# Recomendaciones para eliminar el sesgo de selección en las encuestas de hogares en la coyuntura de la enfermedad por coronavirus (COVID-19)

Mayo de 2020



## Resumen

En el contexto de la pandemia mundial de la enfermedad por coronavirus (COVID-19), la Comisión Económica para América Latina y el Caribe (CEPAL) está elaborando una serie de publicaciones cortas con recomendaciones de política relevantes para este período. En la presente nota se plantean recomendaciones para abordar los problemas de sesgo que pueden presentarse en las encuestas de hogares levantadas durante este período excepcional, como complemento a las sugerencias efectuadas anteriormente sobre los diseños muestrales de este tipo de instrumentos.

## Introducción

En su intento por frenar la velocidad de contagio del COVID-19, los países han impuesto restricciones a la movilidad de las personas, que han impedido la recolección presencial de información de las encuestas de hogares. Para hacer frente a este inconveniente y poder seguir produciendo estadísticas oficiales pertinentes y oportunas, algunas oficinas nacionales de estadística han recurrido a la realización de encuestas por medio del teléfono o por Internet. En el documento "Recomendaciones para la publicación de estadísticas oficiales a partir de encuestas de hogares frente a la coyuntura de la enfermedad por coronavirus (COVID-19)" se propusieron algunas líneas de acción para la conformación de la muestra de hogares a utilizar en las encuestas, con referencia particular al empleo de un panel seleccionado de un período reciente para el que se dispusiera de la información de contacto telefónico (CEPAL, 2020). La presente nota complementa las recomendaciones efectuadas anteriormente, mediante el planteamiento de dos enfoques para minimizar el sesgo generado por la ausencia de respuesta que se va a encontrar en el levantamiento telefónico. Además, se plantea un tercer enfoque pertinente para aquellos casos en que se dificulte la consecución de información auxiliar.

## A. Detección del sesgo

Cambiar el modo de recolección de información de una encuesta de hogares, de un modo presencial a un modo telefónico o a través de la web, puede traer consigo consecuencias indeseadas y, en particular, generar sesgos (de selección, de cobertura y por ausencia de respuesta) de quienes responden a la encuesta. En un escenario en que se está utilizando una muestra de hogares de un período anterior (a la que se denominará "muestra original") y en que se están haciendo todos los esfuerzos por contactar a los hogares seleccionados, el proceso está inevitablemente expuesto a las siguientes contingencias:

- No todos los hogares de la muestra original proveyeron información de su contacto telefónico.
- Algunos hogares proveyeron sus datos de contacto, pero al momento de la entrevista no habitan en la vivienda seleccionada.

- Algunos hogares proveyeron sus datos de contacto, pero al momento de la entrevista han cambiado el número telefónico de contacto.
- No todos los hogares que proveyeron su información de contacto están dispuestos a responder el cuestionario de la encuesta.

A manera de ejemplo, si se supone que la cobertura de la muestra que sí proveyó datos de contacto asciende al 85% y que la probabilidad de que un hogar contactado responda toda la encuesta es del 80%, entonces se contaría solamente con respuestas de un 68% de la muestra original. A estas cuentas habría que sumar el posible efecto de la atrición en el panel —pérdida de participantes a medida que transcurre el panel—, puesto que habrá hogares que dejarán de responder a la encuesta a medida que son contactados de manera reiterativa.

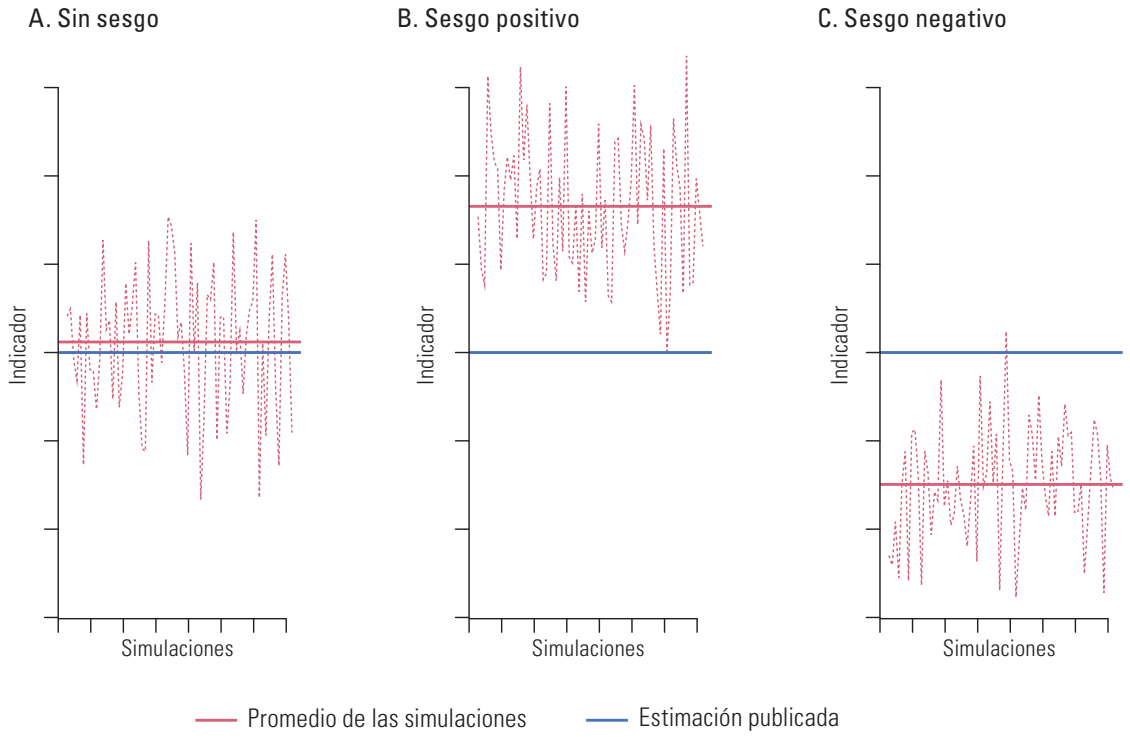
En este escenario, es altamente probable que los hogares respondientes efectivos no tengan características similares a los hogares no respondientes y a los hogares no cubiertos, y que las razones de ausencia de respuesta de los hogares en el levantamiento estén asociadas al fenómeno que se intenta medir (por ejemplo, que en los hogares no respondientes haya más personas desocupadas, o que en los hogares pobres haya más ausencia de respuesta). Ello implica que la información obtenida de los hogares respondientes se encontrará sesgada, por lo que no podrá ser utilizada sin realizar algún tipo de corrección.

Por tanto, el primer paso, una vez que se haya recolectado la información en un determinado período, debería ser la estimación de la magnitud del sesgo. De acuerdo con lo planteado en CEPAL (2020), una posibilidad para ello es utilizar escenarios simulados, sobre la base de la disposición final de los hogares que efectivamente respondieron la encuesta dentro de las unidades primarias de muestreo (UPM) que los contienen. Mediante simulación es posible tratar de predecir cómo hubiese sido el comportamiento de los estimadores en el mes de selección de la muestra si solo se hubiese contado con la información parcial actual. La diferencia entre las estimaciones ya publicadas (insesgadas) y las estimaciones simuladas (eventualmente sesgadas) dará una idea de la magnitud del sesgo.

En el gráfico 1 se presentan tres posibles escenarios para la búsqueda de sesgo. En la imagen de la izquierda se ilustra un caso en que no existe sesgo, mientras que en las imágenes del centro y de la derecha se muestran escenarios correspondientes a un sesgo significativo. La línea horizontal azul corresponde a la estimación publicada en el mes en que se seleccionó la muestra original, mientras que la línea horizontal roja representa el promedio de las simulaciones con la muestra efectiva. Cada uno de los resultados de las simulaciones está representado por las fluctuaciones de las líneas punteadas. Nótese que en los últimos dos escenarios (véanse los gráficos 1.B y 1.C) la mayoría de las simulaciones no cubren la estimación publicada y por ende se puede asegurar que sí existe sesgo.

En general, el intervalo de confianza publicado puede usarse para determinar si existe sesgo y en qué magnitud. Se recomienda revisar todos los escenarios que en promedio estén a más de media desviación de la estimación puntual de la muestra original. En el escenario más optimista, el de ausencia de sesgo, una oficina de estadística estaría en buena posición para replicar los procesos usuales de inferencia. Sin embargo, si se sospecha que existe sesgo, y dependiendo de la información auxiliar disponible, es posible aplicar alguna de las alternativas que se describen en las siguientes secciones.

Gráfico 1
Tres posibles escenarios en la búsqueda del sesgo de selección




Fuente: Comisión Económica para América Latina y el Caribe (CEPAL).

## B. Ajuste por probabilidad de respuesta

En muchos países de la región se han realizado levantamientos de información telefónicos sobre la base de una muestra de períodos anteriores, como alternativa ante las restricciones para la recolección presencial de la información. En CEPAL (2020) se planteó que la mejor estrategia en la conformación del panel era el seguimiento a una muestra completa de meses anteriores (por ejemplo, febrero de 2020), puesto que la selección de una submuestra sobre la unión de las muestras de meses anteriores dará origen a cálculos bastantes complejos de las probabilidades de inclusión de los elementos. Por consiguiente, considerando que se parte de una muestra probabilística que se acoge a las anteriores recomendaciones, es posible realizar ajustes a los factores de expansión de manera diferencial para corregir el sesgo de selección.

Propuesto por Rosenbaum y Rubin (1983), este enfoque es útil para dilucidar la estructura de la ausencia de respuesta y, por consiguiente, corregir el sesgo de cobertura y el sesgo por ausencia de respuesta (Lensvelt-Mulders, Lugtig y Hubregtse, 2009). Para el manejo efectivo de la ausencia de respuesta se consideran las variables dicotómicas $I_k$ y $D_k$, que indican si el hogar pertenece a la muestra original y si ha respondido a la encuesta telefónica, respectivamente. Suponiendo que la distribución de las respuestas efectivas puede ser estimada, la probabilidad de respuesta *(propensity score)* de un hogar en la muestra está dada por:

$$\phi_k = \Pr(D_k = 1 | I_k = 1)$$

Nótese que esta probabilidad es distinta para cada hogar y puede ser estimada usando los datos del panel. Contar con la muestra original, para la cual se obtuvo toda la información del cuestionario en un período anterior, constituye un excelente punto de partida para tratar de eliminar el sesgo, puesto que se tendrá acceso a un conjunto de covariables *x*

para determinar el mejor modelo a fin de estimar el patrón de ausencia de respuesta en la muestra de respondientes efectivos. A manera de ejemplo, las covariables útiles para estimar la probabilidad de respuesta pueden incluir el sexo, la edad, el nivel educativo, el área y la región geográfica de residencia, el estado de ocupación y el ingreso per cápita del hogar en el levantamiento original, entre otras.

Si se asume que la probabilidad de respuesta depende de alguna combinación lineal de las covariables disponibles en la muestra original, es posible ajustar un modelo en que la variable dependiente es $D_k$ y el vector de covariables se representa como $x$. Kim y Riddles (2012) muestran que es posible utilizar un modelo basado en el ajuste de la probabilidad de respuesta de la muestra telefónica mediante la siguiente expresión:

$$\text{logit}(\hat{\phi}_k) = x_k \hat{\beta}$$

donde $\hat{\beta}$ es el vector de coeficientes estimado de la regresión logística. Se debe prestar especial atención a la elección de predictores en el modelo de regresión logística, que debería funcionar bien si las variables de información auxiliar disponibles son relevantes y explicativas de la respuesta telefónica; de otra forma, esta metodología no tendrá ningún beneficio para la reducción del sesgo (y posiblemente lo exacerbará) y dará como resultado errores estándares más grandes.

Teniendo en cuenta que los pesos originales de la encuesta telefónica se representan como $d_k$, y habiendo estimado $\hat{\phi}_k$ para respondientes y no respondientes de la muestra telefónica, entonces el factor de expansión ajustado tomaría la siguiente forma:

$$w_k = \frac{d_k}{\hat{\phi}_k}$$

Utilizar el factor de expansión $w_k$ en el cálculo de los estimadores deseados minimizaría el sesgo de selección que se generó por el cambio de modo de recolección de la información. Los factores asociados con el sesgo de cobertura pueden no ser los mismos que los factores asociados con el sesgo por ausencia de respuesta, por lo que probablemente sería beneficioso modelar estos problemas por separado y luego usar los dos puntajes de propensión como factores de ajuste independientes.

## C. Calibración en dos etapas

Särndal y Lundström (2006) afirman que cuando los estudios por muestreo están afectados por la ausencia de respuesta, es deseable tener un sistema de ponderación que reproduzca la información auxiliar disponible y que sea eficiente al momento de estimar cualquier característica de interés en un estudio multipropósito. Los estimadores de calibración (Deville y Särndal, 1992) satisfacen estas condiciones y bien pueden acomodarse para paliar el sesgo generado por el cambio de modo de recolección de la información.

En principio, se dispone de dos fuentes de información auxiliar. Por un lado, se cuenta con la información que se utiliza usualmente para calibrar los factores de expansión en un levantamiento regular (representada como $x_{1k}$). Por otro lado, se dispone de las variables que fueron medidas en la muestra original (representadas como $x_{2k}$). Ello implica que, después de calcular los pesos para la encuesta telefónica ($s_t$), es posible calibrarlos a nivel de la información auxiliar disponible en la muestra original ($s_m$), a nivel nacional ($u$), o por estratos de interés.

La primera etapa consiste, por tanto, en encontrar un conjunto de pesos calibrados sujetos a la siguiente restricción (Särndal, 2007):

$$\sum_{s_m} w_{1k}\, x_{1k} = \sum_{U} x_{1k}$$

En una segunda etapa se deben usar los pesos intermedios $w_{1k}$ para calcular los pesos finales de calibración $w_k$ de la muestra telefónica, sujetos a la siguiente restricción:

$$\sum_{s_t} w_k \, x_{2k} = \sum_{s_m} w_{1k} \, x_k = \left( \begin{array}{c} \sum_{U} x_{1k} \\ \sum_{s_t} w_{1k} \, x_{2k} \end{array} \right)$$

Para asegurar la coherencia entre las cifras oficiales ya publicadas y las que la encuesta telefónica puede producir, es deseable el uso de los estimadores de calibración. Al aplicar este enfoque se asegura una estructura inferencial robusta en presencia de la información disponible, puesto que se reduce tanto el error de muestreo (aumentando la precisión) como el error debido a la ausencia de respuesta (eliminando el sesgo).

A manera de ejemplo, un procedimiento de calibración en dos etapas podría utilizar las siguientes variables de referencia:

i) En la primera etapa, la calibración de los pesos de la muestra original podría basarse en los totales de edad, región, área y sexo, disponibles en proyecciones demográficas robustas (o en los conteos censales, si el último censo es reciente).

ii) En la segunda etapa, la calibración de los pesos de la muestra telefónica podría basarse en las variables indicadas anteriormente, y además en los totales de ingreso per cápita, condición de ocupación, rama de actividad y escolaridad, obtenidos de la publicación con los resultados de la encuesta original.

El propósito general del proceso de calibración es encontrar un número de restricciones moderado, que permita tener estimaciones aproximadamente insesgadas con una varianza menor que la generada con los factores de expansión originales. En general, los procesos de calibración pueden clasificarse en alguna de las siguientes tres categorías:

i) Calibración con variables continuas, que es el caso en que la calibración se realiza con los totales de variables continuas como ingreso y gasto, entre otras.

ii) Postestratificación con variables categóricas, que es el caso en que la calibración se realiza con los tamaños poblacionales (basados en proyecciones demográficas o registros administrativos) de subgrupos de interés.

iii) Calibración por marginales con variables categóricas (conocido como *raking*), que se define como una calibración sobre los tamaños marginales de tablas de contingencia de subgrupos de interés. A diferencia de los casos anteriores, en esta calibración no se tienen en cuenta los tamaños de los cruces, sino solo los tamaños marginales; por ende, este método induce menos restricciones.

## D. Postestratificación basada en modelos multinivel

En caso de que una oficina de estadística no haya utilizado un panel definido por una muestra probabilística de un período anterior, no se podrá ejecutar ninguna de las opciones anteriores, puesto que no se contará con la información auxiliar necesaria para descifrar el mecanismo de respuesta del operativo telefónico. En estos casos, y como último recurso, es posible realizar algunos ejercicios empíricos basados en modelos predictivos para tener una idea de la magnitud del sesgo y corregirlo.

La regresión multinivel con postestratificación (MRP) es una técnica útil para predecir un parámetro de interés dentro de dominios pequeños mediante el modelado de la media de la variable de interés condicional en los recuentos de postestratificación. Este método fue propuesto inicialmente por Gelman y Little (1997) y ampliado por Park, Gelman y Bafumi (2004). Esta técnica es ampliamente utilizada para corregir el sesgo de selección de las encuestas y su objetivo final es estimar un parámetro de interés (totales, medias o proporciones, entre otros) para todos los estratos (dominios, categorías o subgrupos) en una población finita.

De manera similar al modelo de calibración en dos etapas, para aplicar esta metodología se requiere contar con información auxiliar sobre los totales de personas según las características de interés; por ejemplo, información censal agregada a nivel de regiones sobre el número total de personas para todas las posibles combinaciones de las variables sexo, edad y nivel educativo.

El modelo MRP está compuesto por dos partes: la primera consiste en el ajuste de un modelo de regresión multinivel sobre la base de la encuesta de hogares y la segunda corresponde a la postestratificación, utilizando los conteos censales. A continuación, se describen los pasos de la metodología.

i) Tanto la característica de interés como las covariables (demográficas y geográficas) auxiliares $\boldsymbol{x}$ son observadas en la encuesta. Se puede asumir que las covariables definen un conjunto de $J$ celdas o postestratos ($j=1,\ldots,J$). Por ejemplo, se podría considerar que los postestratos están conformados por los cruces entre 5 categorías de edad, 4 categorías de educación, 2 categorías de área (rural y urbana), 2 categorías de sexo y 20 regiones (provincias, departamentos o estados). Entonces, se tendría: $j=5 \times 4 \times 2 \times 2 \times 20=1.600$ postestratos.

ii) Se ajusta un modelo de regresión multinivel $y\sim(X|Región)$ para obtener una predicción promedio $\hat{\bar{y}}$ de la característica de interés condicional a sus variables demográficas y geográficas. En este paso se debe predecir este valor para cada postestrato; es decir, se debe obtener $\hat{\bar{y}}_j$ para $j=1,\ldots,J$. Nótese que este modelo de regresión es multinivel, por cuanto depende de las 20 regiones definidas anteriormente.

iii) Para cada celda $j$ se dispone de información sobre la población $N_j$, extraída de las proyecciones demográficas. Por ende, es posible estimar el promedio nacional de la variable de interés como un promedio ponderado de las estimaciones $\hat{\bar{y}}_j$ :

$$\hat{\bar{\bar{y}}} = \frac{\sum_j N_j \, \hat{\bar{y}}_j}{\sum_j N_j}$$

En el caso particular de que la variable de interés sea la tasa de desocupación nacional, esta se define como una variable dicotómica $y_{ij}$, que describe el estado de ocupación (ocupado o desocupado) de la persona -ésima en la fuerza de trabajo, que pertenece al postestrato $j$. El objetivo del modelo es relacionar los componentes de $y_{ij}$ con la información auxiliar, $\boldsymbol{x}_{ij}$, lo que puede hacerse mediante un modelo de regresión logística multinivel sobre la probabilidad de estar desocupado $p_{ij}=Pr(y_{ij}=1)$ , definido como:

$$\text{logit}(p_{ij}) = \beta_0^j + \mathbf{x}_i \boldsymbol{\beta}$$

En la expresión anterior, los coeficientes $\boldsymbol{\beta}$ representan los efectos fijos de las variables (edad, educación, área y sexo) sobre las probabilidades del modelo. Por su parte, el primer sumando representa el intercepto aleatorio del modelo que depende de la región:

$$\beta_0^l = \alpha_0 + \alpha_1^l$$

donde los coeficientes $\alpha_0 + \alpha_1^l$ representan el intercepto aleatorio inducido por la región l ($l=1,\ldots,L$). Después de estimar los coeficientes del modelo de regresión multinivel, se puede estimar la probabilidad de que cualquier persona en la fuerza de trabajo esté desocupada, condicionada por su información personal en términos de la región y las demás variables de información auxiliar. Teniendo en cuenta que $n_j$ representa el tamaño de la muestra en el postestrato $j$, la predicción del promedio de individuos desocupados en este cruce corresponderá al promedio de las probabilidades $\hat{p}_{ij}$ que fueron predichas por el modelo en el mismo postestrato $j$. Es decir:

$$\hat{\bar{y}}_j = \frac{\sum_i \hat{p}_{ij}}{n_j}$$

El punto más importante de esta técnica es el ajuste de las diferencias entre la muestra y la población. Para llevar a cabo el proceso de postestratificación se utilizan los conteos censales para cada postestrato, es decir, cuántas personas en la fuerza de trabajo hay en cada una de las 1.600 combinaciones de todos los posibles cruces de las variables auxiliares. Estas cantidades se expresarán como $N_1,\ldots,N_j,\ldots,N_{1600}$ . Finalmente, la estimación de la intención de la tasa de desocupación nacional está dada por:

$$\hat{\bar{\bar{y}}} = \frac{\sum_{j=1}^{1600} N_j \, \hat{\bar{y}}_j}{\sum_{j=1}^{1600} N_j}$$

## E. Conclusiones

En este documento se presenta un enfoque de simulación que puede ser aplicado por las oficinas nacionales de estadística para corregir los posibles sesgos de selección en los levantamientos no presenciales de las encuestas de hogares, en medio de las restricciones a la movilidad que imperan en los países de la región.

En caso de que en el levantamiento se haya utilizado una muestra probabilística de un período anterior (panel), se recomienda usar para la eliminación del sesgo alguna de las dos primeras metodologías mostradas, ajuste por probabilidad de respuesta o calibración en dos etapas. Si, en cambio, no se utilizó un panel, se recomienda optar por el método MRP, como último recurso para proveer al país de cifras orientadoras de políticas públicas.

## Bibliografía


CEPAL (Comisión Económica para América Latina y el Caribe) (2020), "Recomendaciones para la publicación de estadísticas oficiales a partir de encuestas de hogares frente a la coyuntura de la enfermedad por coronavirus (COVID-19)", abril [en línea] https://repositorio.cepal.org/bitstream/handle/11362/45372/4/2000274_es.pdf.

Deville, J. C. y C. E. Särndal (1992), "Calibration estimators in survey sampling", *Journal of the American Statistical Association,* vol. 87, Nº 418.

Gelman, A. y T. Little (1997), "Poststratification into many categories using hierarchical logistic regression", *Survey Methodology*, vol. 27, Nº 2.

Kim, J. K. y M. K. Riddles (2012), "Some theory for propensity-score-adjustment estimators in survey sampling", *Survey Methodology,* vol. 38, Nº 2.

Lensvelt-Mulders, G., P. Lugtig y M. Hubregtse (2009), "Separating selection bias and non-coverage in Internet panels using propensity matching", *Survey Practice,* 2, Nº 6.

OIT (Organización Internacional del Trabajo) (2013), "Resolución sobre las estadísticas del trabajo, la ocupación y la subutilización de la fuerza de trabajo", XIX Conferencia Internacional de Estadísticos del Trabajo, Ginebra, octubre [en línea] http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_234036.pdf.

Park, D., A. Gelman y J. Bafumi (2004), "Bayesian multilevel estimation with poststratification: State-level estimates from national polls", *Political Analysis,* vol. 12, Nº 4.

Särndal, C. E. (2007), "The calibration approach in survey theory and practice", *Survey Methodology,* vol. 33, Nº 2.

Särndal, C. E. y S. Lundström (2006), "Estimation in surveys with nonresponse", *Wiley Series in Survey Methodology,* Wiley.

Rosenbaum, P. R. y D. B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika,* vol. 70, Nº 1.



Este documento es parte de un conjunto de informes elaborados por la Comisión Económica para América Latina y el Caribe (CEPAL) sobre la evolución y los efectos de la pandemia del COVID-19 en América Latina y el Caribe. Fue preparado por la División de Estadísticas, dirigida por Rolando Ocampo, bajo la coordinación general de Alicia Bárcena, Secretaria Ejecutiva de la CEPAL.





CEPAL

Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org