Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Título del Trabajo

Calibración en el muestreo

Autor: Gustavo Cano Posadas

Tutor: Jesús A. Tapia

INDICE:

Introducción

La calibración es una herramienta utilizada a menudo por los profesionales de los estudios sociales con encuesta para mejorar las estimaciones de los parámetros derivados de las preguntas (ítems) de una encuesta realizadas a una muestra de individuos, existiendo una información auxiliar sobre toda la población objetivo del estudio.

La característica clave es la modificación de los factores de elevación (pesos) de los individuos de la muestra encuestada para estimar con la información muestral y la información auxiliar poblacional características poblacionales (parámetros), como totales de población y frecuencias de categoría de ítem de encuesta. Por ejemplo, en una población, la edad y el sexo son variables auxiliares naturales.
La información de la edad y sexo de los individuos de una población está disponible en otras fuentes estadísticas, como un censo o un registro de la población (padrón) y mediante una modificación adecuada de los pesos de los individuos de la muestra, la estructura de la población en estas variables puede ser reproducida exactamente por la muestra.
Si la variable de estudio, es decir, la variable respuesta está correlacionada con la información auxiliar, la mayor precisión en las estimaciones suele obtenerse mediante la aplicación de los nuevos pesos calibrados de los individuos.

La calibración tiene su origen en los procedimientos introducidos por Deming y Stephan (1940). En años posteriores, la técnica de calibración ha recibido considerable atención por parte de las autoridades estadísticas oficiales, destacando:

- Instituto Nacional de Estadística en Reino Unido.
- Instituto Central de Estadística de Irlanda.
- Instituto Nacional de Estadística de España.
- Instituto Nacional de Estadística y Estudios de la Economía de Francia (INSEE).
- Servicio Público Federal de Economía de Bélgica.
- Servicio de Estadística de Canadá.
- Servicio de Estadística de Holanda.

El interés por el enfoque ha crecido ya que Deville y Särndal (1992) mostraron la equivalencia asintótica de la calibración para el estimador de regresión generalizada (Cassel, Särndal y Wretman, 1976), proporcionando así una forma de establecer las propiedades estadísticas de los estimadores calibrados.
A continuación, se explican los principios teóricos que subyacen a los métodos de calibración.

1. Descripción de la técnica de calibración

1.1. Formulación

Desde su incorporación en el muestreo, la información auxiliar se ha utilizado en la etapa de estimación con el objetivo de obtener estimaciones de mayor precisión. Por ejemplo, el estimador de regresión generalizada utiliza variables auxiliares con totales poblacionales conocidos procedentes de fuentes externas. Es bien sabido (Särndal, Swensson y Wretman, 1992) que cuanto más se correlacionen esas variables con la variable de estudio, mejor será la precisión del estimador de regresión generalizada.
Consideremos una población finita *U* de *N* unidades distintas. Podemos denotar la variable de estudio por Y. Ahora supongamos que queremos estimar el total de la variable de estudio, Υ**,** utilizando una muestra *s* de tamaño *n* extraída de U según un determinado diseño muestral.
Para estimar el total de la población en estas circunstancias, se utiliza un estimador lineal:

(1)

$$\hat{\Upsilon} = \sum_{k \epsilon S} \mathbf{d_k} * \mathbf{y_k}$$

Donde d_k es el peso muestral (factor de elevación) del individuo k.

Supongamos ahora que existen *J* variables auxiliares X_1 ... X_j ... X_J, llamadas variables de calibración, con totales poblacionales conocidos (en el caso de variables numéricas) o recuentos marginales (en el caso de variables categóricas). Sin pérdida de generalidad, podemos suponer que todas las variables de calibración son numéricas (de lo contrario, consideramos las variables 0/1 para cada categoría).
Buscamos nuevos pesos de muestreo ω_k que estén "lo más cerca posible" (según lo determinado por una cierta función de distancia) a los pesos iniciales d_k. Estos ω_k están calibrados en los totales X_j de las variables auxiliares *J*; es decir, verifican las ecuaciones de calibración:

(2)

$$\sum_{k \epsilon S} \boldsymbol{\omega_k} * \mathbf{x_{jk}} = \mathbf{X_j} = \sum_{k \epsilon U} \mathbf{x_{jk}} \text{ , j=1..J}$$

Donde S es una muestra tomada de la población U.

La solución a este problema viene dada por:

(3)

$$\boldsymbol{\omega_k} = \mathbf{d_k} * \boldsymbol{F}\left(\mathbf{x'_{jk}} * \boldsymbol{\lambda}\right)$$

Donde:

- $x'_{jk} = (x_{1k}, .., x_{jk}, .., x_{Jk})$
- λ es el vector de J multiplicadores de Lagrange
- F es la función de calibración cuyos términos dependen de la función de distancia que se utiliza.

El vector $\boldsymbol{\lambda}$ está determinado por la solución al sistema no lineal de J ecuaciones con J incógnitas resultantes de las ecuaciones de calibración:

(4)

$$\sum_{k\epsilon S} \mathbf{d_k} * \boldsymbol{F}\left(\mathbf{x'_{jk}} * \boldsymbol{\lambda}\right) * \mathbf{x_{jk}} = \mathbf{X_j}$$

Donde:

- X_j es el vector de dimensión (J, 1) que contiene los totales de las J variables auxiliares.
- x_{jk} es el vector de dimensión (J, k) que contiene los valores de las J variables auxiliares en la unidad muestral k.

Finalmente, el estimador calibrado del total, Υ, para la variable de estudio *Y* es:

(5)

$$\widehat{\boldsymbol{\Upsilon}}_{\mathbf{CAL}} = \sum_{\boldsymbol{k\epsilon S}} \boldsymbol{\omega_k} * \mathbf{y_k}$$

- Por ejemplo, si tomamos los X_j, totales conocidos de las J variables auxiliares, entonces calibramos construyendo pesos ω_k tal que:

(6)

$$\sum_{k\epsilon S} \boldsymbol{\omega_k} * \mathbf{x_{jk}} = \mathbf{X_j}$$

Donde los nuevos pesos ω_k están lo más cerca posible de los viejos pesos d_k. Esto se puede hacer minimizando la siguiente distancia cuadrática Chi-cuadrado:

(7)

$$\sum_{i\epsilon S} \frac{(\boldsymbol{\omega_k} - \mathbf{d_k})^2}{\mathbf{d_k}}$$

Donde d_k y ω_k son los pesos descritos en *(1)* y *(3)*, respectivamente.

1.2. Propiedades estadísticas de los estimadores de calibración.

Deville y Särndal (1992) demostraron que bajo supuestos generales un estimador de calibración es asintóticamente equivalente al estimador de regresión generalizada (GREG) (donde las variables de calibración son las variables de "regresión"), en el sentido de que:

(8)

$$\frac{\left(\hat{\Upsilon}_{CAL} - \hat{\Upsilon}_{GREG}\right)}{N} = O\left(\frac{1}{n}\right)$$

Por lo tanto, un estimador calibrado tiene un sesgo de orden 1/n, que es asintóticamente insignificante. Además, su varianza puede estimarse fácilmente sustituyendo la variable de estudio *y* con los residuos de regresión. En particular, si el modelo de regresión es bueno (es decir, las variables de calibración están correlacionadas con la variable de estudio), la varianza puede reducirse sustancialmente, a expensas de un sesgo muy pequeño.

1.3. Calibración y falta de respuesta.

Aunque originalmente se introdujo como un método de reducción de varianza, la calibración se puede usar tanto para aumentar la precisión de las estimaciones como para reducir el sesgo de falta de respuesta. Contrariamente al enfoque clásico para tratar la falta de respuesta, que consiste en crear grupos de respuesta homogéneos[1], las variables auxiliares deben conocerse únicamente para las unidades que responden. Sin embargo, sus totales de población también deben conocerse, lo cual es bastante restrictivo.

Para resolver este problema de que los totales poblacionales de las variables auxiliares deban ser conocidos, se desarrolló una teoría de calibración "supergeneralizada" (Deville, 2000). La idea es calcular los pesos de la forma:

(9)

$$\omega_k = d_k * F(z'_k * \lambda)$$

Donde:

[1] *Grupos de respuesta homogéneos:*
La muestra se divide en celdas donde se supone que la distribución de la respuesta es uniforme. Entonces, la tasa de respuesta dentro de una celda es una estimación (máximo verosímil) para esa distribución.

$\mathbf{z'_k = (z_{1k} , .. , z_{jk}, .., z_{Jk})}$ son las unidades que responden y son las únicas que es necesario conocer. Estos pesos deben satisfacer las ecuaciones de calibración basadas en las variables de calibración $\mathbf{(X_1 , .. , X_j, .., X_J)}$. *(6)*

$$\sum_{k \epsilon S} \omega_{jk} * x_{jk} = X_j$$

Siendo X_j el total de la variable auxiliar j.

Por lo tanto, no es necesario conocer los totales poblacionales de las variables de falta de respuesta. En particular, las variables obtenidas de las encuestas pueden usarse directamente para corregir la falta de respuesta. Esto es interesante para las encuestas de estudios sociales en las que se espera que la falta de respuesta esté correlacionada con las variables objetivo de la encuesta.

1.4. Elección de la información de calibración.

La elección de la información que se utilizará en la calibración no es tan fácil. Se ha asumido implícitamente en los apartados anteriores que tanto las variables de calibración como los totales de calibración son exactos, es decir, libres de errores. Por supuesto, en la práctica, esta suposición no se cumple. Los errores en la información auxiliar pueden dañar severamente los pesos calibrados, como se muestra en los dos ejemplos siguientes.

Ejemplo 1: Supongamos que los valores de los pesos calibrados, obtenidos a partir de la información auxiliar recopilada durante la encuesta, están subestimados, generando una alta variabilidad. Esto podría suceder, por ejemplo, si la herramienta de medición que se utiliza es defectuosa o no se ajusta correctamente a lo que se quiere medir. Por otro lado, el total de calibración correspondiente, que proviene de una fuente externa, se supone que es exacto. Está claro que la consecuencia de usar dicha variable en la calibración es que los pesos calibrados serán sesgados (sobreestimados).

Ejemplo 2: Supongamos ahora que calibramos los recuentos de frecuencia por estado de actividad. La encuesta recopila el estado de la actividad correspondiente al año de la encuesta. Por otro lado, los recuentos de frecuencia disponibles corresponden a un año anterior y se establecieron sobre la base de otra clasificación de estado de la actividad. La consecuencia del uso de información "inconsistente" sobre el estado de la actividad es que también hace que los pesos calibrados sean sesgados.

En resumen, podemos decir que la calibración es una técnica poderosa para mejorar las propiedades estadísticas de precisión de los estimadores, siempre que se tome precaución con respecto a la información que se utiliza.

Otro caso práctico ocurre cuando los totales de calibración provienen de otra encuesta. Por ejemplo, la Oficina Central de Estadísticas de Irlanda utilizó estimaciones de la EPA (Encuesta de Población Activa) para calibrar los datos de EU-SILC (European Union Statistics on Income and Living Conditions).
El EU-SILC proporciona datos sobre los indicadores sociales como la tasa de pobreza, inclusión social o pensiones entre otros.

El uso de los estimadores de calibración, como el estimador del total, en lugar de los valores "exactos" generalmente es inofensivo, siempre que:

- Estas estimaciones puedan considerarse (casi) imparciales y provengan de una muestra que tenga al menos el mismo tamaño. Si ese tamaño es mayor, se podría esperar una mejor precisión en las estimaciones.
- Ambas encuestas midan la misma información. En otras palabras, que los datos recopilados en ambas encuestas sean "consistentes" (véase el ejemplo 2 anterior).

1.5 Motivos para usar calibración

- Proporciona una forma sistemática de incorporación de información auxiliar disponible.

- Es un medio de obtener estimaciones consistentes de los parámetros que se deseen estimar utilizando para ello variables auxiliares con totales conocidos.

Sin embargo, estos objetivos también son abordados por otros métodos que también utilizan información auxiliar en las estimaciones. Es el caso de los métodos indirectos de regresión y de razón.

A continuación, se procede a explicar las ventajas de utilizar el estimador de calibración frente a los estimadores de regresión y de razón.

Cambios con el incremento y mejora de la información auxiliar

El estimador de razón resultó útil cuando se disponía de una única variable auxiliar. Pero la presencia de más y mejor información obligó a seleccionar otro tipo de estimadores que, además de utilizar las nuevas variables auxiliares, permitieran superar algunos inconvenientes de tipo práctico que se resumen a continuación.

Por un lado, es frecuente que se disponga de los totales poblacionales marginales de cada variable, pero no de los totales conjuntos. Por ejemplo, se puede disponer de los totales de población por grupos de edad y sexo, y también

de los totales de población española y extranjera. Pero puede ocurrir que no se disponga del cruce de nacionalidad por grupos de edad y sexo.

El segundo problema es particular de aquellas encuestas en las que se seleccionan viviendas y, en una etapa posterior, se recoge información de todas las personas que residan en cada vivienda muestral. En este caso, el factor de elevación derivado de las probabilidades de selección, es el mismo para una vivienda y para las personas que habitan en ella, dado que en la vivienda no se realiza ningún proceso de selección. Este hecho resulta útil, pues hay tablas de viviendas y tablas de personas con celdas coincidentes, y si el factor de elevación es el mismo, las estimaciones también lo serán.

El ejemplo tipo es el de la estimación del número de viviendas con una sola persona, y la del total de personas que viven solas. Sería útil emplear una técnica que permitiera realizar una reponderación de las personas de la muestra manteniendo el mismo factor para todas aquellas que residan en la misma vivienda.

Los métodos de calibrado intentan superar los problemas señalados. Esta propuesta metodológica, amplió el campo de actuaciones de tipo práctico, basado en la descripción del estimador de regresión generalizada (GREG) de Särndal, Swensson y Wretman (1992).

En resumen, la mejora en la calidad y facilidad de acceso a información auxiliar procedente de fuentes administrativas, unida al desarrollo de procedimientos informáticos, permite aplicar, de forma relativamente sencilla, técnicas de calibrado, que en general mejoran la precisión de las estimaciones de una encuesta, ayudan a corregir el sesgo introducido por la falta de respuesta, y proporcionan consistencia entre las cifras presentadas y las procedentes de otras fuentes.

No obstante, es necesario mantener un control riguroso de la información auxiliar que se considere, para evitar posibles sesgos derivados de la utilización de datos externos inadecuados.

1.6 Software en calibración

En esta sección introducimos los principales programas y paquetes utilizados en los procesos de calibración.

CALMAR

Es una macro perteneciente al software informático SAS.

Es utilizado en el Instituto Nacional de Estadística en Reino Unido, en el Instituto Central de Estadística de Irlanda, en el Instituto Nacional de Estadística de España, entre muchos otros. Fue desarrollado por el Instituto Nacional de Estadística y Estudios de la Economía de Francia (INSEE).

El éxito en su difusión se debe a que, apoyándose en un desarrollo teórico riguroso, ha sido programado de forma que su utilización no es complicada, y permite aplicar varias opciones de calibrado, tanto sobre variables auxiliares cualitativas como cuantitativas.

Existe otra versión, CALMAR2, desarrollada igualmente por el INSEE, que incluye procedimientos más sofisticados, pero es menos utilizado que CALMAR.

g-CALIB-S

Es un paquete implementado en el programa SPSS.

Fue desarrollado por el Servicio Público Federal de Economía de Bélgica. Es un paquete muy similar a CALMAR.

GES

El Software de Estimación Generalizada (GES) fue desarrollado por el Servicio de Estadística de Canadá. Al igual que el anterior, es un software similar a CALMAR.

BASCULA

Este software fue desarrollado en el Servicio de Estadística de Holanda.

Calib

Calib es una función perteneciente a la librería “sampling” de R.

A diferencia de los anteriores, este es un software libre que no precisa de licencia para su utilización. Ofrece una serie de opciones sencillas de implementar.

En este trabajo se trabajará con R, ya que no se dispone de la licencia necesaria de ninguno de los otros softwares citados, ya que, aunque la macro de CALMAR es gratuita, el software donde se lleva a cabo su ejecución, SAS, no.

2. Métodos de calibración

En esta sección describimos tres métodos de calibración y los criterios de elección de uno u otro.

2.1. Método lineal.

Se basa en la distancia cuadrática Chi-cuadrado explicada en *(7).*
Es el método más utilizado en la práctica debido a su sencillez y a que suele producir buenos resultados.
No obstante, con este método los pesos calibrados pueden tomar valores negativos, lo que no es deseable. Además, los pesos no están acotados y pueden tomar valores extremos.

2.2. Método exponencial.

Está basado en la siguiente distancia:

(10)

$$\omega_k * \log\left(\frac{\omega_k}{d_k}\right) - \omega_k + d_k$$

Al contrario que el método anterior, este siempre conduce a pesos positivos. Sin embargo, al igual que en el método lineal, estos pesos tampoco están acotados y pueden tomar valores bastante grandes.

2.3. Método lineal logit y truncado.

Ambos métodos están "acotados", lo que significa que proporcionan límites inferior y superior en las relaciones de peso $\left(\frac{\omega_k}{d_k}\right)$, generalmente denominadas “g-weights”. Por lo tanto, permiten controlar el rango en el que se calculan los pesos.

Hay que tener en cuenta que la elección de los límites de calibración inferior y superior no es arbitraria y depende de las variables de calibración elegidas: los límites deben ajustarse teniendo en cuenta las diferencias entre las estimaciones basadas en los pesos antiguos y los “valores de referencia” (benchmark) totales que se obtienen con los nuevos pesos.

En la práctica, esos límites están determinados por "adivinar y verificar": se comienza con un pequeño intervalo [inf, sup] y se va ampliando hasta que se consiga una solución.

2.4. Elección del método de calibración.

La elección del método de calibración generalmente se basa en criterios empíricos. De hecho, de acuerdo con 1.2, los estimadores calibrados son asintóticamente equivalentes al estimador de regresión generalizada, por lo que tienen el mismo sesgo y varianza. Sin embargo, aunque todos los métodos de calibración (es decir, todas las funciones de distancia) son en promedio equivalentes, no producen los mismos resultados para una muestra dada.

Por ejemplo, el método lineal puede producir pesos negativos o anormalmente grandes. Las estimaciones basadas en probabilidades de inclusión pequeñas pueden tomar valores inesperados. En este sentido, los métodos "acotados" (logit y lineal truncado) son generalmente preferidos ya que evitan pesos extremos. Sin embargo, vale la pena probar todos los métodos para ver su impacto en el cálculo de los pesos.

3. Aplicación de la técnica de calibración

3.1 Técnicas de calibración en las encuestas de hogares en el INE.

Con el paso del tiempo y gracias a la mejora de los procesos informáticos, al control de los registros administrativos y, en general, a la mayor capacidad técnica de las diferentes organizaciones, se dispone de una mayor variedad de fuentes externas fiables, actualizadas, y relacionadas con las variables objetivo de las encuestas.

Este hecho ha permitido el desarrollo y puesta en práctica de técnicas de estimación más complejas, que son capaces de incorporar información auxiliar multivariante y más desagregada. Todo ello con el fin de mejorar la calidad de las estimaciones obtenidas y, además, como garantía de coherencia de la información publicada.

Entre las fuentes auxiliares habitualmente utilizadas en las encuestas de hogares, se encuentran, al nivel de comunidad autónoma, la población por grupos quinquenales de edad y sexo, la población por nacionalidad y los hogares por tamaño. Y es de esperar que a medio plazo se pueda contar con bastante más información adicional procedente de otras fuentes.

El calibrado como metodología, se viene utilizando en todas las encuestas de hogares y operaciones estadísticas relativas a la población, desde comienzos del presente siglo y con toda seguridad su presencia va a ser cada vez mayor, conforme vaya aumentando la calidad y cantidad de la información auxiliar susceptible de ser utilizada en la mejora de la producción estadística.

El software utilizado en el INE para este propósito es el programa informático CALMAR, que es uno de los más utilizados para llevar a cabo este tipo de reponderaciones, tanto por su facilidad de funcionamiento, como por la solidez teórica de los métodos en los que se sustenta.

3.2 Ejemplo simulado

En este apartado, se procede a realizar un ejemplo práctico utilizando las técnicas de calibración expuestas e implementadas en la librería “sampling”.
La implementación en R de las técnicas de calibración que se van a explicar es muy reciente, concretamente del año 2016-2017, y dado que es muy innovador se ha decidido explicar los ejemplos con el código control implementado en R para una mejor comprensión de los procedimientos.

Los objetivos de este ejemplo son:

- Estimar el total utilizando calibración, teniendo en cuenta información auxiliar (utilizando la función “calibev”).

- Estimar la varianza del estimador del total de calibración (usando las funciones “calibev” y “varest”).

- Estimar el total utilizando Horvitz-Thompson, sin tener en cuenta información auxiliar (usando la función “HTestimator”).

- Estimar la varianza del estimador del total de Horvitz-Thompson (usando la función “varHT”).

- Comprobar la ventaja de utilizar pesos calibrados en las estimaciones.

Comenzamos mostrando como se van a obtener estos estimadores:

1. Estimador de la varianza del estimador del total de calibración usando la función “calibev”.
   Utiliza el método de los residuales, siendo esta estimación calculada de la siguiente forma:

(11)

$$\widehat{Var}(\hat{\Upsilon}_s)=\sum_{k\in s}\sum_{\ell\in s}((\pi_{k\ell}-\pi_k\pi_\ell)/\pi_{k\ell})(w_k e_k)(w_\ell e_\ell)$$

   Donde:

   - e_k denota el residual de la unidad k.
   - π_k y π_ℓ son las probabilidades de inclusión de las unidades k y l, respectivamente.
   - π_{kl} son las probabilidades de inclusión conjunta de las unidades k y l.
   - w_k y w_ℓ son los pesos calibrados de las unidades k y l, respectivamente.

   Todos estos términos se explican más en profundidad en el ejemplo.

2. Estimador de la varianza del estimador del total de calibración usando la función “varest”.
   Utiliza el método de Deville, el cual es el siguiente:

(12)

$$\widehat{Var}(\hat{\Upsilon}_s)=\frac{1}{1-\sum_{k\in s}a_k^2}\sum_{k\in s}(1-\pi_k)\left(\frac{y_k}{\pi_k}-\frac{\sum_{l\in s}(1-\pi_l)y_l/\pi_l}{\sum_{l\in s}(1-\pi_l)}\right)$$

   Donde:

   - $a_k=(1-\pi_k)/\sum_{l\in s}(1-\pi_l)$.
   - π_k y π_ℓ son las probabilidades de inclusión de las unidades k y l, respectivamente.

1º Paso: Simulación de una población artificial.

El tamaño poblacional es N=100 y el tamaño de la muestra escogido es n=20. La información auxiliar va a estar recogida en tres variables, dos dicotómicas que valen 0 o 1 que contienen cada una un total de 100 valores y otra continua que vale 1, 2, ..., 100.

```
X<- cbind(c(rep(1,50), rep(0,50)), c(rep(0,50), rep(1,50)), 1:100)
```

El total poblacional de la información auxiliar por tanto es la suma de cada una de las tres variables.

```
total<- apply(X, 2, "sum")
50     50     5050
```

Se crea a continuación una variable continua Z que va a determinar la probabilidad de inclusión en la muestra, y que toma los valores 150, 151, …, 249.

```
Z<- 150:249
pik<- inclusionprobabilities(Z, 20)
```

Así mismo se crea la matriz con las probabilidades de inclusión conjunta según el algoritmo de Tillé, desarrollado en 1994 y denominado como “Algoritmo del estrato móvil”, el cual está implementado en R en la función “UPtillepi2”.
Este algoritmo devuelve una matriz NxN de la siguiente forma: la diagonal principal contiene las probabilidades de inclusión de primer orden para cada unidad k en la población; los elementos (k, l) son las probabilidades de inclusión conjunta de las unidades k y l, donde k no es igual a l. N es el tamaño de la población.

```
pikl<- UPtillepi2(pik)
```

Finalmente se crea la variable objeto de estudio o variable respuesta Y a partir de las variables ya creadas X y Z:

$$\mathbf{Y_j = 5 * Z_j * \left(\varepsilon_j + \sum_{i=1}^{3} X_{ij} \right)}, \qquad \boldsymbol{\varepsilon_j \sim N\left(0, \frac{1}{3}\right), j = 1, .., 100}$$

```
Y<- 5*Z*(rnorm(100, 0, sqrt(1/3)) + apply(X, 1, "sum"))
```

2º Paso: Calcular las estimaciones mediante 10000 simulaciones.

El estimador de calibración se va a calcular mediante el método lineal. Las simulaciones se llevan a cabo para calcular el estimador del total de calibración y los dos estimadores de varianza del estimador de calibración.
Así mismo también se va a calcular el estimador del total de Horvitz-Thompson, cuya varianza se puede estimar y compararla con los resultados obtenidos mediante calibración.

Dado que se va a emplear el método lineal, el estimador de calibración va a ser el estimador de regresión generalizada. Por lo tanto, se puede calcular una aproximación de la varianza poblacional y utilizarla en la estimación del sesgo de los estimadores de varianza.

Se procede a ejecutar 10000 simulaciones para obtener resultados precisos:

```
nsim<- 10000
c1=c2=c3=c4=c5=c6=numeric(nsim)
for(i in 1:nsim) {
```

Se crea la muestra de tamaño n=20.

```
s<- UPtille(pik)
```

Se calculan las probabilidades de inclusión en la muestra.

```
piks<- pik[s==1]
```

Se crea la matriz con la información auxiliar de la muestra.

```
Xs<- X[s==1, ]
```

Se calculan los g-weights mediante el método lineal según se explica en 2.1.1.

```
g<- calib(Xs, d=1/piks,total,method="linear")
```

Se selecciona la muestra de la variable de interés.

```
Ys<- Y[s==1]
```

Se selecciona la muestra de las probabilidades de inclusión conjunta.

```
pikls<- pikl[s==1, s==1]
```

Se calcula el estimador de calibración y la estimación de la varianza del estimador de calibración mediante el método de los residuales, utilizando la función “calibev”, recogiendo estos estimadores en c1 y c2 respectivamente.

```
cc<- calibev(Ys, Xs, total, pikls, d=1/piks, g, with=FALSE, EPS=1e-6)
c1[i]<- cc$calest
c2[i]<- cc$evar
```

Se calcula el estimador de la varianza del estimador de calibración mediante el método de Deville, utilizando la función “varest”, recogiendo este estimador en c3.

```
c3[i]<- varest(Ys, Xs, pik=piks, w=g/piks)
```

Finalmente se calcula el estimador del total de Horvitz-Thompson, recogiendo este estimador en c4.

```
c4[i]<- HTestimator(Ys, piks)
```

Y el estimador de la varianza del estimador de Horvitz-Thompson, recogiendo este estimador en c5.

```
c5[i]<- varHT(Ys, pikls, 2)
```

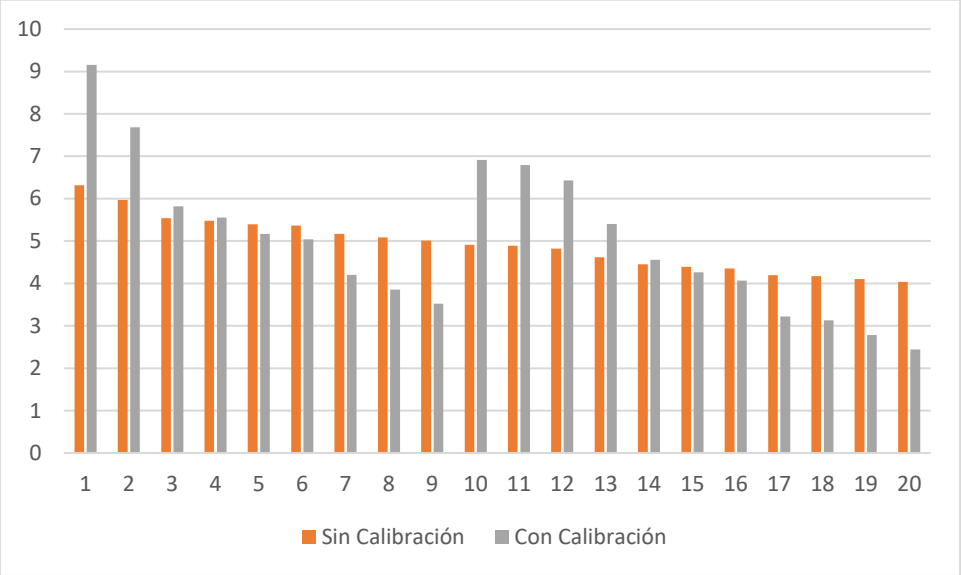
3º Paso: Comparación entre calibrar o no.

Para concluir el ejemplo se procede a comprobar si la calibración mejora la estimación.
Para ver cómo cambian los pesos al utilizar calibración, se muestran a continuación los pesos obtenidos sin y con calibración en una de las iteraciones:

Observación	Pesos sin calibrar	Pesos calibrados
1	6,313291	9,156579
2	5,973054	7,683548
3	5,541667	5,815889
4	5,480769	5,552239
5	5,391892	5,167451
6	5,362903	5,041947
7	5,168394	4,199833
8	5,089286	3,857341
9	5,012563	3,525175
10	4,913793	6,912615
11	4,889706	6,789630
12	4,818841	6,427804
13	4,618056	5,402632
14	4,453125	4,560526
15	4,394273	4,260039
16	4,355895	4,064088
17	4,191176	3,223064
18	4,173640	3,133526
19	4,104938	2,782747
20	4,038462	2,443329
SUMA	**98,28572**	**100**

(13)

En la Tabla (13) se observa cómo cambian los factores de elevación al usar calibración, así como que la suma de los pesos sin calibrar es 98,28572, mientras que al usar calibración la suma de los pesos es exactamente 100, el tamaño poblacional. Esto quiere decir que al no utilizar calibración la estimación final de la variable de interés no va a ser del todo correcta, ya que se está perdiendo información al ajustar mal los pesos, mientras que, al usar calibración, la estimación va a ser más exacta dado que no se pierde nada de información en el proceso de la construcción de los pesos.



El total poblacional es la suma de la variable respuesta Y.

sum(Y) = 5565865

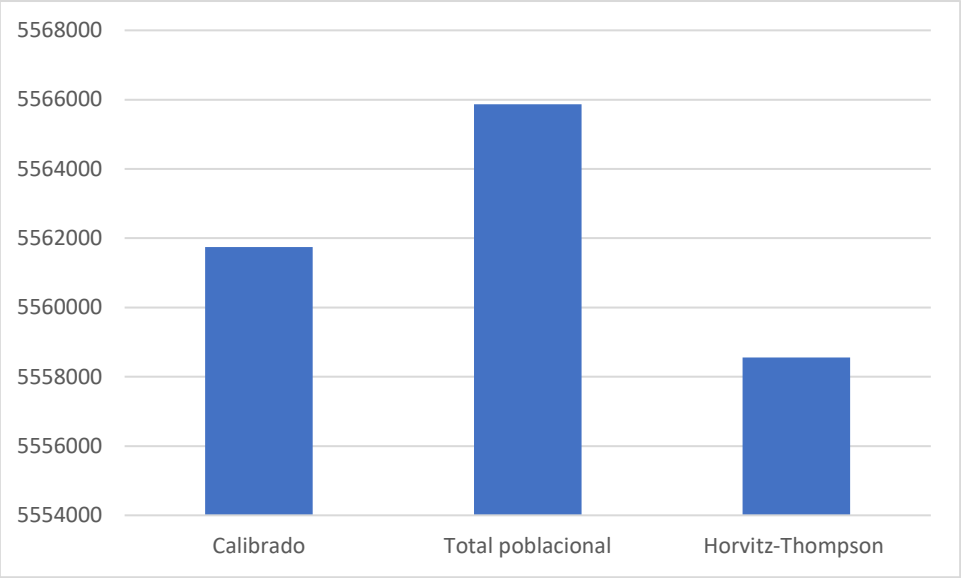
El estimador del total de Horvitz-Thompson y la diferencia de este estimador con respecto al total poblacional es:

mean(c4) = 5558559
abs(sum(Y) - mean(c4)) = 7306.059

El estimador del total de calibración y la diferencia de este estimador con respecto al total poblacional es:

mean(c1) = 5561741
abs(sum(Y) - mean(c1)) = 4124.222

Se puede comprobar como el estimador del total calibrado se encuentra más próximo al total poblacional que el estimador del total de Horvitz-Thompson.



A continuación, se procede estimar la varianza del estimador calibrado.

Estimador de la varianza del estimador de calibración mediante el método de los residuales, utilizando la función “calibev”:

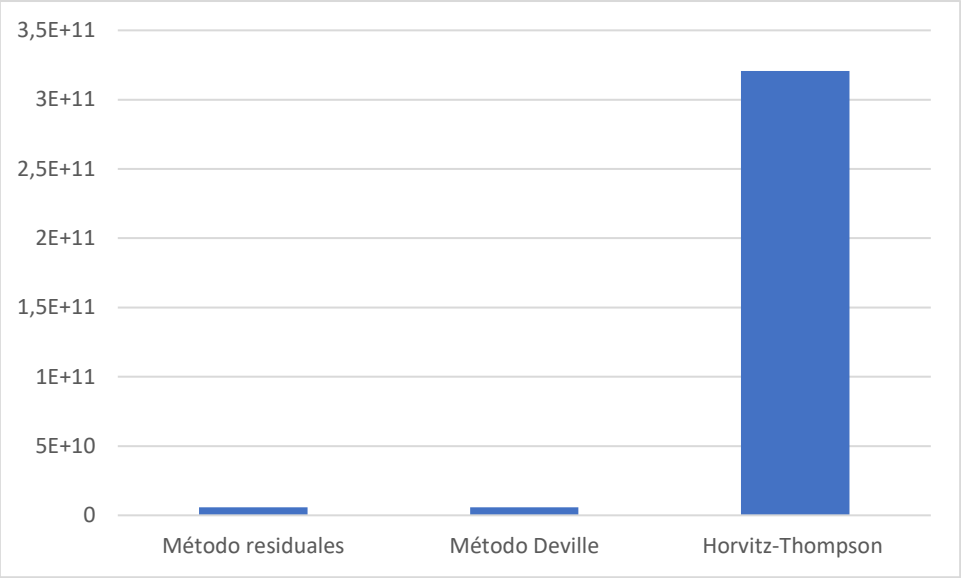
mean(c2) = 5904034353

Estimador de la varianza del estimador de calibración mediante el método de Deville, utilizando la función “varest”:

mean(c3) = 5885447105

Estimador de la varianza del estimador de Horvitz-Thompson:

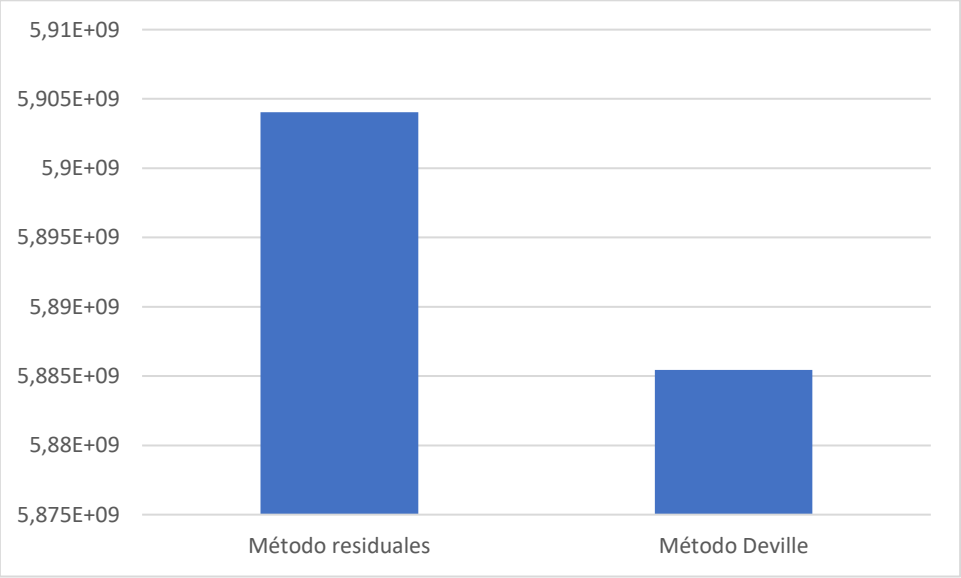
mean(c5) = 320667778996



Se comprueba que la reducción de la varianza utilizando calibración es significativa con respecto a no usarla, la estimación de la varianza de Horvitz-Thomp es 320667778996, mientras que la estimación de la varianza por el método de los residuales y el método de Deville es significativamente más pequeña, 5904034353 y 5885447105.

Por tanto, utilizar calibración reduce el error de estimación.

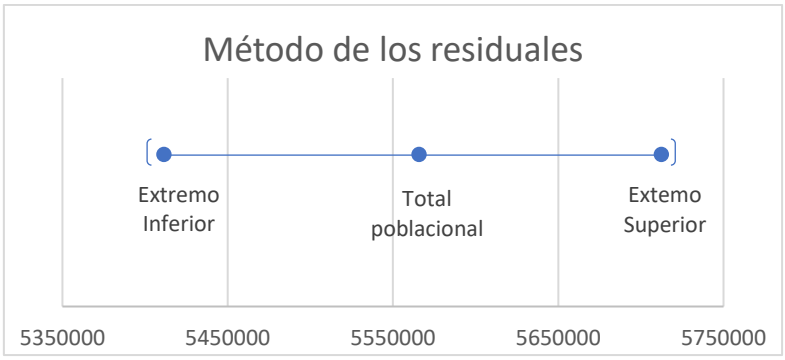
De igual forma, y aunque en menor medida, dentro de la estimación de la varianza del estimador del total por calibración, es mejor el método de los residuales, que a diferencia del método de Deville, tiene en cuenta las probabilidades de inclusión conjunta en la muestra.



Concluimos la simulación estimando, con y sin técnica de calibración, el parámetro total con un intervalo de confianza del 95%:

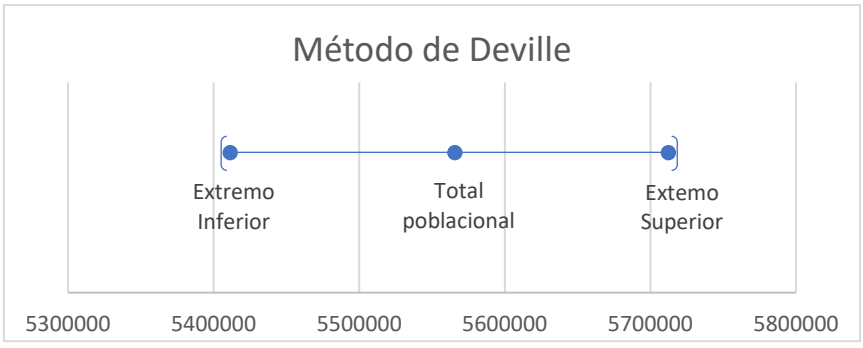
I.C. del estimador de calibración mediante el método de los residuales:

```
ic.c2<- mean(c1) + c(-1,1)*qnorm(0.975)*sqrt(mean(c2))
ic.c2 = [5411142, 5712340]
```



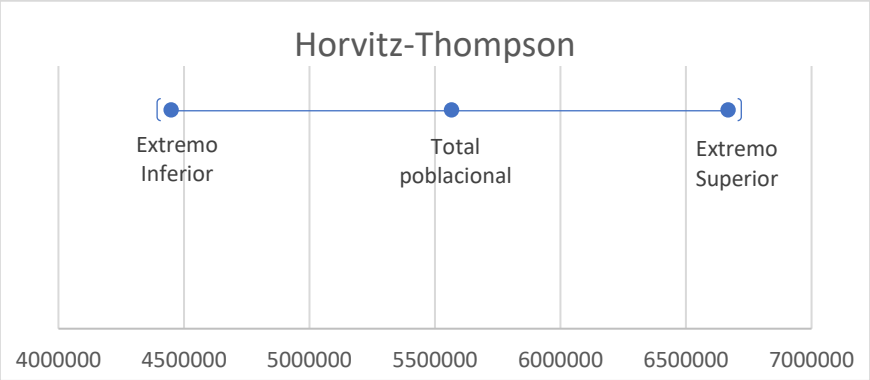
I.C. del estimador de calibración mediante el método de Deville:

```
ic.c3<- mean(c1) + c(-1,1)*qnorm(0.975)*sqrt(mean(c3))
ic.c3 = [5411379, 5712103]
```



I.C. del estimador de Horvitz-Thompson:

```
ic.c5<- mean(c4) + c(-1,1)*qnorm(0.975)*sqrt(mean(c5))
ic.c5 = [4448680, 6668439]
```



Se observa como todos los intervalos contienen al verdadero valor del parámetro, total poblacional, siendo los dos intervalos obtenidos a partir del uso de calibración más precisos, menos amplios, que el obtenido sin utilizar calibración.

3.3 Ejemplo con datos reales procedentes de la EPA

Para finalizar este trabajo se ha decidido aplicar los métodos de calibración explicados en una situación real.
El objetivo es lograr la estimación del número de parados en Castilla y León utilizando calibración para establecer los factores de elevación. Para ello se han recogido los datos desde la página web del INE, http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&secc=1254736030639&idp=1254735976595, en particular de los microdatos que esta entidad pone a disposición de forma pública.
En este ejemplo se utilizan los datos pertenecientes al 1º trimestre de 2018, siendo estos los más actuales disponibles a la fecha de la realización de este trabajo.
Estos datos han sido recogidos por el INE mediante la EPA (Encuesta de Población Activa), la cual recoge un total de 73 variables. En este ejemplo solo se manejan 3 de ellas, que se definirán más adelante.

Cabe decir que lo que se hace en este trabajo ya está hecho por el INE, el cual utiliza, como ya se ha dicho anteriormente, CALMAR. Por lo tanto, el objetivo de este trabajo es lograr una estimación del número de parados en Castilla y León lo más cercana posible a la estimada por el INE.
No se puede lograr unos resultados tan exactos como los obtenidos por el INE dado que esta entidad utiliza un software más sofisticado que posee mayor capacidad computacional, ya que utilizan más variables auxiliares que las utilizadas en este estudio debido a la intensa carga computacional que esto supone.

Para entender bien lo que se va a hacer es necesario explicar qué es la EPA y como es el diseño de la encuesta, a saber:

- Objetivos que tiene la EPA
- Ámbito de la Encuesta
- Marco de la Encuesta
- Diseño muestral utilizado
- Estimadores utilizados

¿Qué es la EPA?

La Encuesta de Población Activa es una encuesta de tipo continuo dirigida a investigar características socioeconómicas de la población, que viene siendo realizada por el INE desde 1964.

Desde su implantación ha sufrido diferentes modificaciones, siempre dirigidas a mejorar la información que proporciona.

Así mismo el INE realiza una evaluación de la calidad de los datos recogidos en la EPA.

Objetivos

La EPA tiene como objetivo principal el conocimiento de la actividad económica del país, en lo relativo al componente humano. Su diseño está orientado a proporcionar información de las principales categorías poblacionales en relación con el mercado de trabajo, así como obtener clasificaciones de estas categorías según distintas variables.

Ámbito de la Encuesta

El ámbito abarcado por la encuesta se desglosa en los tres apartados siguientes:

- **Ámbito poblacional**

  La Encuesta va dirigida a la población que reside en viviendas familiares principales, es decir, las utilizadas toda o la mayor parte del año como residencia habitual o permanente.

- **Ámbito geográfico**

  La encuesta se realiza en todo el territorio nacional.

- **Ámbito temporal**

  La EPA es una encuesta continua con periodicidad trimestral, extendiéndose las entrevistas a lo largo de las trece semanas del trimestre.

Marco de la Encuesta

Para definir el marco de la Encuesta es necesario partir de la división administrativa de España, que aparece de la forma siguiente:

Toda la Nación se encuentra dividida en 17 comunidades autónomas y dos ciudades autónomas. Las comunidades autónomas se dividen a su vez en 50 provincias, de las cuales 47 son peninsulares y 3 insulares. Las provincias se encuentran divididas en municipios y éstos a su vez en distritos municipales.

A partir de lo anterior el INE conjuntamente con los Ayuntamientos hace una nueva subdivisión de los distritos municipales en secciones censales.

Diseño muestral

Tipo de muestreo y unidades muestrales

Se utiliza un muestreo bietápico con estratificación de las unidades de primera etapa.

Las unidades de primera etapa están constituidas por las secciones censales. La muestra de secciones permanece fija indefinidamente con las excepciones siguientes:

1) Salen de la muestra aquellas secciones en las que ya se han visitado todas las viviendas encuestables.

2) Cuando en el proceso de actualización del seccionado a algunas secciones les corresponde salir de la muestra, bien por los cálculos probabilísticos, bien por cambios en la afijación por estratos.

Las unidades de segunda etapa están constituidas por las viviendas familiares principales (ocupadas permanentemente) y los alojamientos fijos (chabolas, cuevas, etc.). No se consideran encuestables las viviendas secundarias o de temporada (ocupadas sólo una parte del año), ni las disponibles para alquiler o venta, ya que no forman parte del ámbito poblacional definido anteriormente.

Dentro de las unidades de segunda etapa no se realiza submuestreo alguno, recogiéndose información de todas las personas que tengan su residencia habitual en las mismas.

Estimadores utilizados

A partir de 2002 se aplican técnicas de reponderación a los estimadores con objeto de ajustar las estimaciones de la encuesta a la información procedente de fuentes externas. Es decir, utilizan factores de elevación calibrados.

Estimación del total de parados en Castilla y León

Para lograr este objetivo la EPA utiliza las siguientes variables auxiliares:

1) Población de 16 y más años por grupos quinquenales de edad y sexo.

2) Población de 16 y más años por nacionalidad española y extranjera.

3) Población de 16 y más años clasificados por tamaño del hogar (1, 2, 3, 4 y 5 o más).

En este trabajo solo se van a tomar las variables auxiliares de grupos quinquenales de edad y sexo, ya que la carga computacional que conlleva utilizar las otras dos utilizadas por el INE es inviable con el software disponible.

Recogida y tratamiento de datos

Los datos han sido tomados desde la sección de “Microdatos” de la EPA, ofrecido de forma pública por el INE. Se han tomado los microdatos pertenecientes al 1º trimestre de 2018.

Este conjunto de datos consta de un total de 160874 individuos y 73 variables.

Estos datos pertenecen a toda España, y dado que en este trabajo solo interesan aquellos individuos pertenecientes a la comunidad autónoma de Castilla y León, se ha aplicado un filtro utilizando para ello la variable CCAA, cuyos valores se corresponden con las posiciones 4 y 5 del conjunto de datos. Cada comunidad autónoma se corresponde con un valor entre 01 y 52, por lo que se ha aplicado el filtro quedándonos con los datos correspondientes al valor 07, que es el indicador de Castilla y León.

Una vez seleccionados los datos de Castilla y León se procede a seleccionar las variables de interés para el estudio, a saber:

- OFEMP: variable indicadora de ser parado o no, clasificando esta variable en valores (1, 0), siendo un 1 si la persona está en paro y un 0 en caso contrario. Es la variable de estudio. A esta variable se la renombrará como Parado.

- EDAD5: variable indicadora del grupo quinquenal de edad al que pertenece el individuo, pudiendo valer esta variable:

  00 = de 0 A 4 años
  05 = de 5 A 9 años
  10 = de 10 A 15 años
  16 = de 16 A 19 años
  20 = de 20 A 24 años
  25 = de 25 A 29 años
  30 = de 30 A 34 años
  35 = de 35 A 39 años
  40 = de 40 A 44 años
  45 = de 45 A 49 años
  50 = de 50 A 54 años
  55 = de 55 A 59 años
  60 = de 60 A 64 años
  65 = 65 o más años

  De este modo, y como en el estudio se va a tratar con personas de edad superior o igual a 16 años, se ha aplicado un filtro excluyendo todos los individuos cuyo valor en esta variable fuese 00, 05 y 10. También se van a agrupar estos valores en los utilizados por el INE para estimar el número de parados, siendo estos:

  16 = de 16 A 19 años
  20 = de 20 A 24 años
  30 = de 25 A 34 años
  40 = de 35 A 44 años
  50 = de 45 A 54 años
  55 = de 55 o más años

  A esta variable se la renombrará como Grupo_Edad.

- SEXO1: variable indicadora de sexo, valiendo un 1 si el individuo es varón y un 6 si el individuo es mujer. Para un cómodo manejo de los datos a la hora de realizar los cálculos, se ha sustituido el valor indicador de mujer, 6, por 0. A esta variable se la renombrará como Sexo.

De este modo el conjunto de datos se ha reducido a 10727, que va a ser el tamaño muestral empleado en el estudio.

El conjunto de datos en esta fase del tratamiento de los datos tiene la siguiente forma:

Parado	Sexo	Grupo_Edad
0	1	40
0	0	30
1	0	20
0	1	55
0	0	16
… …	… …	… …
0	1	40

A continuación, se procede a dividir la variable auxiliar Grupo_Edad en 6 variables auxiliares, denotándolas por cada posible valor de la variable Grupo_Edad y con valores dicotómicos (1, 0), indicando un 1 si el individuo pertenece a su edad correspondiente y un 0 en caso contrario. Esto se hace para poder calibrar correctamente los pesos.

Finalmente, el conjunto de datos con los que se va a trabajar consta de un total de 10727 individuos, la variable de estudio, Parado, y 7 variables auxiliares, teniendo la forma de la siguiente tabla:

Parado	**Sexo**	**Edad_16_19**	**Edad_20_24**	**Edad_25_34**	**Edad_35_44**	**Edad_45_54**	**Edad_55_o_mas**
0	1	0	0	0	0	1	0
0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	1
1	0	0	0	0	0	0	1
… …	… …	… …	… …	… …	… …	… …	… …
0	1	1	0	0	0	0	0

Una vez se tienen los individuos con sus correspondientes valores tanto en la variable de estudio como en las variables auxiliares, para el proceso de calibración es necesario saber los totales poblacionales de las variables auxiliares, ya que son imprescindibles tanto para la construcción de los pesos calibrados como para la posterior estimación del parámetro total de parados.

Estos totales poblacionales también son obtenidos desde el INE, correspondiéndose cada total poblacional con el 1º trimestre de 2018. Estos totales son los siguientes:

Sexo: 1.025.382, siendo este el número de varones, dado que en esta variable el indicador 1 corresponde a varones. El número de mujeres es 1.069.452, aunque este valor no va a ser utilizado.

Edad_16_19: 74.655

Edad_20_24: 103.631

Edad_25_34: 241.436

Edad_35_44: 343.505

Edad_45_54: 378.507

Edad_55_o_mas: 953.100

El total poblacional de individuos de 16 años o más en Castilla y León es

N = 2.094.834.

Para comprobar que los totales de las variables auxiliares han sido recogidos correctamente se comprueba que la suma de varones y mujeres (1.025.382 + 1.069.452 = 2.094.834) es igual a la suma de los individuos pertenecientes a cada edad (74.655 + 103.631 + 241.436 + 343.505 + 378.507 + 953.100 = 2.094.834). A su vez se comprueba que este resultado es N, por lo que la recogida de los totales de las variables auxiliares ha sido correcta.

Obtención de los pesos calibrados y estimación del parámetro

A continuación, se procede a obtener los factores de elevación calibrados, con el objetivo de estimar el número total de parados de Castilla y León. También se va a realizar la comparación del estimador de este parámetro con el estimador que se obtiene sin utilizar información auxiliar, utilizando para ello el estimador de Horvitz-Thompson.

Esta comparación va a hacerse tomando el valor estimado por el INE del número de parados en Castilla y León como el valor “real”.

Así mismo se va a acompañar este estudio con el código control utilizado en R, ayudándonos para ello de las funciones y la ayuda que R ofrece.

Al igual que en el ejemplo explicado en el apartado 6, la librería que se va a utilizar es “sampling”, haciendo uso de las siguientes funciones:

1) calib: para obtener los factores de elevación calibrados.

2) calibev: para obtener el estimador del total calibrado y la estimación de la varianza del estimador del total calibrado mediante el método de los residuales (11).

3) HTestimator: para obtener el estimador del total de Horvitz-Thompson.

4) varHT: para obtener la estimación de la varianza del estimador del total de Horvitz-Thompson.

datos: fichero donde se encuentran todos los individuos, siendo:

La 2º columna la variable de estudio Parados.

La 3º columna la variable auxiliar Sexo.

Las columnas 4 hasta la 9 las variables auxiliares de Grupo de Edad.

Asignación de la variable de estudio, Ys, y las 7 variables auxiliares, Xs:

```
Ys<- datos[,2]
Xs<- datos[,3:9]
```

Total de habitantes en Castilla y León de más de 16 años: 2.094.834

```
N<- 2094834
```

Tamaño de la muestra, los 10727 individuos:

```
n<- dim(datos)[1]
```

Totales de las variables auxiliares, descritos anteriormente:

```
total<- c(1025382, 74655, 103631, 241436, 343505, 378507, 953100)
```

Las probabilidades de inclusión utilizadas por el INE se describen en función del tamaño del hogar en el que vive cada individuo, sin embargo, ante la imposibilidad de incluir esta variable en el estudio, se han tomado las probabilidades de inclusión como si se tratase de muestreo aleatorio simple, teniendo de este modo cada individuo la misma probabilidad de inclusión en la muestra: $\frac{n}{N}$. Al vector que contiene las probabilidades de inclusión en la muestra se lo denomina piks:

```
piks<- rep(n/N, n)
```

Al fijar las probabilidades de inclusión como equiprobables, la matriz que contiene las probabilidades de inclusión conjuntas también va a tener probabilidades equiprobables, teniendo los valores $\frac{n}{N}$ en la diagonal y $(\frac{n}{N})^2$ en el resto de posiciones no diagonales. A esta matriz se la denota como pikls:

```
pikls<- matrix(0, n, n)

diagonal<- n/N
resto<- (n/N)^2

for (i in 1:n) {
  for (j in 1:n) {
    if(i==j) {
      pikls[i, j]<- diagonal
    }
    else {
      pikls[i, j]<- resto
    }
  }
}
```

Por tanto, el vector con las probabilidades de inclusión y la matriz con las probabilidades de inclusión conjuntas tienen la siguiente forma:

piks: [$\frac{n}{N}$, $\frac{n}{N}$, $\frac{n}{N}$, $\frac{n}{N}$, … , $\frac{n}{N}$]

pikls:

Individuo	1	2	3	…	10.726	10.727
1	$\frac{n}{N}$	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$	…	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$
2	$(\frac{n}{N})^2$	$\frac{n}{N}$	$(\frac{n}{N})^2$	…	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$
3	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$	$\frac{n}{N}$	…	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$
…	…	…	…	…	…	…
10.726	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$	…	$(\frac{n}{N})^2$	$\frac{n}{N}$	$(\frac{n}{N})^2$
10.727	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$	$(\frac{n}{N})^2$	…	$(\frac{n}{N})^2$	$\frac{n}{N}$

A continuación, se procede a la construcción de los pesos calibrados. El método utilizado por el INE es el lineal truncado, pero en este trabajo vamos a probar los cuatro métodos de calibración explicados en el apartado 2 para comparar cual funciona mejor:

Método lineal:

```
g<- calib(Xs, d=1/piks, total, method = "linear")
```

Método exponencial:

```
g<- calib(Xs, d=1/piks,total, method = "raking")
```

Método lineal logit:

```
g<- calib(Xs, d=1/piks, total, method = "logit")
```

Método lineal truncado:

```
g<- calib(Xs, d=1/piks, total, method = "truncated", bounds=c(low= 0, upp=10))
```

Para cada método de obtención de los pesos calibrados se va a estimar el total y la varianza del estimador del total, mediante calibrado, guardando estas estimaciones en c1 y c2, respectivamente:

```
cc<- calibev(Ys, Xs, total, pikls, d=1/piks, g, with=FALSE, EPS=1e-6)
c1<- cc$calest
c2<- cc$evar
```

De igual forma se va a obtener un intervalo de confianza con una confianza del 95% para cada método:

```
IC<- c1 + c(-1, 1) * 1.96 * sqrt(c2)
```

Finalmente, también se va a calcular el estimador de Horvitz-Thompson y el estimador de la varianza del estimador de Horvitz-Thompson, guardando estas estimaciones en c4 y c5 respectivamente. Igualmente, se va a calcular también su correspondiente intervalo de confianza con una confianza del 95%.

```
c4<- HTestimator(Ys, piks)
c5<- varHT(Ys, pikls, 1)
ICHT<- c4 + c(-1, 1) * 1.96 * sqrt(c5)
```

Los resultados obtenidos de todas estimaciones son los siguientes:

Método	Estimador del total	Estimador de la varianza del estimador del total	Intervalo de confianza al 95%
Lineal	173.244	29.551.506	[162.589, 183.898]
Exponencial	173.240	29.552.644	[162.584, 183.894]
Lineal logit	173.240	29.552.570	[162.585, 183.895]
Lineal truncado	173.243	29.551.506	[162.589, 183.898]
Horvitz-Thompson	180.444	35.057.830	[168.839, 192.049]

Total de parados en Castilla y León según el INE: 154.800

Se puede observar que no hay diferencia significativa entre los cuatro métodos de calibración, dando todos unas estimaciones similares. Dado que se puede suponer que los cuatro métodos obtienen las mismas estimaciones, se comparará utilizando los resultados del método lineal truncado, que es el utilizado por el INE.

La diferencia del estimador del total de calibración de parados en Castilla y León obtenido en este trabajo de la estimación publicada por el INE es de 18.443.

Este valor dado por el INE no se encuentra dentro de nuestro intervalo de confianza, siendo la diferencia de este valor “real” con el extremo inferior del intervalo de confianza de 7.789.

Esta diferencia se debe a la exclusión en este trabajo de las ya mencionadas variables auxiliares de nacionalidad y tamaño del hogar.
No obstante, se observa como el utilizar calibración mejora la estimación, ya que el estimador no calibrado del total de Horvitz-Thompson es 180.444, que difiere del valor dado por el INE en 25.644 individuos.

Así mismo el estimador de la varianza del estimador del total obtenida mediante Horvitz-Thompson es 35.057.830, mientras que la obtenida mediante calibración es 29.551.506. Se puede observar una clara reducción en la estimación de la varianza al utilizar calibración, lo que supone un intervalo de confianza menos amplio.

La diferencia del valor dado por el INE del total de parados en Castilla y León con respecto al extremo inferior del intervalo de confianza obtenido mediante Horvitz-Thompson es 14.039.

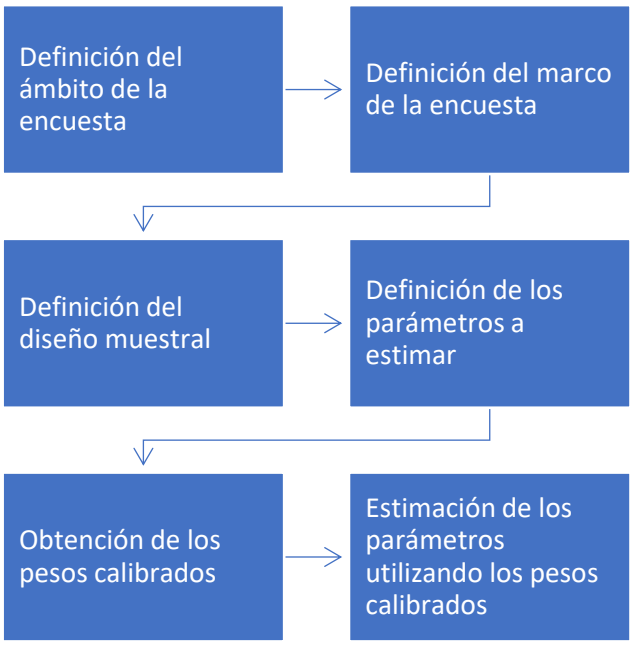
5. Conclusiones

Se ha demostrado como mejora la calidad de las estimaciones haciendo uso de técnicas de calibrado, las cuales ayudan a corregir el sesgo introducido por la falta de respuesta, y proporcionan consistencia entre las cifras presentadas y las procedentes de otras fuentes.

Así mismo se ha visto también la sencillez y facilidad de emplear estas técnicas, ya que, si bien es cierto que gran parte de las instituciones que utilizan técnicas de calibrado lo hacen mediante el software no gratuito CALMAR, perteneciente a SAS, también es posible utilizar estas técnicas de manera sencilla utilizando software libre como lo es R, y así se ha mostrado en este trabajo.

El objetivo de este trabajo se ha visto cumplido, ya que se ha corroborado como la utilización de factores de elevación calibrados ha mejorado las estimaciones, obteniendo una estimación del total de parados en Castilla y León más cercana al valor dado por el INE, así como la reducción de la estimación varianza al utilizar calibración.

Finalmente, y a forma de resumen, los pasos que han de llevarse a cabo si se pretende utilizar pesos calibrados son los siguientes:



6. Código control utilizado

Ejemplo simulado

```
X<- cbind(c(rep(1,50), rep(0,50)), c(rep(0,50), rep(1,50)), 1:100)
total<- apply(X, 2, "sum")
Z<- 150:249
pik<- inclusionprobabilities(Z, 20)
pikl<- UPtillepi2(pik)
Y<- 5*Z*(rnorm(100, 0, sqrt(1/3)) + apply(X, 1, "sum"))
nsim<- 10000
c1=c2=c3=c4=c5=c6=numeric(nsim)
for(i in 1:nsim) {
        s<- UPtille(pik)
        piks<- pik[s==1]
        Xs<- X[s==1, ]
        g<- calib(Xs, d=1/piks,total,method="linear")
        Ys<- Y[s==1]
        pikls<- pikl[s==1, s==1]
        cc<- calibev(Ys, Xs, total, pikls, d=1/piks, g, with=FALSE, EPS=1e-6)
        c1[i]<- cc$calest
        c2[i]<- cc$evar
        c3[i]<- varest(Ys, Xs, pik=piks, w=g/piks)
        c4[i]<- HTestimator(Ys, piks)
        c5[i]<- varHT(Ys, pikls, 2) }
```

```
sum(Y) = 5565865
mean(c4) = 5558559
abs(sum(Y) - mean(c4)) = 7306.059
mean(c2) = 5904034353
mean(c3) = 5885447105
mean(c5) = 320667778996
ic.c2<- mean(c1) + c(-1,1)*qnorm(0.975)*sqrt(mean(c2))
ic.c2 = [5411142, 5712340]
ic.c3<- mean(c1) + c(-1,1)*qnorm(0.975)*sqrt(mean(c3))
ic.c3 = [5411379, 5712103]
ic.c5<- mean(c4) + c(-1,1)*qnorm(0.975)*sqrt(mean(c5))
ic.c5 = [4448680, 6668439]
```

Ejemplo con datos reales procedentes de la EPA

```
Ys<- datos[,2]
Xs<- datos[,3:9]
N<- 2094834
n<- dim(datos)[1]
total<- c(1025382, 74655, 103631, 241436, 343505, 378507, 953100)
piks<- rep(n/N, n)
pikls<- matrix(0, n, n)
diagonal<- n/N
resto<- (n/N)^2
for (i in 1:n) {
  for (j in 1:n) {
    if(i==j) {
      pikls[i, j]<- diagonal
    }
    else {
      pikls[i, j]<- resto
    }
  }
}
g<- calib(Xs, d=1/piks, total, method = "linear")
g<- calib(Xs, d=1/piks,total, method = "raking")
g<- calib(Xs, d=1/piks, total, method = "logit")
g<- calib(Xs, d=1/piks, total, method = "truncated", bounds=c(low= 0, upp=10))
cc<- calibev(Ys, Xs, total, pikls, d=1/piks, g, with=FALSE, EPS=1e-6)
c1<- cc$calest
c2<- cc$evar
IC<- c1 + c(-1, 1) * 1.96 * sqrt(c2)
c4<- HTestimator(Ys, piks)
c5<- varHT(Ys, pikls, 1)
ICHT<- c4 + c(-1, 1) * 1.96 * sqrt(c5)
```

6. Bibliografía

Claes M. Cassel, Carl E. Sarndal and Jan H. Wretman (1976). «Biometrika».

COCHRAN, W.G. (1977) «Sampling Techniques». Wiley.

SARNDAL, C.A. et al. (1991) «Model Assisted Survey Sampling». Springer.

DEVILLE, J.C. and SANDARL, C.E. (1992) «Calibration Estimators in Survey S ampling». JASA.

Deville, J.-C. (2000). «Note sur l'algorithme de Chen, Dempster et Liu».

Jorge Saralegui Gil, INE, (2005). «Problemas de calidad de las encuestas de ho gares en presencia de flujos intensos de migraciones exteriores».

Carlos Pérez Arriero (2012). «Calibrating a household survey by using the Calmar program».

Felipe Jiménez (2014). «Tutorial de muestreo en R».

Kevin McCormack (2015). «The calibration software CALMAR».

Panos M. Pardalos, Anatoly Zhigljavsky, Julius Zilinskas (2016). «Advances in Stochastic and Deterministic Global Optimization».

R-CRAN (2016). Calibration and generalized calibration.

INE (2018). «Encuesta de Población Activa (EPA)».