Bootstrap con Stata

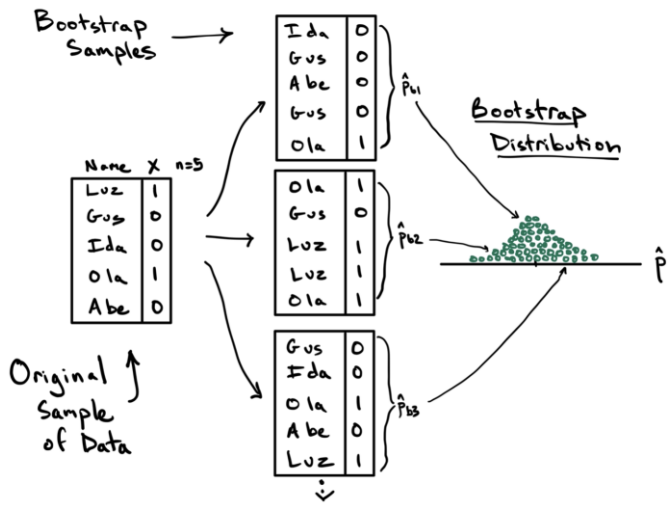
(intervalos de confianza)

Luis Guillen Grados

Chief Data Science

lguillen@geoanalytics.pe

lguilleng@gmail.com



https://www.stat20.org/3-generalization/17-bootstrapping/notes.html

medium.com/@lguilleng

linkedin.com/in/datascientistlg

¿Qué es Bootstrap?

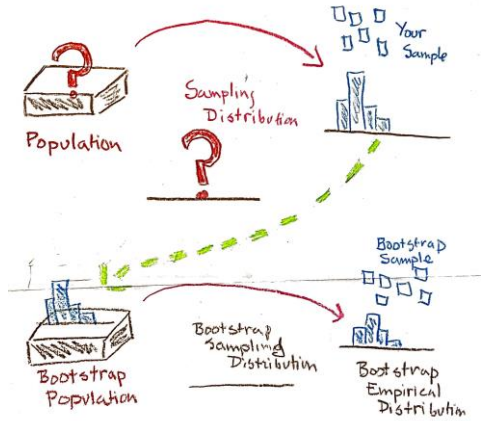
Bootstrap es un método para obtener propiedades (errores estándar, intervalos de confianza y valores críticos) de una distribución muestral de estimadores. Es muy similar a las técnicas de Monte Carlo. Sin embargo, en lugar de especificar completamente el proceso generador de datos (DGP), utilizamos información de la muestra.

El bootstrap se utiliza típicamente para estimadores consistentes pero sesgados. En la mayoría de los casos, conocemos las propiedades asintóticas de estos estimadores. Por lo tanto, podríamos utilizar la teoría asintótica para derivar la distribución muestral aproximada. Eso es lo que hacemos normalmente cuando usamos, por ejemplo, estimadores de máxima verosimilitud.

Pero, ¿qué sucede cuando observa una estadística que no se puede convertir en una suma de variables aleatorias? ¿Qué sucede si el tamaño de su muestra no es grande? ¿De qué otra manera puede aproximar la distribución muestral de una estadística?

Hasta hace varias décadas, la caja de herramientas para responder a estas preguntas era limitada. Sin embargo, con la llegada de computadoras poderosas y una visión brillante y simple de la relación entre la muestra y la población, tenemos una nueva herramienta para evaluar la variabilidad de la muestra. Esa herramienta es el bootstrap.

El bootstrap se basa en la observación de que si su muestra es representativa de la población, entonces la distribución empírica debería ser un buen sustituto de la distribución de la población. Luego, se puede simular el proceso de extraer múltiples muestras de la población extrayendo nuevas muestras (llamadas remuestras) de la distribución empírica.



https://www.stat20.org/3-generalization/17-bootstrapping/notes.html

El algoritmo Bootstrap

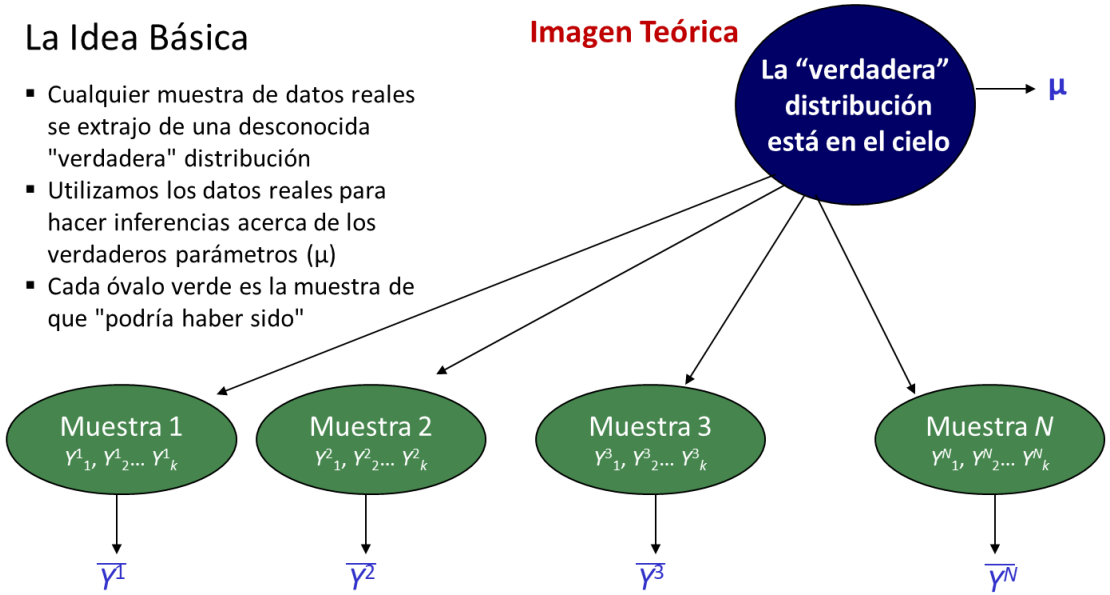
Un procedimiento utilizado para evaluar la variabilidad de muestreo en estadística. Para realizar el bootstrap de una estadística:

1. Tratar la muestra como si fuera la población bootstrap.
2. Extraer una nueva muestra (con reemplazo) de la población bootstrap.
3. Calcular la estadística de interés en la nueva muestra.
4. Repetir los pasos 2 y 3 muchas veces para construir una distribución de muestreo bootstrap.

El nombre del procedimiento se deriva del modismo "levantarse a uno mismo por sus propias botas". Esto ilustra la naturaleza algo milagrosa de este procedimiento. Aunque en realidad solo se obtiene una muestra de la población, el bootstrap permite generar muchas más muestras a través del proceso de muestreo con reemplazo.

La Idea Básica

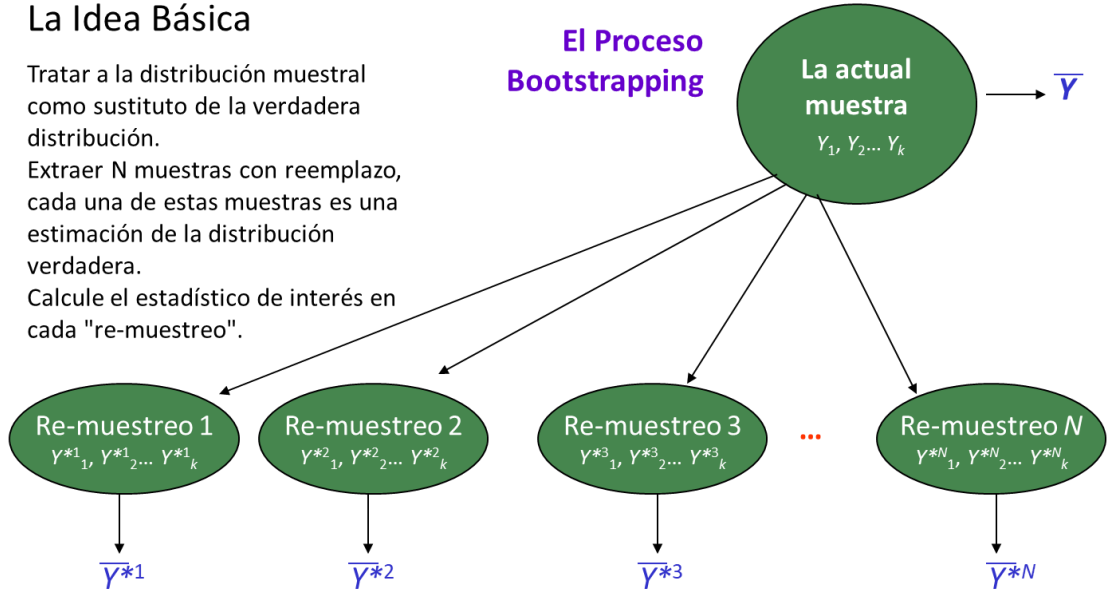
- Cualquier muestra de datos reales se extrajo de una desconocida "verdadera" distribución
- Utilizamos los datos reales para hacer inferencias acerca de los verdaderos parámetros (μ)
- Cada óvalo verde es la muestra de que "podría haber sido"



La distribución de nuestra estimador ($\overline{Y}$) depende tanto de la verdadera distribución y el tamaño (k) de la muestra

La Idea Básica

Tratar a la distribución muestral como sustituto de la verdadera distribución.
Extraer N muestras con reemplazo, cada una de estas muestras es una estimación de la distribución verdadera.
Calcule el estadístico de interés en cada "re-muestreo".



{$\overline{Y}^*$} constituye una estimación de la distribución de $\overline{Y}$.

$$\widehat{se}(\widehat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}^*_b - \widehat{\theta}^*)^2}$$

Intervalos de confianza bootstrap con Stata

```
bootstrap exp_list [, options eform_option] : command
```

Supongamos que deseamos obtener la estimación de bootstrap del error estándar de la mediana de una variable llamada gashog2d. La función de Stata calcula y muestra estadísticas resumidas con el comando "summarize"; calcula medias, desviaciones estándar, sesgo, curtosis y varios percentiles. Entre esos percentiles está el percentil 50: la mediana. Además de mostrar los resultados calculados, summarize los almacena, y al buscar en el manual, descubrimos que la mediana se almacena en r(p50). Para obtener una estimación de bootstrap de su error estándar, todo lo que necesitamos hacer es escribir:

```
bootstrap r(p50), reps(250): summarize gashog2d, detail
(running summarize on estimation sample)

(resultado omitido)

Bootstrap replications (250)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250

Bootstrap results                               Number of obs      =     34,245
                                                Replications       =        250

      command:  summarize gashog2d, detail
        _bs_1:  r(p50)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   20695.92   100.2837   206.37   0.000     20499.37    20892.47
------------------------------------------------------------------------------
```

¿Y si necesitamos estimar el error estándar de la razón entre la media aritmética y mediana de gashog2d?

```
. bootstrap razon=(r(mean)/r(p50)), reps(250): summarize gashog2d, detail
(running summarize on estimation sample)

(resultado omitido)

Bootstrap replications (250)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250

Bootstrap results                               Number of obs      =     34,245
                                                Replications       =        250

      command:  summarize gashog2d, detail
        razon:  r(mean)/r(p50)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       razon |   1.237025    .004802   257.61   0.000     1.227613    1.246436
------------------------------------------------------------------------------
```

Los resultados anteriores, suponen un muestreo aleatorio simple, sin embargo, la ENAHO es una encuesta compleja, por lo que se debe indicar a la técnica de bootstrap el diseño de la muestra (unidad primaria de muestreo, estratificación y ponderación).

```
. bootstrap r(p50), reps(250) cluster(conglome) strata(estrato) force: summarize
> gashog2d [aw=factor07], detail
(running summarize on estimation sample)

(resultado omitido)

Bootstrap replications (250)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250

Bootstrap results

Number of strata   =         8                 Number of obs      =     34,245
                                               Replications       =        250

      command:  summarize gashog2d [aweight= factor07], detail
        _bs_1:  r(p50)

                           (Replications based on 5,359 clusters in conglome)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   22232.19   192.8518   115.28   0.000     21854.21    22610.17
------------------------------------------------------------------------------
```

```
. bootstrap razon=(r(mean)/r(p50)), reps(250) cluster(conglome) strata(estrato)
> force: summarize gashog2d [aw=factor07], detail
(running summarize on estimation sample)

(resultado omitido)

Bootstrap replications (250)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250

Bootstrap results

Number of strata   =         8                 Number of obs      =     34,245
                                               Replications       =        250

      command:  summarize gashog2d [aweight= factor07], detail
        razon:  r(mean)/r(p50)

                           (Replications based on 5,359 clusters in conglome)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       razon |   1.201766   .0069746   172.31   0.000     1.188096    1.215436
------------------------------------------------------------------------------
```

Note que, ahora las replicaciones se basan en el número de unidades primarias de muestreo en la muestra (5,359 conglomerados).

Además, los intervalos de confianza, por defecto se construyen asumiendo que la distribución del estimador en el muestreo sigue una normal con media y varianza obtenida de la muestra (Teorema del Límite Central).

¿Cómo estimar el error estándar en presencia de datos imputados?

```
// Suponiendo que la variable GASHOG2D no ha sido imputada
// creamos una variable GASHOG2D_MISSING con valores perdidos
// con patrón aleatorio

* Generamos una variable aleatoria de 0 y 1
* donde el número 0 aparece aproximadamente el 2% de las veces.
gen aleatorio = rbinomial(1,0.98)

* Creamos una variable con valores perdidos en GASHOG2D
gen gashog2d_missing=gashog2d*aleatorio

* Creamos una variable con valores imputados por la MEDIA
* para GASHOG2D_MISSING
g gashog2d_imputada=gashog2d_missing
summarize gashog2d_missing [aw=factor07]
replace gashog2d_imputada=r(mean) if gashog2d_imputada==!.

* Calculamos error estándar de la MEDIA de las variables
* gashog2d (se supone completa), gashog2d_imputada (por media)
* y gashog2d_missing (con valores perdidos = sin imputar)
svyset conglome [pw=factor07], strata(estrato)
svy: mean gashog2d gashog2d_imputada gashog2d_missing

* Calculamos el error estándar de la variable GASHOG2D_MISSING
* con imputación por media

g auxiliar=gashog2d_missing!=.
program imputacion, rclass
          version 15
                          summarize gashog2d_missing [aw=factor07]
          local media = r(mean)
          replace gashog2d_missing=`media' if gashog2d_missing==.
                          summarize gashog2d_missing [aw=factor07]
                          local media_imputada = r(mean)
                          replace gashog2d_missing=gashog2d_missing*auxiliar
                          return scalar media_imputada = `media_imputada'
  end
```

```
bootstrap r(media_imputada), sav(media_imputada) reps(250) /*
*/ cluster(conglome) strata(estrato): imputacion

use media_imputada, clear
quietly summarize _bs_1
local m=r(mean)
local es=r(sd)
local ls=r(mean)+1.96*r(sd)
local li=r(mean)-1.96*r(sd)

disp _n "Media       Error Estándar         IC(95%)"
disp `m' "     " `es' "      " `li'  "   "  `ls'

use media_imputada, clear
quietly summarize _bs_1

di _n as text "Estimación bootstrap con datos imputados" _n _n /*
   */ "    Media     Err. Estándar           IC(95%)" _n  /*
   */ "{hline 52}" _n %10.2f as result r(mean) "    " %10.2f r(sd) /*
   */ "      " %10.2f r(mean)-1.96*r(sd)  "    "  %10.2f /*
   */ r(mean)+1.96*r(sd) as text _n "{hline 52}"
```

Estimación bootstrap con datos imputados

Media	Err. Estándar	IC(95%)	
26127.39	216.00	25704.02	26550.76