UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

Bootstrap Generalizado especializado en Muestreo Poisson

Mariela Bisso

Directora: Dra. Daniela Rodriguez

22 de Febrero de 2016

Agradecimientos

Quiero agradecer a esas personas que estuvieron conmigo en este camino...

... a mis padres, Graciela y Victorio, por permitirme estudiar y darme todo lo necesario para que hoy esté donde estoy.

... a todo el resto de mi familia y a los amigos de la vida que, cada uno a su manera, estuvieron conmigo en esta etapa.

... a mis compañeros de trabajo por aguantarme hablar del tema todos los dias y en especial a Augusto por confiar en mí.

... a mis compañeros de la facu por hacer que este camino sea mucho más divertido. Gracias Clau, Tati, Agus, Adrián, Lucho, Mariu, Gise, Marie y Estefi por todos los momentos compartidos.

... a mis profesoras del secundario que lograron que ame las matemáticas y a mis profesoras de la facu Daniela y Mariela por estar siempre dispuesta a ayudar y darme ánimos en los momentos que lo necesite, sin ellas tampoco estaría acá.

... y por último, le doy gracias a mi compañero de vida y a su familia. Gracias Mariano por aguantarme en cada examen, por darme fuerzas para ir siempre para delante, sos un gran pilar en mi vida.

gracias!

Índice general

Bootstrap Generalizado, especializado en muestreo Poisson

Resumen

Actualmente la utilización de encuestas por muestreo es de uso habitual en los organismos oficiales de estadística para estimar diversos aspectos relativos a la población, como ser la situación, evolución y futuro de la misma.

El muestreo consiste en la obtención de elementos al azar de una población para estimar determinadas características de ésta cuando no es posible efectuar dicho estudio para cada elemento del universo, es decir cuando no es factible la realización de un Censo de población.

Al analizar cualquier estimación proveniente de una muestra, se necesita saber o aproximar con cierto nivel de confianza cuánto ajusta nuestra estimación al verdadero valor en la población.

Para comenzar, en el Capítulo 1 haremos una introducción a los aspectos fundamentales del muestreo, presentaremos el estimador Horvitz Thompson para el total de una variable de interés, su varianza y su respectivo estimador insesgado para diferentes diseños de muestreo. En el Capítulo 2 se mostrará el funcionamiento del método Bootstrap para poblaciones infinitas, el estimador bootstrap de la varianza e intervalos de confianza para dicha estimación. En el Capítulo 3, transladaremos la idea del método Bootstrap de poblaciones infinitas a poblaciones finitas, que consiste básicamente en la generación de una población a partir de los datos de la muestra observada. De esta población artificial, se extraen B remuestras bootstrap independientes para las cuales se realiza la estimación deseada. Si la cantidad de remuestras es lo suficientemente grande, la distribución del estimador en la muestra bootstrap se interpreta como un estimador de la distribución de muestreo del estadístico en la muestra original. También se discute cual dedería ser el tamaño de las muestras bootstrap para que el estimador de la varianza bootstrap cumpla la condición de ser insesgado bajo los distintos métodos de muestreo.

En el Capítulo 4 detallaremos el método Bootstrap balanceado, utilizados por muchos organismos de estadística, ya que consiste en la generación aleatoria de pesos bootstrap que son incluidos en la base de datos de las encuestas para que el usuario pueda, a partir de ellos, realizar una estimación y estimar su error.

Finalmente en el Capítulo 5, a partir de una base de datos real que tomaremos como nuestra población U, realizaremos la simulación de las muestras para los diferentes diseños de muestreo, estimaremos el total de desempleados de la población y a partir del método bootstrap, su estimación de la varianza.

Capítulo 1

Introducción

El muestreo en poblaciones finitas o encuesta por muestreo surge como consecuencia de la dificultad de medir la totalidad de los individuos y consiste en la selección de una parte de los elementos de una población, con el objetivo de sacar conclusiones de ciertas características de la misma. En este sentido, el muestreo es de suma importancia en la estadística oficial. Asimismo, como ventaja frente a la enumeración completa de una población estadística de interés y a los inconvenientes de una pérdida de exactitud en las conclusiones, el muestreo, evidencia una reducción de los costos implícitos y del tiempo empleado en éste y en el análisis de los datos. Con el propósito de hacer muestreo más eficiente se desarrolla un campo de investigación conocido como *"teoría de muestreo"* que pretende mejorar los métodos para obtener estimaciones suficientemente precisas para el propósito deseado.

Esta estimación se realiza mediante una función de los valores contenidos en la muestra, que se denomina *estimador* y, mientras estemos obteniendo la muestra bajo el muestreo probabilístico, es decir mientras la probabilidad de seleccionar un elemento de la población se conozca o pueda calcularse, esta función es una variable aleatoria. El procedimiento por el cual se selecciona una muestra de unidades de la población se llama *diseño muestral de probabilidad* o *plan muestral de probabilidad*, el cual queda determinado asignando a cada posible muestra $s \subset U$ la probabilidad $P(s)$ de seleccionarla.

1.1. Estimación en poblaciones finitas

Una *población finita* U de tamaño N va a ser un conjunto finito de unidades distintas, que pueden ser numerados como

$$U = \{1, \ldots, N\}.$$

Sobre esta población nos interesará una cierta característica y, o sea, para cada $k \in U$, se define y_k al valor que toma la variable y en el elemento de la población k, también llamados *datos poblacionales*.

Llamaremos *muestra* s de tamaño n a un subconjunto de n unidades de la población U. En lo que sigue, nos centramos en el problema básico de hacer inferencia de una función

de interés de los y_k, llamado parámetro $\Theta = f(y_1, \ldots, y_N)$, a partir de la observación de una muestra seleccionada de acuerdo a un diseño muestral $P(S)$, donde denotamos por S al conjunto formado por todas las muestras extraídas mediante un procedimiento de muestreo determinado. En general la inferencia incluirá la construcción de un intervalo de confianza, o la estimación de alguna medida de precisión, como la varianza del estimador.

Esto implica tres pasos esenciales: elección del diseño muestral, elección del estimador y elección de un estimador de la varianza o construcción de intervalos de confianza.

1.2. Diseño muestral

Como decíamos anteriormente, mediante muestreo probabilístico, es posible asignar a cada muestra una probabilidad conocida de ser seleccionada, de manera que podemos construir una función P definida en el conjunto de todas las muestras S que toma valores en el intervalo $[0,1]$ $(P(.) : S \longrightarrow [0,1])$ y además cumple

$$\sum_{s \subset U} P(s) = 1.$$

Por lo tanto un *diseño muestral* $P(S)$ es una distribución de probabilidad sobre todas las muestras posibles.

Dada una población y dado un diseño muestral determinado, para cualquier muestra $s \in S$, un elemento k de la población puede pertenecer o no a dicha muestra. Para representar la pertenencia o no del elemento k, se define la *variable indicador de pertenencia* $I_k : S \longrightarrow \{0,1\}$ por:

$$I_k(s) = \begin{cases} 1 \quad si \quad k \in s \\ \\ 0 \quad si \quad k \notin s \end{cases}$$

por lo tanto, I_k es una variable aleatoria definida sobre S y su distribución de probabilidad viene dada por

$$P(I_k = 1) = \sum_{s:\, k \in s} p(s) = \pi_k$$

$$P(I_k = 0) = 1 - \pi_k.$$

Definimos la *probabilidad de inclusión de primer orden* π_k, a la probabilidad de que el elemento k de la población este incluido en la muestra elegida, o sea

$$\pi_k = E(I_k) = \sum_{s:\, k \in s} p(s).$$

Por otro lado la *probabilidad de inclusión de segundo orden* π_{kl} indica la probabilidad de que la muestra elegida contenga simultameamente los elementos k y l

$$\pi_{kl} = E\left(I_k I_l\right) = \sum_{k,l \in s} p\left(s\right).$$

Más adelante veremos algunos ejemplos de diseño. A partir de las probabilidades de inclusión de primer y segundo orden, podemos definir la *matriz del diseño muestral* como la siguiente matriz simétrica

$$\pi = \left(\pi_{kl}\right)_{1 \leq k,l \leq N}$$

donde $\pi_{kk} = \pi_k$. Observemos además que la covarianza entre dos variables indicadoras se define:

$$\begin{aligned} \Delta_{kl} &= Cov\left(I_k, I_l\right) = E\left(I_k I_l\right) - E\left(I_k\right) E\left(I_l\right) \\ &= \pi_{kl} - \pi_k \pi_l. \end{aligned}$$

Propiedad: Si el tamaño muestral es fijo, n, entonces se cumple

1. $\sum_{k \in U} \pi_k = n$
2. $\sum_{l \in U} \pi_{kl} = n\pi_k$

Dem:

1. $\sum_{k \in U} I_k = n \;\Rightarrow\; n = \sum_{k \in U} E(I_k) = \sum_{k \in U} \pi_k$
2. $$\begin{aligned} \sum_{k \in U} \Delta_{kl} &= \sum_{k \in U} E(I_k I_l) - \sum_{k \in U} E(I_k) E(I_l) \\ &= E((\underbrace{\sum_{k \in U} I_k}_{n}) I_l) - E(I_l) E(\underbrace{\sum_{k \in U} I_k}_{n}) \\ &= 0 \end{aligned}$$

$$\Rightarrow \sum_{k \in U} \pi_{kl} = \sum_{k \in U} \pi_k \pi_l = n\pi_l$$

1.3. Estimador Horvitz-Thompson

Cualquier función de la variable de interés sobre los elementos de la población se denomina *parámetro poblacional* o simplemente *parámetro* y lo representamos por θ. A partir de los datos observados sobre las unidades extraídas de una muestra de la población podemos construir funciones matemáticas que nos van a ayudar a estimar el valor del parámetro poblacional desconocido.

Recordemos que un *estadístico* es una función de las observaciones de la variable objetivo sobre los elementos de una muestra. Si este estadístico se utiliza para estimar un parámetro, se denomina *estimador* y lo representamos por $\widehat{\theta}$.

De ahora en más, por cuestiones de notación, llamaremos $(Y_1, \ldots, Y_N)$ a los valores observados de la variable para los elementos de la población y $(y_1, \ldots, y_n)$ representarán los valores de la variable de interés en un subconjunto de la población U que llamamos muestra.

Por lo general a uno le interesa estimar el *total de la variable de la población* o la *media poblacional* con lo cual los parámetros poblacionales serían

$$Y = \sum_{k \in U} Y_k$$

y

$$\overline{Y} = \frac{1}{N} \sum_{k \in U} Y_k$$

respectivamente.

Definición: El *estimador Horvitz-Thompson* (H-T) de Y es

$$\widehat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

también se denomina π-*estimador* y el *estimador de la media poblacional* $\overline{Y}$ es

$$\widehat{\overline{Y}}_{HT} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}.$$

El estimador H-T es un estimador directo y no hace uso de ninguna información auxiliar, es decir, utiliza únicamente para su cálculo la información obtenida en la muestra y los pesos de muestreo. Observar que cuanto mayor es la probabilidad de selección, π_k, de la unidad k de la muestra s, menos peso se le da a el dato y_k correspondiente, de ésta manera utiliza la probabilidad para ponderar las respuestas en la estimación del total. Dado un estimador se espera que verifique dos propiedades importantes, una de estas se denomina insesgadez y la otra es que los valores del estimador no se alejen del verdadero valor del parámetro poblacional.

Definición: Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. Se define la *Esperanza* de dicho estimador como la esperaza de la variable aleatoria $\widehat{\theta}$

$$E\left(\widehat{\theta}\right) = \sum_{t \in R} tP\left(\widehat{\theta} = t\right)$$

donde R representa el conjunto de todos los valores posibles del estimador, o sea la imagen de $\widehat{\theta}$ a través de los elementos de S.

Definición: El estimador $\widehat{\theta}$ es *insesgado* para el parámetro poblacional θ si $E\left(\widehat{\theta}\right) = \theta$ para todo valor del vector Y.

En el caso en que $E\left(\widehat{\theta}\right)$ no es igual a θ se dice que el estimador $\widehat{\theta}$ es *sesgado* con respecto a θ. La magnitud de este sesgo en $\widehat{\theta}$ viene dado por

$$B\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$$

El cociente $RB\left(\widehat{\theta}\right) = \frac{B\left(\widehat{\theta}\right)}{\theta}$, se denomina *sesgo relativo* del estimador $\widehat{\theta}$.

Propiedad: El estimador H-T es insesgado de Y, si $\pi_k > 0 \quad \forall k \in U$.

Dem:

$$\begin{aligned} E\left(\widehat{Y}_{HT}\right) &= E\left(\sum_{k\in s}\frac{y_k}{\pi_k}\right) = E\left(\sum_{k\in U}\frac{Y_k}{\pi_k}I_k\right) \\ &= \sum_{k\in U}\frac{Y_k}{\pi_k}E\left(I_k\right) = \sum_{k\in U}\frac{Y_k}{\pi_k}\pi_k \\ &= \sum_{k\in U}Y_k = Y. \end{aligned}$$

Definición: Se define la *varianza* de $\widehat{\theta}$ y se la denota por $Var\left(\widehat{\theta}\right)$, a la siguiente expresión

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= E\left(\widehat{\theta} - E\left(\widehat{\theta}\right)\right)^2 \\ &= \textstyle\sum_{t\in R}\left(t - E\left(\widehat{\theta}\right)\right)^2 P\left(\widehat{\theta} = t\right) \\ &= E\left(\widehat{\theta}^2\right) - E\left(\widehat{\theta}\right)^2. \end{aligned}$$

Es decir, la varianza es una medida que cuantifica la concentración de las estimaciones alrededor de su valor medio.

Propiedad: La varianza del estimador H-T es

$$Var\left(\widehat{Y}_{HT}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}.$$

Dem:

$$\begin{aligned} Var\left(\widehat{Y}_{HT}\right) &= E\left(\widehat{Y}_{HT}^2\right) - E\left(\widehat{Y}_{HT}\right)^2 \\ &= E\left(\left(\sum_{i=1}^{N} I_i \frac{Y_i}{\pi_i}\right)^2\right) - E\left(\sum_{i=1}^{N} I_i \frac{Y_i}{\pi_i}\right)^2 \\ &= E\left(\sum_{i=1}^{N}\sum_{j=1}^{N} I_i I_j \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\right) - E\left(\sum_{i=1}^{N} I_i \frac{Y_i}{\pi_i}\right)^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\right) - \left(\sum_{i=1}^{N}\pi_i\frac{Y_i}{\pi_i}\right)^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}. \end{aligned}$$

Si el tamaño de la muestra es fijo se puede ver que

$$Var\left(\widehat{Y}_{HT}\right) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}-\pi_i\pi_j\right)\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2$$

y se puede ver también que esta expresión es equivalente a

$$Var(\widehat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2.$$

Definición: Dado un estimador $\widehat{\theta}$ se define el *error de muestreo* o *error de estimación del estimador*, σ como su desviación típica, es decir, la raíz cuadrada de su varianza. Por lo tanto,

$$\sigma\left(\widehat{\theta}\right) = +\sqrt{Var\left(\widehat{\theta}\right)}.$$

En muchas ocasiones, la varianza y el error muestral del estimador $\widehat{\theta}$ no son prácticos de calcular, hasta incluso imposible, debido a que sus valores dependen de los valores de la variable en estudio para todos los miembros de la población y generalmente estos datos no están disponibles. Por lo que es necesario estimar los valores $Var\left(\widehat{\theta}\right)$ y $\sigma\left(\widehat{\theta}\right)$ a partir de los datos muestrales. Sus estimadores se denotan $\widehat{Var}\left(\widehat{\theta}\right)$ y $\widehat{\sigma}\left(\widehat{\theta}\right)$, donde $\widehat{\sigma}\left(\widehat{\theta}\right)$, denominado *estimación del error estándar del estimador* $\widehat{\theta}$, es la raíz cuadrada positiva de $\widehat{Var}\left(\widehat{\theta}\right)$. De este modo,

$$\widehat{\sigma}\left(\widehat{\theta}\right) = +\sqrt{\widehat{Var}\left(\widehat{\theta}\right)}.$$

Propiedad: En el caso particular del estimador H-T el estimador de la varianza tiene la siguiente forma

$$\widehat{V}ar\left(\widehat{Y}_{HT}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

que será un estimador insesgado de la varianza del estimador H-T si $\pi_{ij} > 0$ para todo $i, j \in \{1, \ldots, n\}$.

Si el plan es de tamaño fijo,

$$\widehat{V}ar\left(\widehat{Y}_{HT}\right) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

que también será un estimador insesgado siempre que $\pi_{ij} > 0$ para todo $i, j \in \{1, \ldots, n\}$. Se puede ver que la siguiente expresión es equivalente al estimador insesgado de la varianza para un plan de tamaño fijo

$$\widehat{V}ar(\widehat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

Definición: Se define el *error relativo de muestreo o coeficiente de variación* del estimador $\widehat{\theta}$ como el cociente entre su desviación típica y su valor esperado. Su expresión viene dada por

$$CV\left(\widehat{\theta}\right) = \frac{\sigma\left(\widehat{\theta}\right)}{E\left(\widehat{\theta}\right)}$$

a diferencia del error de muestreo, es una medida adimensional lo que nos va a permitir comparar estimadores entre sí sin tener en cuenta las unidades de medida.

Definición: Se define el *error cuadrático medio* $ECM\left(\widehat{\theta}\right)$ como la diferencia entre el estimador y lo que se estima . Su expresión viene dada por,

$$\begin{aligned} ECM\left(\widehat{\theta}\right) &= E[\left(\widehat{\theta}-\theta\right)^2] \\ &= \textstyle\sum_{t\in R}(t-\theta)^2 P\left(\widehat{\theta}=t\right). \end{aligned}$$

El error cuadrático medio y la varianza muestral se relacionan mediante la siguiente expresión

$$ECM\left(\widehat{\theta}\right) = \sigma\left(\widehat{\theta}\right)^2 + B\left(\widehat{\theta}\right)^2$$

donde $B\left(\widehat{\theta}\right)$ es el sesgo del estimador $\widehat{\theta}$. De este modo para un estimador insesgado, el error cuadrático medio y la varianza de un estimador son equivalentes.

1.3.1. Muestreo Bernoulli

Un diseño extremadamente simple es el *Muestreo Bernoulli* (BE). Dado un orden de los elementos del universo $U = 1, \ldots, N$, sea π una constante fija tal que $0 < \pi < 1$ y sean $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ distribuídos uniformemente en el $(0, 1)$ ($\sim Unif(0, 1)$) independientes. La selección o no del k-ésimo elemento es debido por la siguiente regla: si $\epsilon_k < \pi$ entonces el elemento es seleccionado, en otro caso no ($k = 1, \ldots, N$). Claramente la probabilidad de selección de primer orden para el elemento k es

$$\pi_k = P(\epsilon_k < \pi) = \pi \quad \forall k \in [1, \ldots, N]$$

y para cada $k \neq l$ los eventos, k es seleccionado y l es seleccionado son independientes por lo que las probabilidades de segundo orden cumplen

$$\pi_{kl} = \pi_k \pi_l \quad \forall k, l \in [1, \ldots, N] \quad k \neq l.$$

Por lo tanto, si n_s el tamaño de s, el diseño muestral obtenido, es

$$p(s) = \underbrace{\pi \ldots \pi}_{n_s veces} \underbrace{(1-\pi) \ldots (1-\pi)}_{N-n_s veces} = \pi^{n_s}(1-\pi)^{N-n_s}.$$

En el muestreo Bernoulli n_s no es el mismo para todas las muestras posibles, la probabilidad de que la muestra aleatoria tenga exactamente tamaño n_s esta dado por:

$$\binom{N}{n_s} \pi^{n_s}(1-\pi)^{N-n_s} \quad para \quad n_s = 0, \ldots, N$$

entonces el tamaño de la muestra es binomialmente distribuída, con $E(n_s) = N\pi$ y $V(n_s) = N\pi(1-\pi)$.

Propiedad: Bajo el diseño BE el estimador H-T del total tiene la siguiente forma

$$\widehat{Y}_{HT,BE} = \frac{1}{\pi} \sum_{k \in s} y_k$$

la varianza está dada por

$$Var(\widehat{Y}_{HT,BE}) = \left(\frac{1}{\pi} - 1\right) \sum_{k \in U} Y_k^2$$

y un estimador insesgado para la varianza es

$$\widehat{V}ar(\widehat{Y}_{HT,BE}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_{k\in s} y_k^2.$$

1.3.2. Muestreo aleatorio simple sin reemplazo

Un diseño muestral es *aleatorio simple* (*MAS*) si todas las muestras de igual tamaño tiene la misma probabilidad de ser seleccionadas.

Una de las maneras de llevar a cabo el muestreo aleatorio simple sin reemplazo (*MAS sr*) es por medio del siguiente esquema:

1. Seleccionar el primer elemento con igual probabilidad, $1/N$, sobre todos los miembros de la población.

2. Seleccionar el segundo elemento con igual probabilidad, $1/(N-1)$, sobre todos los miembros de la población menos el que ya fue seleccionado.

$\vdots$

n. Seleccionar el elemento n-ésimo con igual probabilidad, $1/(N-n+1)$, sobre todos los miembros de la población que aún no fueron seleccionados.

Es decir, el diseño muestral es

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & si \qquad \#s = n \\ \\ 0 \qquad\quad & si \quad en\ caso\ contrario \end{cases}$$

donde

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

exactamente $\binom{N-1}{n-1}$ muestras s incluyen al elemento k y exactamente $\binom{N-2}{n-2}$ muestras s incluyen los elementos k y l (con $k \neq l$). Entonces, la probabilidad de inclusión de primer orden para todo $k \in U$ viene dada por

$$\pi_k = \sum_{k\in s} p\,(s) = \sum_{k\in s} \binom{N}{n}^{-1} = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N} = f$$

donde f se llama *fracción muestral* y la probabilidad de inclusión de segundo orden para todo $k \neq l \in U$ viene dada por

$$\pi_{kl} = \sum_{k,l \in s} p(s) = \sum_{k,l \in s} \binom{N}{n}^{-1} = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}.$$

Luego tenemos,

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l = \begin{cases} \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)} & si \quad k \neq l \\ \\ \frac{n}{N}(1 - \frac{n}{N}) = \frac{n(N-1)}{N^2} & si \quad k = l \end{cases}.$$

Bajo un diseño *MAS sr* se tiene que $\widehat{Y}_{HT,MASsr}$ es un estimador insesgado de Y y tiene la siguiente forma

$$\widehat{Y}_{HT,MASsr} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} y_k \frac{N}{n} = \frac{N}{n} \sum_{k \in s} y_k = \frac{1}{f} \sum_{k \in s} y_k$$

$$\widehat{\overline{Y}}_{HT,MASsr} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{N} \sum_{k \in s} y_k \frac{N}{n} = \frac{1}{n} \sum_{k \in s} y_k$$

y su varianza es

$$Var\left(\widehat{Y}_{HT,MASsr}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2$$

donde,

$$S^2 = \frac{1}{2}\frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}(Y_i - Y_j)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

entonces, el estimador de la varianza viene dado por

$$\widehat{Var}\left(\widehat{Y}_{HT,MASsr}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) s^2$$

donde

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

frecuentemente $\left(1 - \frac{n}{N}\right)$ es llamado el *término de corrección de la población finita*. Este término puede ser ignorado si el ratio muestral $\frac{n}{N}$ es pequeño ($\leq 0{,}05$).

1.3.3. Muestreo aleatorio simple con reemplazo

En un diseño de muestreo aleatorio simple con reemplazo ($MAS\ cr$) todas las muestras tienen la misma probabilidad de ser selecionadas. La cantidad de muestra de tamaño n de una población de tamaño N, teniendo en cuenta que es con reposición, es N^n. Por lo tanto se tiene el siguiente diseño muestral

$$p(s) = \begin{cases} (N^n)^{-1} & si \quad la\ cantidad\ de\ elementos\ en\ s = n \\ \\ 0 & si \qquad\qquad en\ caso\ contrario \end{cases}$$

donde exactamente nN^{n-1} muestras s incluyen al elemento k y exactamente n^2N^{n-2} muestras s incluyen a los elementos k y l $(k \neq l)$, entonces las probabilidades de primer y segundo orden de inclusión son respectivamente

$$\pi_k = \sum_{k \in s} p\,(s) = \sum_{k \in s} (N^n)^{-1} = nN^{n-1}(N^n)^{-1} = \frac{n}{N}$$

y

$$\pi_{kl} = \sum_{k,l \in s} p\,(s) = \sum_{k,l \in s} (N^n)^{-1} = n^2N^{n-2}(N^n)^{-1} = \frac{n^2}{N^2}$$

luego

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l = \begin{cases} \frac{n^2}{N^2} - \frac{n^2}{N^2} = 0 & si \quad k \neq l \\ \\ \frac{n}{N} - \frac{n^2}{N^2} = \frac{n}{N}\left(1 - \frac{n}{N}\right) = f(1-f) & si \quad k = l \end{cases}.$$

$\widehat{Y}_{HT,MAScr}$ es un estimador insesgado de Y y tiene la siguiente forma

$$\widehat{Y}_{HT,MAScr} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} y_k \frac{N}{n} = \frac{N}{n} \sum_{k \in s} y_k = \frac{1}{f} \sum_{k \in s} y_k$$

$$\widehat{\overline{Y}}_{HT,MAScr} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} = \frac{1}{N} \sum_{k \in s} y_k \frac{N}{n} = \frac{1}{n} \sum_{k \in s} y_k$$

y su varianza es

$$Var\left(\widehat{Y}_{HT,MAScr}\right) \quad = \quad \tfrac{N^2}{n}\sigma^2$$

donde,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

entonces, el estimador de la varianza viene dado por

$$\widehat{V}ar\left(\widehat{Y}_{HT,MAScr}\right) = \frac{N^2}{n}s^2$$

donde

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Notemos que s^2 es un estimador insesgado de S^2 y de σ^2.

1.3.4. Muestreo con probabilidad proporcional al tamaño con reemplazo

A diferencia del muestreo aleatorio simple (MAS), las distintas muestras obtenidas a partir del diseño con probabilidad proporcional al tamaño (pps) no tienen la misma probabilidad de ser seleccionadas.

Se dispone de una variable auxiliar X, donde se supone que sus valores son conocidos para toda la población, $X = (X_1, \ldots, X_N)$. En lo que sigue veremos uno de los posibles métodos de selección de este diseño.

Sea $t_X = \sum_{i=1}^{N} X_i$ y supongamos que se tiene una lista ordenada de los elementos de la población. Se define los totales acumulados de la variable X de la siguiente manera

$$\begin{aligned} ta_1 &= X_1 \\ ta_2 &= X_1 + X_2 \\ \vdots &= \vdots \\ ta_N &= t_X. \end{aligned}$$

Se toma un número aleatorio r, $(0 < r \leq t_X)$ y se elige el elemento i-ésimo de la población si $ta_{i-1} < r \leq ta_i \Longrightarrow p_i = probabilidad\ de\ seleccionar\ el\ elemento\ i\ como\ primer\ elemento = \frac{X_i}{t_X}$. Luego para seleccionar una muestra de tamaño n se repite este procedimiento n veces.

Entonces la probabilidad de inclusión de primer orden y segundo orden vienen dados respectivamente por

$$\pi_k = np_k = \frac{nX_k}{t_X}$$

y

$$\pi_{kl} = \pi_k \pi_l = \frac{nX_k}{t_X} \frac{nX_l}{t_X} = \left(\frac{n}{t_X}\right)^2 X_k X_l.$$

Luego para un diseño *pps* el estimador del total queda dado por

$$\widehat{Y}_{HT,pps} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{np_k}$$

y su varianza y estimador insesgado de la varianza quedan respectivamente

$$Var(\widehat{Y}_{HT,pps}) = \frac{1}{n} \sum_{k=1}^{N} p_k (Z_k - Y)^2$$

y

$$\widehat{Var}(\widehat{Y}_{HT,pps}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (z_k - \widehat{Y}_{HT,pps})^2$$

donde

$$Z_k = \frac{Y_k}{p_k} \; con \; k = (1, \ldots, N)$$

y

$$z_k = \frac{y_k}{p_k} \; con \; k = (1, \ldots, n).$$

Recordemos que los Y_k representa un valor de la variable de interés en el elemento k de la población, miestras que y_k representa un valor de la variable de interés en el elemento k de la muestra.

1.3.5. Muestreo con probabilidad proporcional al tamaño sin reemplazo

Este método de selección, al igual que el método anterior, dispone de las probabilidades de selección de primer orden proporcionales a una variable de tamaño auxiliar X

$$\pi_k = \frac{nX_k}{t_X}$$

donde $t_X = \sum_U X_k$.

La ventaja de los métodos sin reemplazo es que, generalmente, son más eficientes y precisan menor tamaño de muestra para cometer el mismo error que los métodos con reemplazo. Pero estos métodos suelen ser más complicados ya que no tienen una expresión simple para las probabilidades de inclusión de segundo orden.

Luego para la estimación del total Y de una variable de interés se tiene la expresión clásica de la varianza

$$Var(\widehat{Y}_{HT,\pi ps}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2.$$

y su estimador insesgado viene dado por

$$\widehat{V}ar(\widehat{Y}_{HT,\pi ps}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

1.3.6. Muestreo Sistemático

Nos concentraremos en el *muestreo sistemático* (SI) en su forma básica. Tenemos un universo de N elementos en una lista, se elige un número entero a tal que $N = na + c$ donde c es un entero que cumple $0 \leq c < a$ y con igual probabilidad seleccionamos aleatoriamente un elemento r donde $r \in \{1, 2, \ldots, a\}$. El número entero positivo a se fija de antemano y se llama el *intervalo de muestreo.*

La muestra finalmente queda

$$s = \begin{cases} \{r, r+a, r+2a, \ldots, r+(n-1)a\} = s_r & si \quad c < r \leq a \\ \{r, r+a, r+2a, \ldots, r+na\} = s_r & si \quad 1 \leq r \leq c \end{cases}$$

y por lo tanto el tamaño de la muestra queda determinada por:

$$n_s = \begin{cases} n & si \quad c < r \leq a \\ n+1 & si \quad 1 \leq r \leq c \end{cases}.$$

El conjunto de las posibles muestras, denotado por S, consiste de los a diferentes conjuntos que pueden ser obtenidos

$$S = \{s_1, \ldots, s_a\}.$$

El diseño de muestreo para este esquema viene dado por:

$$p(s) = \begin{cases} 1/a & si\ s \in S \\ 0 & para\ cualquier\ otra\ muestra\ s. \end{cases}$$

Como todas las muestras posibles son disjuntas se tiene que para cada elemento k hay solo una muestra que lo contiene, luego

$$\pi_k = 1/a$$

mientras que para todo $k \neq l \in U$,

$$\pi_{kl} = \begin{cases} 1/a & si\ k\ y\ l\ pertenecen\ a\ una\ muestra\ s \\ o & en\ otro\ caso. \end{cases}$$

Para el diseño SI, con un intervalo de muestreo a, el π estimador de la población total Y queda determinado por

$$\widehat{Y}_{HT,SI} = a \sum_{k \in s} y_k$$

donde $s \in \{s_1, \ldots, s_a\}$, la varianza está dada por

$$Var\left(\widehat{Y}_{HT,SI}\right) = a \sum_{r=1}^{a} (Y_{s_r} - \bar{Y})^2$$

donde $Y_{s_r} = \sum_{k \in s_r} y_k$ y $\bar{Y} = \sum_{r=1}^{a} Y_{s_r}/a = Y/a$. La varianza puede también escribirse como

$$Var\left(\widehat{Y}_{HT,SI}\right) = a(a-1)S_t^2$$

donde

$$S_t^2 = \frac{1}{a-1} \sum_{r=1}^{a} (Y_{s_r} - \bar{Y})^2.$$

La propiedad deseable de que los $\pi_k > 0$ no se verifica en este caso, por lo que la fórmula general del estimador de la varianza no es insesgada en el muestreo Sistemático.

1.3.7. Muestreo Poisson

Muestreo Bernoulli, muestreo aleatorio simple y muestreo sistemático son diseños de igual probabilidad, es decir, para todo $k \in S$ las probabilidades de inclusión de primer orden π_k son iguales, lo cual conduce a estimadores simples. Pero esto no es lo que ocurre normalmente en el muestreo de encuestas. La mayoría de los diseños utilizados en la práctica son de probabilidad desigual ya que suelen ser más eficientes.

Un ejemplo de diseño de probabilidades desiguales es el *muestreo Poisson* (PO). Este diseño de tamaño aleatorio es una generalización del muestreo Bernoulli.

Sea π_k un determinado valor positivo de probabilidad de inclusión para el elemento k-ésimo, donde $k = 1, \ldots, N$. Entonces la variable indicadora I_k cumple lo siguiente:

$$P(I_k = 1) = \pi_k,\ P(I_k = 0) = 1 - \pi_k$$

$k = 1, \ldots, N$. El diseño muestral PO es tal que la probabilidad de que la muestra s sea seleccionada es la siguiente

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U-s} (1 - \pi_k)$$

donde $s \in S$ y S es el conjunto de todos los 2^N subconjuntos de U. Y, además, por la independencia, se tiene que las probabilidades de segundo orden cumplen, $\pi_{kl} = \pi_k \pi_l$ para todo $k \neq l$.

Dadas ciertas probabilidades de inclusión $\pi_1, \ldots, \pi_N$ y $\epsilon_1, \ldots, \epsilon_N \sim Unif(0, 1)$ variables aleatorias independientes. Si $\epsilon_k < \pi_k$, el elemento k $(k = 1, \ldots, N)$ es seleccionado, en otro caso no.

En el muestreo Poisson el tamaño de la muestra n_s es aleatorio, con media

$$E_{PO}(n_s) = \sum_U \pi_k,$$

pues $n_s = \sum_{k=1}^N I_k$, y varianza

$$V_{PO}(n_s) = \sum_U \pi_k(1 - \pi_k).$$

Bajo el diseño PO, el π estimador de la población total sigue el resultado general dado por

$$\widehat{Y}_{HT,PO} = \sum_s \frac{y_k}{\pi_k}.$$

La varianza está dada por

$$Var(\widehat{Y}_{HT,PO}) = \sum_U \pi_k(1 - \pi_k)\frac{Y_k^2}{\pi_k^2} = \sum_U \left(\frac{1}{\pi_k} - 1\right) Y_k^2$$

luego el estimador insesgado de la varianza es

$$\widehat{Var}(\widehat{Y}_{HT,PO}) = \sum_s (1 - \pi_k)\frac{y_k^2}{\pi_k^2}.$$

Observación: $Var(\widehat{Y}_{HT,PO})$ puede ser extremadamente grande debido a la variabilidad del tamaño de la muestra. En el muestreo Bernoulli, el diseño es completamente especificado tan pronto como hemos fijado el tamaño de la muestra esperada (asumiendo que conocemos N). Por el contrario, en el muestreo Poisson hay una serie de opciones para la elección de los π_k dado un tamaño de muestra esperado fijo.

Entonces, ¿Cuál es la mejor elección de los π_k?

Una respuesta se obtiene minimizando la varianza para un tamaño esperado de muestra fijo $n = \sum_U \pi_k$. Esto es equivalente a minimizar el producto

$$\left(\sum_U \frac{Y_k^2}{\pi_k}\right)\left(\sum_U \pi_k\right)$$

pero por la desigualdad de Cauchy-Schwartz se tiene

$$\left(\sum_U \frac{Y_k^2}{\pi_k}\right)\left(\sum_U \pi_k\right) \geq \left(\sum_U Y_k\right)^2$$

donde la igualdad vale si o sólo si $Y_k/\pi_k = \lambda$ con λ una constante. Asumiendo que $Y_k > 0$ para todo k, tenemos $\pi_k = Y_k/\lambda$. Finalmente, siendo $n = (\sum_U \pi_k)$ obtenemos

$$\pi_k = \frac{nY_k}{\sum_U Y_j}$$

$k = 1, \ldots, N$, asumiendo que se cumple $Y_j \leq \sum_U Y_k/n$ para todo j pues si $Y_j > \sum_U Y_k/n$ tenemos que $\pi_j > 1$, pues

$$\pi_k = \frac{Y_k}{\lambda} \Rightarrow n = \sum_U \pi_k = \frac{\sum_U Y_k}{\lambda} \Rightarrow \lambda = \frac{\sum_U Y_k}{n} \Rightarrow \pi_k = \frac{Y_k}{\frac{\sum_U Y_j}{n}}.$$

Pero los y_k no son conocidos para toda la población, por lo que esta solución no es factible en la práctica. Sin embargo, en algunas encuestas tenemos acceso a una o más variables auxiliares, es decir, variables cuyos valores son conocidos para toda la población. Supongamos $X_1, \ldots, X_N$ valores positivos conocidos de una variable auxiliar X. Puede haber también razón para asumir que Y es aproximadamente proporcional a X. En este caso podemos decir que π_k es proporcional a X_k conocido, es decir para todo $k = 1, \ldots, N$

$$\pi_k = \frac{nX_k}{\sum_U X_j}$$

$k = 1, \ldots, N$, asumiendo que se cumple $X_k \leq \sum_U X_k/n$ para todo k por la misma razón que antes. Las probabilidades de inclusión definidas de esta forma se llaman *probabilidades de inclusión proporcional al tamaño*. Si Y_k/X_k es casi constante, el estimador H-T resultante tendrá una varianza pequeña.

Aunque este razonamiento es correcto, hay un inconveniente que el muestreo Poisson comparte con el muestreo Bernoulli y es que tienen tamaño de muestra aleatorio. Para mostrar este inconveniente supongamos que es posible elegir π_k de manera óptima, es decir, $\pi_k = nY_k/\sum_U Y_k$ con $n = \sum_U \pi_k$ el tamaño de la muestra esperado. En este caso extremo, el π estimador se convierte en

$$\widehat{Y}_{HT,PO} = \sum_s \frac{y_k}{\pi_k} = \left(\frac{n_s}{n}\right)\sum_U Y_k = \left(\frac{n_s}{n}\right) Y.$$

Entonces, la variación muestra a muestra de $\widehat{Y}_{HT,PO}$ se limitará a la variación en el tamaño n_s de la muestra.

Este argumento nos lleva a esperar que el π estimador es bueno si fuese posible construir un diseño de tamaño fijo con probabilidades de inclusión π_k que sean cercanas a ser proporcionales a Y_k. Si los π_k son exactamente proporcionales a los Y_k en un diseño de tamaño fijo, el π estimador tendría varianza mínima.

Capítulo 2

Bootstrap para poblaciones infinitas

Bootstrap como método fue conceptualizado y descrito por Efron (1979). Se trata de un método general a partir del cual pueden calcularse diferentes propiedades de ciertos estimadores cuya distribución es desconocida, es un tipo de técnica de remuestreo de datos que permite resolver problemas relacionados con la estimación de errores estandar e intervalos de confianza. En esencia, veremos que el método permite aproximar la distribución de un estadístico y de sus propiedades mediante un procedimiento muy simple: Crear un gran número de muestras con reposición de los datos observados. En lo que sigue veremos la idea principal de esta técnica.

Partiendo de que tenemos una función de distribución F desconocida e $Y = \{Y_1, \ldots, Y_n\}$ una muestra aleatoria de dicha distribución, nos interesa estimar un cierto parámetro de interés $\theta = T(F)$, donde T es alguna función definida sobre el espacio de funciones de distribución. Sea $\widehat{\theta} = T_n = h(Y_1, \ldots, Y_n)$ un estadístico con una cierta función de distribución H que depende de h y de F. El método boostrap tiene tres pasos esenciales:

1. Construir un estimador $\tilde{F}_n$ de la función de distribución F, aquí consideraremos la distribución empírica.

2. Generar a partir de $\tilde{F}_n$ una nueva muestra $Y^* = \{Y_1^*, \ldots, Y_n^*\}$ que se denominará muestra bootstrap.

3. Aproximar la distribución $H(h(Y), F)$ por la distribución $\tilde{H}(h(Y^*), \tilde{F}_n)$ en el remuestreo (que en gran parte de los casos prácticos se estimará, a su vez, mediante el método de Montecarlo).

Ahora, supongamos que queremos aproximar una propiedad de T_n bajo la distribución H, el método de simulación nos permite estimar dichas propiedades a partir de la misma propiedad bajo la distribución $\tilde{H}$, por ejemplo, $Var_H(T_n)$ se estimará $Var_{\tilde{H}}(T_n)$.

Veamos en que consiste el método de simulación y después en más detalle como se estima la varianza y los intervalos de confianza a partir del método bootstrap.

2.1. Simulación, método de Montecarlo

Supongamos que tenemos una muestra $X_1, \ldots, X_B$ independientes e identicamente distribuída con distribución G. Por la ley de los grandes números

$$\bar{X} = \frac{1}{B}\sum_{j=1}^{B} X_j \xrightarrow{P} \int X dG(X) = E(X)$$

cuando $B \to \infty$. De manera que si elegimos una muestra de tamaño grande podemos utilizar la media muestral para aproximar el valor $E(X)$ (y hacer que la diferencia entre $\bar{X}$ y $E(X)$ sea despreciable). De manera más general, dada una función h cualquiera, se tiene

$$\frac{1}{B}\sum_{j=1}^{B} h(X_j) \xrightarrow{P} \int h(X) dG(X) = E(h(X))$$

cuando $B \to \infty$. En particular, podemos ver los siguientes ejemplos

$$\frac{1}{B}\sum_{j=1}^{B}(X_j - \bar{X})^2 = \frac{1}{B}\sum_{j=1}^{B}(X_j)^2 - \left(\frac{1}{B}\sum_{j=1}^{B} X_j\right)^2$$

$$\xrightarrow{P} \int X^2 dG(X) - \left(\int X dG(X)\right)^2 = Var(X)$$

por lo tanto, podemos utilizar la varianza de los valores simulados para aproximar $Var(X)$.

También se puede estimar cuantiles de la distribución de la siguiente manera

$$\frac{1}{B}\sum_{j=1}^{B} I(X_j \leq t) \xrightarrow{P} E(I(X_j \leq t)) = P(X \leq t) = G(t).$$

2.2. Estimación de la varianza Bootstrap

Como dijimos anteriormente, vamos a aproximar $Var_H(T_n)$ a partir del método de Montecarlo.

La idea es:

- generar $Y^* = \{Y_1^*, \ldots, Y_n^*\}$ de la distribución $\tilde{F}_n$.

- calcular $T_n^* = h(Y_1^*, \ldots, Y_n^*)$.

Ahora, ¿Cómo podemos generar $Y_1^*, \ldots, Y_n^*$ de $\tilde{F}_n$?

Elegir una observación con distribución $\tilde{F}_n$ es equivalente a elegir un punto al azar del conjunto de datos original. Por lo tanto para simular $Y^* = \{Y_1^*, \ldots, Y_n^*\}$ de $\tilde{F}_n$ basta elegir n observaciones con reemplazo de los datos originales $Y_1, \ldots, Y_n$.

Una vez obtenida la muestra bootstrap $Y^* = \{Y_1^*, \ldots, Y_n^*\}$, calculamos $T_n^* = h(Y_1^*, \ldots, Y_n^*)$ y obtendremos una observación de $\tilde{H}$. Si volvemos a repetir este procedimiento B veces obtendremos $T_{n,1}^*, \ldots, T_{n,B}^*$ y podremos a partir de Montecarlo estimar $Var_{\tilde{H}}(T_n)$ es decir

$$\widehat{V}ar_{\tilde{H}}(T_n) = \frac{1}{B}\sum_{j=1}^{B}(T_{n,j}^*)^2 - \left(\frac{1}{B}\sum_{j=1}^{B}T_{n,j}^*\right)^2.$$

A esta estimación de la varianza se la nota como v_{boot}. Si queremos estimar el error estándar, calculamos $\sqrt{v_{boot}}$ y lo notamos por $\widehat{se}_{boot}$.

2.3. Intervalos de confianza Bootstrap

Unos de los problemas más comunes es el de obtener intervalos de confianza para un parámetro estimado. Existe más de un método para la construcción de los intervalos de confianza Bootstrap de un estadístico. En lo que sigue describiremos tres métodos distintos para generar estos intervalos.

La idea para la construcción de un intervalo de confianza para θ de nivel $1-\alpha$ es encontrar un intervalo $IC_\alpha(\widehat{\theta})$ que satisfagan $P(\theta \in IC_\alpha(\widehat{\theta})) = 1-\alpha$.

2.3.1. Método 1: Intervalo de confianza Normal

Esta aproximación es la más simple para la construcción de un intervalo de confianza, aunque no necesariamente la mejor. Sea $\widehat{\theta}$ un estimador del parámetro θ con un error estándar se. Si el tamaño muestral es grande se tiene que por el Teorema Central del límite

$$Z = \frac{\widehat{\theta} - E(\widehat{\theta})}{se}$$

se aproxima a una Normal estándar.

Luego, resulta

$$\left[\widehat{\theta} - z_{\alpha/2}.se, \widehat{\theta} + z_{\alpha/2}.se\right]$$

donde $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, con Φ la función de distribución de la normal estándar, es un intervalo de confianza para θ de nivel $1-\alpha$

Como deciamos, este es el método más simple para la construcción de un intervalo de confianza, pero dicho intervalo no es exacto a menos que la distribución del estadístico sea normal. Además, el parámetro *se* se trató como un parámetro conocido, aunque en el método bootstrap es estimado por la desviación estándar muestral de las réplicas bootstrap de $\widehat{\theta}$. Por lo que, su expresión viene dada por

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{se}_{boot}$$

donde, como dijimos antes $\widehat{se}_{boot}$ representa la estimación bootstrap del error estandar.

2.3.2. Método 2: Intervalo de confianza a partir de un pivote

Dados θ y $\widehat{\theta}_n$, definamos $R_n = \widehat{\theta}_n - \theta$ y H su distribución, es decir

$$H(r) = P_F(R_n \leq r).$$

Sea $C = (a, b)$ el intervalo tal que $a = \widehat{\theta}_n - H^{-1}(1-\alpha/2)$ y $b = \widehat{\theta}_n - H^{-1}(\alpha/2)$ entonces se cumple que

$$\begin{aligned} P(a \leq \theta \leq b) &= P(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\ &= P(\widehat{\theta}_n - b \leq \widehat{\theta}_n - \theta \leq \widehat{\theta}_n - a) \\ &= P(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\ &= H(\widehat{\theta}_n - a) - H(\widehat{\theta}_n - b) \\ &= H(H^{-1}(1-\alpha/2)) - H(H^{-1}(\alpha/2)) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha. \end{aligned}$$

Luego $C = (a, b)$ es un intervalo de confianza exacto de nivel $1-\alpha$ para θ, pero a y b dependen de la función de distribución H desconocida. Por lo tanto, a partir de lo visto podemos dar una estimación Bootstrap de H del siguiente modo:

Sean $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ las replicaciones del método Bootstrap de $\widehat{\theta}_n$ y $R^*_{n,b} = \widehat{\theta}^*_{n,b} - \widehat{\theta}_n$ entonces

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R^*_{n,b} \leq r).$$

Sea r^*_β el cuantil β de $\left\{R^*_{n,1}, \ldots, R^*_{n,B}\right\}$ y sea θ^*_β el cuantil β de $\left\{\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}\right\}$. Notar que $r^*_\beta = \widehat{\theta}^*_\beta - \widehat{\theta}_n$. Luego es claro que podemos estimar a y b como

$$\begin{aligned} \widehat{a} &= \widehat{\theta}_n - \widehat{H}^{-1}(1-\alpha/2) = \widehat{\theta}_n - r^*_{1-\alpha/2} = 2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2} \\ \widehat{b} &= \widehat{\theta}_n - \widehat{H}^{-1}(\alpha/2) = \widehat{\theta}_n - r^*_{\alpha/2} = 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}. \end{aligned}$$

Por lo tanto, definimos un *intervalo de confianza boostrap a partir de un pivote* como

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}).$$

2.3.3. Método 3: Intervalo de confianza a partir de percentiles

Este método utiliza la distribución empírica de las réplicas bootstrap como distribución de referencia. Los cuantiles de la distribución empírica bootstrap son estimadores de los cuantiles de la distribución muestral de $\widehat{\theta}$. El intervalo se define como

$$C_n = (\widehat{\theta}^{*,(\alpha/2)}, \widehat{\theta}^{*,(1-\alpha/2)})$$

donde $\widehat{\theta}^{*,(\beta)}$ es el cuantil β de la distribución Bootstrap. Como en la práctica la distribución de $\widehat{\theta}^*$ es desconocida, este intervalo no se puede obtener de manera exacta, pero es natural considerar la siguiente aproximación del intervalo percentil Bootstrap dada por,

$$C_n = (\widehat{\theta}^*_{\alpha/2}, \widehat{\theta}^*_{1-\alpha/2})$$

donde, como antes, $\widehat{\theta}^*_\beta$ es el cuantil β de $\left\{\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}\right\}$. Si la distribución Bootstrap es aproximadamente Normal, los intervalos de confianza de este método con el del método 1 no deberán ser tan distintos.

Supongamos que existe una transformación monótona m tal que $\widehat{\Omega} = m(\widehat{\theta})$ es aproximadamente Normal ($\widehat{\Omega} \sim N(\phi, c^2)$, donde $c > 0$ es constante y $m(\theta) = \phi$). Si definimos $\widehat{\Omega}^*_b = m(\widehat{\theta}^*_{n,b})$, sea ω^*_β el cuantil β de $\left\{\widehat{\Omega}^*_1, \ldots, \widehat{\Omega}^*_B\right\}$ que por ser $\widehat{\Omega}$ una transformación monótona preserva los cuantiles y por lo tanto se tiene que $\omega^*_\beta = m(\widehat{\theta}^*_\beta)$. Como $\widehat{\Omega} \sim N(\phi, c^2)$ el cuantil $\alpha/2$ de $\widehat{\Omega}$ es $\phi - z_{\alpha/2}c$ entonces $\omega^*_{\alpha/2} = \widehat{\phi} - z_{\alpha/2}c$ y por lo tanto se tiene que $\omega^*_{1-\alpha/2} = \widehat{\phi} + z_{\alpha/2}c$. Luego

$$\begin{aligned}
P(\widehat{\theta}^*_{\alpha/2} \leq \theta \leq \widehat{\theta}^*_{1-\alpha/2}) &= P(m(\widehat{\theta}^*_{\alpha/2}) \leq m(\theta) \leq m(\widehat{\theta}^*_{1-\alpha/2})) \\
&= P(\omega^*_{\alpha/2} \leq \phi \leq \omega^*_{1-\alpha/2}) \\
&= P(\widehat{\phi} - z_{\alpha/2}c \leq \phi \leq \widehat{\phi} + z_{\alpha/2}c) \\
&= P(z_{\alpha/2} \leq \frac{\widehat{\phi} - \phi}{c} \leq -z_{\alpha/2}) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

Capítulo 3

Bootstrap para poblaciones finitas

3.1. Introducción

Como en el resto de los capítulos nuestro problema es estimar la distribución de una muestra estadística o la distribución de muestreo de un estadístico $\widehat{\theta}$ de un parámetro de interés θ . En este problema, la población finita U de N elementos toma el rol de la función de distribución desconocida para el caso de una población infinita. El método Bootstrap para poblaciones finitas se considera una extensión natural de la técnica para el muestreo en poblaciones infinitas. La idea es generar una población artificial, llamada *población bootstrap*, a partir de los datos de la muestra observada s de tamaño n. A partir de esa población, se extraen B remuestras del mismo tamaño que la muestra original, dentro de cada una de esas remuestras $s_1, \ldots, s_B$ se calcula el estimador $\widehat{\theta}_b$ del parámetro θ de la misma manera que se calcula el estadístico $\widehat{\theta}$ para la muestra s $(b = 1, \ldots, B)$. Para B grande, la distribución de $\widehat{\theta}_b$ se interpreta como una estimación de la distribución de muestreo $\widehat{\theta}$. Por lo tanto, la varianza $Var(\widehat{\theta})$ de $\widehat{\theta}$ se puede estimar por el método de Monte Carlo de la siguiente manera

$$\widehat{V}ar(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b - \bar{\theta})^2$$

donde

$$\bar{\theta} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b$$

3.2. Población Bootstrap

En este capítulo veremos distintas maneras de generar la población bootstrap que luego será usada para la extracción de la B remuestras $s_1, \ldots, s_B$ de tamaño n. Veremos los casos en donde los pesos de diseño son enteros y donde no lo son y cuáles serían las condiciones

mínimas que uno espera para obtener una estimación optima de la varianza bajo distintos métodos de muestreo.

3.2.1. Población Bootstrap para muestreo aleatorio simple

El primero en adaptar el método bootstrap original para el caso de una muestra aleatoria simple (MAS) sin sustitución fue Gross (1980). De esta manera, dada una muestra s de tamaño n se genera una población bootstrap U_G^* de tamaño $N_G^* = N$ de la verdadera población U de tamaño N mediante la replicación de cada valor de la muestra y_k exactamente N/n veces, en el caso que $N/n \in \mathbb{N}$, proporcionando una variable y^* denotando a los clones de los valores de la muestra.

De esta manera, una vez generada la población U_G^*, B remuestras de tamaño n se extraen de U_G^* siguiendo el método de muestreo original. Es decir, cada uno de los n valores en la muestra $y_1, \ldots, y_n$ tiene la misma probabilidad $\frac{N/n}{N} = \frac{1}{n}$ de ser seleccionado como primer elemento del remuestreo.

Luego del primer elemento seleccionado , el valor ya elegido tiene una probabilidad de

$$\frac{\frac{N}{n} - 1}{N - 1} = \frac{N - n}{n(N-1)}$$

de volver a ser seleccionado, mientras que los demás no seleccionados como el primer elemento tienen una probabilidad de ser seleccionados como segundo elemento de la remuestra de

$$\frac{N/n}{N - 1} = \frac{N}{n(N-1)}$$

y así sucesivamente. En general, un valor y_k observado en s tiene una probabilidad

$$\frac{N - n.h_{k,j-1}}{n\,(N - j + 1)}$$

de ser seleccionado en la j-ésima etapa de una selección de remuestreo ($j = 1, \ldots, n$). Donde $h_{k,j-1}$ representa el número de veces que el valor y_k fue seleccionado en los primeros $(j-1)$-ésimos pasos del remuestreo ($h_{k,0} = 0\ \forall k \in s, h_{k,j} < \frac{N}{n}\ \forall k \in s\ \forall j = 1, \ldots, n$).

Esta población no tiene por que ser generada en realidad, sino que también puede llevarse a cabo mediante la aplicación del mecanismo de probabilidad que se ha descripto anteriormente directamente a la muestra s. Estas remuestras forman la base para la estimación de la distribución de muestreo del estimador $\widehat{\theta}$ para el parámetro θ en el muestreo aleatorio simple.

Si $N/n \notin \mathbb{N}$, o sea, si los pesos de diseño no son enteros, la pseudo población U_{HT}^* no sólo contiene $\lfloor N/n \rfloor$ unidades enteras con cada valor y_k de la variable y, si no que también contine $(N/n - \lfloor N/n \rfloor)$ partes de cada unidad ($\forall k \in s$), por ejemplo, si tengo una población de tamaño $N = 2600$ y una muestra de tamaño $n = 400$ ($\frac{N}{n} = \frac{2600}{400} = 6,5$) implica que

la población bootstrap U^*_{HT} esta compuesta por 6 unidades enteras de cada valor en la muestra y además media unidad de cada valor y_k $(k = 1, \ldots, n)$, pero en este caso la población boostrap no puede ser calculada realmente y necesita llevarse a cabo mediante el cálculo de los pesos bootstrap.

3.2.2. Población Bootstrap para muestreo con probabilidad proporcional al tamaño

Otro diseño muestral básico es el muestreo con probabilidades proporcionales al tamaño muestral sin reemplazo (πps), en el que las probabilidades de inclusión son proporcionales al tamaño de una variable auxiliar X llamada variable auxiliar de tamaño. En este caso la selección se lleva a cabo, por ejemplo, mediante la selección sistemática de elementos ordenados aleatoriamente en una lista. Este método de muestreo es eficiente para el estimador H-T de un total de la variable Y ($\widehat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$ donde π_k es la probabilidad de inclusión de primer orden), cuando la variable auxiliar de tamaño X y la variable de estudio Y están relacionadas proporcionalmente, pero el cálculo de la estimación de la varianza puede ser muy engorroso. Holmberg (1998) propuso un enfoque bootstrap para estimar la varianza para el muestreo general πps.

Notamos al total de la variable tamaño X en U como t_X bajo la siguiente restricción $X_k.n \leq t_X \ \forall k \in U$, donde X_k es la medida de tamaño para la k-ésima unidad en la población.

Luego el peso de diseño para la unidad $k \in U$ se define como

$$\frac{1}{\pi_k} = \frac{t_X}{X_k.n}$$

que a su vez se descompone en una parte entera $\left\lfloor \frac{t_X}{X_k.n} \right\rfloor$ y el resto $\frac{t_X}{X_k.n} - \left\lfloor \frac{t_X}{X_k.n} \right\rfloor$.

Para generar la población bootstrap U^*_H, se replican los valores de y_k y x_k de cada unidad k de la muestra, $\left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ veces e independientemente de estas una más con probabilidad $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor$. Este proceso crea una población U^*_H de tamaño N^*_H tal que $E(N^*_H) = N$. Una vez generada la población, las probabilidades de inclusión de la muestra π_k deben ser recalculadas de acuerdo con la variable X^* que consiste de los valores replicados de X. Luego como antes, generamos a partir de la población B muestras de tamaño n con un diseño πps, y se realiza la estimación en cada una de ellas de igual manera que se calculó para la muestra original y se procede a estimar la varianza a partir del método de simulación.

Barbiero y Mecatti (2010) proponen hacer un uso más completo de la información auxiliar con el objetivo de simplificar el procedimiento presentado por Holmberg para muestreo πps. Sostienen que es natural pedir que la población bootstrap cumpla los siguientes requisitos:

- Dadas una muestra s original y una población bootstrap U^*, el total de la variable

tamaño X^* en U^* debe ser igual al total de la variable tamaño X en U, es decir $t_{X^*} = t_X$.

- El total de la variable Y^* en U^* debe ser igual al total de la variable Y estimado a partir del estimador H-T en la muestra original s, es decir $t_{Y^*} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \widehat{Y}_{HT}$.

- Para s dado, sobre todas las B remuestras, $s_1, \ldots, s_B$ el total estimado de la variable Y^* en s_b (b=1, ..., B) debe tener la misma esperanza que el total de la variable Y estimado a partir del estimador H-T en la muestra original s, es decir $E\left(\sum_{k=1}^{n} \frac{y_k^{*,b}}{\pi_k^*}\right) = E\left(\sum_{k=1}^{n} \frac{y_k}{\pi_k}\right) = E(\widehat{Y}_{HT})$.

Estas propiedades son ideales y, desafortunadamente, para los diferentes métodos bootstrap relacionados con la creación de un pseudo-población, estas tres propiedades sólo se cumplen cuando se tiene $\frac{1}{\pi_k} \in \mathbb{N}\ \forall k \in s$. Por esta razón Barbiero y Mecatti, propusieron un método πps X-balanceado donde, después de remuestrar cada unidad k de la muestra $\left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ veces en la población bootstrap U^*_{BM}, se remuestrean más unidades de una lista ordenada de forma decreciente de sus valores $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor$, se agregan hasta que se consiga la mínima diferencia entre t^*_X y t_X. De esta manera, los elementos con mayor parte entera $\left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ tienen más probabilidad de ser incluido en U^*_{BM} con respecto a los elementos que presentan una parte entera inferior. Luego de que la población fue generada, las probabilidades π_k tienen que volver a calcularse antes de que comience la selección de las B replicas de tamaño n.

Ninguno de los métodos vistos hasta ahora para obtener la población bootstrap garantiza un tamaño de la población $N^* = N$ cuando los pesos de diseño no son enteros. En lo que sigue se propone un procedimiento para obtener la población bootstrap U^*_{HTB} basado en el estimador Horvitz-Thompson para el problema de pesos de diseño no enteros, que permite además un número de repeticiones no enteras de los valores de la muestra de Y y X. Para cada k en la muestra se replica exactamente $\frac{1}{\pi_k} = \frac{t_X}{x_k.n}$ veces. Es decir, la población bootstrap U^*_{HTB} se genera con $\left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ unidades enteras de y_k y también por $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ partes de una unidad cuando $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor > 0\ \forall k \in s$. De esta manera, U^*_{HTB} tiene un tamaño esperado N^*_{HTB} de $E(N^*_{HTB}) = \sum_s \frac{1}{\pi_k} = N$. Para este método de remuestreo, toda unidad k perteneciente a la población bootstrap tiene una probabilidad de inclusión de remuestreo proporcional a su valor X original, miestras que para $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ esta probabilidad es proporcional a $\frac{t_X}{x_k.n} - \left\lfloor \frac{t_X}{x_k.n} \right\rfloor$ veces X. Por lo tanto, después de la generación de la población bootstrap U^*_{HTB}, los pesos de diseño π_k no tienen que volver a ser calculados.

Entonces, el valor y_k de la muestra original ($k = 1, \ldots, n$) tiene una probabilidad

$$\frac{t_X - n.h_{k,j-1}.x_k}{n.\left(t_X - \sum_{s_{b_{j-1}}} x_i\right)}$$

para ser seleccionada en la b-ésima remuestra en el j-ésimo paso de elección de las n unidades de remuestreo ($j = 1, \ldots, n$) cuando $\frac{t_X}{x_k.n} - h_{k,j-1}.x_k > 0$, de lo contrario su probabilidad de inclusión es cero. Donde $h_{k,j-1}$ denota el número de veces que y_k fue elegido dentro de los primeros $(j-1)$ pasos de la selección de n unidades de la remuestra s_b y $s_{b_{j-1}}$ denota el subconjunto de la remuestra s_b después de la $(j-1)$-ésima elección.

El uso de este mecanismo de probabilidad en el proceso de remuestreo puede reemplazar la generación física de la población bootstrap U^*_{HTB}. Para la técnica HTB propuesta, en relación con las tres propiedades deseadas para una estimación de la varianza, se tiene:

- El total t_{X^*} de la variable tamaño X^* en U^*_{HTB} esta dada por: $t_{X^*} = \sum_{k=1}^{n} x_k.\frac{1}{\pi_k} = t_X$.
- El total t_{Y^*} de la variable Y^* en U^*_{HTB} esta dada por: $t_{Y^*} = \sum_{k=1}^{n} y_k.\frac{1}{\pi_k} = \widehat{Y}_{HT}$.
- El valor esperado para el estimador H-T del total t_{Y^*} de Y^* en U^*_{HTB} es: $E^*\left(\sum_{k=1}^{n} \frac{y^*_k}{\pi^*_k}\right) = t_{Y^*} = \widehat{Y}_{HT}$. Donde E^* denota la esperanza sobre todas las remuestras.

Obviamente para la aplicación de esta técnica bootstrap a la práctica esta idea debe extenderse a métodos de muestreo en general, asegurando una población bootstrap con la misma estructura que la población original. En lo que sigue vamos a estudiar el estimador bootstrap para los diseños muestrales que vimos anteriormente y en particular para el estimador Horvitz Thompson y un estimador de su varianza.

En las secciones siguientes, sólo por el hecho de simplificar la notación, notaremos al estimador Horvitz Thompson para el total de una variable de interés Y como $\widehat{Y}_{HT}$ independientemente del diseño elegido.

3.3. Muestreo aleatorio simple con reemplazo

Supongamos que se quiere estimar un total poblacional Y de una variable en una población finita U de tamaño N. Se seleccionan n unidades de la población mediante muestreo aleatorio simple con reemplazo. Tenemos entonces una muestra original $(y_1, \ldots, y_n)$. Luego estimamos el total Y mediante el estimador Horvitz Tompson,

$$\widehat{Y}_{HT} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

por lo visto en el Capítulo 1 sabemos que la varianza y el estimador usual del estimador vienen dados por

$$Var(\widehat{Y}_{HT}) = \frac{N^2}{n}\sigma^2$$

y

$$\widehat{V}ar(\widehat{Y}_{HT}) = \frac{N^2}{n}s^2$$

donde

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

y

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Ahora, una vez obtenida la muestra bootstrap con reemplazo de tamaño n^* de la muestra original, notemosla por $(y_1^*, \ldots, y_{n^*}^*)$, tenemos el correspondiente estimador Horvitz Thompson para el total

$$\widehat{Y}_{HT}^* = \frac{N}{n^*}\sum_{i=1}^{n^*} y_i^*$$

dado que la muestra se extrajo con reposición se tiene que, dado $i = (1, \ldots, n)$

$$P[y_j^* = y_i] = \frac{1}{n} \quad \forall j = 1, \ \ldots, n^*$$

entonces la esperanza y la varianza vienen dados por

$$E_*(y_j^*) = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\widehat{Y}_{HT}}{N}$$

$$Var_*(y_j^*) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n-1}{n}s^2$$

respectivamente, dado que las observaciones bootstrap son independientes e idénticamente distribuídas, la media y la varianza de $\widehat{Y}_{HT}^*$ condicionadas a la muestra original vienen dadas por

$$E_*(\widehat{Y}_{HT}^*) = \frac{N}{n^*}\sum_{j=1}^{n^*} E(y_j^*) = \widehat{Y}_{HT}$$

y

$$Var_*(\widehat{Y}_{HT}^*) = \left(\frac{N}{n^*}\right)^2 \sum_{j=1}^{n^*} Var_*(y_j^*) = \frac{N^2}{n^*}\frac{n-1}{n}s^2.$$

Luego, se tiene que el estimador bootstrap de la varianza en general no coincide con el estimador usual de la varianza y no es un estimador insesgado a menos que $n^* = n - 1$.

Si $n^* = n$ se tiene que la estimación es sesgada, y su sesgo viene dado por

$$\begin{aligned} sesgo(Var_*(\widehat{Y}^*_{HT})) &= E_*(Var_*(\widehat{Y}^*_{HT})) - Var(\widehat{Y}_{HT}) \\ &= \left(\frac{N}{n}\right)^2 (n-1)\,\sigma^2 - \frac{N^2}{n}\,\sigma^2 \\ &= \left(-\frac{1}{n}\right) Var(\widehat{Y}_{HT}) \end{aligned}$$

notese que si n es grande el sesgo es despreciable, pero podría ser importante si n es pequeño.

3.4. Muestreo aleatorio simple sin reemplazo

De la misma manera que en el caso anterior dado

$$\widehat{Y}^*_{HT} = \frac{N}{n^*} \sum_{i=1}^{n^*} y^*_i$$

tenemos que

$$E_*(y^*_j) = \frac{1}{n} \sum_{i=1}^{n} y_i = N\ \widehat{Y}_{HT}$$

$$Var_*(y^*_j) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{n-1}{n} s^2$$

ya que estos valores no son afectados por la muestra original, si no por la muestra bootstrap. Entonces como antes tenemos que la varianza bootstrap del estimador está dada por

$$Var_*(\widehat{Y}^*_{HT}) = \frac{N^2}{n^*} \frac{n-1}{n} s^2.$$

La diferencia con el muestreo anterior viene dada por el tipo de muestreo utilizado para la muestra original, por lo cual tenemos la varianza y su correspondiente estimador de la varianza del estimador Horvitz Thompson para el muestreo aleatorio simple sin sustitución dados en el Capítulo 1 por

$$Var(\widehat{Y}_{HT}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S^2$$

y su estimador

$$\widehat{Var}(\widehat{Y}_{HT}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) s^2.$$

Notemos que el estimador bootstrap de la varianza no coincide con el estimador insesgado usual y su sesgo viene dado por

$$\begin{aligned} sesgo(Var_*(\widehat{Y}^*_{HT})) &= E_*(Var_*(\widehat{Y}^*_{HT})) - Var(\widehat{Y}_{HT}) \\ &= \left(\frac{N^2}{n^*}\right)\frac{n-1}{n}S^2 - \frac{N^2}{n}(1-f)S^2 \\ &= \left(\frac{n-1}{n^*} - (1-f)\right)\frac{N^2}{n}S^2 \end{aligned}$$

donde $f = \frac{n}{N}$. Si tomamos $n^* = n-1$ el sesgo del estimador bootstrap de la varianza se transforma en

$$\begin{aligned} sesgo(Var_*(\widehat{Y}^*_{HT})) &= f\,\frac{N^2}{n}S^2 \\ &= f\;E(Var_*(\widehat{Y}^*_{HT})) \end{aligned}$$

Si f es pequeño, el sesgo es desperciable. Si el sesgo no es despreciable, hay algunas variantes para el método bootstrap en el caso que sea un muestreo aleatorio simple sin reemplazo.

3.4.1. Variante del método bootstrap por factor de corrección

Para el caso en que $n^* = n-1$, un estimador insesgado de la varianza viene dado por

$$Var_{*,FC}(\widehat{Y}^*_{HT}) = (1-f)Var_*(\widehat{Y}^*_{HT}).$$

3.4.2. Variante del método bootstrap por reescalado

Definimos las siguientes observaciones reescaladas

$$y^{\#}_j = \bar{y} + \sqrt{1-f}\sqrt{\frac{n^*}{n-1}}(y^*_j - \bar{y})$$

la total bootstrap de la variable reescalada se transforma en

$$\begin{aligned} \widehat{Y}^{\#}_{HT} &= \frac{N}{n^*}\sum_{j=1}^{n^*} y^{\#}_j \\ &= N\bar{y} + \frac{N}{n^*}\sqrt{1-f}\sqrt{\frac{n^*}{n-1}}\sum_{j=1}^{n^*}(y^*_j - \bar{y}) \end{aligned}$$

y por lo tanto el estimador bootstrap de la varianza queda

$$\begin{aligned} Var_*(\widehat{Y}^{\#}_{HT}) &= (1-f)\frac{n^*}{n-1}Var_*(\widehat{Y}^*_{HT}) \\ &= (1-f)\frac{n^*}{n-1}\frac{N^2}{n^*}\frac{n-1}{n}s^2 \\ &= (1-f)\frac{N^2}{n}s^2 = \widehat{V}ar(\widehat{Y}_{HT}) \end{aligned}$$

Notar que $Var_*(\widehat{Y}^{\#}_{HT})$ coincide con el estimador de la varianza del total de la variable Y en un muestreo aleatorio simple sin reemplazo.

Luego si se considera $n^* = n$ los elementos bootstrap reescalados quedan

$$y_j^{\#} = \bar{y} + \sqrt{1-f}\sqrt{\frac{n}{n-1}}(y_j^* - \bar{y}).$$

Entonces, en el caso de que f no es lo suficientemente chico para despreciar el sesgo de la varianza, después de cada muestra bootstrap seleccionada antes de evaluar el estadístico que nos interesa estudiar, calculamos los valores reescalados y evaluamos el estadístico con estos nuevos valores. Para luego, a partir de las B muestras bootstrap calcular nuestro estimador de la varianza mediante el método de Monte Carlo.

3.4.3. Variante del método bootstrap por reemplazo

Una elección acertada del n^* elimina el sesgo de la varianza bootstrap. A partir de la fórmula del sesgo de la varianza bootstrap

$$sesgo(Var_*(\widehat{Y}^*_{HT})) = \left(\frac{n-1}{n^*} - (1-f)\right)\frac{N^2}{n}S^2$$

observamos que si elegimos

$$n^* = \frac{n-1}{1-f}$$

el sesgo se anula y la varianza bootstrap queda igual que el estimador usual de la varianza.

Para n^* definida de esta manera, probablemente en la práctica, no resulte ser un número entero. Por lo que podemos tomar $n^* = \left\lceil \frac{n-1}{1-f} \right\rceil$.

3.5. Muestreo con probabilidad proporcional al tamaño con reemplazo

Como ya vimos anteriormente, en el muestreo con probabilidad proporcional al tamaño, las probabilidades de inclusión de primer orden se construyen a partir de una variable auxiliar X (conocida para el total de la población U de tamaño N) de la siguiente manera

$$\pi_k = \frac{n\, X_k}{t_X}$$

donde $t_X = \sum_U X_k$. Sea $(y_1, \ldots, y_n)$ la muestra *pps wr* original, en este caso el estimador usual Horvitz Thompson para el total está dado por

$$\widehat{Y}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{1}{n} \sum_{k=1}^{n} z_k$$

donde $z_k = \frac{y_k}{x_k/t_X}$. Su varianza y estimador de la varianza son respectivamente

$$Var(\widehat{Y}_{HT}) = \frac{1}{n} \sum_{i=1}^{N} p_i (Z_i - Y)^2$$

donde $p_i = \frac{X_i}{t_X}$ y

$$\widehat{Var}(\widehat{Y}_{HT}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (z_k - \widehat{Y}_{HT})^2.$$

Sean $(z_1^*, \ldots, z_{n^*}^*)$ una muestra bootstrap obtenida mediante muestreo aleatorio simple con reemplazo de la muestra $(z_1, \ldots, z_n)$, entonces el estimador del total de la variable Y queda de la siguiente manera

$$\widehat{Y}_{HT}^* = \frac{1}{n^*} \sum_{k=1}^{n^*} z_k^*$$

donde z_k^* son variables aleatorias independientes e idénticamente distribuídas con esperanza y varianza dadas por

$$E_*(z_k^*) = \frac{1}{n} \sum_{k=1}^{n} z_k = \widehat{Y}_{HT}$$

y

$$Var_*(z_k^*) = \frac{1}{n} \sum_{k=1}^{n} (z_k - \widehat{Y}_{HT})^2$$

$\forall k = 1, \ldots, n$ entonces

$$E_*(\widehat{Y}_{HT}^*) = \frac{1}{n^*} \sum_{k=1}^{n^*} E_*(z_k^*) = E_*(z_k^*) = \widehat{Y}_{HT}$$

y

$$\begin{aligned} Var_*(\widehat{Y}^*_{HT}) &= \frac{1}{n^{*2}}\sum_{k=1}^{n^*} Var_*(z^*_k) = \frac{1}{n^*}\frac{1}{n}\sum_{k=1}^{n}(z_k - \widehat{Y}_{HT})^2 \\ &= \frac{n-1}{n^*}\widehat{Var}(\widehat{Y}_{HT}) \end{aligned}$$

luego el estimador bootstrap de la varianza de $\widehat{Y}_{HT}$ es igual a $\frac{n-1}{n^*}$ veces el estimador usual de la varianza bajo muestreo con probabilidad proporcional al tamaño con reemplazo. Si $n^* = n-1$ el estimador queda insesgado y si se toma $n^* = n$ el estimador queda con $sesgo = \frac{n-1}{n}$, que es despreciable si n es lo suficientemente grande.

3.6. Muestreo con probabilidad proporcional al tamaño sin reemplazo

Al igual que en el caso anterior, muestreo *pps cr*, tenemos que el estimador del total para muestreo πps es

$$\widehat{Y}_{HT} = \sum_{k=1}^{n}\frac{y_k}{\pi_k} = \frac{1}{n}\sum_{k=1}^{n} z_k$$

pero la varianza y el estimador de la varianza vienen dados por

$$Var(\widehat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

y

$$\widehat{Var}(\widehat{Y}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

en general ninguna variante bootstrap va a poder construir un estimador insesgado de la varianza, por lo que se suele considerar que la muestra fue obtenida mediante muestreo *pps cr* como una buena aproximación si la fracción de muestreo $f = \frac{n}{N}$ es chica.

Sean $(z^*_1, \ldots, z^*_{n^*})$ una muestra bootstrap obtenida mediante muestreo aleatorio simple con reemplazo de la muestra $(z_1, \ldots, z_n)$, entonces el estimador del total queda de la siguiente manera

$$\widehat{Y}^*_{HT} = \frac{1}{n^*}\sum_{k=1}^{n^*} z^*_k$$

donde z^*_k son variables aleatorias independientes e idénticamente distribuídas con esperanza y varianza dadas por

$$E_*(z_k^*) = \frac{1}{n}\sum_{k=1}^{n} z_k = \widehat{Y}_{HT}$$

y

$$Var_*(z_k^*) = \frac{1}{n}\sum_{k=1}^{n}(z_k - \widehat{Y}_{HT})^2$$

$\forall k = 1, \ldots, n$ entonces

$$Var_*(\widehat{Y}_{HT}^*) = \frac{1}{n^*}Var_*(z_k^*) = \frac{1}{n^*}\frac{1}{n}\sum_{k=1}^{n}(z_k - \widehat{Y}_{HT})^2$$

luego para $n^* = n - 1$ tenemos un estimador insesgado para la varianza en le caso de un muestreo *pps cr*.

Así el método bootstrap tiende a sobreestimar la varianza en muestreo con probabilidad proporcional al tamaño sin reemplazo. El sesgo suele ser despreciable si, como nombramos anteriormente, $f = \frac{n}{N}$ es chica.

Capítulo 4

Bootstrap generalizado especializado en Muestreo Poisson

El método Bootstrap para la estimación de la varianza es recomendable para encuestas que son llevadas a cabo por organismos nacionales de estadística que publican sus bases de datos ya que el usuario, a partir de ciertos pesos bootstrap generados con el método, puede fácilmente realizar su estimación del error. En este capítulo, nos centramos en explicar en que consiste el método bootstrap generalizado y luego se discuten temas como la elección de la distribución utilizada para generar los pesos bootstrap, la elección del número de repeticiones bootstrap y la posible aparición de pesos bootstrap negativos. La desventaja que presenta este método comparado a otros métodos de estimación de la varianza es que sufre un error de simulación y este puede no ser insignificante si el número de repeticiones bootstrap no es suficientemente grande.

Luego de algunas adaptaciones de la idea original de Efron vistos en el capítulo 3, nos centramos en la extensión que realizaron Bertail y Combris (1997) llamado *Boostrap Generalizado* o *Bootstrap ponderado*, que consiste en la generación aleatoria de pesos bootstrap de manera que los primeros dos (o más) momentos de error de diseño de muestreo son seguidos por los correspondientes momentos bootstrap. Con este enfoque, los pesos bootstrap se generan usando una distribución adecuada, sin requerir la creación real de una población bootstrap.

4.1. Boostrap generalizado

Como se comentó anteriormente la idea básica del método bootstrap generalizado consiste en generar los pesos bootstrap a fin de capturar los primeros dos (o más) momentos de error de diseño de muestreo. Esta técnica puede ser utilizada en la mayoría de los diseños de muestreo mientras que exista un estimador Horvitz Thompson de la varianza escrito de forma cuadrática semi definida positiva, y que según el estimador, sea posible calcular o aproximar con cierta precisión, las pobabilidades de inclusión de primer y segundo orden.

Otra característica importante del método bootstrap generalizado es que los pesos generados, mencionados anteriormente, no son creados para una variable en particular, lo que implica que una vez creados y puestos en la base de datos, el usuario puede utilizarlos para cualquier variable de interés.

Dada una población finita U de tamaño N, supongamos que queremos estimar el total de una cierta variable de interés, $Y = \sum_{k \in U} Y_k$, donde Y_k representa el valor observado de la variable Y en el k-ésimo elemento de la población. Sea s una muestra de tamaño n que se selecciona de dicha población de acuerdo con un diseño de probabilidades $P(S)$ y consideramos como estimador, el estimador Horvitz Thompson del total $\widehat{Y}_{HT} = \sum_{k \in s} w_k y_k$ donde $w_k = \frac{1}{\pi_k}$ y $\pi_k > 0$ es la probabilidad de selección de primer orden de la unidad k.

Sea j un entero positivo, se define el j-ésimo momento de error de diseño de muestreo al estimar Y por $\widehat{Y}_{HT}$

$$m_j = E_p(\widehat{Y}_{HT} - Y)^j$$

donde E_p significa que la esperanza se calcula con respecto al diseño de muestreo. A partir de la definición, es fácil ver que $m_1 = 0$, ya que $\widehat{Y}_{HT}$ es un estimador insesgado de Y y que m_2 es la varianza de diseño de $\widehat{Y}_{HT}$.

Consideramos el estimador insesgado $\widehat{m}_2$ de la varianza Horvitz Thompson para m_2. Entonces se tiene que

$$\widehat{m}_2 = \sum_{k \in s} \sum_{l \in s} \sigma_{kl} \, \check{y}_k \, \check{y}_l$$

donde, π_{kl} es la probabilidad de selección de segundo orden para las unidades k y l ($\pi_{kl} > 0 \; \forall \; k, l \in s$), y donde

$$\sigma_{kl} = \begin{cases} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} & si \; k \neq l \\ (1 - \pi_k) & si \; k = l \end{cases}$$

y

$$\check{y}_k = w_k y_k = \frac{y_k}{\pi_k}$$

esta estimación se puede reescribir de la forma cuadrática semi definida positiva de la siguiente manera

$$\widehat{m}_2 = \check{Y}' \, \Sigma \, \check{Y}$$

donde $\check{Y} = (\check{Y}_1, \ldots, \check{Y}_n)$ y Σ es la matriz simétrica de $n \times n$ que contiene σ_{kl} en la k-ésima fila y en la l-ésima columna.

El estimador Horvitz Thompson para la varianza puede utilizarse para cualquier diseño en el cual se puedan calcular o aproximar las probabilidades de seleccion de primer y segundo orden, sin embargo este estimador podría tomar valores negativos para algunos diseños, por

lo que Σ no es necesariamente semi definida positiva. Una alternativa para los diseños de tamaño fijo es el estimador de la varianza Sen-Yates-Grundy, que se puede escribir de la forma cuadrática como

$$\sigma_{kl} = \begin{cases} \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} & si\ k \neq l \\ (1-\pi_k) - \sum_{i\in s} \frac{\pi_{ki}-\pi_k\pi_i}{\pi_{ki}} & si\ k = l \end{cases}$$

donde ahora la matriz Σ es semi definida positiva si se cumple $\sigma_{kl} \leq 0\ \forall k \neq l$. Como tiene la restricción de que el diseño tiene que ser de tamaño fijo, no nos sirve para el caso de un diseño de muestreo Poisson, pero afortunadamente la varianza del estimador Horvitz Thompson es siempre positiva para este diseño.

En el caso Poisson como la elección de las unidades en la muestra se realizan de forma independiente una de la otra, tenemos que $\pi_{kl} = \pi_k\pi_l$ para todo $k \neq l, (k, l = 1, \ldots, n)$ y por lo tanto

$$\sigma_{kl} = \begin{cases} 0 & si\ k \neq l \\ (1-\pi_k) & si\ k = l \end{cases}$$

Entonces la matriz Σ es siempre semi definida positiva para este diseño.

Denotamos como peso bootstrap para la unidad k a $w_k^* = w_k\ a_k$, con a_k una variable aleatoria cuyas propiedades explicitaremos mas adelante. Luego, se define el estadístico bootstrap $\widehat{Y}_{HT}^* = \sum_{k\in s} w_k^* y_k$ y el error de muestreo bootstrap como $(\widehat{Y}_{HT}^* - \widehat{Y}_{HT})$. Luego queremos que se cumplan las siguientes condiciones

$$E_*(\widehat{Y}_{HT}^* - \widehat{Y}_{HT})^j = \widehat{m}_j \qquad j = 1, \ldots, J$$

con J un número entero positivo, en general $J = 2$ o $J = 3$. E_* indica que los momentos se evaluan con respecto a la distribución de los a_k, $k \in s$, condicionada a la muestra s. La distinción entre los diferentes procediminetos bootstrap se encuentran sólo en la elección de la distribución de a_k.

Sea a el vector que contiene en la posición k la variable bootstrap a_k y sea 1_n el vector de dimensión n que contiene 1 en todos sus elementos. Luego podemos escribir

$$\widehat{Y}_{HT}^* - \widehat{Y}_{HT} = \sum_{k\in s} w_k^* y_k - \sum_{k\in s} w_k y_k = \sum_{k\in s} (a_k - 1) w_k y_k = \sum_{k\in s} (a_k - 1)\check{y}_k = (a - 1_n)'\check{y}$$

por lo tanto, reescribiendo las condiciones anteriores se tiene que

$$E_*\left[(a - 1_n)'\,\check{y}\right] = 0 \ \Rightarrow\ E_*(a) = 1_n$$

y

$$E_*\left[(a - 1_n)'\,\check{y}\right]^2 = \check{y}'\,\Sigma\,\check{y} \ \Rightarrow\ E_*\left[(a - 1_n)\,(a - 1_n)'\right] = \Sigma.$$

Por lo cual, la elección de la distribución para los ajustes bootstrap tiene que satisfacer estas dos condiciones

- $E_* (a) = 1_n$.
- $E_* \left[(a - 1_n) (a - 1_n)' \right] = \Sigma$.

Un ejemplo de distribución bootstrap fue generado por Bertail y Combris (1997), quienes proponen al vector a como

$$a = 1_n + \Sigma^{1/2} \tilde{a}$$

donde $\tilde{a}$ es un vector que contiene n variables aleatorias independientes entre sí, con media 0 y varianza 1 y la matriz $\Sigma^{1/2}$ se obtiene a partir de la descomposición espectral de Σ, es decir, $\Sigma^{1/2} = \Gamma \Lambda^{1/2} \Gamma'$. Donde Λ es la matriz diagonal de autovalores de Σ y Γ es la correspondiente matriz ortonormal de autovalores.

Una opción simple para generar los $\tilde{a}_k$ $(k = 1, \ldots, n)$ es a partir de la distribución normal estándar, entonces el vector a sigue una distribución normal multivariante $N(1_n, \Sigma)$. Otra opción es utilizar la distribución

$$\tilde{a}_k = \begin{cases} -\epsilon & \text{con} \quad P(\tilde{a}_k = -\epsilon) = 1/1 + \epsilon^2 \\ \frac{1}{\epsilon} & \text{con} \quad P(\tilde{a}_k = \frac{1}{\epsilon}) = \epsilon^2/1 + \epsilon^2 \end{cases}$$

donde $\epsilon > 0$ es una constante a ser elegida por el estadístico. Si $\epsilon = 1$, la distribución de los $\tilde{a}_k$ queda simétrica.

4.2. Estimación de la varianza

Vamos a estimar la varianza $\widehat{Y}_{HT}$ a partir de un estimador de la varianza de $\widehat{Y}^*_{HT}$ a través del método de Monte Carlo, mediante la generación de B vectores independientes $a^{(b)}$con $(b = 1, \ldots, B)$ con la misma distribución que el vector a. El elemento de orden k de $a^{(b)}$ se nota por $a_k^{(b)}$ y su peso asociado $w_k^{*(b)} = w_k a_k^{(b)}$.

Luego la estimación de la varianza bootstrap $\widehat{m}_2 = E_*(\widehat{Y}^*_{HT} - \widehat{Y}_{HT})^2$ es

$$Var_B(\widehat{Y}_{HT}) = \frac{1}{B} \sum_{b=1}^{B} (\widehat{Y}^{*(b)}_{HT} - \widehat{Y}_{HT})^2 = \check{y}' \, \Sigma_B \, \check{y}$$

donde $\widehat{Y}^{*(b)}_{HT} = \sum_{k \in s} w_k^{*(b)} y_k$, $(b = 1, \ldots, B)$ y

$$\Sigma_B = \frac{1}{B} \sum_{b=1}^{B} (a^{(b)} - 1_n)(a^{(b)} - 1_n)'$$

la matriz Σ_B es la versión Monte Carlo de Σ.

4.3. Enfoque pseudo-población

Una pseudo-población, como nombramos en el Capítulo 3, se crea mediante la replicación de cada unidad de la muestra w_k veces (suponiendo que w_k son todos enteros).

A continuación, se selecciona una muestra bootstrap por muestreo Poisson utilizando las probabilidades de selección originales, este procedimiento es equivalente a la generación de los a_k de forma independiente, para $k \in s$, a partir de la distribución binomial $Bin(w_k, \pi_k)$. Es fácil ver que este ajuste bootstrap satisface las condiciones necesarias.

Desafortunadamente, el peso w_k por lo general no son números enteros. Para solucionar este problema, se sugeriere que se extienda este enfoque de pseudo-población replicando B veces los tres pasos siguientes

1. Se genera aleatoriamente

$$w_k^r = \begin{cases} \lfloor w_k \rfloor & \textit{con probabilidad } \lfloor w_k \rfloor + 1 - w_k \\ \lfloor w_k \rfloor + 1 & \textit{con probabilidad } w_k - \lfloor w_k \rfloor . \end{cases}$$

   donde $\lfloor w_k \rfloor$ es el número entero más grande menor o igual a w_k. Es fácil ver que la esperanza de w_k^r es igual a w_k.

2. Crear una pseudo-población mediante la replicación de cada unidad w_k^r veces. El estimador $\widehat{Y}_{HT}$ de esta pseudo-población es $\widehat{Y}_{HT}^r = \sum_{k \in s} w_k^r y_k$.

3. Desde la pseudo-población, generar una muestra bootstrap utilizando el muestreo Poisson y las probabilidades de selección de primer orden originales.

Los últimos dos pasos del procedimiento anterior equivalen a la generación de los ajustes bootstrap a_k^r de forma independiente a partir de la distribución binomial $Bin(w_k^r, \pi_k)$, por lo que la generación física de la pseudo-población no es necesaria.

El estimador bootstrap resulta $\widehat{Y}_{HT}^{r*} = \sum_{k \in s} w_k^{r*} y_k$, donde $w_k^{r*} = w_k a_k^r$ y el error bootstrap es $\widehat{Y}_{HT}^{r*} - \widehat{Y}_{HT}^r$. Podemos ver facilmente que este error se puede reescribir como

$$\begin{aligned} \widehat{Y}_{HT}^{r*} - \widehat{Y}_{HT}^r &= \sum_{k \in s} w_k y_k (a_k^r - \pi_k w_k^r) \\ &= \sum_{k \in s} w_k y_k (a_k - 1) \\ &= \widehat{Y}_{HT}^* - \widehat{Y}_{HT} \end{aligned}$$

donde $a_k = a_k^r + (1 - \pi_k w_k^r)$.

Es fácil comprobar que el vector bootstrap a_k satisface las condiciones necesarias. Desafortunadamente a_k puede ser negativo para algún $k \in (1, \ldots, n)$ pero es siempre mayor que -1.

Capítulo 5

Aplicación a datos reales

Para este Capítulo, a modo de ilustración, tomamos una base de personas del censo 2010 pertenecientes a un aglomerado de una cierta provincia, que por cuestiones de confidecialidad no diremos de cual de trata, que tomaremos como nuestro universo de población.

Dicha base consta de 652 registros. Tomaremos como nuestra variable de interés el total de desempleados, sumando la variable para toda la población tenemos un total de 2478 personas desempleadas y además tomarémos como nuestra variable auxiliar de tamaño la cantidad de personas en el aglomerado con un total de 66915. A partir de esta población vamos a ver distintos diseños de muestreo, para luego estimar el total de desempleados en la población que llamaremos θ y estimar la varianza a partir del método Bootstrap comparandola con la verdadera varianza del diseño.

A la hora de hacer inferencia a partir de una muestra hay varias cuestiones que uno debe tener en cuenta, una de ellas es el tipo de diseño y el tamaño de muestra extraída de la población, para que se pueda obtener un equilibrio entre el costo y la rapidez del estudio realizado. Estas cuestiones son necesarias para que los datos obtenidos sean representativos, es decir, para poder estimar con un cierto nivel de confianza.

La simulación se realizó a partir de funciones implementadas en R, que permiten hacer la selección según el diseño y el tamaño de muestreo, calcular la varianza real del estimador y realizar la estimación bootstrap con la cantidad de réplicas deseadas para cada diseño.

En la siguiente tabla, fijado un diseño MAS con reemplazo, iremos variando el tamaño de muestra y realizaremos la estimación del total de desempleados correspondiente para 1000 réplicas bootstrap.

Tamaño muestral	Fracción de muestreo	Estimación muestra original	Estimación bootstrap	Varianza muestra original	Varianza bootstrap	Coeficiente de variación real (%)	Coeficiente de variación (%)	Intervalo de confianza 95%	
								Lim. Inferior	Lim. Superior
50	0,08	2295	2288	143024	137420	15,3	16,2	1617	3104
75	0,12	3347	3327	95349	186933	12,5	13,0	2547	4190
100	0,15	2484	2476	71512	66398	10,8	10,4	2008	2986
125	0,19	2582	2580	57210	58466	9,7	9,4	2138	3072
150	0,23	2647	2652	47675	51688	8,8	8,6	2238	3117
175	0,27	2224	2231	40864	28537	8,2	7,6	1904	2545
200	0,31	2813	2810	35756	53234	7,6	8,2	2370	3277
225	0,35	2275	2279	31783	31458	7,2	7,8	1933	2646
250	0,38	2626	2627	28605	29289	6,8	6,5	2311	2984
275	0,42	2613	2607	26004	30155	6,5	6,7	2266	2971
300	0,46	2356	2362	23837	22109	6,2	6,3	2071	2658
325	0,50	2435	2435	22004	21270	6,0	6,0	2153	2714
350	0,54	2776	2777	20432	24976	5,8	5,7	2474	3085
375	0,58	2549	2546	19070	20219	5,6	5,6	2288	2832
400	0,61	2443	2438	17878	18254	5,4	5,5	2176	2714

Tabla 1:Total de desocupados para un diseño MAS cr para distintos tamaños de muestras.

A partir de la *Tabla* 1 cabe pensar que a mayor tamaño de muestra inicial, fijado el tamaño de réplicas bootstrap, se obtienen menor coeficiente de variación y menor longitud del intervalo de confianza de la estimación bootstrap. Claro está que esta tabla fue construida a partir de una muestra inicial para cada uno de los tamaños.

Sean un parámetro de interés θ, un estimador $\widehat{\theta}$ y las B estimaciones bootstrap $\left(\widehat{\theta}_1, \ldots, \widehat{\theta}_B\right)$. Llamamos coeficiente de variación bootstrap de la estimación $\widehat{\theta}$ a la siguiente expresión

$$CV_B\left(\widehat{\theta}\right) = \frac{\sqrt{Var_B\left(\widehat{\theta}\right)}}{E\left(\widehat{\theta}\right)}$$

donde

$$Var_B\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{B}^{b=1}\left(\widehat{\theta}_b - \bar{\theta}\right)^2$$

$$\bar{\theta} = \frac{1}{B}\sum_{B}^{b=1}\widehat{\theta}_b$$

Como en nuestro caso conocemos el verdadero valor del parámetro de interés en lugar de poner en el denominador de la expresión del coeficiente de variación $E\left(\widehat{\theta}\right)$ ponemos θ. En lo que sigue verémos un ejemplo de como funciona el método Bootstrap.

Fijemos una muestra MAS con reemplazo y un tamaño de muestra inicial de 100 elementos de la población. Veamos el método bootstrap para 200 réplicas.

El siguiente gráfico muestra la distribución de las 200 muestras bootstrap junto con dos rectas que representan el límite inferior y el límite superior del intervalo de confianza de nivel 95 %.

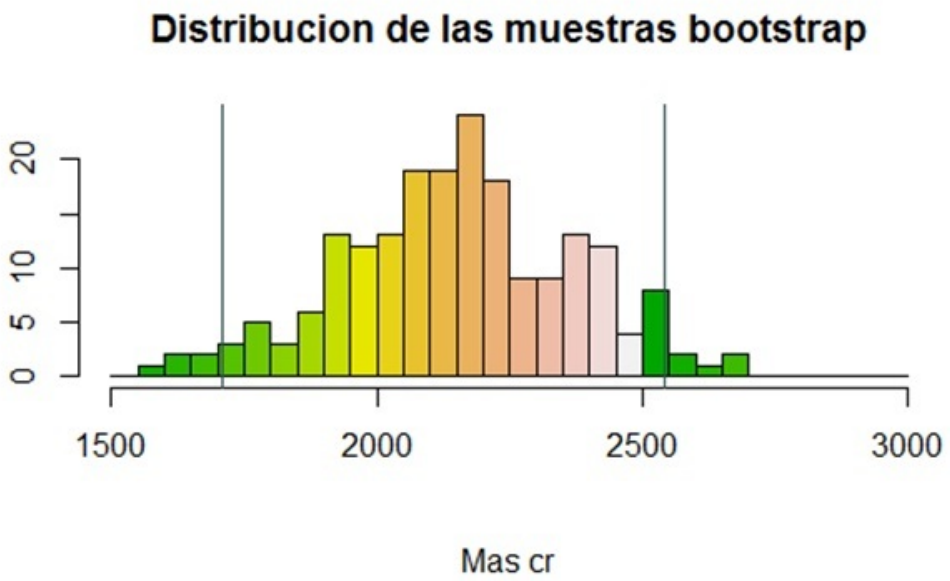

Gráfico 1: Distribución de las estimaciones para 200 muestras bootstrap.

Ahora veamos una tabla donde se muestra cada una de las 200 muestras bootstrap con su respectiva estimación

Número de réplica bootstrap	Estimación	Número de réplica bootstrap	Estimación	Número de réplica bootstrap	Estimación	Número de réplica bootstrap	Estimación
1	2210	51	1976	101	2295	151	2328
2	2536	52	2445	102	2249	152	2126
3	2054	53	2399	103	2165	153	2406
4	2412	54	1943	104	2204	154	1708
5	1584	55	2315	105	2373	155	2112
6	1904	56	2184	106	2165	156	2243
7	2262	57	2165	107	2034	157	1741
8	1904	58	1878	108	2139	158	1956
9	2530	59	1734	109	2126	159	2289
10	2021	60	2458	110	2132	160	1643
11	2262	61	2086	111	1780	161	2249
12	2289	62	2184	112	2308	162	2360
13	2171	63	2002	113	2191	163	2093
14	2165	64	1858	114	2360	164	1865
15	2517	65	2243	115	2132	165	2191
16	2373	66	2158	116	2478	166	1930
17	2119	67	2106	117	2452	167	2601
18	1963	68	2112	118	2165	168	2380
19	2328	69	2399	119	2132	169	2380
20	2067	70	2262	120	2119	170	2119
21	1839	71	2236	121	2197	171	2073
22	2054	72	2080	122	2158	172	2152
23	2419	73	2132	123	2191	173	2073
24	1982	74	2008	124	2073	174	2412
25	1695	75	2034	125	2236	175	2119
26	1995	76	2008	126	2041	176	1936
27	2204	77	1989	127	2093	177	2491
28	2543	78	1995	128	2445	178	1897
29	2419	79	2047	129	2054	179	2204
30	2158	80	2126	130	2582	180	2347
31	1910	81	1963	131	2445	181	2236
32	2223	82	2406	132	2158	182	2099
33	1982	83	2223	133	2223	183	2347
34	2158	84	2171	134	2425	184	1852
35	1910	85	1904	135	2139	185	2158
36	2419	86	2517	136	2067	186	2315
37	2380	87	2152	137	1963	187	1826
38	2523	88	2667	138	2243	188	2047
39	2373	89	2262	139	1904	189	1917
40	1786	90	2556	140	2693	190	2021
41	1773	91	2191	141	2249	191	2015
42	2328	92	2425	142	2093	192	2523
43	2047	93	1786	143	1963	193	2171
44	1989	94	2360	144	2080	194	1949
45	2112	95	2054	145	2054	195	1949
46	2132	96	2347	146	2080	196	2217
47	2256	97	1617	147	2054	197	1793
48	2262	98	2517	148	1917	198	1689
49	2360	99	1839	149	2139	199	2171
50	1865	100	2354	150	2249	200	2021

Tabla 2: Estimación de total de desempleados. $b = 200$ de una muestra inicial Mas cr, $n = 100$.

Luego como resumen de la Tabla 2 tenemos que

$$\widehat{Y}^*_{HT} = \frac{1}{200}\sum_{b=1}^{200}\widehat{Y}^{*,b}_{HT} = 2152,93$$

donde $\widehat{Y}^{*,b}_{HT}$ es el estimador Horvitz Thompson evaluado en la b-ésima muestra bootstrap, entonces el estimador de la varianza bootstrap es

$$Var_B(\widehat{Y}^*_{HT}) = \frac{1}{200}\sum_{b=1}^{200}(\widehat{Y}^{*,b}_{HT} - \bar{\widehat{Y}}^*_{HT})^2 = 47020,03$$

obteniendo un coeficiente de variación de

$$CV_B(\widehat{Y}^*_{HT}) = \frac{\sqrt{47020,03}}{2478} = 0,09.$$

Para todos los diseños presentados se realizaron los calculos verdaderos de la varianza del estimador, salvo para el caso del diseño PPS sr en donde no es posible, a los efectos prácticos, calcularla. De este modo se procedió a realizar una estimación de la misma a partir de simulación, donde se realizaron 25000 muestras originales y se calculó el estimador para cada una de esas muestras y luego su varianza. El siguiente gráfico muestra cual sería la varianza del estimador hasta la muestra número x (donde x representa el valor en el eje de las abscisas)

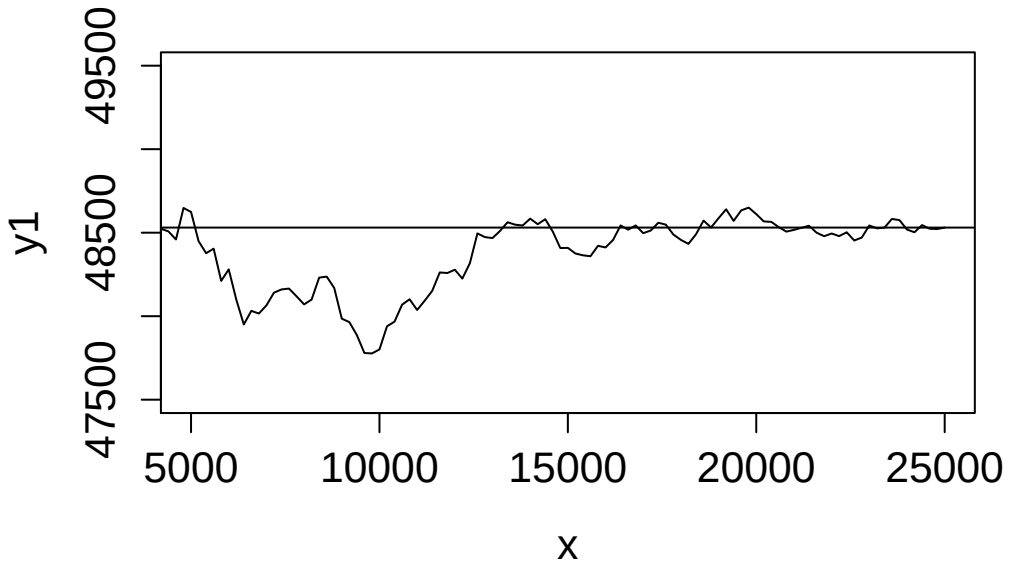

Gráfico 2: Estimación de la varianza para un diseño PPS sr a partir de 25000 muestras de tamaño $n = 100$.

Otro detalle a tener en cuenta al realizar una estimación bootstrap es el tamaño de réplicas. En la siguiente tabla vemos un resumen de los métodos para diferentes tamaños de réplicas bootstrap.

Diseño	b=500	b=1000	b=1500	b=2000	b=2500	b=3000
MAS cr	12,38	11,14	11,73	11,92	11,96	11,84
MAS sr FC	8,98	9,66	9,30	9,61	9,24	9,45
MAS sr REES	10,86	10,12	10,23	10,36	10,24	10,18
PPS cr	10,64	10,10	10,23	10,00	10,17	10,03
PPS sr	9,94	10,08	9,83	9,64	9,64	9,98
Poisson	10,98	10,74	10,69	11,13	10,80	10,65

Tabla 3: Coeficiente de variación estimado en % para diferentes tamaños de réplicas boostrap.

Ahora fijado el tamaño de la muestra , $n = 100$, y la cantidad de replicaciones bootstrap, $b = 1000$, veamos como se comportan los distintos diseños de muestreo para la estimación del total de desempleados

Diseño	Estimación	Desvío estándar	Coeficiente de variación en %	Intervalo de confianza	
				Lim. Inferior	Lim. Superior
MAS cr	2660	310,92	12,5	2106	3325
MAS sr FC	2595	249,27	10,1	2114	3155
MAS sr REES	2595	265,03	10,7	2094	3109
PPS cr	2423	186,99	7,5	2048	2800
PPS sr	2663	274,52	11,1	2154	3241
Poisson	3102	233,76	9,4	2659	3551

Tabla 4: Estimación del total de desempleados , a partir de una muestra de tamaño 100 y 1000 replicaciones bootstrap para distintos diseños de muestreo.

La *Tabla* 4 fue realizada a partir de una muestra extraída de cada diseño, lo cual no tiene mucha representatibidad a la hora de comparar, por lo que los siguientes gráficos nos muestran la distribución del coeficiente de variación estimados a partir de 3000 muestras originales de cada diseño, además mostramos como se distribuyen alrededor del verdadero coeficiente de variación.

Diseño	Coeficiente de variación real en %
MAS cr	7,63
MAS sr	6,36
PPS cr	6,29
PPS sr	4,84
Poisson	7,51

Tabla 5: Coeficiente de variación real en % del total de desocupados a partir de una muestra de tamaño 200 para los distintos diseños de muestreo.

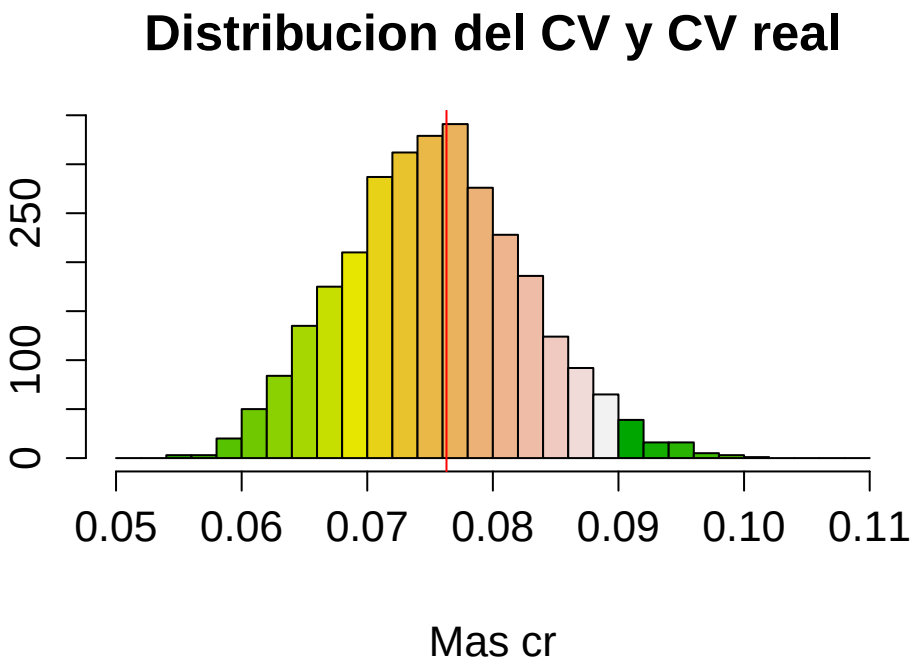

Gráfico 3: Distribución del coeficiente de variación bootstrap a partir de 3000 muestras de un diseño MAS sin reemplazo de tamaño 200.

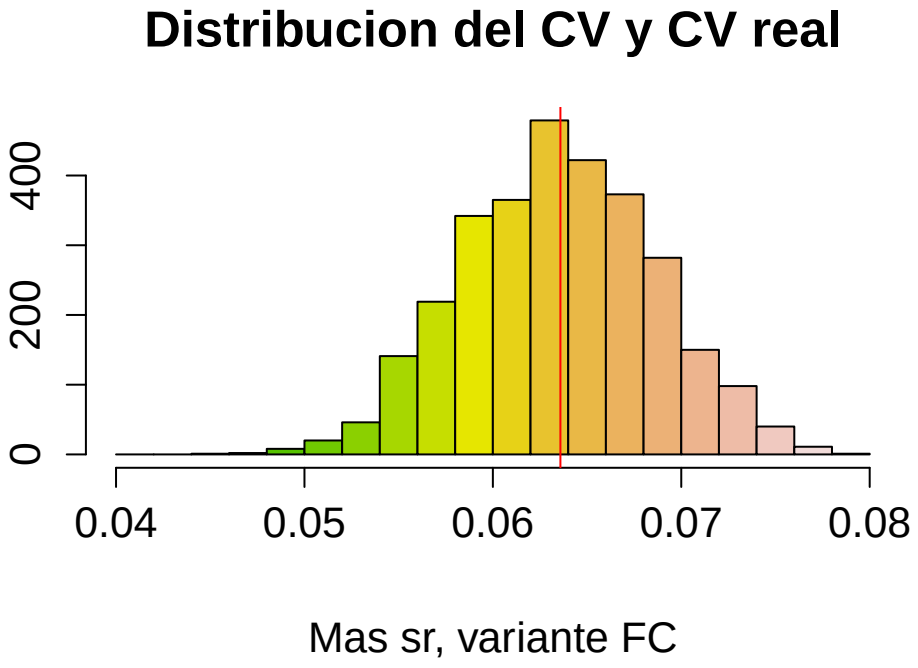

Gráfico 4: Distribución del coeficiente de variación bootstrap con variante FC a partir de 3000 muestras de un diseño MAS sin reemplazo de tamaño 200.

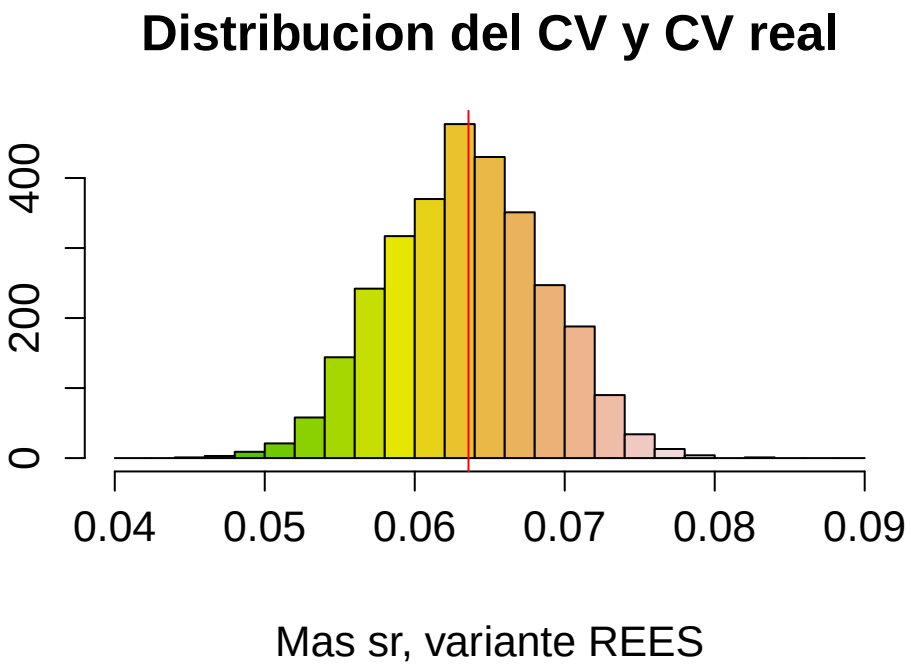

Gráfico 5: Distribución del coeficiente de variación bootstrap con variante REES a partir de 3000 muestras de un diseño MAS sin reemplazo de tamaño 200.

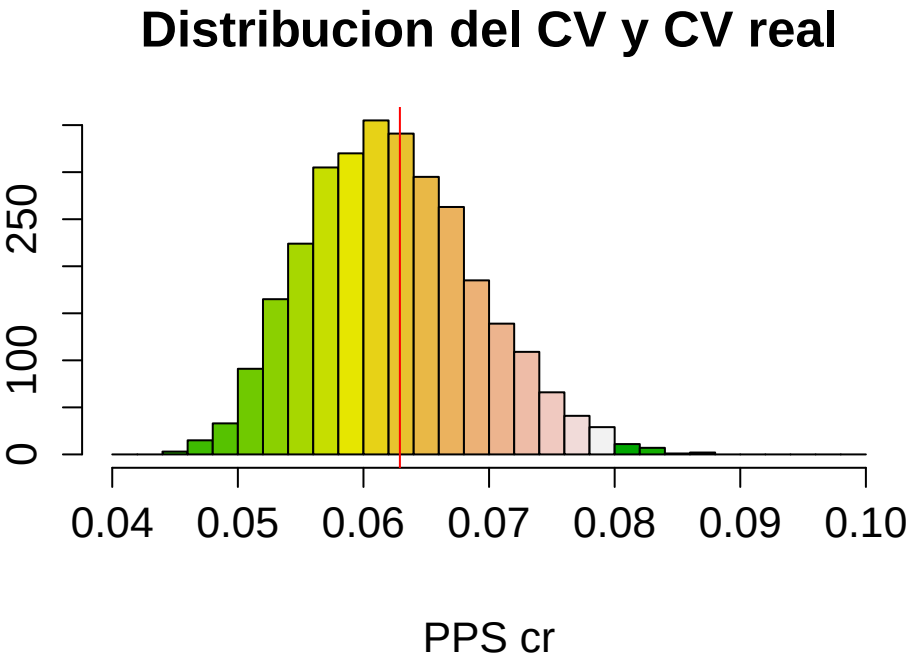

Gráfico 6: Distribución del coeficiente de variación bootstrap a partir de 3000 muestras de un diseño PPS con reemplazo de tamaño 200.

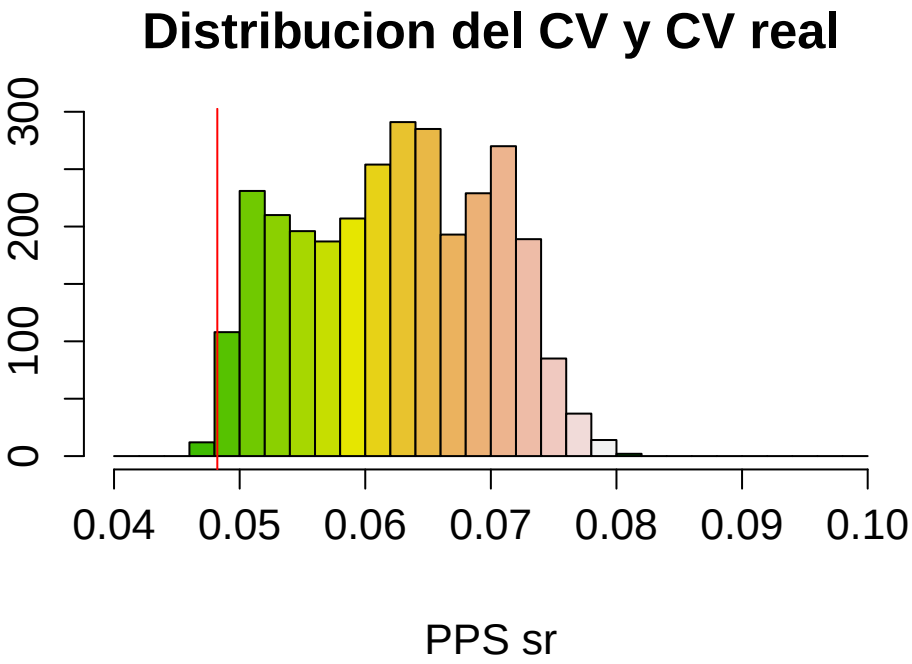

Gráfico 7: Distribución del coeficiente de variación bootstrap a partir de 3000 muestras de un diseño PPS sin reemplazo de tamaño 200.

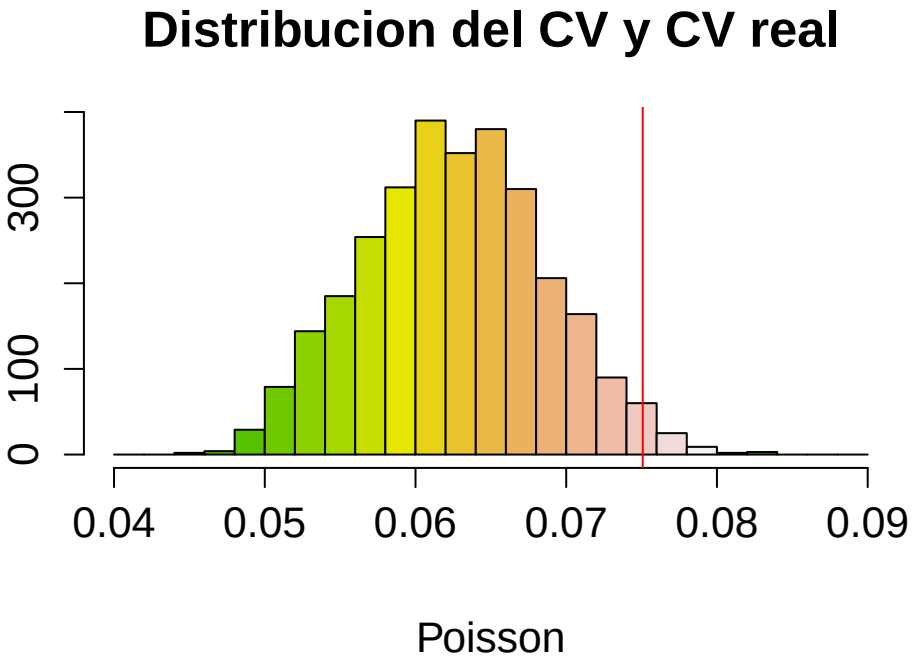

Gráfico 8: Distribución del coeficiente de variación bootstrap a partir de 3000 muestras de un diseño Poisson de tamaño 200.

A partir de los gráficos vemos que las estimaciones bootstrap del coeficiente de variación ajustan bien para los diseños Mas con y sin reemplazo y PPS con reemplazo planteados en este documento, no así para el diseño PPS sin reemplazo y para el diseño Poisson. Recordar que el diseño PPS sin reemplazo se aproximaba bastante bien a un diseño PPS con reemplazo si la fracción de muestreo es lo suficientemente chica, no es nuestro caso como muestra el *gráfico* 7 ya que tenemos $n = 200$ y $N = 652$, unas de las posibles soluciones para este problema sería disminuir el n para poder lograr una fracción de muestreo más pequeña, que los diseños sean más parecidos y por lo tanto que bootstrap aproxime mejor en ese caso, pero correríamos el riesgo de que al ser una muestra más pequeña deje de ser representativa a la población. Veamos como se distribuye las estimaciones Bootstrap con un tamaño de muestra original de $n = 100$.

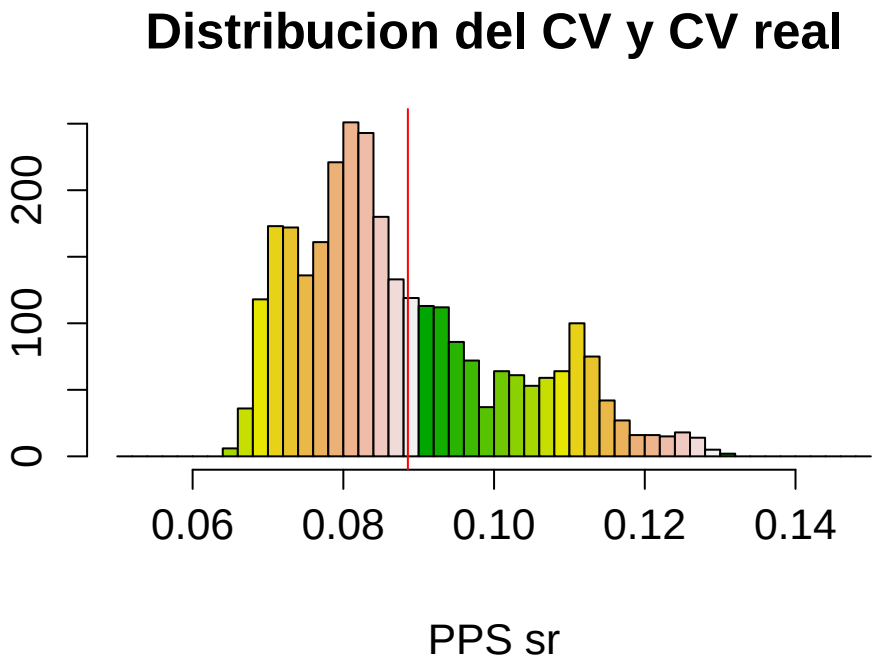

Gráfico 9: Distribución del coeficiente de variación bootstrap a partir de 3000 muestras de un diseño PPS sin reemplazo de tamaño 100.

Vemos que hay una leve mejoría pero aún no es suficiente para obtener una buena aproximación bootstrap. En la práctica por lo general se cuenta con una población de estudio mucho más grande que en este documento, lo que permite jugar un poco más con el tamaño de muestra original elegido. Para el caso de diseño Poisson veremos en el siguiente gráfico que la variante generalizada del método Bootstrap soluciona el problema para este diseño.

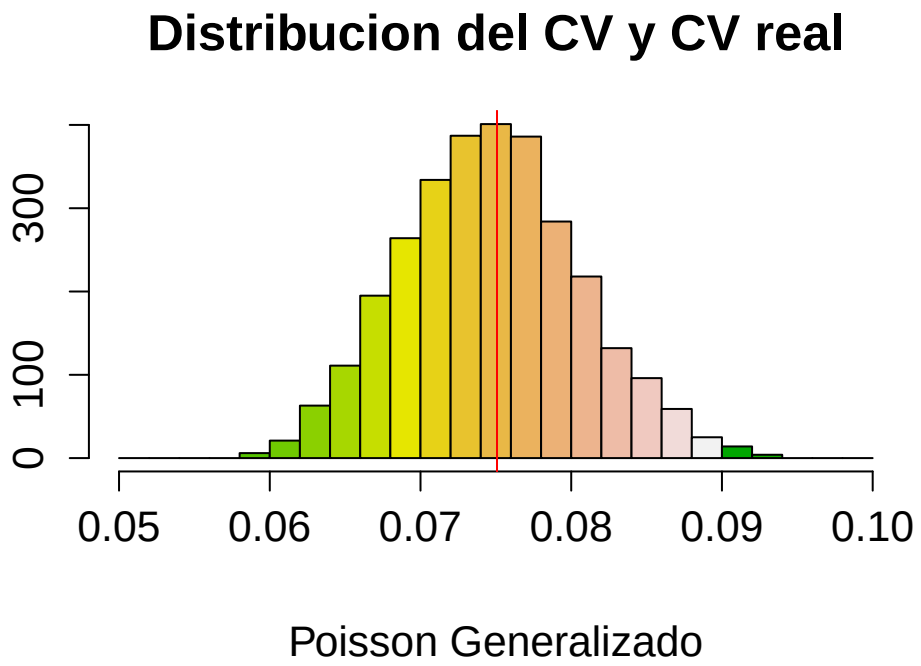

Gráfico 10: Distribución del coeficiente de variación bootstrap Generalizado a partir de 3000 muestras de un diseño Poisson.

Por lo tanto tenemos que el método Bootstrap generalizado ajusta mucho más para el caso de un diseño Poisson que el método bootstrap original.

Apéndice A

Funciones en R

Las funciones que se utilizaron para la realización de este trabajo fueron de realización propia, en lo que sigue se mostrará en código en *R*.

La primera función realizada es la funcion selecciono, que realiza la selección del diseño y el tamaño deseado.

```
library(sampling)
library(pps)
library(TeachingSampling)
library(samplingVarEst)
```

```
# data= base de la población
# tipo=tipo de seleccion (MAS, pps, poisson)
# n=tamaño de la muestra
# rep=si 0 no (si es "SI" es con reposición, si es "NO" es sin reposición)
# variable=variable de estudio
# var_tamaño= variable auxiliar de tamaño
```

```
selecciono<-function (data, tipo, n, rep, variable,var_tamaño){

  N <- NROW(data)
  if (tipo=="MAS"){
    if (rep=="si"){

      # genero la matriz de las probabilidades de inclusión
```

```
    data$pik<-n/N

    #selecciono la muestra
    s<-S.WR(N,n)
    muestra<-data[s,]

  }else{

    # genero la matriz de las probabilidades de inclusión
    data$pik<-n/N

    #selecciono la muestra
    s<-S.SI(N,n)
    muestra<-data[s,]

  }
}

if (tipo=="pps"){
  if (rep=="si"){

    # genero la matriz de las probabilidades de inclusión
    sum<-sum(var_tamaño)
    data$pik<-n*var_tamaño/sum

    #selecciono la muestra
    s<-S.PPS(n,var_tamaño)
    muestra<-data[s[,1],]

  }else{

    # genero la matriz de las probabilidades de inclusión
    sum<-sum(var_tamaño)
    data$pik<-n*var_tamaño/sum

    #selecciono la muestra
    s<-UPsystematic(data$pik)
    muestra<-getdata(data,s)

  }
}
if (tipo=="poisson"){

  # genero la matriz de las probabilidades de inclusión
```

```
    sum<-sum(var_tamaño)
    data$pik<-n*var_tamaño/sum

    #selecciono la muestra
    s<-S.PO(N,data$pik)
    muestra<-data[s,]

  }

  out<-muestra
  invisible(out)
}
```

La siguiente función realiza para cada diseño de muestreo y el tamaño de la muestra el cálculo de la verdadera varianza del estimador.

```
library(sampling)




# data= base de la población
# tipo=tipo de seleccion (MAS, pps, poisson)
# n=tamaño de la muestra
# rep=si 0 no (si es "SI" es con reposición, si es "NO" es sin reposición)
# variable=variable de estudio
# var_tamaño= variable auxiliar de tamaño




var_real<-function (data, tipo, n, rep, variable, var_tamaño){

  N <- NROW(data)
  if (tipo=="MAS"){
    if (rep=="si"){

      # varianza real del estimador para este diseño
      var<- (N^2/n)*((N-1)/N)*var(variable)

    }else{

      # varianza real del estimador para este diseño
```

```
    var<-(N^2/n)*(1-(n/N)) *var(variable)

  }
}

if (tipo=="pps"){
  if (rep=="si"){

    # genero la matriz de las probabilidades de inclusión
    sum<-sum(var_tamaño)
    data$pik<-n*var_tamaño/sum

    # varianza real del estimador para este diseño
    data$pk<-var_tamaño/sum
    data$zk<-variable/data$pk
    toty<-sum(variable)
    data$s<-data$pk*(data$zk-toty)^2
    tots<-sum(data$s)
    var<-tots/n

  }else{

    sum<-sum(data$personas)
    data$pik<-n*data$personas/sum

    # varianza real del estimador para este diseño
    y<-numeric(25000)
    for(i in 1:25000){
      s<-UPsystematic(data$pik)
      muestra<-getdata(data,s)
      y[i]<-sum(muestra$desoc*1/muestra$pik)
    }
    var<-var(y)
    x<-seq(1000,25000,by=200)
    y1<-numeric(length(x))
    for(j in 1:length(x)){
      z<-y[0:x[j]]
      y1[j]<-var(z)
    }
    c1<-var-1000
    c2<-var+1000
    plot(x, y1, type="l",ylim=c(c1,c2),xlim=c(5000,25000))
    abline(h=var)
  }
```

```
  }

  if (tipo=="poisson"){

    # genero la matriz de las probabilidades de inclusión
    sum<-sum(var_tamaño)
    data$pik<-n*var_tamaño/sum

    # varianza real del estimador para este diseño
    data$tot<-variable*variable*(1/data$pik-1)
    var<-sum(data$tot)
    data$tot <-NULL
  }

  out<-var
  invisible(out)

}
```

La siguiente función realiza las muestras bootstrap a partir de la muestra original y el tamaño de las réplicas, devolviendo el estimador en la muestra original, un vector con las B estimaciones bootstrap, la varianza bootstrap, el desvio estándar y el intervalo de confianza bootstrap a partir del percentil.

```
library(TeachingSampling)
```

```
# muestra
# tipo=tipo de seleccion (MAS, pps, poisson)
# rep=si 0 no (si es "SI" es con reposición, si es "NO" es sin reposición)
# variante= fc,rees (solo se toma en cuenta si diseño=MAS y rep =no)
# n=tamaño de la muestra original
# N=tamaño de la población
# estadistico
# m=tamaño de las muestras boostrap
# b=cantidad de muestras boostrap
# a=nivel de significación del intervalo de confianza (0.05 o 0.10)
# variable=variable de interés
```

```
bootstrap<-function (muestra, tipo, rep, variante, n, N, estadistico, m, b, a, variable){

  t <- numeric(b)
  muestra$w<-1/muestra$pik
  t0<-estadistico(muestra$desoc,muestra$w)

  if (tipo=="MAS"){
    if (rep=="si"){

      for(i in 1:b){
        s<-S.WR(n,m)
        m_boot<-muestra[s,]
        t[i]=estadistico(m_boot$desoc,m_boot$w*n/m)
      }
      var<-((b-1)/b)*var(t)
      sd<-sqrt(var)

      a1<-a/2
      a2<-1-a/2
      C1<-quantile (t, prob = a1)
      C2<-quantile (t, prob = a2)
      IC<-c(C1,C2)

    }else{

      if (variante=="fc"){

        for(i in 1:b){
          s<-S.WR(n,m)
          m_boot<-muestra[s,]
          t[i]=estadistico(m_boot$desoc,m_boot$w*n/m)
        }
        var<-(1-n/N)*((b-1)/b)*var(t)
        sd<-sqrt(var)

        a1<-a/2
        a2<-1-a/2
        C1<-quantile (t, prob = a1)
        C2<-quantile (t, prob = a2)
        IC<-c(C1,C2)
      }

      if (variante=="rees"){
```

```
      muestra$y_mean=mean(muestra$desoc)
      muestra$var2=muestra$y_mean+sqrt(1-n/N)*sqrt(m/(n-1))*(muestra$desoc-muestra$y_mean)
      for(i in 1:b){
        s<-S.WR(n,m)
        m_boot<-muestra[s,]
        t[i]=estadistico(m_boot$var2,m_boot$w*n/m)
      }
      var<-((b-1)/b)*var(t)
      sd<-sqrt(var)

      a1<-a/2
      a2<-1-a/2
      C1<-quantile (t, prob = a1)
      C2<-quantile (t, prob = a2)
      IC<-c(C1,C2)
    }
  }
}

if (tipo=="pps"){

  for(i in 1:b){
    s<-S.WR(n,m)
    m_boot<-muestra[s,]
    t[i]=estadistico(m_boot$desoc,m_boot$w*n/m)
  }
  var<-((b-1)/b)*var(t)
  sd<-sqrt(var)

  a1<-a/2
  a2<-1-a/2
  C1<-quantile (t, prob = a1)
  C2<-quantile (t, prob = a2)
  IC<-c(C1,C2)
}

if (tipo=="poisson"){
  for(i in 1:b){
    s<-S.WR(n,m)
    m_boot<-muestra[s,]
    t[i]=estadistico(m_boot$desoc,m_boot$w*n/m)
  }
  var<-((b-1)/b)*var(t)
```

```
    sd<-sqrt(var)

    a1<-a/2
    a2<-1-a/2
    C1<-quantile (t, prob = a1)
    C2<-quantile (t, prob = a2)
    IC<-c(C1,C2)
  }

  out<-list(t=t, t0=t0, var=var, sd=sd, IC=IC)
  invisible(out)
}
```

Por último la siguiente función realiza las muestras bootstrap generalizadas a partir de la muestra original con diseño Poisson, devolviendo los B vectores de ajuste, el estimador en la muestra original, un vector con las B estimaciones bootstrap, la varianza bootstrap, el desvio estándar y el intervalo de confianza bootstrap a partir del percentil.

```
# muestra
# b=cantidad de muestras boostrap
# estadistico
# c=nivel de significación del intervalo de confianza (0.05 o 0.10)



bootstrap_generalizado<-function(muestra,b,estadistico,c){

  muestra$w<-1/muestra$pik
  n1<-NROW(muestra)
  t0=estadistico(muestra$desoc,muestra$w)
  a<-matrix(0,n1,b)
  y<-numeric(b)

  sigma_sum<-matrix(0,n1,n1)
  sigma_raiz<-matrix(0,n1,n1)

  for(i in 1:n1){
    sigma_raiz[i,i]<-sqrt(1-muestra$pik[i])
  }

  unos<-rep(1, n1)
```

```
  for(j in 1:b){

    a1<-rbinom(n1,1,1/2)

    for(i in 1:n1){
      if (a1[i]==0) {
        a1[i]<- -1
      }
    }

    a[,j]<-unos+sqrt(unos-muestra$pik)*a1
    y[j]<-estadistico(muestra$desoc,muestra$w*a[,j])
    sigma_sum<-sigma_sum+t(t(a[,j]-unos))%*%(a[,j]-unos)

  }
  sigma_sum<-sigma_sum/b

  var<-(muestra$desoc*muestra$w) %*% (sigma_sum %*% t(t(muestra$desoc*muestra$w)))
  sd<-sqrt(var)

  a1<-c/2
  a2<-1-c/2
  C1<-quantile (y, prob = a1)
  C2<-quantile (y, prob = a2)
  IC<-c(C1,C2)

  out<-list(t=y, t0=t0, var=var, sd=sd, a=a, IC=IC)
  invisible(out)

}
```

Bibliografía

[1] Barbiero, A. and Mecatti, F. (2010) Bootstrap Algorithms for Variance Estimation in Sampling. *P Mantovan, P Secchi (eds.), Complex Data Moderling and Computationally Intensive Statistical Methods, Spriner, Milan.*pp. 57-69

[2] Bertail, P. and Combris, P. (1997) Bootstrap généralisé d'un sondage. *Annales d'économie et de Statistique.*

[3] Cochran, G. W. (1971) Técnicas de muestreo. *CECSA.*

[4] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *the Annals of Statistics .*

[5] Efron, B. (1981) Censored Data and the Bootstrap. *Journal of the American Statistical Association.*

[6] Gross, S. (1980) Median Estimation in Sample Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,*pp. 181-184

[7] Holmberg, A. (1998) A Bootstrap Approach to Probability -to- size Sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association,pp. 378-383

[8] Särndal, E. C.; Swensson, B. and Wretman, J. (1997) Model Assisted Survey Sampling. *New York, Springer-Verlag.*

[9] Wolter, M. K. (1985) Introduction to Variance Estimation.*New York, Springer-Verlag, XII.*