

Nota Metodológica #007

Factores de expansión en los datos del Barómetro de las Américas

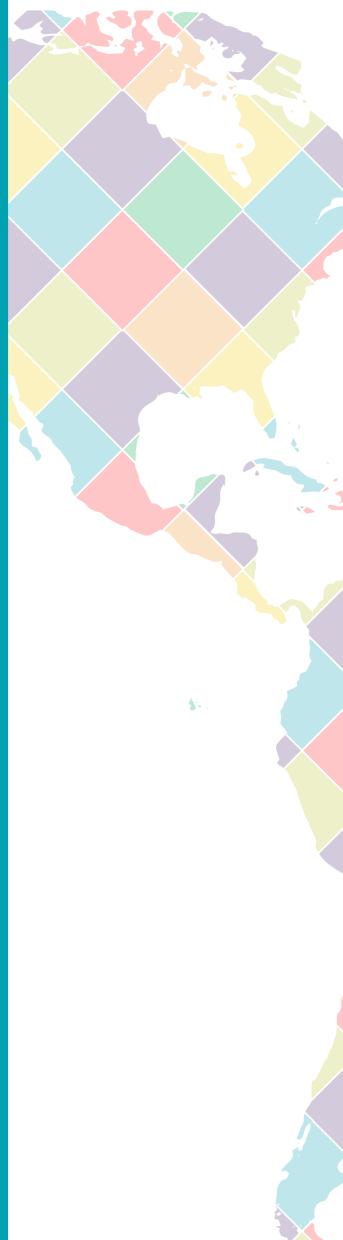
Oscar Castorena

Vanderbilt University

26 de enero de 2021

Hallazgos principales:

- Se describe los factores de expansión incluidos en los datos del Barómetro de las Américas, junto con explicaciones de los tipos de ajustes que esos factores de expansión efectuan a los datos en bruto de la encuesta.
- Se proporciona recomendaciones sobre cómo usar correctamente los diferentes factores de expansión en los datos dado los diferentes análisis que un investigador podría realizar.
- Se utiliza un análisis ilustrativo para resaltar las implicaciones de omitir factores de expansión en el análisis de encuesta, lo que puede afectar tanto las estimaciones puntuales como los errores estándar.



Los factores de expansión son una característica común de las bases de datos de encuestas. Estos permiten a los investigadores hacer inferencias sobre poblaciones de interés que de otra forma presentarían fallas si se realizaran análisis simples de la información de las encuestas sin ponderar. Puesto que tales inferencias son el propósito de la investigación por muestreo en opinión pública, es importante entender el propósito y la implementación apropiada de los factores de expansión, así como de otras características de un diseño de muestra complejo. Sin embargo, un meta-análisis reciente encuentra errores analíticos (p.ej.: análisis que ignoran características complejas del diseño, tales y como los factores de expansión, la estratificación o el muestreo por conglomerados) como una característica recurrente en el análisis secundario de la información de opinión pública¹. Más aún, como argumentan West, Sakshaug y Kim (2017), estos errores analíticos limitan los considerables recursos que se invierten en reducir las múltiples fuentes de error en la recolección de encuestas de opinión pública como lo resume el paradigma del error total de la encuesta². Esta Nota Metodológica tiene el objetivo de prevenir tales errores analíticos al ofrecer una revisión general de los factores de expansión incluidos en los datos del Barómetro de las Américas. Esta revisión clasifica las ponderaciones en las bases de datos y discute los ajustes que realizan a la información de las encuestas sin ponderar. En el apéndice de esta nota se incluye una tabla resumiendo el uso de los factores de expansión para todos los países y las rondas del Barómetro de las Américas. Asimismo, esta nota ofrece un ejemplo ilustrativo de las implicaciones que tiene para el análisis el ignorar el diseño complejo de las encuestas del Barómetro de las Américas.

Descripción de los factores de expansión del Barómetro de las Américas

En esta sección, se describen los diferentes tipos de factores de expansión usados en la base de datos del Barómetro de las Américas en términos de los ajustes que se deben hacer. Se usan encuestas de Ecuador

como ejemplo de diferentes factores de expansión/ajustes en la información del Barómetro de las Américas. El material complementario en el apéndice de esta nota incluye el código de Stata para reproducir estos ejemplos. Ecuador es un caso útil, pues a lo largo del tiempo, los factores de expansión han sido usados con distintos propósitos. La serie histórica de Ecuador también incluye muestras auto-ponderadas.

Los factores de expansión incluidos en las bases de datos del Barómetro de las Américas generalmente buscan ajustar los datos en una de tres formas en y, dependiendo del análisis, una combinación de las tres. El primer tipo de factor de expansión es el *ajuste de post-estratificación*. En este caso, la muestra final obtenida a través de la encuesta se desvía del diseño de la muestra en características sociodemográficas importantes. La mayoría de las encuestas del Barómetro de las Américas usa emparejamiento de frecuencias para la selección de los entrevistados al interior de los hogares, lo que produce muestras que usualmente no requieren ajustes de post-estratificación³. Algunas muestras, sin embargo, usan una selección aleatoria dentro de cada hogar. Como resultado, la distribución en la muestra obtenida puede que no coincida con la distribución de la población. Un ejemplo común es la desviación en la composición de género de la muestra pues los entrevistados varones tienen una menor probabilidad de estar en casa durante el horario de trabajo de campo⁴. Estas desviaciones requieren el uso de factores de expansión para lograr que las proporciones ajustadas de las características sociodemográficas básicas (usualmente el sexo, grupo de edad, y/o ubicación urbana/rural) concuerden con las proporciones de la población. En el Barómetro de las Américas, la post-estratificación se implementa a nivel del estrato primario (la variable estratopri en la base de datos) de tal forma que las proporciones ajustadas coincidan con las proporciones en la población al nivel del estrato primario. Esto es importa porque las encuestas del Barómetro de las Américas están diseñadas para ser representativas a nivel del estrato primario.

Las encuestas del Barómetro de las Américas de Ecuador para las rondas 2004 y 2006 incluyen factores de expansión que realizan ajustes de post-estratificación. La recolección de información a través de esas encuestas

utilizó selección aleatoria dentro del hogar y, en consecuencia, la muestra final presenta desbalances por género. La Tabla 1 resume los porcentajes de cada categoría de género en las encuestas sin ponderar de Ecuador (2004-2012). En las dos primeras rondas, existe una sobrerepresentación de mujeres entrevistadas. Las encuestas realizadas después de la ronda 2006 usaron un emparejamiento de frecuencias a nivel del hogar y, por lo tanto, no presentan tales desbalances en la muestra final. La Tabla 2 resume la distribución por género usando los factores de expansión incluidos en las bases de datos. Vale la pena notar que los ajustes post-estratificación de los factores de expansión llevan las distribuciones de género de las dos primeras rondas a coincidir con las proporciones de referencia en la población. Cualquier análisis de la distribución de las variables de las encuestas de Ecuador de 2004 a 2006 que ignore los factores de expansión puede resultar en estimaciones sesgadas si la variable de interés está correlacionada con el género.

Tabla 1: Distribución del género en las encuestas de Ecuador sin ponderar

	2004	2006	2008	2010	2012
Hombre	40.47	42.53	50.00	49.22	49.93
Mujer	59.53	57.47	50.00	50.78	50.07

Nota: 2004 N = 3,000; 2006 N = 2,925; 2008 N = 3,000; 2010 N = 2,999; 2012 N = 1,500.

Tabla 2: Distribución del género usando factores de expansión en las encuestas de Ecuador

	2004	2006	2008	2010	2012
Hombre	49.00	49.00	49.00	49.00	49.93
Mujer	51.00	51.00	51.00	51.00	50.07

Nota: 2004 N = 3,000; 2006 N = 2,925; 2008 N = 3,000; 2010 N = 2,999; 2012 N = 1,500.

El segundo tipo de factor de expansión común en los datos de la encuesta del Barómetro de las Américas es uno destinado a ajustar por cualquier sobremuestreo que haya sido parte del diseño de muestra. A diferencia de los ajustes por posestratificación descritos arriba, este tipo de factor

de expansión se construye de manera planificada en las bases de datos finales antes de iniciar trabajo de campo. El Laboratorio LAPOP ocasionalmente incluye en sus encuestas sobremuestras de subpoblaciones de interés. Estas pueden ser necesarias si LAPOP o sus socios locales están interesados en estudiar las actitudes de subpoblaciones cuya representación en una muestra típica representativa a nivel nacional puede ser muy pequeña para obtener estimaciones precisas de las opiniones o relaciones entre variables. Un ejemplo de este tipo de sobremuestra es la encuesta 2010 en Chile, la cual incluyó una sobremuestra de las áreas afectadas por un terremoto. Puesto que dichas áreas afectadas por el terremoto estarían sobrerepresentadas, un factor de expansión es necesario para cualquier análisis que busque hacer inferencias sobre Chile a nivel nacional. Específicamente, ese factor de expansión otorga menos peso a las respuestas de las áreas afectadas por el terremoto (la sobremuestra) y da más peso a otras áreas.

La encuesta del Barómetro de las Américas para Ecuador entre 2004 y 2010 también incluye factores de expansión que ajustan los datos por sobremuestras en el diseño de la muestra. En este caso, la región oriental, la cual incluye la Amazonía, contó con una sobremuestra. Debido a que esta zona es muy poco habitada, una muestra representativa a nivel nacional usualmente incluirá muy pocos entrevistados como para poder hacer inferencias estadísticas con precisión sobre esta subpoblación. En la serie histórica del Barómetro de las Américas, la región oriental incluyó una sobremuestra entre 2004 y 2010 para permitir hacer tales inferencias. La Tabla 3 resume las distribuciones sin ponderar de las encuestas en Ecuador (2004-2012). Contrastá con esta distribución con los porcentajes ponderados en la Tabla 4, los cuales ajustan la distribución de acuerdo con los valores de referencia en la población. Se debe tener en cuenta que las estimaciones para 2012 (que no incluyen una sobremuestra y, por tanto, corresponden a una muestra auto-ponderada) son idénticas para las distribuciones ponderadas y sin ponderar. Cualquier análisis de las distribuciones de las variables entre 2004 y 2010 de las encuestas en Ecuador que ignore los factores de expansión puede resultar en estimaciones sesgadas si la variable de interés está correlacionada con la región

geográfica.

Tabla 3: Distribución regional de las encuestas de Ecuador sin ponderar

	2004	2006	2008	2010	2012
Costa	44.23	44.21	44.23	43.88	50.33
Montaña	39.83	39.45	39.83	39.71	45.00
Este	15.93	16.34	15.93	16.41	4.67

Nota: 2004 N = 3,000; 2006 N = 2,925; 2008 N = 3,000; 2010 N = 2,999; 2012 N = 1,500.

Tabla 4: Distribución de las encuestas de Ecuador usando factores de expansión

	2004	2006	2008	2010	2012
Costa	50.91	50.58	49.58	49.20	50.33
Montaña	45.08	45.39	46.31	46.66	45.00
Este	4.00	4.03	4.11	4.14	4.67

Nota: 2004 N = 3,000; 2006 N = 2,925; 2008 N = 3,000; 2010 N = 2,999; 2012 N = 1,500.

Cada uno, o la combinación de los dos anteriores ajustes componen las muestras clasificadas como ponderadas en la información técnica de LAPOP para cada encuesta. La variable ponderadora que debe usarse para análisis de un solo país-año es la variable “wt” en cada base de datos. La mayoría de las encuestas del Barómetro de las Américas son auto-ponderadas. Esto significa que no requieren ninguno de los ajustes discutidos anteriormente y todas las observaciones en dichos datos tienen una variable “wt” con un valor de uno (todos los entrevistados tienen el mismo peso)⁵.

El tipo final de factor de expansión incluido en los datos del Barómetro de las Américas está dirigido a análisis de múltiples encuestas (ya sean varios países o diferentes rondas). La variable ponderadora que debe usarse para este tipo de análisis es la variable “weight1500” en cada base de datos. Puesto que algunas encuestas (gracias al sobremuestreo) pueden tener mucho más de 1.500 entrevistas, estas encuestas pueden tener una mayor influencia que otras al realizar inferencias sobre la región en

su conjunto. El ajuste “weight1500” trata cada encuesta como si tuviera 1.500 entrevistados al realizar el análisis agregado⁶. Es importante notar que este ponderador ajusta los análisis regionales o agregados para que cada país tenga el mismo peso y no pondera cada país de acuerdo con al tamaño de su población. Por ejemplo, cuando se usa “weight1500” para ejecutar análisis agregados, la muestra de Uruguay tiene la misma influencia que la muestra de Brasil⁷.

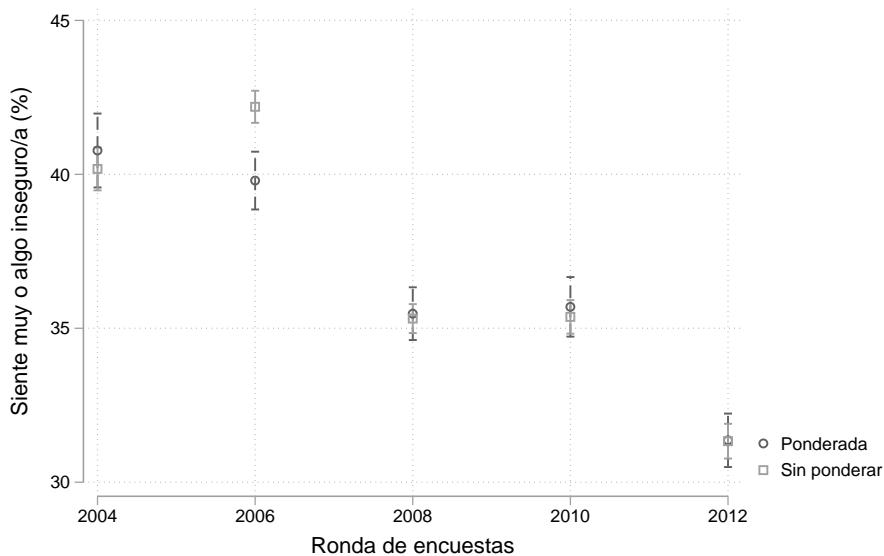
Se recomienda que cuando un investigador trabaje con la base de datos agregada de cualquier tipo (i.e., una base de datos que contiene más de un estudio ya sea de varios países o de diferentes rondas), este utilice “weight1500” como la variable de ponderación. Esto incluye instancias donde el investigador está usando la base de datos agregada para analizar un estudio en particular. En tales casos, “weight1500” no afectará ninguna de las inferencias que se hagan a partir de un análisis de un estudio en particular. Este es el caso porque “weight1500” se aplica uniformemente a todos los entrevistados de un estudio dado, multiplicando la variable “wt” por un valor constante. Por lo tanto, cuando se analiza un único estudio, utilizar “weight1500” será equivalente a usar “wt”⁸.

¿Qué implica (No)-usar factores de expansión en el análisis de encuestas de opinión pública?

La anterior sección ofreció una descripción de la naturaleza y el propósito del uso de factores de expansión en los datos del Barómetro de las Américas. Continuando con el ejemplo de Ecuador, esta sección ilustrará las implicaciones para el análisis de no tomar en cuenta los factores de expansión en los datos y, de forma más general, el diseño complejo de encuesta del estudio del Barómetro de las Américas. Dado que el Barómetro de las Américas incluye un conjunto básico de preguntas en cada ronda que tiene sus orígenes en 2004 en algunos países, estos datos son especialmente útiles para los investigadores interesados en los patrones de la opinión pública a lo largo del tiempo. Por esta razón, se realizan dos análisis de una pregunta a lo largo del tiempo. Uno tiene en cuenta los

factores de expansión y el diseño complejo de las encuestas⁹ y otro que trata la información como si se tratara de una muestra aleatoria simple.

Para tal efecto, se analiza una pregunta sobre la inseguridad en el barrio (**AOJ11**) que se viene incluyendo desde 2004. La pregunta dice “Hablando del lugar o el barrio donde usted vive y pensando en la posibilidad de ser víctima de un asalto o robo, ¿usted se siente muy seguro(a), algo seguro(a), algo **inseguro(a)** o muy **inseguro(a)**?” El Gráfico 1 muestra el porcentaje de entrevistados que responden ya sea muy inseguro o algo inseguro a lo largo de las distintas rondas de encuestas (acá hay un enfoque en el período entre 2004 y 2012). Los símbolos circulares representan las estimaciones de un análisis basado en el diseño (sin ponderar) y los símbolos cuadrados representan estimaciones de un análisis que ignora el diseño complejo de la muestra.



Fuente: El Barómetro de las Américas, LAPOP, 2004-2012

Gráfico 1: Percepción de la inseguridad en el barrio, Ecuador 2004-2012

El primer patrón para notar en el Gráfico es que entre 2004 y 2010 las estimaciones sin ponderar son menores que las estimaciones pondera-

das. Como se discutió anteriormente, las encuestas en Ecuador entre 2004 y 2010 incluyen una sobremuestra de la región oriental del país. Puesto que esta zona del país está sobrerepresentada en estas rondas de la encuesta, cualquier análisis de variables donde existan diferencias significativas entre regiones puede resultar en estimaciones sesgadas. Esto parece explicar la tendencia descendiente en las estimaciones de inseguridad en el barrio en el Gráfico 1. La región oriental, que incluye la Amazonía y que es en general más rural que el resto del país, tiene niveles considerablemente menores de inseguridad. Para el período entre 2004 y 2012, 30.5 % de los entrevistados en la región oriental reportaron sentirse inseguros, comparados con 42.6 % para el resto del país. La sobrerepresentación de los entrevistados de esta región produce un sesgo hacia abajo en las estimaciones del análisis sin ponderaciones.

Un segundo aspecto a destacar en el gráfico son los intervalos de confianza del 95 % alrededor de cada estimación (ponderada y no ponderada). La diferencia entre ambos valores es más evidente en las estimaciones para la ronda 2012. En este caso, la encuesta es autoponderada, por lo que los análisis ponderados y sin ponderar producen los mismos resultados. Sin embargo, es también claro que dichos análisis resultan en intervalos de confianza diferentes. Las razones de ello es que el análisis sin ponderación trata la información como si se fuera una muestra aleatoria simple, mientras que las estimaciones ponderadas se derivan de un análisis que tiene en cuenta el diseño complejo del Barómetro de las Américas, el cual incluye estratos, muestreo por conglomerados y en algunos casos factores de expansión¹⁰. Este análisis basado en el diseño tiene en cuenta los efectos de este diseño de la muestra en la estimación de varianzas y generación de los intervalos de confianza¹¹. Estas diferencias en las varianzas y en los intervalos de confianza pueden afectar significativamente las inferencias que los investigadores intenten llevar a cabo a partir de la información de las encuestas.

Conclusión

Esta nota ofrece una revisión general de los diferentes tipos de factores de expansión incluidos en los datos del Barómetro de las Américas, incluyendo factores de expansión pensados como ajustes de post-estratificación, factores de expansión con el propósito de ajustar sobreencuesteos y factores de expansión con el fin de ajustar los datos para los análisis entre países y/o a lo largo del tiempo. Estos factores de expansión se resumen en el material complementario del apéndice para todos los países y rondas del Barómetro de las Américas. Usando el ejemplo de las encuestas de Ecuador, esta nota demuestra las implicaciones de un análisis que ignora los factores de expansión, así como el diseño complejo de la encuesta. El ejemplo de la inseguridad en el barrio en Ecuador para el período entre 2004 y 2012 demuestra que los análisis sin ponderar pueden resultar en estimaciones sesgadas. Asimismo, las diferencias en los intervalos de confianza de las estimaciones de la ronda 2012 demuestran la importancia de tener en cuenta los efectos de diseño incluso cuando se trataba con muestras autoponderadas. Este último punto es especialmente importante si se tiene en cuenta que la gran mayoría de encuestas del Barómetro de las Américas son autoponderadas. Incluso al trabajar con esas muestras autoponderadas, los investigadores deben ejecutar análisis que incorporen los efectos del diseño muestral.

Tabla 5: Resumen de los factores de expansión en los datos del Barómetro de las Américas

Country	2004	2006	2008	2010	2012	2014	2016	2018
México	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Guatemala	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
El Salvador	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Honduras	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	ponderada ⁺	autoponderada	autoponderada
Nicaragua	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	ponderada ⁺	autoponderada	autoponderada
Costa Rica	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Panamá	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	ponderada ⁺	autoponderada	autoponderada
Colombia	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Ecuador	ponderada ⁺⁺	autoponderada	autoponderada	autoponderada				
Bolivia	ponderada ⁺⁺	autoponderada	autoponderada					
Perú	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	ponderada ⁺
Paraguay	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Chile	—	autoponderada	autoponderada	autoponderada	ponderada ⁺	ponderada*	autoponderada	autoponderada
Uruguay	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Brasil	—	autoponderada	autoponderada	autoponderada	ponderada ⁺	ponderada ⁺	ponderada ⁺	ponderada ⁺
Venezuela	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	—
Argentina	—	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Rep. Dom.	ponderada*	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Haití	—	autoponderada	autoponderada	autoponderada	ponderada ⁺	ponderada ⁺	ponderada ⁺	—
Jamaica	—	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada	autoponderada
Guyana	—	autoponderada	—	ponderada ⁺	autoponderada	autoponderada	ponderada ⁺	—
Trinidad y Tobago	—	—	—	—	ponderada ⁺	ponderada ⁺	autoponderada	—
Belice	—	—	—	autoponderada	autoponderada	autoponderada	autoponderada	—
Estados Unidos	—	autoponderada	ponderada*	ponderada*	ponderada*	ponderada*	ponderada*	ponderada*
Canadá	—	autoponderada	ponderada*	ponderada*	ponderada*	ponderada*	ponderada*	ponderada*
Surinam	—	—	—	—	—	—	—	—
Bahamas	—	—	—	—	—	—	—	—
Barbados	—	—	—	—	—	—	—	—
Granada	—	—	—	—	—	—	—	—
Santa Lucia	—	—	—	—	—	—	—	—
Dominica	—	—	—	—	—	—	—	—
Antigua y Barbuda	—	—	—	—	—	—	—	—
San Vincente y las Granadinas	—	—	—	—	—	—	—	—
San Cristóbal y Nieves	—	—	—	—	—	—	—	—

* ajuste por post-estratificación ⁺ factores de expansión para ajustar sobremuestra en el diseño

Apéndice: Código de Stata para usar los factores de expansión en los datos del Barómetro de las Américas

Sintaxis para los resultados en la Nota Metodológica

Stata incluye un conjunto de comandos y características que facilitan trabajar con información de encuestas complejas. Con el fin de usar estas características, primero se debe declarar el diseño muestral de la base de datos. Esto se hace usando el comando `svyset`¹². Cuando se usan los datos agregados (i.e., datos de múltiples países y/o varios años), como en los análisis para esta nota, la sintaxis específica y los resultados de Stata son los siguientes:

```
. svyset upm [pw=weight1500], strata(strata)
pweight: weight1500
VCE: linearized
Single unit: missing
Strata 1: strata
SU 1: upm
FPC 1: <zero>
```

Aunque el lector debe consultar la documentación de Stata para obtener detalles sobre el comando `svyset`, existen algunos aspectos de la sintaxis arriba que vale la pena mencionar. Primero, “`upm`” declara la variable en la base de datos que identifica las unidades primarias de muestreo en los datos del Barómetro de las Américas. Recuerde que, puesto que el Barómetro de las Américas (2004-2019) se lleva a cabo encuesta cara-a-cara, las encuestas se recogen en conglomerados geográficos. Esta parte del comando `svyset` identifica dichos conglomerados. Segundo, como se recomienda en esta nota para las bases de datos agregadas, “`weight1500`” se define como el factor de expansión de la muestra (`[pw=weight1500]`).

Finalmente, la opción `strata()` se usa para especificar el estrato primario en el diseño de la muestra, el cual se identifica en este caso con la variable llamada “strata” en los datos del Barómetro de las Américas. Como se discute en la nota, estas tres características del diseño muestral (factores de expansión, conglomerados y estratos) tienen un impacto en el cálculo tanto de estimaciones puntuales como de los errores estándar.

Las bases de datos descargadas de LAPOP deben venir con los parámetros del diseño muestral predefinidos. Uno puede usar el comando `svydescribe` para verificar que las características del diseño (tales como la variable ponderadora y el estrato primario) son adecuadas para los datos y el análisis.

Una vez que se declare la estructura de los datos de la encuesta en Stata, podemos comenzar a usar `svy`, el comando prefijo para las encuestas, para realizar análisis que tomen en cuenta el diseño muestral. Sin embargo, primero mostramos la sintaxis de un análisis “ingenuo” en la Tabla 1 que refleja la distribución de género sin ponderar a lo largo de las rondas de encuestas en Ecuador. Nótese que la sintaxis abajo restringe el análisis a las observaciones de Ecuador (el cual se identifica con el valor 9 en la variable “país” en el conjunto de datos) y para la ronda 2012 y anteriores (esto se hace usando la condición `if` después del comando de tabulación). Estas restricciones son necesarias en la sintaxis porque estamos trabajando con los datos agregados que incluyen todos los países y rondas.

```
.tab q1 wave if pais == 9 & wave <= 2012, col nofreq
```

Como se discute en el texto principal, la sintaxis arriba trata los datos como si se tratara de una muestra simple aleatoria en lugar de una encuesta polietápica con factores de expansión, conglomerados y muestreo estratificado. La sintaxis abajo produce los resultados de la Tabla 2, la cual tiene en cuenta los efectos de diseño sobre la distribución del género entre rondas de la encuesta en Ecuador. Nótese que la única diferencia con la anterior sintaxis es la adición del prefijo `svy` antes del comando `percent` (el cual ejecuta el análisis con los efectos de diseño) y la inclusión

de la opción `percent` (la cual produce porcentajes en la tabla de resultados en vez de las proporciones usuales cuando se usa `svy: tab`). El prefijo `svy` simplifica considerablemente la ejecución en Stata de análisis con datos de encuestas que incorporen los efectos de diseño.

```
. svy q1 wave if pais == 9 & wave <= 2012, col percent
```

La sintaxis de abajo reproduce los resultados de la Tabla 3 en la nota, los cuales reflejan la distribución sin ponderar de los estratos regionales en las encuestas de Ecuador para las distintas rondas.

```
. tab estratopri wave if pais == 9 & wave <= 2012, col nofreq
```

Las distribuciones correctas (i.e., ponderadas) de las tres regiones en la Tabla 4 se producen con la sintaxis de abajo, la cual usa el prefijo `svy`.

```
. svy: tab estratopri wave if pais == 9 & wave <= 2012, col
```

Con el fin de reproducir los resultados en el Gráfico 1 de la nota, uno debe primero recodificar la variable “aoj11” del Barómetro de las Américas para pasar a ser una variable dicotómica como se describe en el texto principal. Esto se realiza usando la sintaxis de abajo, la cual genera una nueva variable llamada “aoj11_recode”.

```
. recode aoj11 (1 2 =0) (3 4=100), gen(aoj11_recode)
```

Las estimaciones sin ponderar del Gráfico 1 se generan usando el comando `mean`. La sintaxis de abajo ignora la necesidad de los factores de expansión de la encuesta, así como la estructura de los datos que incluye conglomerados y estratificación. También se presenta el resultado al ejecutar el comando.

```
. mean aoj11_recode if pais == 9 & wave <= 2012, over(wave)
```

Mean estimation

Number of obs = 13,324

```
_subpop_1: wave = 2004
_subpop_2: wave = 2006
_subpop_3: wave = 2008
```

```
_subpop_4: wave = 2010
_subpop_5: wave = 2012
```

Over		Mean	Std. Err.	[95 % Conf. Interval]
------	--	------	-----------	-----------------------

aoj11_recode				
_subpop_1	34.96456	.8761865	33.24711	36.68201
_subpop_2	44.7269	.9217111	42.92021	46.53358
_subpop_3	40.30852	.898408	38.54751	42.06953
_subpop_4	37.50419	.8868644	35.76581	39.24257
_subpop_5	38.13046	1.259984	35.66072	40.60021

Con el objetivo de ejecutar un análisis que incorpore los efectos de diseño (p.ej.: ponderado) de las tendencias a lo largo del tiempo de la inseguridad en el barrio, uno simplemente necesita agregar el prefijo `svy` como en la sintaxis y el resultado de abajo.

```
. svy: mean aoj11_recode if pais == 9 & wave <= 2012, over(wave)
(running mean on estimation sample)
```

```
Survey: Mean estimation
Number of strata = 15                               Number of obs = 13,286
Number of PSUs = 449                                Population size = 7,422.2153
Design df = 434
```

```
_subpop_1: wave = 2004
_subpop_2: wave = 2006
_subpop_3: wave = 2008
_subpop_4: wave = 2010
```

```
_subpop_5: wave = 2012
```

Linearized

Over		Mean	Std. Err.	[95% Conf. Interval]
------	--	------	-----------	----------------------

aoj11_recode				
_subpop_1	38.84717	2.040934	34.83582	42.85851
_subpop_2	47.98135	2.150574	43.75451	52.20819
_subpop_3	42.56732	1.468479	39.6811	45.45354
_subpop_4	42.93966	1.370067	40.24687	45.63245
_subpop_5	38.13046	1.75099	34.68899	41.57194

Consecuencias de usar el factor de ponderación incorrecto

En el texto principal de esta nota metodológica se recomienda que los investigadores siempre usen “weight1500” como la variable ponderadora al usar las bases de datos agregadas del Barómetro de las Américas. El siguiente ejemplo ilustra las consecuencias de usar la variable de ponderación equivocada (en este caso “wt” en vez de “weight1500”). Para hacerlo, se usa una gráfica referenciada en la nota.

En la discusión del Gráfico 1 en el texto principal, se mencionó que en el periodo entre 2004 y 2012, el 30.5 % de los entrevistados de la región oriental reportaron sentirse inseguros en comparación con el 42.6 %

en el resto del país. Estos porcentajes fueron calculados mediante un análisis que incorpora los efectos de diseño al usar el prefijo `svy` en Stata. Antes de ejecutar el análisis, una variable dicotómica para la región este es creada mediante la siguiente sintaxis.

```
. gen east = 1 if estratopri == 903 & pais == 9
. replace east = 0 if estratopri != 903 & pais == 9
```

La siguiente sintaxis y resultados producen las cifras 30.5% y 42.6% reportadas en el texto central. Nótese que debido a que la variable de ponderación es “weight1500”, cada ronda de la encuesta de Ecuador tiene el mismo peso al calcular el promedio del periodo. Asimismo, esto es especialmente necesario puesto que la encuesta de Ecuador antes de 2012 tenía sobremuestras grandes.

```
. svy: mean aoj11_recode if pais == 9 & wave <= 2012, over(east)
(running mean on estimation sample)
```

```
Survey: Mean estimation
Number of strata = 15                               Number of obs = 13,286
Number of PSUs = 449                                Population size = 7,422.2153
Design df = 434
```

```
0: east = 0
1: east = 1
```

Linearized

Over		Mean	Std. Err.	[95% Conf. Interval]
------	--	------	-----------	----------------------

```
aoj11_recode |
 0 | 42.59942 .8310237 40.96609 44.23275
 1 | 30.46638 2.507745 25.53754 35.39522
```

Para ilustrar cómo el uso de los factores de expansión equivocados en el análisis de los datos agregados puede afectar los resultados, se usa el comando `svyset` para definir “wt” como la variable ponderadora en los datos. La sintaxis y los resultados aparecen a continuación:

```
. svyset upm [pw=wt], strata(strata)
pweight: wt
VCE: linearized
Single unit: missing
Strata 1: strata
SU 1: upm
FPC 1: <zero>
```

Ahora volvemos a ejecutar los análisis, pero con la ponderación equivocada y se obtiene el resultado de abajo. En vez de obtener 30.5 % como en el análisis anterior, este análisis resulta en que 31.5 % de los entrevistados de la región oriental reportaron sentirse inseguros en su barrio en el periodo 2004-2012. La diferencia aquí resulta del hecho de que las rondas anteriores con un mayor número de entrevistados tienen más influencia al calcular el promedio para el período en comparación con un análisis que use “weight1500”, el cual asigna el mismo peso a todas las rondas.

```
. svy: mean aoj11_recode if pais == 9 & wave <= 2012, over(east)
(running mean on estimation sample)

Survey: Mean estimation
Number of strata = 15                               Number of obs = 13,286
Number of PSUs = 449                                Population size = 13,281.871
Design df = 434
```

```
0: east = 0
1: east = 1
```

Linearized

Over		Mean	Std. Err.	[95 % Conf. Interval]
------	--	------	-----------	-----------------------

aoj11_recode				
0		42.97838	.8478024	41.31207 44.64469
1		31.46284	2.57176	26.40819 36.5175

Aunque la diferencia puede no ser grande en términos sustantivos en este ejemplo en particular, sí ilustra que usar la variable de ponderación equivocada puede llevar a estimaciones diferentes del valor de interés. Para los investigadores de la opinión pública que quieran minimizar la posibilidad de errores analíticos, es importante tener en cuenta el papel y el propósito de las variables de ponderación para cada tipo de análisis.

Notas

1. West, Sakshaug y Aurelien (2016).
2. Weisberg (2009).
3. Para obtener mayor información sobre la selección de los entrevistados, ver https://www.vanderbilt.edu/lapop/ab2018/AmericasBarometer_2018-19_Technical_Report_W_102919.pdf.

4. Ver Silver y col. (2019) para una discusión de los retos en las encuestas cara a cara.
5. La tabla en el apéndice de esta nota incluye un resumen de los factores de expansión usados en cada estudio del Barómetro de las Américas. Para mayor información sobre el diseño de cada estudio, e información sobre cómo especificar correctamente los factores de expansión y declarar la información de la muestra para el análisis en Stata, ver el documento de información técnica que acompaña a cada estudio.
6. Para obtener mayor información sobre cómo se calcula weight1500, ver https://www.vanderbilt.edu/lapop/docs/AmericasBarometer_weighting_scheme_all_years_of_AB_v2.pdf.
7. Otra alternativa sería ponderar los entrevistados en cada país de manera proporcional al tamaño de su población. Así, continuando con el ejemplo, las respuestas de Uruguay tendrían un peso mucho menor en comparación con las encuestas de Brasil. LAPOP considera que los países son unidades de análisis relevantes y por lo tanto el estándar para análisis regionales es ponderar cada país con el mismo peso.
8. Para acceder a un ejemplo de cómo usar “wt” en lugar de “weight1500” puede afectar los resultados, ver el material complementario en el apéndice.
9. El diseño complejo hace referencia a las características tales como la estratificación y el muestreo por conglomerados que distingue a las muestras del Barómetro de las Américas de muestras aleatorias simples, las cuales no son viables en encuestas cara a cara. Esto se incorpora usando el comando `svy` en Stata; ver la siguiente nota al pie.
10. Estos análisis se realizaron usando Stata 14.2. Un análisis que incorpora los efectos de diseño en Stata implica usar el prefijo `svy` junto con los comandos de estimación (ver ayuda del comando `help svy_estimation` en Stata para mayor información).
11. Pese a que la estratificación usada en la selección de muestras del Barómetro de las Américas reduce la varianza muestral de las estimaciones, el muestreo por conglomerados la aumenta. El efecto general en las varianzas depende del diseño muestral particular, así como a partir de los pesos que la muestra incluye, lo cual puede aumentar la variación muestral.
12. Ver <https://www.stata.com/manuals13/svysvyset.pdf> para información sobre `svyset`.

Referencias

- Silver, Laura, Martha McRoy, Kat Devlin y Patrick Moynian. 2019. «Who's Home and Who Isn't? The Challenges of Conducting Face-to-Face Interviews in Jordan.» Pew Research Center: Decoded. <https://medium.com/pew-research-center-decoded/whos-home-and-who-isn-t-the-challenges-of-conducting-face-to-face-interviews-in-jordan-95a2a7081ded>.
- Weisberg, Herbert F. 2009. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago, IL: University of Chicago Press.
- West, Brady T., Joseph W. Sakshaug y Guy Alain S. Aurelien. 2016. «How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data?» PLOS ONE 11 (6): 1-29.
- West, Brady T., Joseph W. Sakshaug y Yumi Kim. 2017. «Analytic Error as an Important Component of Total Survey Error: Results from a Meta-Analysis.» En *Total Survey Error in Practice*, editado por P. Biemer, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, B. West, E. de Leeuw y N. C. Tucker. Hoboken, NJ: Wiley.

Oscar Castorena (oscar.castorena@vanderbilt.edu) es un investigador postdoctoral en LAPOP.

Este reporte fue editado por la Dra. Elizabeth J. Zechmeister y la Dra. Carole J. Wilson. Este reporte fue traducido por el Dr. Juan Camilo Plata, Mariana V. Ramírez Bustamante y el Dr. J. Daniel Montalvo. La auditoría de este reporte fue hecha por Mariana V. Ramírez Bustamante. El formato, la producción, la revisión, los gráficos y la distribución del reporte fueron manejados por Rubí Arana y Laura Sellers. Nuestros datos e informes están disponibles para su descarga gratuita en el sitio web del proyecto. Síganos en Twitter o Facebook para mantenerse en contacto.

Como miembro fundador de la iniciativa de transparencia de la Asociación Americana para la Investigación de la Opinión Pública (AAPOR), LAPOP se compromete a la divulgación rutinaria de nuestros procesos de recopilación de datos y presentación de informes. Se puede encontrar más información sobre los diseños de la muestra del Barómetro de las Américas en vanderbilt.edu/lapop/core-surveys.

Esta *Nota Metodológica* ha sido posible gracias al apoyo del pueblo de Estados Unidos de América a través de la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID) y Vanderbilt University. Las opiniones expresadas en esta *Nota Metodológica* corresponden al autor y LAPOP y no reflejan necesariamente el punto de vista de USAID, del Gobierno de los Estados Unidos o cualquiera de las otras instituciones financieras. Las encuestas del Barómetro de las Américas de LAPOP son financiadas principalmente por USAID y Vanderbilt University. La ronda 2018/19 también tuvo el apoyo del BID, el PNUD, la Open Society Foundations y varios socios académicos e investigadores a lo largo de las Américas.