Factores de expansión

Arturo Maldonado

20/02/2021

INTRODUCCIÓN

En este documento se explicará cómo introducir el factor de expansión y las ponderaciones en los cálculos estadísticos usando los datos del Barómetro de las Américas. Seguiremos lo que se trabajó en el documento de "Manipulación de datos", disponible aquí.

SOBRE LA BASE DE DATOS

Los datos que vamos a usar deben citarse de la siguiente manera: Fuente: Barómetro de las Américas por el Proyecto de Opinión Pública de América Latina (LAPOP), wwww.LapopSurveys.org. Pueden descargar los datos de manera libre aquí En este enlace, se pueden registrar o entrar como "Free User". En el buscador, se puede ingresar el texto "merge". Ahí se tendrá acceso a la base de datos completa "2004-2018 Grand Merge Free" en versión para STATA. Se descarga la base de datos en formato zip, la que se descomprime en formato .dta. Una vez descargada y guardada en el directorio de trabajo, se tiene que leer la base de datos como un objeto dataframe en R. En este documento se carga una base de datos recortada. Esta base de datos se encuentra alojada en el repositorio "materials_edu" de la cuenta de LAPOP en GitHub.

```
library(rio)
lapop18 <- import("https://raw.github.com/lapop-central/materials_edu/main/LAPOP_AB_Merge_2018_v1.0.sav
```

SOBRE EL USO DE LOS FACTORES DE EXPANSIÓN

Cuando un investigador abre una base de datos en cualquier programa estadístico, el software asume que los datos provienen de un muestro simple aleatorio. Cuando se trabaja con datos de opinión pública, como los datos del Barómetro de las Américas, el diseño muestral generalmente no es un muestreo simple aleatorio, sino un diseño complejo, con estratificaciones, segmentaciones, cuotas, en diferentes etapas. Como indica el reporte técnico de la ronda 2018/19 del Barómetro de las Américas, disponible aquí, las muestras en cada país fueron diseñadas usando un diseño probabilístico multietapa (con cuotas al nivel de hogares para la mayoría de países), y fueron estratificadas por regiones principales en el país, tamaño de la municipalidad y por áreas urbanas y rurales dentro de las municipalidades. Este diseño muestral complejo se tiene que incorporar generalmente en los cálculos. En ciertas ocasiones si no se incopora puede llevar a diferencias en los resultados. Una explicación más detallada sobre el uso de los factores de expansión y las potenciales consecuencias de no usarlos con los datos del Barómetro de las Américas puede ser leída en la Nota Metodológica 007 (Castorena, 2021), disponible aquí. Esta Nota Metodológica describe tres escenarios de usos de factores de expansión:

1. Ajuste post estratificación: cuando la muestra no es autoponderada y se desvía de ciertas características sociodemográficas importantes. En la base de datos, esta característica se ajusta con la variable "estratopri".

2. Ajuste por sobremuestreo: cuando la muestra incluye una sobremuestra de subpoblaciones de interés. En la base de datos, esta característica se ajusta con la variable "wt".
3. Ajuste de múltiples encuestas: cuando se analiza datos de varios países o diferentes rondas. La variable ponderadora en la base de datos es "weight1500", que estandariza las muestras de cada país a 1,500 observaciones.

CONSECUENCIAS DE NO USAR FACTORES DE EXPANSIÓN

Como indica la Nota metodológica, "los análisis sin ponderar pueden resultar en estimaciones sesgadas" (p.9). Por ejemplo, en el documento sobre manipulación de datos replicamos los resultados sobre el apoyo a la democracia en Honduras (45%) y Uruguay (76.2%), para lo que se calculó la variable recodificada y se describió.

```
lapop18$ing4rec <- car::recode(lapop18$ing4, "1:4=0; 5:7=1")
table(lapop18$ing4rec)
```

```
##
##     0     1
## 12261 17828
```

En este dataframe se puede calcular la distribución del apoyo a la democracia en estos dos países y se puede reportar los porcentajes redondeado.

```
round(prop.table(table(lapop18$ing4rec[lapop18$pais==4]))*100, 1)
```

```
##
##  0  1
## 55 45
```

```
round(prop.table(table(lapop18$ing4rec[lapop18$pais==14]))*100, 1)
```

```
##
##    0    1
## 23.8 76.2
```

Se observa que estos resultados son iguales a los que aparecen en el Gráfico 1.2 del reporte "El Pulso de la Democracia" (p.12), disponible aquí. Esto es esperable porque, como indica la Tabla 5 de la Nota Metodológica, ambos países tienen un diseño muestral autoponderado, por lo que estos cálculos aquí, que no incluyen el diseño, coinciden con los del reporte, que sí incluyen el factor de expansión. Un caso diferente es el de Brasil que, según la Nota Metodológica, tiene un diseño muestral ponderado, por lo que sí requeriría usar el factor de expansión para ajustar la sobremuestra en el diseño. Si se calcula el descriptivo del apoyo a la democracia en Brasil sin incluir el factor de expansión se obtiene un resultado distinto al del reporte.

```
round(prop.table(table(lapop18$ing4rec[lapop18$pais==15]))*100, 1)
```

```
##
##    0    1
## 40.2 59.8
```

En este cálculo obtenemos 59.8%, mientras que en el Gráfico 1.2 del reporte se observa 60.0%. Esta diferencia es debida a que el comando `table`, y luego `prop.table`, no incluyen el factor de expansión.

INCLUYENDO EL FACTOR DE EXPANSIÓN

Algunos comandos en R permiten la inclusión de una variable de expansión en los cálculos. El paquete `descr`, por ejemplo, incluye varios comandos, como `compmeans` o `crosstab` que permiten esta inclusión del factor de expansión. Para reproducir los datos que se observan en el Gráfico 1.2 del reporte, se puede usar el comando `compmeans` que permite calcular la media de una variable (como ing4rec, cuya media es igual a la proporción) por grupos de una variable factor, como "pais", ponderando los resultados por una variable, como "weight1500". Se agrega la especificación `plot=FALSE` para desactivar la producción del gráfico.

```
library(descr)
compmeans(lapop18$ing4rec, lapop18$pais, lapop18$weight1500, plot=FALSE)
```

```
## Mean value of "La democracia es mejor que cualquier otra forma de gobierno"
## according to "País"
##            Mean     N Std. Dev.
## 1     0.6272307  1436 0.4837099
## 2     0.4888451  1432 0.5000501
## 3     0.5856655  1454 0.4927762
## 4     0.4501005  1436 0.4976772
## 5     0.5153743  1451 0.4999359
## 6     0.7235940  1457 0.4473736
## 7     0.5380612  1479 0.4987179
## 8     0.5978999  1460 0.4904899
## 9     0.5443122  1479 0.4982010
## 10    0.4914110  1454 0.5000983
## 11    0.4926471  1475 0.5001155
## 12    0.5121786  1463 0.5000225
## 13    0.6387097  1419 0.4805438
## 14    0.7619359  1451 0.4260454
## 15    0.5999750  1470 0.4900697
## 17    0.7110368  1468 0.4534353
## 21    0.5922659  1458 0.4915818
## 23    0.5118871  1334 0.5000461
## 40    0.7173120  1496 0.4504565
## 41    0.7430692  1497 0.4370869
## Total 0.5928825 29072 0.4913055
```

De acuerdo a estos resultados, vemos que Brasil, pais 15, tiene un apoyo a la democracia de 0.599975. Si transformamos este número en porcentaje, aproximando a 1 decimal, reproducimos el valor de 60% que se observa en el Gráfico 1.2 del reporte. No solo eso, además, se observa que para el resto de países, los datos se replican. Por ejemplo, para México, país 1, esta tabla muestra un apoyo a la democracia de 0.6272307, o, en porcentaje aproximado a 1 decimal, 62.7%, igual al dato del reporte.

Otra forma de replicar los resultados incorporando el efecto de diseño es usando el paquete `survey`, paquete especialmente desarrollado para trabajar con diseños muestrales complejos. La Nota Metodológica incluye un apéndice con el código de STATA para usar los factores de expansión en los datos del Barómetro de las Américas. Aquí haremos lo mismo en R, para lo cual usaremos el comando `svydesign` (similar al comando svyset en STATA). Con este comando se crea un nuevo objeto llamado "lapop.design", que guarda la información de las variables contenidas en el dataframe, incluyendo en los cálculos el factor de expansión. Por tanto, si luego se creara una nueva variable, se tendría que correr nuevamente esté comando para que este objeto "lapop.design" incluya esta nueva variable.

```
library(survey)
lapop.design<-svydesign(ids =~upm, strata =~ estratopri, weights = ~weight1500, nest=TRUE, data=lapop18)
```

Una vez creado los datos con el factor de expansión en el objeto "lapop.design", se puede usar los comandos nativos del paquete `survey` para realizar cálculos. Por ejemplo, para calcular la media de la variable "ing4rec" (apoyo a la democracia) en toda la base de datos de la ronda 2018/19, se puede usar el comando `svymean`.

```
svymean(~ing4rec, lapop.design, na.rm=T)
```

```
##            mean    SE
## ing4rec 0.59288 0.003
```

De esta manera se reproduce el valor de la última fila de resultados del comando `compmeans`, que corresponde al promedio de toda la muestra. Es decir, de ambas maneras se está encontrando el mismo resultado. Para reproducir los resultados por país, se puede usar el comando `svyby` que permite hallar resultados (como la media, usando `svymean`) de una variable ("ing4rec"), por valores de otra variable ("pais").

```
svyby(~ing4rec, ~pais, design=lapop.design, svymean, na.rm=T)
```

```
##    pais   ing4rec         se
## 1     1 0.6272307 0.01245940
## 2     2 0.4888451 0.01358318
## 3     3 0.5856655 0.01267273
## 4     4 0.4501005 0.01197688
## 5     5 0.5153743 0.01419558
## 6     6 0.7235940 0.01512205
## 7     7 0.5380612 0.01372306
## 8     8 0.5978999 0.01212261
## 9     9 0.5443122 0.01357881
## 10   10 0.4914110 0.01374835
## 11   11 0.4926471 0.01337323
## 12   12 0.5121786 0.01624846
## 13   13 0.6387097 0.01161029
## 14   14 0.7619359 0.01240878
## 15   15 0.5999750 0.01556882
## 17   17 0.7110368 0.01415857
## 21   21 0.5922659 0.01050698
## 23   23 0.5118871 0.01325745
## 40   40 0.7173120 0.01261872
## 41   41 0.7430692 0.01189836
```

En este caso, vemos que esta tabla es exactamente igual a la reportada con `compmeans`, pues ambas usan el factor de expansión. De esta manera, hemos visto dos maneras de incorporar el efecto de diseño muestral en los cálculos básicos con los datos del Barómetro de las Américas. Más adelante, se verá la inclusión del factor de expansión en otros cálculos más complejos, como el cálculo de intervalos de confianza o de regresiones. En estos documentos se trabajará la versión simple, sin incluir estos efectos y con los comandos más básicos de R, y luego la versión compleja, incluyendo el factor de expansión en los cálculos.