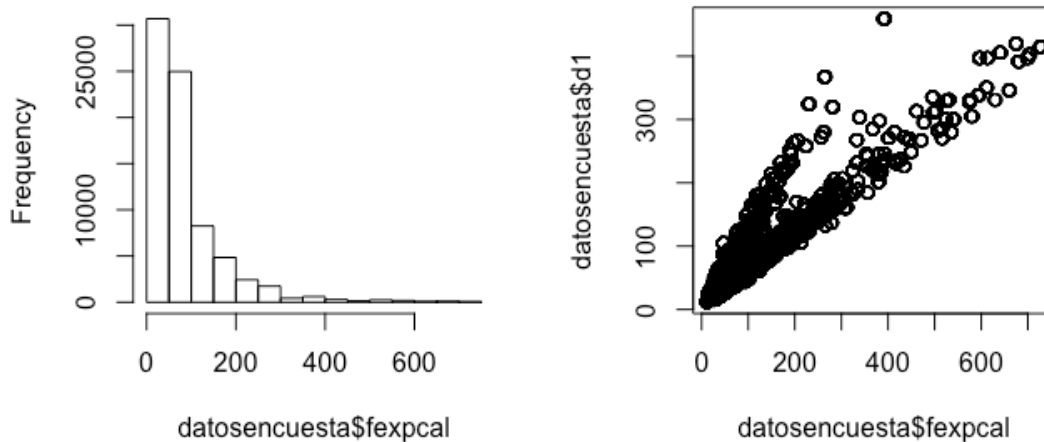


Figura 5. Distribución de los pesos calibrados (izquierda) y relación de ajuste con respecto a los pesos iniciales (derecha).

3.6. Trimming de los pesos

Un inconveniente que se genera debido a la ausencia de respuesta y la desactualización de los marcos de muestreo y la estructura de la calibración es que, si bien el estimador resultante tendrá un sesgo cercano a cero, la distribución de los pesos calibrados muestra datos extremos que hacen que la varianza del estimador crezca y que, por ende, la precisión de la inferencia decrezca. La figura 6 muestra la distribución de los factores de expansión calibrados con la presencia de datos extremos en algunos departamentos de El Salvador.

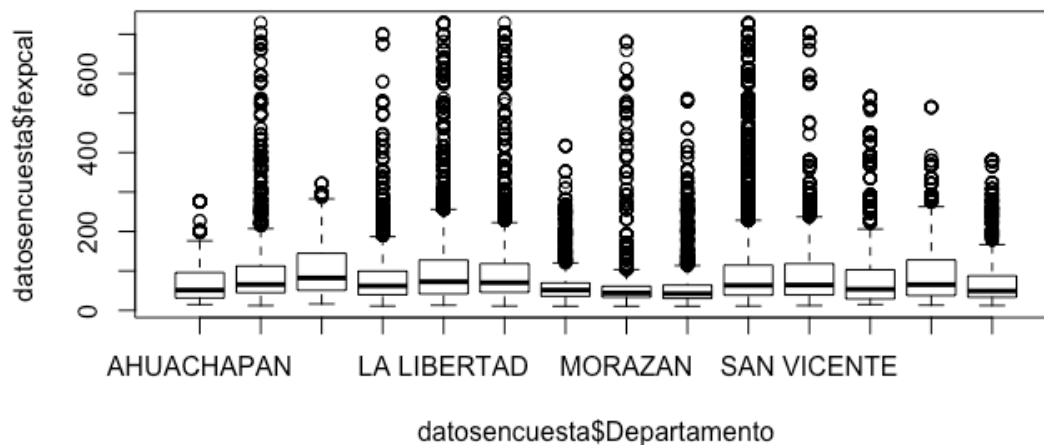


Figura 6. Distribución de los pesos de calibración para los 14 departamentos.

Para hacerle frente a este problema, es posible programar un procedimiento de trimming siguiendo las recomendaciones de Valliant, Dever, y Kreuter (2013, sec. 14.4), que puede ser resumido de la siguiente manera:

1. Recortar cualquier peso mayor a 3.5 veces la mediana de la distribución de pesos calibrados. Por tanto,

$$U = 3.5 \times \text{mediana}(\mathbf{w}).$$

2. Cualquier peso con magnitud superior a U se trunca de la siguiente manera

$$w_k^* = \begin{cases} U, & \text{si } w_k \geq U \\ w_k, & \text{en otro caso} \end{cases}$$

3. Determinar la cantidad neta perdida debido al recorte de pesos extremos

$$K = \sum_s (w_k - w_k^*)$$

4. Distribuir K equitativamente entre las unidades que no fueron recortadas.
5. Iterar hasta que todos los nuevos pesos calibrados estén por debajo de U .

La figura 7 muestra los resultados finales del proceso de *trimming*. Se nota que los datos extremos en los factores de expansión han sido correctamente manejados y que la distribución general de los pesos se mantuvo en los departamentos.

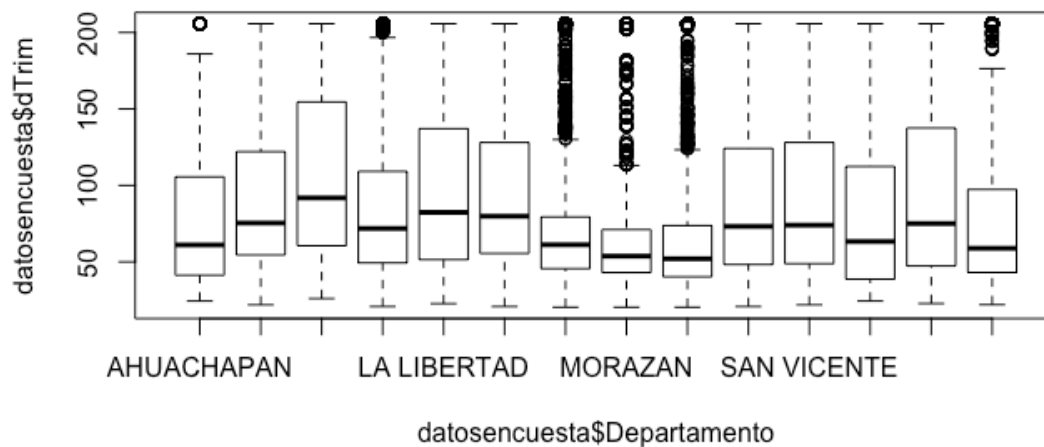


Figura 7. Distribución de los pesos finales después del trimming.



Referencias

- Gutiérrez, Andrés. 2018. «Rediseño de la Encuesta de Hogares de Propósitos Múltiples (EHPM) - Informe de Misión a la DIGESTYC». Comisión Económica para América Latina y el Caribe (CEPAL).
- Särndal, Carl-Erik, y Sixten Lundström. 2006. Estimation in surveys with nonresponse. Repr. Wiley series in survey methodology. Wiley.
- Silva, PL. d N. 2004. «Calibration estimation: when and why, how much and how». Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística.
- Valliant, Richard, y Jill A. Dever. 2017. Survey Weights: A Step-by-step Guide to Calculation. 1 edition. Stata Press.
- Valliant, Richard, Jill A. Dever, y Frauke Kreuter. 2013. Practical Tools for Designing and Weighting Survey Samples. Springer New York. <https://doi.org/10.1007/978-1-4614-6449-5>.